Speech Modeling and Robust Estimation for Diagnosis of Parkinson's Disease

Shi, Liming

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

SPEECH MODELING AND ROBUST ESTIMATION FOR DIAGNOSIS OF PARKINSON'S DISEASE

BY
LIMING SHI

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

Speech Modeling and Robust Estimation for Diagnosis of Parkinson's Disease

Ph.D. Dissertation

Liming Shi

Dissertation submitted September, 2019

Dissertation submitted:	October 2019
PhD supervisor:	Prof. Mads Græsbøll Christensen Aalborg University
Assistant PhD supervisor:	Assoc. Prof. Jesper Rindom Jensen Aalborg University
PhD committee:	Associate Professor Troels Pedersen (chairman) Aalborg University
	Senior Researcher, PhD Phillip Garner Idiap Research Institute
	Professor Jonathon A. Chambers University of Leicester
PhD Series:	Technical Faculty of IT and Design, Aalborg University
Department:	Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-531-4

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2020

Curriculum Vitae

Liming Shi

Liming was born on August 11, 1989 in Henan, China. He received the B.Eng. degree in electronics and information engineering from Henan University of Technology, Henan, China and the M.Eng. degree in information and communication engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2012 and 2015, respectively. He is currently a PhD student at the Audio Analysis Lab, Department of Architecture, Design and Media Technology, Aalborg University, Aalborg, Denmark. His research interests include adaptive filtering and speech analysis.

Curriculum Vitae

Abstract

According to the Parkinson's Foundation, more than 10 million people worldwide suffer from Parkinson's disease (PD). The common symptoms are tremor, muscle rigidity and slowness of movement. There is no cure available currently, but clinical intervention can help alleviate the symptoms significantly. Recently, it has been found that PD can be detected and telemonitored by voice signals, such as sustained phonation /a/. However, the voiced-based PD detector suffers from severe performance degradation in adverse environments, such as noise, reverberation and nonlinear distortion, which are common in uncontrolled settings.

In this thesis, we focus on deriving speech modeling and robust estimation algorithms capable of improving the PD detection accuracy in adverse environments. Robust estimation algorithms using parametric modeling of voice signals are proposed. We present both segment-wise and sample-wise robust pitch tracking algorithms using the harmonic model. The first order Markov chain is used to impose smoothness prior for the pitch. In segment-wise pitch tracking, we have proposed a method to track the pitch, harmonic order and voicing state jointly based on Bayesian tracking framework. In sample-wise pitch tracking, to deal with colored noise, the noise is modeled as time-varying autoregressive process. The proposed algorithms are compared with the state-of-the-art pitch estimation algorithms and are evaluated on the Parkinson's disease database. Apart from extracting pitch information, we have also looked into the problem of autoregressive moving average (ARMA) modeling of voiced speech and its parameters estimation. Due to the sparse nature of the excitation signal for the voiced speech, both least 1-norm criterion and sparse Bayesian learning are applied to improve the ARMA coefficients estimation. The proposed ARMA estimation methods are shown to perform better than the least squares based method in terms of spectral distortion. We have also proposed a dictionary-based speech enhancement algorithm using non-negative matrix factorization, where the dictionary items for both speech and noise are parameterized by AR coefficients. Finally, we investigated on the performance of a vast number of speech enhancement and dereveberation algorithms for diagnosis of PD with degraded speech signals.

Resumé

Ifølge Parkinson's Foundation er der mere end 10 millioner mennesker over hele verden der lider af Parkinsons sygdom. De typiske symptomer er rysten, muskelstivhed og langsomme bevægelser. Aktuelt er der ingen kur, men klinisk indgriben kan hjælpe med at mindske symptomerne betragteligt. For nylig fandt man ud af at Parkinsons sygdom kan detekteres og telemonitorereres via talesignaler såsom vedvarende fonation af lyden /a/. Desværre lider denne detektor alvorlige tab i ydelse i kritiske miljøer med støj, rumklang og ulineær forvrængning, hvilket er typisk i ukontrollerede omgivelser.

I denne afhandling fokuserer vi på at udlede robuste talemodeller og estimeringsalgoritmer, som er i stand til at forbedre detektionen af Parkinsons sygdom i kritiske miljøer. Til dette formål foreslås robuste estimatorer baseret på parametrisk modellering af talesignaler. Der præsenteres både segment- og samplevise robuste algoritmer til sporing af grundfrekvensen baseret på den harmoniske model. Disse anvender en førsteordens Markovkæde til at indføre forhåndsviden omkring grundfrekvensens jævne udvikling. Til segmentvis sporing af grundfrekvensen, er der foreslået en metode til at spore både grundfrekvensen, den harmoniske modelorden, og stemmetilstanden samtidigt baseret på en Bayesiansk fremgangsmåde. I den samplevise metode til grundfrekvenssporing, benyttes en tidsvarierende autoregressiv proces til at modellere støjen for a kunne håndtere farvet støj. De foreslåede algoritmer er sammenlignet med de nyeste algoritmer til grundfrekvensestimering og er evalueret på en database med reelle taleoptagelser fra Parkinsons patienter.

Ud over ekstrahering af information vedrørende grundfrekvensen, kigges der også på autoregressiv glidende gennemsnit (ARMA) modellering a stemt tale og estimering af tilhørende parametre. På grund af eksiteringssignalernes sparsomme natur ved stemt tale anvendes der både et 1-norms kriterie og sparsom Bayesiansk læring til at forbedre estimeringen af ARMA koefficienterne. De foreslåede ARMA estimeringsmetoder har vist sig at være mere nøjagtige end mindste kvadrater metoden i forhold til spektral forvrængning. Der er også foreslået en kodebogsbaseret taleforbedringsalgoritme basered på ikke-negativ matrixfaktorisering, hvor kodebogselementerne for både tale

og støj er parameteriseret ved hjælp af autoregressive koefficienter. Til sidst undersøges ydeevnen for en række metoder til taleforbedring og reducering af rumklang i forhold til diagnose af Parkinsons sygdom ud fra forringede taleoptagelser.

List of publications

The main body of this thesis consists of the following publications:

[A] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust Bayesian pitch tracking based on the harmonic model," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 11, pp. 1737–1751, Nov 2019.

[B] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "A Kalman-based fundamental frequency estimation algorithm," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct 2017, pp. 314–318.

[C] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Instantaneous Bayesian pitch tracking in colored noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, submitted, Sep 2019.

[D] L. Shi, J. R. Jensen, J. K. Nielsen, and M. G. Christensen, "Multipitch estimation using block sparse Bayesian learning and intra-block clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 2018, pp. 666–670.

[E] L. Shi, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A variational EM method for pole-zero modeling of speech with mixed block sparse and Gaussian excitation," in *Proc. European Signal Processing Conf.*, Aug 2017, pp. 1784–1788.

[F] L. Shi, J. R. Jensen, and M. G. Christensen, "Least 1-norm pole-zero modeling with sparse deconvolution for speech analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, March 2017, pp. 731–735.

[G] A. H. Poorjam, M. S. Kavalekalam, L. Shi, Y. P. Raykov, J. R. Jensen, M. A. Little, and M. G. Christensen, "Automatic quality control and enhancement for voice-based remote Parkinson's disease detection," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, submitted, 2019.

[H] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric NMF for speech enhancement," in *Proc. European Signal Processing Conf.*, Sep. 2018, pp. 2320–2324.

Contents

Preface

This thesis is submitted to the Technical Faculty of IT and Design at Aalborg University in partial fulfilment of the requirements for the degree of doctor of philosophy. The thesis consists of two parts: the first part is an introduction to the field of speech modeling and robust estimation and summarises the contributions of the Ph.D. project. The introduction is followed by the scientific papers that have been published through different outlets.

The work was carried out in the period of June 2016 to June 2019 in the Audio Analysis Lab at Department of Architecture, Design and Media Technology at Aalborg University. The project was made possible by a research grant from the Danish Council for Independent Research under Grant DFF 4184-00056. First of all, I would like to express my sincere gratitude to my supervisor Mads Græsbøll Christensen for his guidance and motivating me throughout my Ph.D. work. I would like to thank him for his ideas, and giving me the freedom to explore the vast field of speech signal processing. I am also grateful to my co-supervisor Jesper Rindom Jensen for his supervision and commenting on my papers in spite of his busy schedule. Finally, I would like to thank Jesper Kjær Nielsen for his valuable time and guidance.

I thank my colleagues at the Audio Analysis Lab and friends for making life at Aalborg University a pleasant one and for all the social gatherings we had during the past three years. I would like to thank my parents for supporting me and standing with me through my ups and downs. Finally, I would like to thank my wife Jing Pi for her endless love and bearing with my busy schedule.

Liming Shi
Aalborg University, November 1, 2019

Part I

Introduction

Introduction

1 Background

Parkinson's disease (PD) is one of most common neurological disorders, and it affects millions of people around the world. The symptoms include tremor, rigidity and loss of muscle control. Although there is no cure, it is important to develop reliable monitoring tools for clinical intervention and medication treatment to alleviate the symptoms as early as possible. It was shown in early research that 90% of Parkinson patients suffer from vocal impairment [1, 2]. More recent studies showed that PD can be tele-monitored by noninvasive speech analysis of sustained phonation of vowels [3–8]. Using speech signals for the diagnosis of PD is promising and attractive because of the saving of time and travel cost for people to visit clinics physically. The block diagram of a general voice-based Parkinson's disease detection system can be found in Fig. 1. In a general voice-based PD detection system, the voice signal from a speaker is first recorded. Based on the recorded voice signal, features are extracted and fed to a trained classifier. Finally, a decision is made based on the classifier's output. For example, in Fig. 1, if the output of the classifier is one/zero, the speaker will be diagnosed with/without the Parkinson's disease. In [3], a variety of speech features extracted from relatively clean and sustained /a/ phonations have been tested for PD detection with a kernel support vector machine (KSVM). The results show that the overall correct classification performance can reach to 91.4%. However, one problem is that, in large-scale clinical application, we may only have access to the telephone quality speech signal, where the additive noise [9–11], reverberation [12, 13] and nonlinear distortion [14] severely affect the quality of speech signal and the classification performance of the PD detection system. Another problem is that some traditional feature extraction algorithms are based on assumptions that may not be valid for the sound signals that are used in the PD detection application, such as white Gaussian excitation assumption for linear prediction coding for the sustained /a/ sound [15]. Next, we first present some of the existing speech features used for PD detection in Subsection 1.1, briefly describe the speech degradations happened

during voice recording stage in Subsection 1.2, and present the objectives and structures of this thesis in Subsection 1.3.

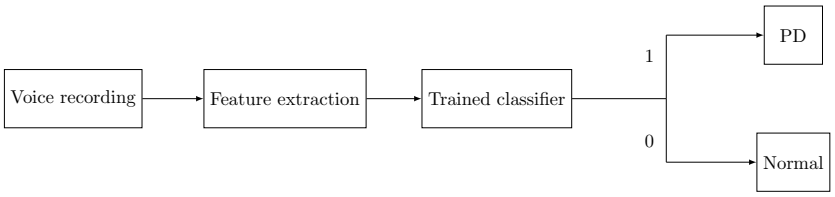


Fig. 1: The block diagram of a general voice-based Parkinson's disease (PD) detection system.

1.1 Speech features

As one of the main components of a PD detection system shown in Fig. 1, the feature extraction converts the raw sound signal into some selected speech features. The speech features to be used and the performance of the feature extraction algorithms heavily influence the performance of the classifier [3]. In this subsection, we focus on some of the physical and physiological meaningful speech features used for the diagnosis of PD. For patients with PD, the voice production becomes affected due to the deterioration of neurological control of the muscles. Therefore, differences in physiology and pathology phonation, such as the vocal fold vibration pattern, vocal tremor, placement of articulators etc., between PD and healthy people can be expected. There have been extensive studies of speech features for PD [3, 16, 17]. These features can be broadly categorized as vocal fold vibration-related and vocal tract-related features.

Vocal fold vibration-related features

Voiced sound is quasi-periodic and can thus be accurately modeled as a sum of harmonics (harmonic decomposition) with frequencies related to the fundamental frequency (a.k.a., pitch), which can be viewed as the vocal fold vibration frequency [18, 19]. A time domain signal from a segment of sustained /a/ and its power spectrum are shown in Fig. 2. As can be seen, the signal is quasi-periodic with pitch around 69 Hz. As explained in [17], many dysphonia features are defined based on the pitch information. For example, the feature that quantifies the pitch perturbation is referred to as jitter, which can be computed using the pitch contour [20, 21]. The pitch period entropy [3] can be seen as a feature that takes into account both the healthy, smooth vibrato and microtremor, and the logarithmic nature of speech pitch in speech production. The spectrograms and pitch estimates of sustained /a/

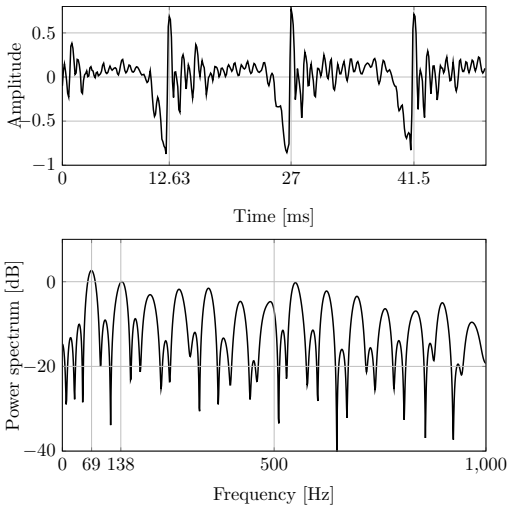


Fig. 2: Time-domain signal and power spectrum from a segment of a sustained /a/ signal with a pitch around 69 Hz.

from a speaker with PD and a healthy speaker are shown in Fig. 3. Using F_n to denote a pitch estimate at the n^{th} time instant, the jitter values, defined as [21]

$$\text{jitter} = \frac{1}{N-1} \sum_{n=1}^{N-1} |F_n - F_{n+1}| (F_n / F_{n+1})|, \tag{1}$$

for the sustained /a/ in Fig. 3 from the speaker with PD and the healthy speaker are 42×10^{-3} and 2.4×10^{-3}, respectively. As can be seen, in this example, the patient with PD has a larger jitter value than the healthy speaker. Another vocal fold vibration-related feature is the harmonic to noise ratio [22], a degree of hoarseness measure. It is computed as the ratio between the energy of the tonal component and the noise.

Vocal tract-related features

The vocal tract consists of the nose, mouth, tongue and lips. The interaction between the vocal folds and the vocal tract is often referred to as the source-filter coupling in phonation [23]. The speech signals are considered to be the result of a linear convolution of the vocal fold signal with the impulse response of the vocal tract. The spectral peaks of the vocal tract are usually referred to as the formants. Popular models for the vocal tract impulse response are the autoregressive (AR) model [24, 25] and the autoregressive moving average (ARMA) model [26]. The performance of using estimated AR coefficients for PD detection has been investigated in [27]. Another vocal

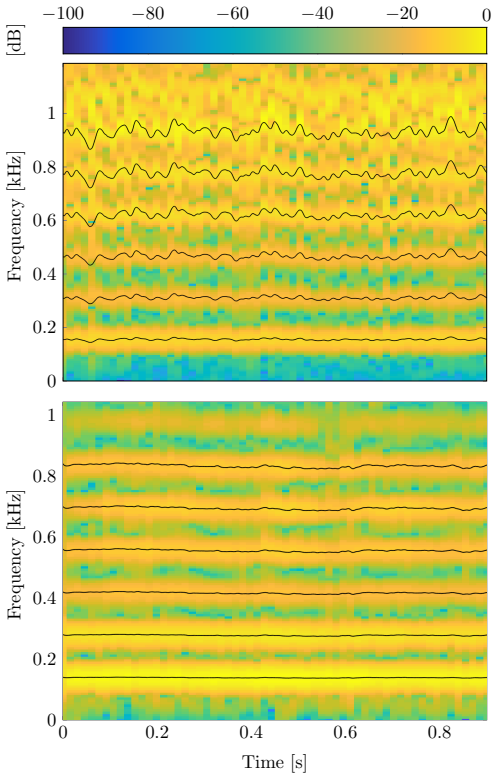


Fig. 3: Spectrograms and pitch estimates of sustained /a/ from a speaker with PD and a healthy speaker.

tract-related feature is the Mel-Frequency Cepstral Coefficients (MFCCs), reference standard feature for speaker identification [28] and automatic speech recognition [29]. They are aimed at detecting subtle changes in the motion of the articulators (tongue, lips) [30]. MFCCs have also been used for PD detection [27, 30].

1.2 Speech degradations

The voice recording, another main component of a PD detection system shown in Fig. 1, is a process of recording sound signals using a single or multiple microphones. In this thesis, we only focus on the single microphone case. As we mentioned earlier, when the speech signal is recorded remotely, it may suffer from different types of degradations, such as additive noise, reverberation and nonlinear distortion. Next, we explain these types of degradations in detail:

Additive noise

Additive background noise is one of the most common types of degradations in the process of speech acquisition. The background noise may be produced by interfering speakers as in a cocktail party scenario (a.k.a., babble noise), or is simply due to the presence of traffic noise, wind noise or home appliance noise etc. The acoustic noise encountered in real life usually has non-stationary and non-Gaussian properties. The speech intelligibility [9] and the performance of speech recognizer [31] may be reduced due to the presence of the additive noise. The spectrogram of the degraded sustained /a/ signal from a speaker with PD (same signal as in Fig. 3) in 10 dB babble noise is shown in Fig. 4. It can be seen that the noise masks the pitch of the target speech.

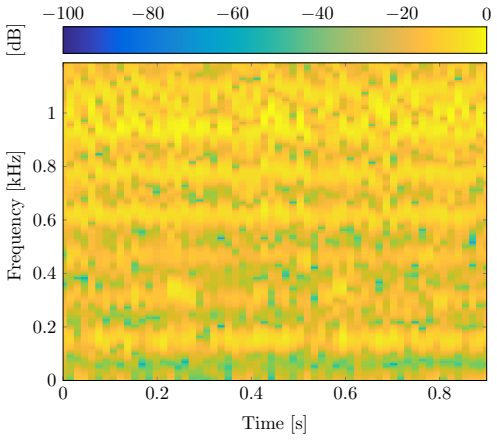


Fig. 4: Spectrogram of babble noise corrupted sustained /a/ from a speaker with PD (same signal as in Fig. 3) with SNR= 10 dB.

Reverberation

Reverberation is the process of multi-path propagation and occurs when the acoustic signals are recorded in an acoustically enclosed space. In such scenarios, except for the sound signal that travels directly from the speaker to the listener, the listener may receive multiple delayed and attenuated versions of the sound signal. The attenuation is due to the absorption characteristic of the walls or furniture. It is shown in [32] that reverberation has a detrimental effect on listeners' ability to perceptually separate voices with normally intonated pitch contours. Moreover, reverberation also disrupts listeners' ability to exploit differences in the spatial location of competing voices/sounds. A commonly used metric to measure the reverberation is the reverberation time

(RT60) [12]. RT60 is the amount of time for the reflected sound signal to undergo a decay of 60 dB [33, 34]. Fig. 5 shows the spectrogram of the sustained /a/ signal used in Fig. 3 and Fig. 4 with RT60 $= 0.5$ s. It can be seen from the figure that the pitch contour is smeared in the sense that the harmonic series is less clearly defined [32].

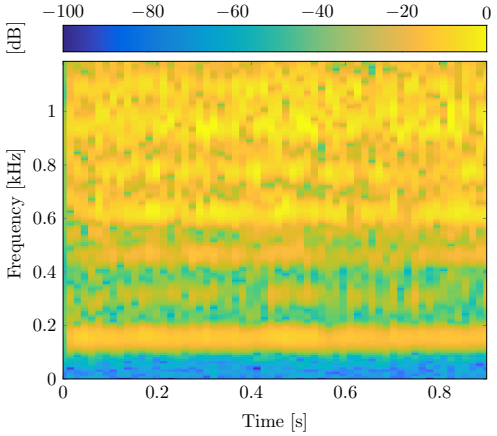


Fig. 5: Spectrogram of reverberated sustained /a/ from a speaker with PD (same signal as in Fig. 3) with RT60 $= 0.5$ s.

Nonlinear distortion

Aside from degradations due to the additive noise and reverberation, in the process of speech acquisition, the analogue-to-digital (AD) converter and speech coding schemes used for speech transmission may also introduce degradation, which is usually nonlinear. One example of nonlinear distortion for speech signal is peak clipping, which happens when the amplitude of a speech signal exceeds the dynamic range of the AD converter. For more details on different types of nonlinear distortion and their effect on sound quality and features, we refer to [14, 35].

1.3 Objectives and structures

Due to these speech degradations described in Subsection 1.2, the performance of some feature extraction algorithms may be reduced, leading to a low classification accuracy. There are at least three types of methods to deal with these degradations. The first type is using feature extraction algorithms that are robust against these degradations [36–39]. This type of methods is usually designed based on the individual feature of interest and its characteristics. And thus, different robust feature extraction algorithms should be

used to obtain different features. The second type is using speech cleaning up methods, such as speech enhancement, on the degraded signal [40, 41] prior to extracting features. One advantage of using this type of methods is that one single speech cleaning method can be used for obtaining a variety of features. However, some of the features may be altered significantly due to the speech cleaning up stage. The third type is data-driven approaches using deep neural network [42, 43]. The advantage of using deep neural network-based approaches is that the neural network can be trained for different kinds of distortions. However, to obtain a good performance, a large number of data is required. Moreover, to obtain results under a different parameter setting, the deep neural network has to be re-trained, which is usually computational complex and time-consuming. In order to keep the feature characteristics and have a low computational complexity, in this thesis, we focus on deriving robust feature extraction algorithms for PD detection, especially against the additive noise degradation described in Subsection 1.2. Two types of features, i.e., pitch, AR and ARMA coefficients (described in Subsection 1.1), will be considered. The motivation for focusing on these two types of features is that they have physical and physiologically meanings as we discussed earlier. Also, it is shown in [3] that a significant classification performance increase for PD detection can be gained using the pitch period entropy, which requires the information of pitch. Many algorithms have been proposed for pitch, AR and ARMA coefficients estimation [18, 26, 44–55]. However, in state-of-the-art parametric model-based segment-wise pitch estimation work [18, 48, 49], the noise signal is usually assumed to be white Gaussian, which is not realistic in real life applications. Moreover, the small variations of the pitches for patients with PD may not be captured well using segment-wise pitch estimation methods. Furthermore, the temporal information of the pitches, harmonic orders and voicing states has not been exploited jointly in pitch estimation algorithms. In state-of-the-art AR and ARMA coefficients estimation work, the vocal fold excitation signal is usually assumed to be white Gaussian (except [52, 53]), which is not valid for the sustained voiced signal used for PD detection. Therefore, using the prior information, such as the non-white or non-Gaussian properties of the noise signal and the vocal fold excitation signal, the temporal smoothness and the non-stationary characteristics of the pitches, robust feature extraction algorithms for the diagnosis of PD application may be obtained. Apart from this, we also intend to investigate on the performance of speech cleaning up methods for PD detection in adverse environments.

We first explain the mathematical models based on the pitch, AR and ARMA coefficients in Section 2. These models will be used across the entire thesis. An overview of the state-of-the-art pitch estimation algorithms, AR and ARMA coefficients estimation algorithms will be given in Section 3 and Section 4, respectively. In Section 5, we briefly give an overview of some

speech clean up methods.

2 Speech modeling

The famous quote "all models are wrong, but some are useful" [56] can also be applied to speech modeling. The speech signal is essentially a non-stationary and non-Gaussian process. No ground truth model is available for modeling the speech signal. Therefore, any speech model is a simplification of reality. However, speech modeling can be useful in many perspectives. For example, some speech models explicitly contain features that we are interested in as parameters, such as the pitch in the harmonic model [57]. Some speech models can be used for reducing the dimensions, and thus lowering the bit rate, such as the source-filter model [58, 59]. In this section, we present some popular used models for speech signals, and highlight the harmonic model in Subsection 2.1, and source-filter model using AR and ARMA processes in Subsection 2.2.

Consider the following general signal observation model:

$$y_n = s_n + v_n, \quad n \in \mathcal{Z}, \tag{2}$$

where $\mathcal{Z}$ denotes the set of integers, and y_n, s_n and v_n denote the observation signal, speech signal and noise at the n^{th} time instance, respectively. Various number of signal models, both in time-domain and frequency domain, have been proposed for modeling the speech signal s_n. Time-domain models include the harmonic model, source-filter model etc. Frequency-domain models include the complex Gaussian model with zero-mean and time-varying variance [60–62], supergaussian model [63] etc.

2.1 Harmonic model

As we mentioned earlier, a voiced speech signal, such as the sustained /a/, is quasi-periodic for the sample range $1 \leq n \leq N$. The most commonly used signal model for a periodic signal is the harmonic model, i.e.,

$$s_n = \sum_{k=1}^{K} (a_k \cos(k\omega n) + b_k \sin(k\omega n)), \tag{3}$$

where a_k and b_k are the weights of the k^{th} harmonic, ω is the pitch, K is the total number of harmonics. The advantage of using the harmonic model (3) is that the voiced speech signal is explicitly expressed as a function of the pitch ω, which makes this model convenient for fundamental frequency estimation. However, the underlying assumption of the model (3) is that both the weights and fundamental frequency are constant for N samples. Due to the

non-stationarity characteristic of speech signals, we would like the segment size N to be as small as possible. On the other hand, for periodic signals, the more cycles we have (a larger N), the more accurate fundamental frequency estimate in a segment we can obtain [64]. To resolve this contradiction, instead of using a fixed ω, the harmonic chirp model, in which the fundamental frequency is either a linearly increasing or decreasing function w.r.t. n, has been introduced [65–67]. Another variant of the harmonic model (3) is the quasi-harmonic model, where both the linear weights and fundamental frequency are modulated [68].

2.2 Source-filter model

The harmonic model presented in Subsection 2.1 is only useful for modeling voiced speech signal, but not for unvoiced speech signal. In the case of the voiced speech, the air forced out of the lungs travels through periodically vibrating vocal folds to form the excitation signal that is periodic in nature [69], whereas the excitation signal in the case of unvoiced speech is more noise like. A popular used model for both cases are the source-filter model, which is commonly used for speech reproduction [59, 70]. In this model, a linear filter is used to model the properties of the vocal tract and the spectral shaping of the vocal source. The excitation signal to the filter is set to either white noise or pulse train with delta functions located at pitch period intervals based on whether the sound is voiced or not. Both all-pole filters and pole-zero filters have been proposed as speech production model [54, 58, 71, 72]. For the all-pole speech production filter model, a sample of speech can be written as the AR process, i.e.,

$$s_n = -\sum_{k=1}^{K} a_k s_{n-k} + e_n, \tag{4}$$

where a_k denotes the linear prediction coefficient, K is the linear prediction order, e_n denotes the excitation signal. In contrast, for the pole-zero speech production filter model, a sample of speech signal can be written as the following ARMA process:

$$s_n = -\sum_{k=1}^{K} a_k s_{n-k} + \sum_{l=0}^{L} b_l e_{n-l}, \tag{5}$$

where a_k and b_k denotes the coefficients of the pole-zero model, K and L denote the order. Clearly, if we set $L = 0$ and $b_0 = 1$, (5) will reduce to (4). Therefore, the AR model can be seen as a special case of the ARMA model. The advantage of using the more complicated ARMA model than the AR model is that the nasals (e.g., /m/ and /n/), fricatives, or laterals sounds contain zeros on the spectrum, which are easier to be fitted with the ARMA model [54, 71, 72].

3 Pitch estimation

As noted before, pitch is one of the features related to vocal fold vibrations. The pitch information can help improve the performance of a voice-based PD detection system significantly [3]. Therefore, deriving robust pitch estimation algorithms for degraded speech signals is important for large-scale clinical trials of a practical voice-based PD detection system. In this section, we present an overview of the state-of-the-art pitch estimation algorithms. A widely used assumption in common pitch estimation methods is that pitch is an unknown but determined value over a fixed signal length signal (e.g., 15-40 ms for running speech [64, 73, 74]). A variety of segment-by-segment pitch estimation approaches have been proposed. These methods can be broadly categorized as non-parametric and parametric approaches. Non-parametric pitch estimation approaches are usually computational cheap but they suffer from performance degradation in low SNR scenarios. Examples of non-parametric approaches include the Kalman-based approach [75], YIN [44], Kaldi [45], SWIPE [46] and PEFAC [47]. On the other hand, parametric methods (e.g., harmonic model-based pitch estimators [18, 48, 49]) are more robust against noise [64, 76], but usually more computationally expensive [48]. Next, we briefly explain some of the non-parametric pitch estimation algorithms in Subsection 3.1, and the harmonic model-based pitch estimation algorithm in Subsection 3.2.

3.1 Non-parametric pitch estimation

Non-parametric pitch estimation approaches are based on the correlation method. Fig. 6 shows the autocorrelation function of the signal from Fig. 2. If the pitch for speech signal is constrained to the range 50 to 500 Hz (i.e., 2 to 20 ms in time), the pitch value can be determined as the largest peak index in the range 2 to 20 ms, that is 14.5 ms (i.e., 69 Hz). The correlation-based pitch estimation methods gained popularity due to the low computational complexity. In this subsection, we briefly describe the non-parametric pitch estimators YIN [44], SWIPE [46] and PEFAC [47], which are used as comparison algorithms for the proposed methods in this thesis.

YIN [44]: A cumulative mean normalized difference function, that is less sensitive to amplitude change than the autocorrelation function, is used as the cost function.

SWIPE [46]: A correlation-based frequency-domain pitch estimator. The normalized inner product between the spectrum of the input signal and a modified cosine is used as the cost function.

PEFAC [47]: The PEFAC estimates the pitch by convolving the power spectrum in the log-frequency domain with a filter that sums the energy of the pitch harmonics.

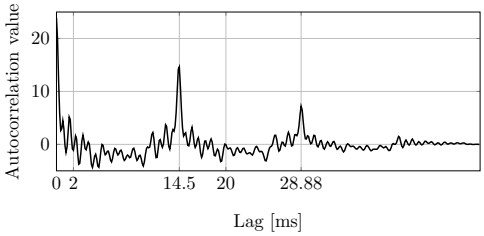


Fig. 6: Autocorrelation function of a segment of the sustained /a/

3.2 Harmonic model-based pitch estimation

In this subsection, we present the harmonic model-based pitch estimator using nonlinear least squares (NLS), which is directly related to the proposed methods in this thesis. Collecting N observation samples into a vector and writing (2) and (3) in matrix form yields

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{v}_n, \tag{6}$$

where $\mathbf{y}_n = [y_1, y_2, \cdots, y_N]^T$ and $\mathbf{v}_n$ is defined in the same way as $\mathbf{y}_n$, $\mathbf{a} = [a_1, b_1, a_2, b_2, \cdots, a_K, b_K]^T$, and $\mathbf{Z} = [\mathbf{z}(1), \mathbf{c}(1), \mathbf{z}(2), \mathbf{c}(2), \cdots, \mathbf{z}(K), \mathbf{c}(K)]$ with $\mathbf{z}(k)$ and $\mathbf{c}(k)$ defined as $\mathbf{z}(k) = [\cos(k\omega), \cos(k\omega 2), \cdots, \cos(k\omega N)]^T$ and $\mathbf{c}(k) = [\sin(k\omega), \sin(k\omega 2), \cdots, \sin(k\omega N)]^T$, respectively. As can be seen from (6), the weight vector $\mathbf{a}$ is a collection of linear parameters, while the pitch ω is a nonlinear parameter. With fixed ω and K and assuming the noise is a white Gaussian process, the maximum likelihood (ML) estimate of $\mathbf{a}$ can be written as $\hat{\mathbf{a}} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}$. Replacing $\mathbf{a}$ in (6) with the ML estimate $\hat{\mathbf{a}}$, the ML estimator of the pitch is

$$\begin{aligned}\hat{\omega} =& \arg\min_{\omega} \left\| \mathbf{y} - \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y} \right\|_2^2 \\ =& \arg\max_{\omega} \mathbf{y}^T\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}.\end{aligned} \tag{7}$$

Because the cost function in (7) is not convex w.r.t. ω, no closed form solution for ω can be obtained. We resort to the grid search method, where we uniformly discretize the pitch $\omega \in \{\omega^f, 1 \le f \le F\}$ over the range $[\omega_{\min}, \omega_{\max}]$, where $\omega_{\min}$ and $\omega_{\max}$ denote the lowest and highest pitches in the searching space, respectively. Prior information can be used to set $\omega_{\min}$ and $\omega_{\max}$. For example, pitch is usually between 50 to 500 Hz for speech signals. The harmonic order K can be estimated using model selection criteria, such as the Akaike information criterion or the Bayesian information criterion [77, 78]. However, due to the matrix inversion operation in (7), when we have a fine pitch grid and the maximum allowed number of harmonics is high (e.g.,

$K = 20$), the naïve implementation of the above NLS pitch estimator is computational complex. Recently, the authors in [79] found that the linear parameters $\hat{\mathbf{a}}$ can be computed recursively w.r.t. the harmonic order K. Fig. 7 shows the cost function (7) of the signal from Fig. 2 using the fast NLS algorithm in [79]. As can be seen, the cost function has the largest value around 69 Hz. In order to reduce the number of large errors in the harmonic model-based pitch estimators, smoothness prior on the pitch is imposed in [80] by using first-order Markov chain.

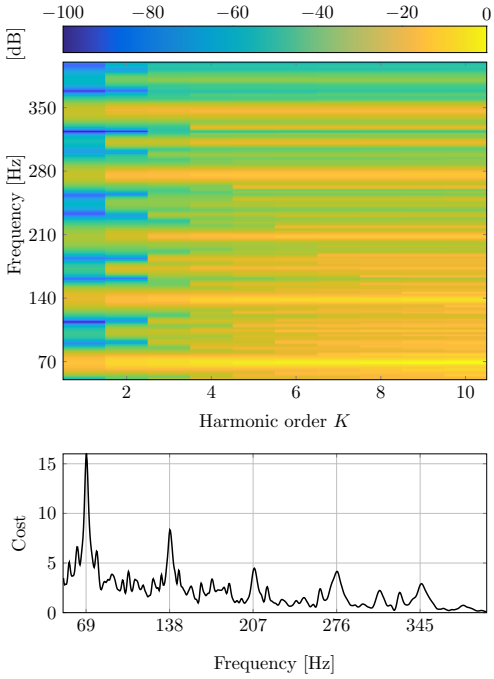


Fig. 7: The NLS cost w.r.t. (ω, K) for of a segment of the sustained /a/, and the cost when $K = 10$

4 AR and ARMA coefficients estimation

Except for extracting the vocal fold-related feature, i.e., pitch, we also investigated on obtaining the vocal tract-related feature, that is the AR and ARMA coefficients using the source-filter modeling of speech (see Subsection 2.2). It is a common practice to assume a white Gaussian noise excitation for the AR model (4) and ARMA model (5) due to the mathematical tractability and simplicity of the resulting algorithms [26]. Using the white Gaussian noise assumption on the excitation signal and the maximum likelihood estimation method, the likelihood functions w.r.t. the AR and ARMA coefficients can be

written as least squares (LS) cost functions. The coefficients estimation algorithms using the least squares cost function perform well for unvoiced speech and voiced speech with a low pitch. However, for high pitch voiced speech, using the least squares method for AR coefficients estimation leads biased estimates [50, 51]. To mitigate this problem, an AR modeling approach with a distortionless response at frequencies of harmonics is proposed in [50]. Another AR modeling method using regularisation to penalize rapid changes in the spectral envelope is proposed in [81]. In [52], a short-time energy weighted AR modeling approach is presented. In fact, as we mentioned earlier, for the voiced speech, the excitation signal is quasi-periodic and can be better modeled as impulse train instead of white Gaussian noise. Motivated by this, a least 1-norm criterion based AR coefficients estimator is proposed for voiced speech analysis [53], where sparse prior on the excitation signal is imposed by using the Laplace distribution. Next, we present the AR coefficients estimation methods using both the least squares and least 1-norm criterion in Subsection 4.1, and the ARMA coefficients estimation method using the least squares cost function in Subsection 4.2.

4.1 AR coefficients estimation

In this subsection, we review the least squares and least 1-norm AR coefficients estimation methods which are directly related to the proposed methods in this thesis. Using (4) and collecting the samples from N_1 to N_2, the speech signal can be written in the following matrix form:

$$\mathbf{s} = -\mathbf{S}\mathbf{a} + \mathbf{e}, \tag{8}$$

where $\mathbf{a} = [a_1, \cdots, a_K]^T$, $\mathbf{s} = [s(N_1), s(N_1+1), s(N_2)]^T$, $\mathbf{e}$ is defined similarly to $\mathbf{s}$, and

$$\mathbf{S} = \begin{bmatrix} s(N_1-1) & \cdots & s(N_1-K) \\ \vdots & & \vdots \\ s(N_2-1) & \cdots & s(N_2-K) \end{bmatrix}.$$

Assuming the excitation time sequence e_n is an i.i.d. zero-mean Gaussian process, the maximum likelihood estimate for the AR coefficient vector $\mathbf{a}$ can be written as the following least squares form:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{s} + \mathbf{S}\mathbf{a}\|_2^2 . \tag{9}$$

The sample indices N_1 and N_2 can be chosen in various ways, leading to different approaches. For example, setting $N_1 = 1$ and $N_2 = N + K$ leads to the autocorrelation method, and setting $N_1 = K + 1$ and $N_2 = N$ leads to the covariance method [26]. Using he autocorrelation method, a stable all-pole filter (all the poles lie inside the unit circle) can be obtained. Once the

estimate AR coefficients are obtained, the excitation signal can be estimated using $\hat{\mathbf{e}} = \mathbf{s} + \mathbf{S}\hat{\mathbf{a}}$.

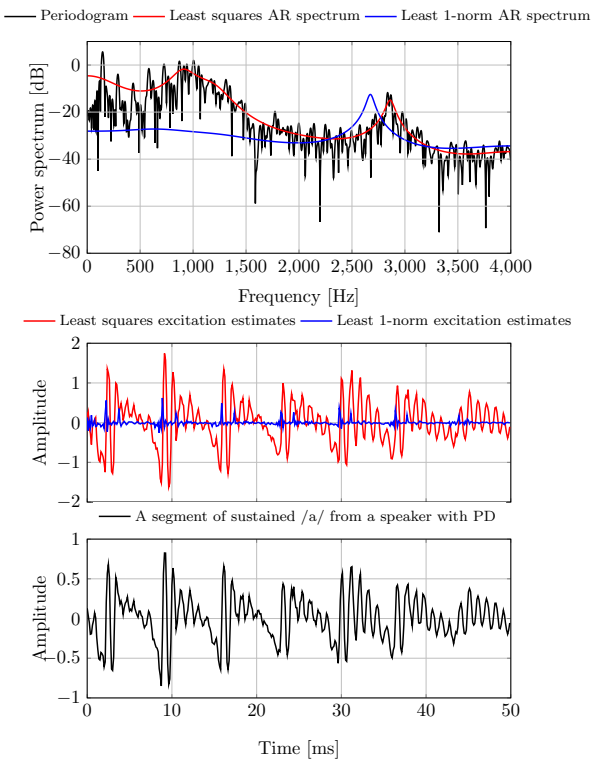


Fig. 8: AR spectrum and excitation estimates for a segment of sustained /a/ using the least squares and least 1-norm AR with order 12.

For voiced speech signal, the white Gaussian assumption on the excitation e_n may be violated. Considering the sparse characteristic of the excitation signal, the zero-mean Laplace distribution is used to model e_n [53], leading to the following least 1-norm form cost function:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{s} + \mathbf{S}\mathbf{a}\|_1 . \tag{10}$$

The estimated AR spectrum and excitation signal using the least squares and least 1-norm AR with order 12 based on the autocorrelation method for a segment of sustained /a/ from a speaker with PD is shown in Fig. 8. As can be seen from Fig. 8, the excitation estimates of the least 1-norm-based method is more sparse than the least squares-based approach. Moreover, the AR spectrum estimate of the least 1-norm-based method is less influenced by the harmonics than the least squares-based approach.

4.2 ARMA coefficients estimation

The ARMA modeling of speech has the advantage of fitting nasal sound [54, 55] easier than the AR modeling. Next, we present the two-stage least squares ARMA coefficients estimation method [26]. In the first stage, using the AR coefficients estimation method (9) with a sufficiently high-order K', we can obtain a coarse estimate of the excitation $\hat{\mathbf{e}} = \mathbf{s} + \mathbf{S}\hat{\mathbf{a}}$. In the second stage, replace $e(n), (K'+1 \leq n \leq N)$ in (5) by $\hat{e}(n)$ obtained in the first stage, and solve the following least squares problem:

$$\min_{\mathbf{z}} \left\| \mathbf{x}' + \mathbf{X}'\mathbf{z} \right\|_2^2 \tag{11}$$

where $\mathbf{x}' = [x\left(N_1'\right), x\left(N_1'+1\right) \cdots x\left(N_2'\right)]^T$, $\mathbf{X}' = [\bar{\mathbf{X}}, -\widehat{\mathbf{E}}]$ and

$$\bar{\mathbf{X}} = \begin{bmatrix} x\left(N_1'-1\right) & \cdots & x\left(N_1'-K\right) \\ \vdots & & \vdots \\ x\left(N_2'-1\right) & \cdots & x\left(N_2'-K\right) \end{bmatrix}$$

$$\widehat{\mathbf{E}} = \begin{bmatrix} \hat{e}\left(N_1'-1\right) & \cdots & \hat{e}\left(N_1'-L\right) \\ \vdots & & \vdots \\ \hat{e}\left(N_2'-1\right) & \cdots & \hat{e}\left(N_2'-L\right) \end{bmatrix}$$

and $\mathbf{z} = [a_1 \cdots a_K, b_1 \cdots b_L]^T$. N_1' and N_2' are usually set to $K^{'} + L + 1$ and N, respectively.

5 Speech clean up methods

As we mentioned before, speech degradation due to additive noise, reverberation and nonlinear distortion leads to a reduction of the performance of some feature extraction algorithms and result in a low PD detection accuracy. Except for developing robust feature extraction algorithms, we can also clean up the speech signals and then extract features from the enhanced signals. In this thesis, we only focus on removing the effect of additive noise and reverberation, and leave removing the effect of nonlinear distortion for future study. Following [9, 12], we refer the techniques for removing additive noise and reverberation as speech enhancement and speech dereverberation, respectively. Many methods have been proposed to perform speech enhancement [9, 82–84] and speech dereveberation [12, 85–91]. The speech enhancement methods can be broadly categorised into unsupervised and supervised methods. Most of the unsupervised methods rely on Wiener filtering and noise spectrum estimation techniques. Some of the major classes of noise spectrum estimation algorithms are 1) Minimum statistics [92] 2) Minima controlled recursive averaging (MCRA) [93, 94] 3) Minimum mean square

error (MMSE) power spectral estimation [95]. For these noise spectrum estimators, the underlying assumption is that the speech signal is nonstationary, while the noise signal is relatively stationary. However, for the diagnosis of PD application, the signal of interest in this thesis, sustained /a/, is a stationary signal. Therefore, we may not obtain good noise reduction performance using these noise spectrum estimators. Some of popular classes of supervised speech enhancement methods are 1) Dictionary-based methods [96–100] 2) deep learning-based methods [101, 102]. Similarly, the speech dereverberation methods can also be broadly categorised into unsupervised and supervised methods. Unsupervised speech dereverberation methods include 1) Late reverberation spectral variance estimation [85] 2) Delayed linear prediction based methods (DLP) [86, 88] 3) Nonnegative matrix factorization (NMF)-based method [89]. Supervised speech dereveberation methods are mainly based on deep learning [90, 91]. Next, we mainly focus on supervised methods for speech clean up since these methods usually outperform unsupervised approaches.

Dictionary-based speech enhancement

We first explain the dictionary-based supervised speech enhancement method. This method requires the speech and noise dictionaries training and estimates of the weighting parameters. The enhanced signal is usually estimated using a Wiener filter. The performance of this method depends heavily on the amount of data used for training. Consider the following speech observation model:

$$\mathbf{y}_n \approx \mathbf{s}_n + \mathbf{m}_n, \tag{12}$$

where $\mathbf{y}_n = [y_{n,1}, \cdots, y_{n,F}]^T$, $\mathbf{s}_n$ and $\mathbf{m}_n$ are defined similarly to $\mathbf{y}_n$, $y_{n,f}$, $s_{n,f}$, and $m_{n,f}$ denote the magnitude of the Short Time Fourier Transform (STFT) of the observed noisy signal, clean speech and noise, respectively, the subscripts n ($1 \leq n \leq N$) and f ($1 \leq f \leq F$) are the time frame and frequency bin indices. There are two stages for dictionary-based speech enhancement: training stage and enhancement stage. In the training stage, the speech basis matrix $\overline{\mathbf{W}}$ and noise basis matrix $\ddot{\mathbf{W}}$ are trained separately with clean training speech and noise based on different clustering schemes [103–105]. In the enhancement stage, a spectral gain $\mathbf{g}_n^{\text{NMF}}$ is obtained based on Wiener filtering, i.e.,

$$\mathbf{g}_n^{\text{NMF}} = \frac{\overline{\mathbf{W}}\,\overline{\mathbf{h}}_n}{\overline{\mathbf{W}}\,\overline{\mathbf{h}}_n + \ddot{\mathbf{W}}\,\ddot{\mathbf{h}}_n}, \tag{13}$$

where the division is element-wise, $\overline{\mathbf{h}}_n$ and $\ddot{\mathbf{h}}_n$ denote the weighting vectors for the speech and noise dictionaries, respectively. Fig. 9 shows the spectrogram of the learned dictionary by performing Itakura-Saito distance based

NMF on the magnitude spectrums of a database of 130 sustained /a/ from speakers with PD [106] into 40 basis vectors. The sampling frequency is set to 16000 Hz, and the segment length is 1024. Several methods have been pro-

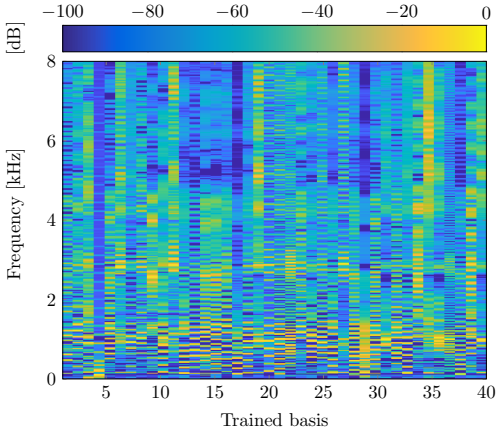


Fig. 9: The spectrogram of the learned dictionary using NMF for sustained /a/, the sampling frequency is 16000 Hz.

posed to improve the performance of conventional dictionary-based speech enhancement approaches by imposing sparsity prior [100, 107, 108] and temporal continuity prior [96, 109–111].

Deep learning-based speech enhancement or dereverberation

Deep learning-based speech enhancement or dereverberation methods formulate the speech clean up as a learning problem where the system uses the training data to learn discriminative patterns of clean speech signal. These algorithms can be divided into two components: 1) training targets that specifies the target that is to be achieved, such as the log magnitude spectrogram of the clean signal in [90], the time-frequency mask in [91] 3) acoustic features which are fed as an input in to the neural network, such as the log magnitude or power spectrogram in [90, 101], the amplitude modulation spectrogram and MFCC in [91].

6 Contributions

This thesis mainly deals with the robust feature extraction and investigates on speech clean up methods for diagnosis of Parkinson's disease. In this thesis, we have proposed both robust pitch estimation, sparse ARMA coefficients estimation and dictionary-based speech enhancement algorithms. The

main body of this thesis is constituted by papers A-H. The papers A and D deal with segment-wise pitch estimation algorithms whereas papers B and C deal with sample-wise pitch estimation algorithms. In the papers E and F, we have proposed ARMA coefficients estimation algorithms using sparsity prior. In paper H, we propose a supervised speech enhancement algorithm based on NMF and AR modeling of speech signal. In paper G, we show that the accuracy of the PD detector for degraded speech can be improved by using the dictionary-based speech enhancement and the deep learning based speech dereverberation algorithms.

Paper A The first paper in this thesis deals with robust segment-wise pitch tracking in noisy condition. A joint pitch, harmonic order and voicing state tracking algorithm based on the harmonic model and the maximum a posterior (MAP) criterion is proposed. Smoothness priors on the pitch, harmonic order and voicing state are imposed by using the first order Markov process models. The proposed pitch tracking algorithm is able to reduce the number of large pith errors. Moreover, combing with prewhitening approach, the proposed method is robust against different types of noise.

Paper B The segment-wise pitch tracking algorithm in Paper A was based on the assumption that the pitch and harmonic amplitudes are constant within a segment. This assumption may not always hold true and it was shown in [112], that the pitch may have nonstationary characteristic. In this paper, a sample-wise pitch tracking algorithm based on the time-varying harmonic model and the extended Kalman filter is proposed. Both the pitch and harmonic amplitudes are modeled as time-varying processes using the first Markov chains. Due to the nonlinearity characteristic of the observation equation, extended Kalman filter is applied. The pitch and amplitudes estimates using this approach have continuous and high temporal resolution properties.

Paper C The sample-wise pitch tracking algorithm proposed in Paper B assumes that the noise is white Gaussian, which may be invalid in real life applications. In this paper, we extend the method proposed in Paper B by further applying a time-varying AR model on the noise signal. The prior knowledge on the pitch range and stability of the AR model is also imposed. Due to the non-Gaussian characteristic of the state noise, the nonlinearity of the state and observation equations, the unscented Kalman filter and particle filters are proposed to obtain the pitch and amplitudes estimates. In addition to the good feature we described for the algorithm in Paper B, the proposed algorithm also shows high robustness against colored noise.

Paper D In this paper, a segment-wise multi-pitch estimation algorithm based on a pitch dictionary and the block sparsity prior is proposed. To reduce the pitch halving errors and counter the problem of unknown harmonic order, a intra-block clustering method is introduced. The estimator is based on the block sparse Bayesian learning. The proposed algorithm is robust against noise and has less pitch halving errors.

Paper E Apart from extracting the pitch feature from voiced speech signal based on the harmonic model, an algorithm for extracting the ARMA coefficients for voiced speech is proposed in this paper. Instead of assuming the excitation signal for the ARMA process is white Gaussian noise or impulse train, in the proposed method, the excitation signal is assumed to be a block sparse signal. This is motivated by quasi-periodic and temporal-correlated properties of the glottal flow signal. The variational expectation maximization approach is applied to obtain the estimates of the ARMA coefficients. As a result, the proposed estimator is less influenced by the vocal fold signal when estimating the ARMA coefficients in voiced speech scenarios.

Paper F In this paper, an ARMA coefficients estimation algorithm is proposed for voiced speech signal using the prior information that the excitation signal is sparse. A least 1-norm cost function is applied, instead of the traditional least squares cost function. The proposed method has the advantage of obtaining smoother spectral envelope and more sparse excitation estimates.

Paper G In this paper, the performance of sustained /a/-based Parkinson's disease (PD) detection system in additive noise and reverberation is investigated. Given that the specific degradation is known, we explore the effectiveness of a variety of speech enhancement and dereverberation algorithms in compensating these degradations and improving the PD detection accuracy.

Paper H In this paper, we propose a supervised speech enhancement method based on the NMF technique, and AR modeling of speech and noise. The dictionary items for both the speech and noise are parameterised by the AR coefficients. The NMF technique is used to estimate the excitation variance. The Wiener filtering approach is applied to obtain the enhanced speech signal. This approach is faster than the traditional AR dictionary-based speech enhancement method. Moreover, compared with the traditional NMF-based speech enhancement method, the dictionary size is reduced by using the AR coefficients.

7 Conclusion and directions for future research

The main outcome of this thesis was the proposal of robust estimation methods based on parametric modeling of speech signals. The models used here consisted of the harmonic model and its variants, the AR model and the ARMA model. The methods proposed in this thesis can be categorised into robust feature extraction and speech enhancement methods. We have shown by means of objective experiments, the robustness of using harmonic modeling of speech signal and AR modeling of colored noise for pitch extraction, the benefits of imposing the smooth prior to reduce the pitch halving errors. Moreover, in ARMA modeling of speech signal, we have shown by imposing the sparsity prior, the spectral distortion can be reduced. Furthermore, we have also proposed a supervised method for speech enhancement based on NMF and AR modeling of speech and noise. Apart from these, we have investigated on the effectiveness of a variety of speech enhancement and dereverberation algorithms in compensating the speech degradations and improving the PD detection accuracy. One area of further research would be to investigate the performance of the proposed robust feature extraction methods for PD detection. In paper E, we had investigated on how we can separate the vocal tract and vocal fold signal based on the ARMA model and the block sparse prior. We believe that, instead of using a fixed block size, clustered sparsity prior can help improve the vocal fold and vocal tract separation performance. Moreover, we would also like to mention the possibility of taking into account the temporal information while estimating the AR and ARMA coefficients. Furthermore, we would also like to derive fast frame-wise pitch tracking algorithm robust against colored noise using the concepts proposed in Paper A and C. Finally, we would like to remark that the computational complexity of these algorithms have to be analysed and further reduced.

References

[1] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[2] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[3] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkin-

son's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, April 2009.

[4] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, April 2010.

[5] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 350–367, 2011.

[6] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[7] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *J. Acoust. Soc. Am.*, vol. 139, no. 1, pp. 481–500, 2016.

[8] S. Arora, L. Baghai-Ravary, and A. Tsanas, "Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 2871–2884, 2019.

[9] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[10] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[11] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[12] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[13] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 7, pp. 1236–1246, 2018.

[14] C.-T. Tan, B. C. Moore, and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *Journal of the Audio Engineering Society*, vol. 51, no. 11, pp. 1012–1031, 2003.

[15] J. D. Markel and A. J. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013, vol. 12.

[16] D. A. Rahn III, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis," *Journal of Voice*, vol. 21, no. 1, pp. 64–71, 2007.

[17] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, Jan 2014.

[18] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[19] D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003.

[20] V. L. Heiberger and Y. Horii, "Jitter and shimmer in sustained phonation," in *Speech and language*. Elsevier, 1982, vol. 7, pp. 299–332.

[21] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth annual conference of the international speech communication association*, 2007.

[22] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[23] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. 1902–1915, 2008.

[24] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[25] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109–118, 1992.

[26] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005, vol. 452.

[27] A. Tsanas, "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning," Ph.D. dissertation, Oxford University, UK, 2012.

[28] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, 2005.

[29] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, 2012.

[30] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012.

[31] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 7398–7402.

[32] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2871–2876, 2003.

[33] M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[34] H. Kuttruff, *Room acoustics*. CRC Press, 2016.

[35] A. Poorjam, J. Jensen, M. Little, and M. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in *Proc. Interspeech*, 2017, pp. 289–293.

[36] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, no. 1-3, pp. 133–147, 1998.

[37] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621–624, 2005.

[38] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2010, pp. 4574–4577.

[39] ——, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, 2016.

[40] J. H. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," vol. 39, no. 4, pp. 795–805, 1991.

[41] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[42] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2263–2276, 2016.

[43] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 5220–5224.

[44] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[45] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[46] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[47] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[48] B. Quinn and P. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, 1991.

[49] J. Sward, H. Li, and A. Jakobsson, "Off-grid fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 2, pp. 296–303, Feb. 2018.

[50] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 221–239, 2000.

[51] T. Drugman and Y. Stylianou, "Fast inter-harmonic reconstruction for spectral envelope estimation in high-pitched voices," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1418–1422, 2014.

[52] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1295–1313, 2013.

[53] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, 2012.

[54] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 237–248, 2009.

[55] T. Kobayashi and S. Imai, "Design of iir digital filters with arbitrary log magnitude function by wls techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 247–252, 1990.

[56] G. E. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[57] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 1990, pp. 249–252.

[58] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[59] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. Elsevier Science Inc., 1995.

[60] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[61] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[62] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[63] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[64] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2042–2056, Oct 2013.

[65] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2014, pp. 1400–1404.

[66] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement and noise statistics estimation for non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 4, pp. 645–658, 2016.

[67] T. L. Jensen, J. K. Nielsen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "A fast algorithm for maximum-likelihood estimation of harmonic chirp parameters," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5137–5152, 2017.

[68] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM–FM signal decomposition with application to speech analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 2, pp. 290–300, 2010.

[69] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001, vol. 1.

[70] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 7. IEEE, 1982, pp. 614–617.

[71] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 229–234, 1977.

[72] K. Song and C. Un, "Pole-zero modeling of speech based on high-order pole model fitting and decomposition method," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1556–1565, 1983.

[73] K. Paliwal and K. Wójcicki, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Process. Lett.*, vol. 15, pp. 785–788, 2008.

[74] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2354–2367, 2016.

[75] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 102–105, 2012.

[76] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[77] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[78] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[79] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.

[80] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004.

[81] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 1, pp. 65–73, 2007.

[82] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Springer Science & Business Media, 2009, vol. 2.

[83] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[84] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[85] E. A. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, 2009.

[86] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[87] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[88] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[89] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2009, pp. 45–48.

[90] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[91] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[92] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[93] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[94] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[95] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[96] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 3, pp. 882–892, 2007.

[97] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 1, pp. 163–176, 2005.

[98] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2008, pp. 4029–4032.

[99] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[100] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[101] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2013.

[102] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[103] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[104] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 6, pp. 1698–1712, 2012.

[105] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 1, pp. 99–113, 2019.

[106] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–2901, 2014.

[107] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[108] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse nmf via bilevel optimization with applications to

speech enhancement," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 2014, pp. 11–15.

[109] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[110] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2011, pp. 17–20.

[111] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian nmf with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 4561–4564.

[112] H. Hajimolahoseini, R. Amirfattahi, H. Soltanian-Zadeh, and S. Gazor, "Instantaneous fundamental frequency estimation of non-stationary periodic signals using non-linear recursive filters," *IET Signal Processing*, vol. 9, no. 2, pp. 143–153, 2015.

Part II

Papers

Paper A

Robust Bayesian Pitch Tracking Based on the Harmonic Model

Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen,
Max A. Little and Mads Græsbøll Christensen

The paper has been published in the
IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019

Abstract

Fundamental frequency is one of the most important characteristics of speech and audio signals. Harmonic model-based fundamental frequency estimators offer a higher estimation accuracy and robustness against noise than the widely used autocorrelation-based methods. However, the traditional harmonic model-based estimators do not take the temporal smoothness of the fundamental frequency, the model order, and the voicing into account as they process each data segment independently. In this paper, a fully Bayesian fundamental frequency tracking algorithm based on the harmonic model and a first-order Markov process model is proposed. Smoothness priors are imposed on the fundamental frequencies, model orders, and voicing using first-order Markov process models. Using these Markov models, fundamental frequency estimation and voicing detection errors can be reduced. Using the harmonic model, the proposed fundamental frequency tracker has an improved robustness to noise. An analytical form of the likelihood function, which can be computed efficiently, is derived. Compared to the state-of-the-art neural network and non-parametric approaches, the proposed fundamental frequency tracking algorithm has superior performance in almost all investigated scenarios, especially in noisy conditions. For example, under 0 dB white Gaussian noise, the proposed algorithm reduces the mean absolute errors and gross errors by 15% and 20% on the Keele pitch database and 36% and 26% on sustained /a/ sounds from a database of Parkinson's disease voices. A MATLAB version of the proposed algorithm is made freely available for reproduction of the results[1].

1 Introduction

The problem of estimating the fundamental frequency or pitch information from noisy sound signals occurs in many applications, such as speech synthesis [1], voice disorder detection [2], and automatic speech recognition [3]. Fundamental frequency is a physical feature defined as the lowest frequency component of a periodic signal, while pitch is a perceptual feature, related to human listening [4]. Our objective is to estimate fundamental frequency. But, following [5, 6], we do not distinguish between fundamental frequency and pitch and use them interchangeably. Pitch is usually estimated using a segment of sound signals (a.k.a., frame) with a fixed segment length (e.g., 15-40 ms for speech signals [7–9]). Numerous pitch estimation algorithms have been proposed in the last fifty years, which can be categorized as unsupervised and supervised approaches. Unsupervised pitch estimation methods can be further categorized as non-parametric and parametric approaches. Examples of non-parametric approaches include the YIN [10], RAPT [11],

[1] An implementation of the proposed algorithm using MATLAB may be found in `https://tinyurl.com/yxn4a543`

SWIPE [12] and PEFAC [5] methods. YIN and RAPT compute autocorrelation functions from short frames of sound signals in the time domain. However, they are not robust against noise [13] and suffer from pitch octave errors (that is, a rational multiple of the true pitch) [3]. To reduce the pitch octave errors, SWIPE uses the cross-correlation function against a sawtooth signal combined with the spectrum of the signal, and exploits only the first and prime harmonics of the signal. PEFAC estimates the pitch in the log-frequency domain by convolving each frame's power spectrum with a filter that sums the energy of the pitch harmonics. Dynamic programming is used to obtain a smooth estimate of the pitch track. Due to the filtering and built-in spectral normalization methods, PEFAC is claimed to work in high levels of noise. However, a long frame length (e.g., 90.5 ms in PEFAC by default) is required to obtain good pitch estimation accuracy which is not practical in many real-time applications. More recently, a single frequency filtering approach based pitch estimation algorithm is proposed, which exploits the high SNR frequency component to overcome the effects of degradations in speech signal [14].

By contrast, parametric methods (e.g., harmonic model-based pitch estimators [6, 15, 16]) have also been proposed for pitch estimation. Compared with non-parametric approaches, harmonic model-based pitch estimators work with a short frame length (e.g., 20 ms), and show higher robustness to additive noise, fewer octave errors, and better time-frequency resolution [7, 17]. Recently, a computationally efficient pitch estimator based on a harmonic model has been proposed, which is referred to as the fast NLS [13]. However, one problem with most of the harmonic model based pitch estimators is that they do not take the temporal smoothness of the pitch, the harmonic order, and voicing into account as they process each frame independently. As a result, outliers, due to octave errors or voicing detection errors, occur. A sample-by-sample Kalman filtering-based pitch tracking algorithm using a time-varying harmonic model is proposed in [18] by assuming that the pitch and weights follow first-order Markov chains. A particle filtering-based pitch tracking algorithm based on the source-filter speech model combining with the harmonic modelling of input source is introduced in [19]. However, the good performance of the algorithms in [18] and [19] requires careful initializations. Moreover, it is difficult to integrate the time-varying model order into these algorithms, see [20] as an example of combing discrete and continuous state spaces. With either a known or estimated model order, a maximum a posteriori (MAP) pitch estimator based on the harmonic model has been developed to exploit the temporal dynamics of the pitch [21]. The model weights and observation noise variance are estimated by maximizing the maximum likelihood function (i.e., a frequentist perspective). Smooth pitch estimates are obtained, and thus the pitch octave errors are reduced. An additional voicing state is further considered in [22]

for estimating the pitch and obtaining the voiced-unvoiced decision jointly. However, the pitch tracking approach in [21] and [22] has many drawbacks. First, the assumption of a fixed harmonic order for multiple frames is not valid. In fact, in audio signals, the harmonic order often changes from frame to frame [9]. Second, matrix inversions are required to be stored for each candidate pitch to reduce the computational complexity. Third, errors can be found in transition frames where the voicing changes, because the past pitch information is not exploited when an unvoiced frame occurs. Finally, it is well-known that estimating parameters from a frequentist's perspective leads to over-fitting [23].

More recently, neural network based supervised pitch tracking algorithms were proposed [24–26], which show robustness against noise. The method proposed in [25] uses deep stacking network for joint speech separation and pitch estimation. The CREPE [26] discretises the pitch in logarithmic scale and uses a deep convolutional neural network to produce a pitch estimate. However, the unvoiced/silent state is not considered in the model. The maximum value of the output of the neural network is used as a heuristic estimate of the voicing probability. Moreover, to satisfy user's demand for different frequency resolution or frame length, the whole system is required to be retrained, which is usually time-consuming.

In this paper, we propose a fully Bayesian harmonic model-based pitch tracking approach. By using the harmonic model, as opposed to non-parametric methods, improved robustness against background noise and octave errors can be obtained. First-order Markov processes are used to capture the temporal dynamics of pitch, harmonic order, and voicing. By using information from previous frames, the rate of octave errors and the voicing detection errors can be further reduced. Compared to [21] and [22], we not only consider the temporal dynamics of pitch and voicing, but also of the harmonic order, which enables us to detect if any pitch is present, and estimate the pitch and harmonic order jointly and accurately. Moreover, past information on pitch is exploited to improve robustness against temporal voicing changes. Furthermore, by adopting a fully Bayesian approach to model weights and observation noise variances, the overfitting can be avoided. By assigning a proper transition pdf for the weights, fast NLS [13] can be easily incorporated into the proposed algorithm, leading to low computational and storage complexities.

The rest of the paper is organized as follows. In Section 2, we briefly review general Bayesian tracking theory. In Section 3 and Section 4, we present the proposed harmonic observation and state evolution models, respectively. In Section 5, the proposed pitch tracking algorithm is derived based on the harmonic observation and state evolution models. In Section 6, we briefly review the prewhitening step for dealing with non-Gaussian noise. Simulation results are given in Section 6.1, and the conclusions given in Section 7.

Notation: Boldface symbols in lowercase and uppercase letters denote column vectors and matrices, respectively.

2 Bayesian tracking

In this section, we briefly review Bayesian tracking in general, which forms the fundamental structure of the proposed pitch tracking algorithm. Consider the problem of estimating the state sequence $\{\mathbf{x}_n\}, 1 \leq n \leq N$ from noisy observations $\{\mathbf{y}_n\}, 1 \leq n \leq N$, related by

$$\mathbf{y}_n = h(\mathbf{x}_n, \mathbf{v}_n), \tag{A.1}$$

where $h(\cdot)$ denotes a mapping function between the state and observation vectors, $\mathbf{v}_n$ denotes an i.i.d. observation noise sequence, and n denotes the time index. The state sequence follows a first-order Markov process:

$$\mathbf{x}_n = f(\mathbf{x}_{n-1}, \mathbf{m}_n), \tag{A.2}$$

where $f(\cdot)$ denotes a mapping function between the current and previous states, and $\mathbf{m}_n$ denotes an i.i.d. state noise sequence. The elements in the state vector $\mathbf{x}_n$ can either be continuous or discrete. Assume that the posterior pdf $p(\mathbf{x}_{n-1}|\mathbf{Y}_{n-1})$ is available with the initial pdf being defined as $p(\mathbf{x}_0)$, where $\mathbf{Y}_{n-1}$ denotes a collection of observation vectors from the first observation vector up to the $(n-1)^{\text{th}}$ observation vector, i.e.,

$$\mathbf{Y}_{n-1} = [\mathbf{y}_1, \cdots, \mathbf{y}_{n-1}].$$

The objective of Bayesian tracking is to obtain a posterior distribution over the state vector $\mathbf{x}_n$ based on the current and previous observations recursively, i.e., $p(\mathbf{x}_n|\mathbf{Y}_n)$. The posterior $p(\mathbf{x}_n|\mathbf{Y}_n)$ can be obtained in two stages: predict and update.

In the prediction stage, we obtain the prediction pdf $p(\mathbf{x}_n|\mathbf{Y}_{n-1})$ by using the transition pdf $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ from (A.2), i.e.,

$$\begin{aligned} & p(\mathbf{x}_n|\mathbf{Y}_{n-1}) \\ = & \int p(\mathbf{x}_n|\mathbf{x}_{n-1}) p(\mathbf{x}_{n-1}|\mathbf{Y}_{n-1}) d\mathbf{x}_{n-1}, \quad 2 \leq n \leq N, \\ & p(\mathbf{x}_1) = \int p(\mathbf{x}_1|\mathbf{x}_0) p(\mathbf{x}_0) d\mathbf{x}_0, \quad n = 1, \end{aligned} \tag{A.3}$$

which is known as the Chapman-Kolmogorov equation. Note that if the elements in $\mathbf{x}_n$ are all discrete variables, the integration operator should be replaced with the summation operator.

In the update stage, combining (A.1) and the prediction pdf from the prediction stage, Bayes' rule can be applied to obtain the posterior, i.e.,

$$p(\mathbf{x}_n|\mathbf{Y}_n) = \frac{p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{Y}_{n-1})p(\mathbf{x}_n|\mathbf{Y}_{n-1})}{p(\mathbf{y}_n|\mathbf{Y}_{n-1})}, \quad 2 \leq n \leq N,$$
$$p(\mathbf{x}_1|\mathbf{Y}_1) = \frac{p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1)}{p(\mathbf{y}_1)}, \quad n = 1, \tag{A.4}$$

where $p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{Y}_{n-1})$ and $p(\mathbf{y}_1|\mathbf{x}_1)$ are the likelihood functions and $p(\mathbf{y}_n|\mathbf{Y}_{n-1})$ and $p(\mathbf{y}_1)$ are the normalization factors, respectively. Closed form solutions can be obtained for (A.3) and (A.4) in at least two cases. In the first case, when both $\mathbf{v}_n$ and $\mathbf{m}_n$ are drawn from Gaussian distributions with known parameters, and both $h(\mathbf{x}_n, \mathbf{v}_n)$ and $f(\mathbf{x}_{n-1}, \mathbf{m}_n)$ are linear functions over the variables, (A.3) and (A.4) reduce to the well-known Kalman-filter [23]. In the second case, when the state space is discrete and has a limited number of states, (A.3) and (A.4) reduce to the forward step of the forward-backward algorithm for hidden Markov model (HMM) inference [23]. In other cases, the inference of the posterior $p(\mathbf{x}_n|\mathbf{Y}_n)$ can be approximated using Monte Carlo approaches, such as particle filtering [27]. Next, we define the mapping function $h(\cdot)$ and formulate the observation equation (A.1) based on the harmonic model in Section 3, and then explain the state evolution model (A.2) for the proposed pitch tracking algorithm in Section 4.

3 Harmonic observation model

3.1 The harmonic observation model

Consider the general signal observation model given by

$$\mathbf{y}_n = \mathbf{s}_n + \mathbf{v}_n, \tag{A.5}$$

where the observation vector $\mathbf{y}_n$ is a collection of M samples from the n^{th} frame defined as

$$\mathbf{y}_n = [y_{n,1}, \cdots, y_{n,M}]^T,$$

the clean signal vector $\mathbf{s}_n$ and noise vector $\mathbf{v}_n$ are defined similarly to $\mathbf{y}_n$, M is the frame length in samples and n is the frame index. We assume that $\mathbf{v}_n$ is a multivariate white noise processes with zero mean and diagonal covariance matrix $\sigma_n^2\mathbf{I}$, σ_n^2 is the noise variance, $\mathbf{I}$ is the identity matrix. When voiced speech or music is present, we assume that the pitch, model weights and model order are constant over a short frame (typically 15 to 40 ms for speech signals and longer for music signals) and $s_{n,m}$ (i.e., the m^{th} element of $\mathbf{s}_n$)

follows the harmonic model, i.e.,

$$\mathrm{H}_1 : s_{n,m} = \sum_{k=1}^{K_n} \left[\alpha_{k,n}\cos(k\omega_n m) + \beta_{k,n}\sin(k\omega_n m)\right], \tag{A.6}$$

where $\alpha_{k,n}$ and $\beta_{k,n}$ are the linear weights of the k^{th} harmonic, $\omega_n = 2\pi f_n / f_\text{s}$ is the normalized digital radian frequency, f_s is the sampling rate, and K_n is the number of harmonics. When voiced speech/music is absent (unvoiced or silent), a null model is used, i.e.,

$$\mathrm{H}_0 : \mathbf{y}_n = \mathbf{v}_n. \tag{A.7}$$

Note that, based on the source-filtering model of speech generation, the unvoiced speech can be modelled as a coloured Gaussian process [28]. The observation noise in practice may have non-stationary and non-Gaussian properties, such as babble noise. However, we can deal with this by prewhitening the observation signals [9], which will be described in Section 6. Writing (C.6) in matrix form and combining (C.5) and (C.6) yields

$$\mathrm{H}_1 : \mathbf{y}_n = \mathbf{Z}(\omega_n, K_n)\mathbf{a}_{K_n} + \mathbf{v}_n, \tag{A.8}$$

where

$$\begin{aligned}
\mathbf{Z}(\omega_0, K_n) &= [\mathbf{c}(\omega_n), \cdots, \mathbf{c}(K_n\omega_n), \mathbf{d}(\omega_n), \cdots, \mathbf{d}(K_n\omega_n)], \\
\mathbf{c}(\omega_n) &= [\cos(\omega_n 1), \cdots, \cos(\omega_n M)]^T, \\
\mathbf{d}(\omega_n) &= [\sin(\omega_n 1), \cdots, \sin(\omega_n M)]^T, \\
\mathbf{a}_{K_n} &= [\alpha_{1,n}, \cdots, \alpha_{K_n,n}, \beta_{1,n}, \cdots, \beta_{K_n,n}]^T.
\end{aligned}$$

We can further write (A.7) and (A.8) together by introducing a binary voicing indicator variable u_n, i.e.,

$$\mathbf{y}_n = u_n\mathbf{Z}(\omega_n, K_n)\mathbf{a}_{K_n} + \mathbf{v}_n, \tag{A.9}$$

where $u_n \in \{0, 1\}$. When $u_n = 0$ and $u_n = 1$, (A.9) reduces to the unvoiced and voiced models (A.7) and (A.8), respectively.

We write the state vector as $\mathbf{x}_n = [\mathbf{a}_{K_n}, \sigma_n^2, \omega_n, K_n, u_n]^T$. Comparing (A.9) and (A.1), we can conclude that the mapping function $h(\cdot)$ is a nonlinear function w.r.t. the state vector $\mathbf{x}_n$. Moreover, the state vector $\mathbf{x}_n$ contains continuous variables $\mathbf{a}_{K_n}$, σ_n^2, ω_n and discrete variables K_n and u_n. However, due to the non-linear characteristics of (A.9) w.r.t. ω_n, uniform discretisation over the pitch ω_n is commonly used [13]. An off-grid estimate of ω_n can be obtained by pitch refinement algorithms, such as gradient descent [29]. Our target is to obtain estimates of the fundamental frequency ω_n, the harmonic order K_n, and the voicing indicator u_n, that is a subset of $\mathbf{x}_n$ defined as $\ddot{\mathbf{x}}_n = [\omega_n, K_n, u_n]^T$, from the noisy observation $\mathbf{y}_n$.

4 The state evolution model

In this section, we derive the state evolution model (A.2) or more generally the transition probability density/mass function (pdf/pmf) $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{Y}_{n-1})$ for continuous/discrete states of the proposed model. Following the fast NLS pitch estimation approach [13], we uniformly discretize the pitch $\omega_n \in \{\omega^f, 1 \leq f \leq F\}$ over the range $[\omega_{\min}, \omega_{\max}]$, where $\omega_{\min}$ and $\omega_{\max}$ denote the lowest and highest pitches in the searching space, respectively. Prior information can be used to set $\omega_{\min}$ and $\omega_{\max}$. For example, pitch is usually between 70 to 400 Hz for speech signals. The grid size is set to

$$\left\lfloor F\frac{\omega_{\max}}{2\pi} \right\rfloor - \left\lceil F\frac{\omega_{\min}}{2\pi} \right\rceil + 1,$$

where F denotes the DFT size for computing the likelihood function (see Section 5 and [13] for further details), $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the flooring and ceiling operators, respectively. It is also shown that the optimal choice of F depends on the frame length and the harmonic order [13]. However, for simplicity and fast implementation, in this paper, we set $F = 2^{14}$. The state space for the discrete variables can be expressed as $\{\mathcal{M}(n) : [\omega_n = \omega^f, K_n = k, u_n = 1]^T, 1 \leq f \leq F, 1 \leq k \leq K^{\max}\} \cup \{\mathcal{M}_0(n) : u_n = 0\}$. The prediction pdf $p(\mathbf{x}_n|\mathbf{Y}_{n-1})$ defined in (A.3) can be factorized as

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{Y}_{n-1}) =& p(\mathbf{a}_{K_n}|\sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \times \\ & p(\sigma_n^2|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) p(\ddot{\mathbf{x}}_n|\mathbf{Y}_{n-1}). \end{aligned} \tag{A.10}$$

We first explain the transition pdfs for the continuous variables σ_n^2 and $\mathbf{a}_{K_n}$, and then discuss the transition pmfs for the discrete variables ω_n, K_n and u_n. The selection of a state evolution model is a trade-off between being physically accurate and ending up with a practical solution.

4.1 Transition pdfs for the noise variance and weights

To obtain the prediction pdf for the noise variance $p(\sigma_n^2|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$, the transition pdf for the noise variance $p(\sigma_n^2|\sigma_{n-1}^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$ should be defined. A reasonable assumption for the noise variance is that it changes slowly from frame to frame. For example, the unknown parameter σ_n^2 can be assumed to evolve according to an inverse Gamma distribution [30], i.e.

$$p(\sigma_n^2|\sigma_{n-1}^2) = \mathcal{IG}(\sigma_n^2|c, d\sigma_{n-1}^2). \tag{A.11}$$

where $\mathcal{IG}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp(-\frac{\beta}{x})$ and $\Gamma(\cdot)$ denotes the gamma function. With this transition pdf, an analytical form of the posterior distribution on $\mathbf{x}_n$ cannot be derived. A sequential Monte Carlo approach can be

used to approximate the posterior numerically [31]. However, the major drawback of any Monte Carlo filtering strategy is that sampling in high-dimensional spaces can be inefficient [32]. A Rao-blackwellized particle filtering approach [33], which marginalises out some of the variables for statistical variance reduction, can be used to deal with this problem. However, we do not pursue this approach any further in this paper, and leave it for future work. Instead, for simplicity, we assume independence between σ_n^2 and σ_{n-1}^2, and use the Jeffery's prior, i.e.,

$$p(\sigma_n^2|\sigma_{n-1}^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \propto 1/\sigma_n^2, \sigma_n^2 > 0. \tag{A.12}$$

As can be seen, the Jeffery's prior (A.12) is a limiting case of (A.11) with $c \to 0$ and $d \to 0$.

Similarly, we define the transition pdf for the weights as $p(\mathbf{a}_{K_n}|\mathbf{a}_{K_{n-1}}, \sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$. Imposing smoothness dependency on the weight time evolution can reduce pitch octave errors [34]. However, in order to use the fast algorithm [13], we assume that the model weights between consecutive frames are conditionally independent given previous observations and the rest of unknown variables. Following [35], we use the hierarchical prior

$$\begin{aligned} &p(\mathbf{a}_{K_n}|\mathbf{a}_{K_{n-1}}, \sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}, g_n) \\ =&\mathcal{N}(\mathbf{a}_{K_n}|0, g_n\sigma_n^2 \left[(\mathbf{Z}(\omega_n, K_n)^T \mathbf{Z}(\omega_n, K_n)\right]^{-1}), \end{aligned} \tag{A.13}$$

$$p(g_n|\delta) = \frac{\delta - 2}{2}(1 + g_n)^{-\delta/2}, g > 0, \tag{A.14}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that the vector $\mathbf{x}$ has the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The prior distribution for the weights (A.13) is known as Zellner's g-prior [36]. As can be seen from (A.13), given ω_n and K_n, the prior covariance matrix is a scaled version of the Fisher information matrix. With Zellner's g-prior, a closed-form calculation of the marginal likelihood can be obtained [37]. Moreover, the fast algorithm in [13] for computing the marginal likelihood can be applied (see Section 5 for detail).

The graphical model for the proposed method is shown in Fig. A.1. Note that, instead of obtaining point estimates of the noise variance and weight parameters using maximum likelihood [21], a Bayesian approach is used to represent the full uncertainty over these parameters.

4.2 Transition pmfs for ω_n, K_n and u_n

In [21], to reduce octave errors, a first-order Markov model is used for the pitch evolution provided that the harmonic order is fixed and

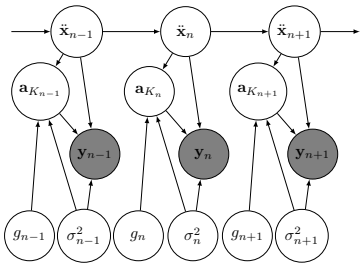


Fig. A.1: A graphical model of the proposed method with shaded nodes indicating observed variables.

known/estimated for multiple frames. Another voicing evolution model is further considered in [22] by imposing the so-called "hang-over" scheme [38]. Although in some cases, the harmonic order may not be of interest, it is still necessary to estimate it to obtain correct pitch estimates [17]. In fact, considering the temporal dynamics of the model order helps reducing the octave errors, which will be verified by the simulation results. Moreover, using priors for the model order is also necessary for model comparison [35]. In this paper, we propose to track the pitch ω_n, the harmonic order K_n and the voicing indicator u_n jointly. More specifically, we impose smoothness constraints on ω_n and K_n, and hang-over on voicing state using first-order Markov processes. The transition probability for the n^{th} frame to be voiced with pitch ω_n and harmonic order K_n when the previous frame is also voiced with ω_{n-1} and K_{n-1} can be expressed as

$$\begin{aligned} &p(\mathcal{M}(n)|\mathcal{M}(n-1)) \\ =&p(\omega_n, K_n|\omega_{n-1}, K_{n-1}, u_{n-1}=1, u_n=1)\times \\ &p(u_n=1|u_{n-1}=1). \end{aligned} \tag{A.15}$$

We assume that the pitch ω_n and harmonic order K_n evolve according to their own, independent dynamics given $u_n = 1$ and $u_{n-1} = 1$, i.e.,

$$\begin{aligned} &p(\omega_n, K_n|\omega_{n-1}, K_{n-1}, u_n=1, u_{n-1}=1) \\ =&p(\omega_n|\omega_{n-1}, u_n=1, u_{n-1}=1)\times \\ &p(K_n|K_{n-1}, u_n=1, u_{n-1}=1), \end{aligned} \tag{A.16}$$

which means when both time frame $n-1$ and n are voiced, the pitch and harmonic order only depend on their previous states. In fact, this assumption is only true when the product of the maximum allowed harmonic order and the pitch is less than half of the sampling frequency. However, by using a Bayesian approach, a model with a larger harmonic order is more penalized than with a smaller one. Even if a large value is used for the maximum

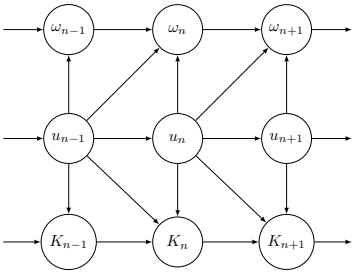


Fig. A.2: A graphical model specifying conditionally independence relations for the discrete variables.

allowed harmonic order in the proposed approach, the posterior model probability with a large harmonic order can be small [39]. In [40], an infinite number of harmonics is used, and the non-parametric prior distribution is used to penalize the models with large harmonic orders. By assuming the pitch and harmonic order are conditionally independent given $u_n = 1$ and $u_{n-1} = 1$, the Bayesian inference of the model posterior, shown in Section 5, can be simplified. The transition probability for the n^{th} frame to be voiced with pitch ω_n and harmonic order K_n when the previous frame is unvoiced/silent can be expressed as

$$\begin{aligned} &p(\mathcal{M}(n)|\mathcal{M}_0(n-1)) \\ =&p(\omega_n, K_n|u_n = 1, u_{n-1} = 0)p(u_n = 1|u_{n-1} = 0). \end{aligned} \tag{A.17}$$

The priors from an unvoiced frame to a voiced frame $p(\omega_n, K_n|u_n = 1, u_{n-1} = 0)$ are set to $p(\omega_m, K_m|\mathbf{Y}_m, u_m = 1)$, which can be calculated as

$$p(\omega_m, K_m|\mathbf{Y}_m, u_m = 1) = \frac{p(\omega_m, K_m, u_m = 1|\mathbf{Y}_m)}{1 - p(u_m = 0|\mathbf{Y}_m)}, \tag{A.18}$$

where m is the closest frame index to n that satisfies the constraint $p(u_m = 0|\mathbf{Y}_m) < 0.5$ (m^{th} frame is voiced). In fact, if the previous frame is not voiced, we exploit the information from the latest frame that is voiced as the prior for the pitches and harmonic orders. The motivation for this choice is that the pitch and harmonic order usually do not change abruptly after a short segment of unvoiced/silent frames. Using the past information as the prior, robustness against the voicing state changes can be improved. The graphical model for the evolution of $\ddot{\mathbf{x}}(n)$ is shown in Fig. A.2. Assuming the Markov processes are time-invariant, we can express the transition matrices for the pitch, harmonic order and voicing as $\mathbf{A}^{\omega}$, $\mathbf{A}^{K}$ and $\mathbf{A}^{u}$, respectively.

5 Pitch tracking

In this section, a joint pitch and harmonic order tracking, and voicing detection algorithm is derived based on the Bayesian tracking formulas (A.3) and (A.4). First, note that, by assuming that σ_n^2 and σ_{n-1}^2 are conditionally independent given $\ddot{\mathbf{x}}_n$ and $\mathbf{Y}_{n-1}$, and $\mathbf{a}_{K_n}$ and $\mathbf{a}_{K_{n-1}}$ are conditionally independent given σ_n^2, $\ddot{\mathbf{x}}_n$ and $\mathbf{Y}_{n-1}$, the prediction pdfs are equal to the transition pdfs, i.e.,

$$p(\sigma_n^2|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) = p(\sigma_n^2|\sigma_{n-1}^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}), \tag{A.19}$$

$$\begin{aligned} &p(\mathbf{a}_{K_n}|\sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \\ &= \int p(\mathbf{a}_{K_n}|\mathbf{a}_{K_{n-1}}, \sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}, g_n) p(g_n; \delta) dg_n. \end{aligned} \tag{A.20}$$

Based on (A.3), prediction pmfs for discrete variables $p(\ddot{\mathbf{x}}(n)|\mathbf{Y}_{n-1})$ can be expressed as

$$\begin{aligned} &p(\mathcal{M}(n)|\mathbf{Y}_{n-1}) \\ &= \sum_{\mathcal{M}(n-1)} p(\mathcal{M}(n)|\mathcal{M}(n-1)) p(\mathcal{M}(n-1)|\mathbf{Y}_{n-1}) + \\ &\quad p(\mathcal{M}(n)|\mathcal{M}_0(n-1)) p(\mathcal{M}_0(n-1)|\mathbf{Y}_{n-1}), \end{aligned} \tag{A.21}$$

$$\begin{aligned} &p(\mathcal{M}_0(n)|\mathbf{Y}_{n-1}) \\ &= \sum_{h=0}^{1} p(u_n = 0|u_{n-1} = h) p(u_{n-1} = h|\mathbf{Y}_{n-1}) \\ &= p(u_n = 0|u_{n-1} = 0) p(\mathcal{M}_0(n-1)|\mathbf{Y}_{n-1}) + \\ &\quad p(u_n = 0|u_{n-1} = 1)(1 - p(\mathcal{M}_0(n-1)|\mathbf{Y}_{n-1})). \end{aligned} \tag{A.22}$$

With the prediction pdfs and pmfs in hand, we can obtain the update equation based on (A.4). In order to obtain the posteriors for the pitch, harmonic order and voicing indicators, the weights and noise variance can be integrated out from the update equation, i.e.,

$$\begin{aligned} &p(\ddot{\mathbf{x}}_n|\mathbf{Y}_n) \\ &\propto \int p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{Y}_{n-1}) p(\mathbf{x}_n|\mathbf{Y}_{n-1}) d\mathbf{a}_{K_n} d\sigma_n^2 \\ &= p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) p(\ddot{\mathbf{x}}_n|\mathbf{Y}_{n-1}), \end{aligned} \tag{A.23}$$

where $p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$ denotes a marginal likelihood, defined as

$$\begin{aligned} p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) = \int p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{a}_{K_n}|\sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \times \\ p(\sigma_n^2|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) p(g_n; \delta) d\mathbf{a}_{K_n} d\sigma_n^2 dg_n. \end{aligned} \tag{A.24}$$

Using (A.9), (A.12), (A.13), (A.14), (A.19) and (A.20), a closed-form marginal likelihood can be obtained, i.e.,

$$\begin{aligned} & p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \\ = & \left[\frac{(\delta-2)}{2K_n+\delta-2}{}_2F_1\left[\frac{M}{2}, 1; \frac{2K_n+\delta}{2}; R^2(\omega_n, K_n)\right]\right]^{u_n} \times \\ & m_M(\mathbf{y}_n), \end{aligned} \tag{A.25}$$

where

$$m_M(\mathbf{y}_n) = \frac{\Gamma(\frac{M}{2})}{(\pi||\mathbf{y}_n||_2^2)^{\frac{M}{2}}}, \tag{A.26}$$

$$R^2(\omega_n, K_n) = \frac{\mathbf{y}_n^T\mathbf{Z}(\omega_n, K_n)\hat{\mathbf{a}}_{K_n}}{\mathbf{y}_n^T\mathbf{y}_n}, \tag{A.27}$$

$$\hat{\mathbf{a}}_{K_n} = (\mathbf{Z}(\omega_n, K_n)^T\mathbf{Z}(\omega_n, K_n))^{-1}\mathbf{Z}(\omega_n, K_n)\mathbf{y}_n, \tag{A.28}$$

$m_M(\mathbf{y}_n)$ denotes the null model likelihood (i.e., $p(\mathbf{y}_n|u_n = 0)$) and ${}_2F_1$ denotes the Gaussian hypergeometric function [41]. To compute $R^2(\omega_n, K_n)$ for all the candidate pitches and harmonic orders, the fast algorithm [13] can be applied. Moreover, from a computational point of view, a Laplace approximation of (A.24) can be derived as an alternative instead of marginalizing w.r.t. g_n analytically [35]. Note that, for the discrete vector $\ddot{\mathbf{x}}_n$, it should satisfy the normalisation constraint,

$$\begin{aligned} 1 = & \sum_{\ddot{\mathbf{x}}_n} p(\ddot{\mathbf{x}}_n|\mathbf{Y}_n) \\ = & p(\mathcal{M}_0(n)|\mathbf{Y}_n) + \sum_{\mathcal{M}(n)} p(\mathcal{M}(n)|\mathbf{Y}_n). \end{aligned} \tag{A.29}$$

Finally, estimates of the pitch and harmonic order and the voiced/unvoiced state can be jointly obtained using the maximum a posterior (MAP) estimator. More specifically, the n^{th} frame is labeled as voiced if $p(u_n = 0|\mathbf{Y}_n) < 0.5$, and the pitch and harmonic order are obtained as

$$(\hat{\omega}_n, \hat{K}_n) = \max_{\omega_n, K_n} p(\omega_n, K_n, u_n = 1|\mathbf{Y}_n). \tag{A.30}$$

The proposed Bayesian pitch tracking algorithm is shown in Algorithm 1. To make inferences, we need to specify the transition matrices for the pitch $p(\omega_n|\omega_{n-1}, u_n = 1, u_{n-1} = 1)$, the harmonic order $p(K_n|K_{n-1}, u_n = 1, u_{n-1} = 1)$ and $p(u_n|u_{n-1})$. Following [21], we set $p(\omega_n|\omega_{n-1}, u_n = 1, u_{n-1} = 1) = \mathcal{N}(\omega_n|\omega_{n-1}, \sigma_\omega^2)$. The transition probability for the model order is chosen as $p(K_n|K_{n-1}, u_n = 1, u_{n-1} = 1) = \mathcal{N}(K_n|K_{n-1}, \sigma_K^2)$. Smaller σ_ω^2 and σ_K^2 lead to smoother estimates of the pitch and harmonic order while larger values make

the inference less dependent on the previous estimates. The matrix $\mathbf{A}^u$ is controlled by $p(u_n = 1|u_{n-1} = 0)$ and $p(u_n = 0|u_{n-1} = 1)$. In order to reduce the false negative (wrongly classified as unvoiced when a frame is voiced) rate, we set $p(u_n = 1|u_{n-1} = 0) = 0.4$, $p(u_n = 0|u_{n-1} = 1) = 0.3$, respectively, that is, the transition probability from unvoiced to voiced is higher than from voiced to unvoiced. Note that each row of $\mathbf{A}^\omega$, $\mathbf{A}^K$, and $\mathbf{A}^u$ is normalised to ensure they are proper pmfs. By setting $\sigma_\omega^2 \to \infty$, $\sigma_K^2 \to \infty$, $p(u_n = 1|u_{n-1} = 0) = 0.5$ and $p(u_n = 0|u_{n-1} = 1) = 0.5$, the proposed algorithm reduces to the fast NLS algorithm [13]. Moreover, using (A.16), (A.18), and the definitions of $\mathbf{A}^\omega$, $\mathbf{A}^K$ and $\mathbf{A}^u$, an MAP estimator that maximizes the joint posterior $p(\ddot{\mathbf{x}}_1, \cdots, \ddot{\mathbf{x}}_N|\mathbf{Y}_N)$, instead of marginal posterior $p(\ddot{\mathbf{x}}_n|\mathbf{Y}_n)$ in (A.23), can also be derived, which is known as the Viterbi algorithm [23]. Although the Viterbi algorithm may help obtaining better pitch estimates by using future data, it has high storage complexity. In this paper, we only focus on the online pitch tracking in Algorithm 1.

Algorithm 1 The proposed Bayesian pitch tracking

1: Initiate the harmonic order $K^{\max}$, transition matrices $\mathbf{A}^\omega$, $\mathbf{A}^K$ and $\mathbf{A}^u$, and the initial probability $p(u_0|\mathbf{y}_0)$ and $p(\omega_0, K_0, u_0 = 1|\mathbf{y}_0)$ satisfying the constraint $p(u_0 = 0|\mathbf{y}_0) + \sum_{\omega_0,K_0} p(\omega_0, K_0, u_0 = 1|\mathbf{y}_0) = 1$
2: **for** $n = 1, 2, \cdots$ **do**
3: *Prediction step:*
4: Obtain $p(\mathcal{M}(n)|\mathbf{Y}_{n-1})$ based on (A.21), (A.15) and (A.17).
5: Obtain $p(\mathcal{M}_0(n)|\mathbf{Y}_{n-1})$ based on (A.22).
6: *Update step:*
7: Calculate $p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$ using the fast weight estimation algorithm [13] and (A.25).
8: Calculate the unnormalised posteriors $p(\mathcal{M}(n)|\mathbf{Y}_n)$ and $p(\mathcal{M}_0(n)|\mathbf{Y}_n)$ based on (A.23).
9: Normalise the posteriors based on the constraint (A.29).
10: *MAP estimation:*
11: **if** $p(\mathcal{M}_0(n)|\mathbf{Y}_n) > 0.5$ **then**
12: The n^{th} frame is labeled as unvoiced/silent.
13: **else**
14: The n^{th} frame is labeled as voiced.
15: Estimating $\hat{\omega}_n$ and $\hat{K}_n$ based on (A.30).
16: Update $p(\omega_m, K_m|\mathbf{Y}_m, u_m = 1)$ based on (A.18).
17: **end if**
18: **end for**

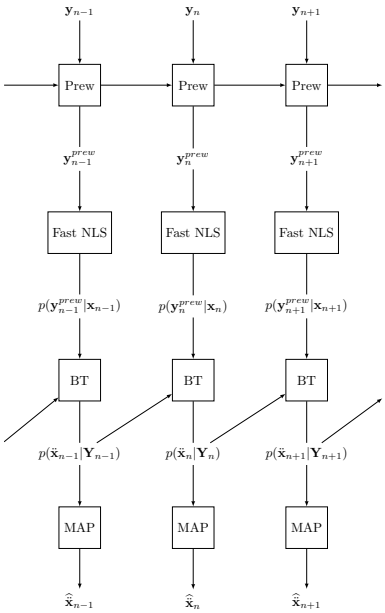


Fig. A.3: A block diagram of the proposed algorithm with prewhitening for colored noise, where Prew, and BT are abbreviations for prewhitening, Bayesian tracking, respectively.

6 Prewhitening

The fast NLS and proposed pitch tracking algorithm are derived under the assumption of white Gaussian noise. However, this assumption is usually violated in practice, for example, babble noise in a conference hall. Therefore, a prewhitening step is required to deal with the inconsistency between the white Gaussian noise model assumption and real life noise model. A linear prediction (LP) based prewhitening step is applied to each frame to deal with the non-white Gaussian noise (see [9, 42] for detail). The power spectral density (PSD) of the noise given noisy signals is estimated using the minimum mean-square error (MMSE) estimator [43]. We refer to the fast NLS and proposed algorithm with prewhitening step as Prew-Fast NLS and Prew-Proposed, respectively. Combing the prewhitening step and Algorithm 1, a block diagram for the proposed pitch tracker in colored noise scenarios is shown in Fig. A.3, where $\mathbf{y}_n^{prew}$ denotes the prewhitened observation vector and $\widehat{\ddot{\mathbf{x}}}_{n-1}$ denotes an estimate of $[\omega_n, K_n, u_n]^T$.

7 Simulation

In this section, we test the performance of the proposed harmonic model-based pitch tracking algorithm on real speech signals.

7.1 Databases

The databases used for evaluating the performance of different algorithms are as follows:

MIS database: containing 300 audio files from 6 different instrument classes: piano, violin, cello, flute, bassoon, and soprano saxophone, at a sampling rate of 44.1 kHz[2].

Keele pitch database: containing 10 spoken sentences from five male and five female speakers at a sampling rate of 20 kHz [44]. The "ground truth" pitch estimates are extracted from electroglottography with 10 ms time frame increment and 25.6 ms frame length. In fact, there are many spikes and wrong estimates in the "ground truth" pitch values, especially in the transient frames. However, we present the results for the Keele database to facilitate comparison with other pitch estimation algorithms that use this database.

Parkinson's disease database: containing 130 sustained /a/ phonations from patients with Parkinson's disease [45] at a sampling rate of 44.1 kHz. Each of the phonations is in one second length. The estimated "ground truth" pitches in 10 ms time frame increment are extracted from electroglottography (EGG).

7.2 Performance measures

Three performance measures are considered:

Total error ratio (TER) [22]: voicing detection performance measure. It is calculated based on the ratio between the number of incorrect voicing detection (false positive and true negative) estimates and the number of total estimates.

Gross error ratio (GER) [12]: accuracy measure of pitch estimates. It is computed based on the ratio between the number of pitch estimates that differ by more than 20 percents from the ground truth and the number of total estimates. The unvoiced frames from the ground truth are excluded and the pitch value of the voiced frame that is wrongly labeled as unvoiced frames by different pitch estimation algorithms is set to 0.

Mean absolute error (MAE) [45]: accuracy measure of pitch estimates. It is computed based on mean of the absolute errors between the ground truth and estimates. The unvoiced frames from the ground truth are excluded and the oracle voicing detector is used for all the algorithms.

[2]Audio files available in `http://theremin.music.uiowa.edu`

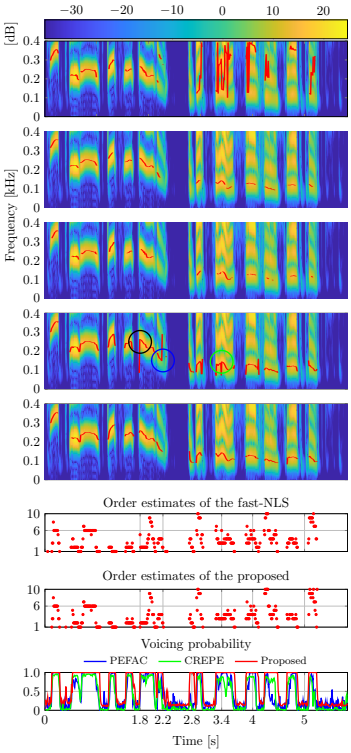


Fig. A.4: Pitch estimates from PEFAC, CREPE, YIN, fast NLS and the proposed, the order estimates of the fast NLS and the proposed, and the voicing probabilities for real speech signals in 0 dB white Gaussian noise (from top to bottom).

7.3 Experimental results on speech and audio samples

In this subsection, the experimental results of different pitch estimation algorithms for one speech and one audio sample, are presented in the first and second experiments, respectively.

First, the proposed approach is tested on concatenated speech signals uttered by a female speaker first, male speaker second, sampled at 16 kHz [3]. The spectrogram of the clean speech signals, pitch estimates, order estimates and the voicing detection results for PEFAC, CREPE, YIN, fast NLS and the proposed algorithm are shown in Fig. C.1. The time frames of the spectrograms without red lines on top are unvoiced or silent frames. The variances for the transition matrices σ_ω^2 and σ_K^2 are set to $\frac{16\pi^2}{f_s^2}$ and 1, respectively. The SNR for white Gaussian noise is set to 0 dB. The candidate pitch ω_0 is constrained to the range $2\pi\,[70, 400]\,/\,f_s$ for PEFAC, YIN, fast NLS and the pro-

[3]The example speech signal file is available in `https://tinyurl.com/yxn4a543`

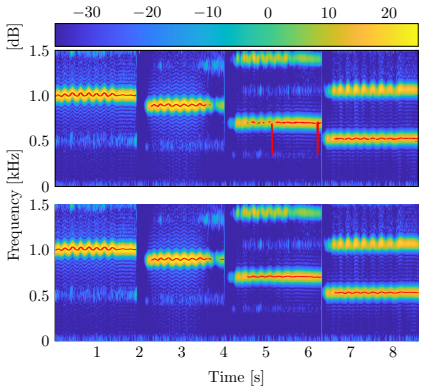


Fig. A.5: Pitch estimates of fast NLS and the proposed algorithm for musical sounds in -5 dB white Gaussian noise (from top to bottom).

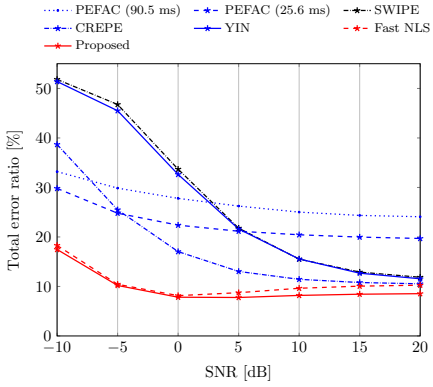


Fig. A.6: Total error ratio in different SNRs for the Keele pitch database in white Gaussian noise

posed algorithm. However, the results for the neural network based approach CREPE is based on the model with the pitch range $2\pi\,[32.7\ 1975.5]\,/f_{\mathrm{s}}$ provided by the authors [26]. To change the settings for CREPE, re-training of the neural network model is required. The maximum allowed harmonic order for the proposed and fast NLS is set to 10. The frame length is $M = 400$ samples (25 ms) with 60% overlap (10 ms time frame increment). As can be seen from Fig. C.1, the voicing detection results of both the fast NLS and the proposed algorithm are better than those of YIN, PEFAC and CREPE. For example, the frames around 2.8 s are correctly classified as voiced by the fast NLS and the proposed, but wrongly labeled as unvoiced by YIN, PEFAC and CREPE. Fast NLS suffers from octave errors, and has outliers particularly in the transition frames where voicing decisions change. In the transition frame around 1.8 s, the pitch and number of harmonics are wrongly estimated to

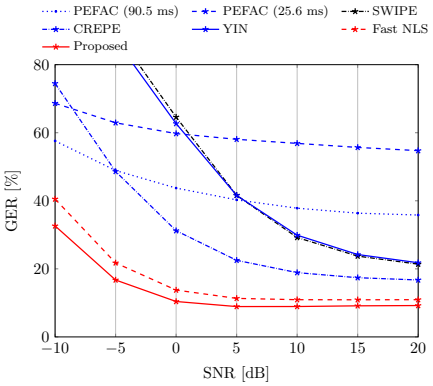


Fig. A.7: Gross error ratio in different SNRs for the Keele pitch database in white Gaussian noise

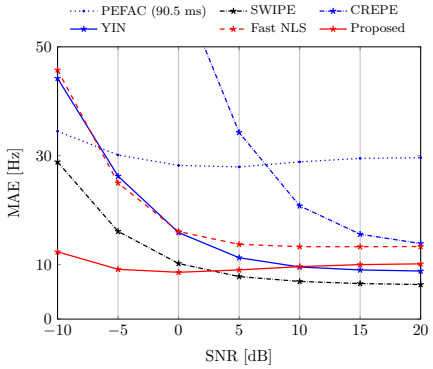


Fig. A.8: Mean absolute error in different SNRs for the Keele pitch database with oracle voicing detector in white Gaussian noise

84.8 Hz and five, respectively, by the fast NLS. In contrast, they are estimated to 248.8 Hz and one, respectively, by the proposed. Clearly, the estimates of the proposed fit better to the spectrogram than the estimates of the fast NLS. The reason for the robustness against transient frames of the proposed algorithm is that the pitch and harmonic order information of the latest voiced frame is used as the prior, i.e. (A.18). The harmonic order of the frame in 2.2 s is estimated to two by both the fast NLS and the proposed. However, the pitch is wrongly estimated to 288.8 Hz by the fast NLS, but correctly estimated to 150.4 Hz by the proposed. By imposing temporal smoothness prior on the pitch using the Markov process model $p(\omega_n|\omega_{n-1}, u_n = 1, u_{n-1} = 1)$, smoother estimates of the pitches are obtained. An octave error is produced by the fast NLS in the frame around 3.4 s. The pitch and harmonic order are estimated to 72 and six, respectively, by the fast NLS, but 143.2 and three,

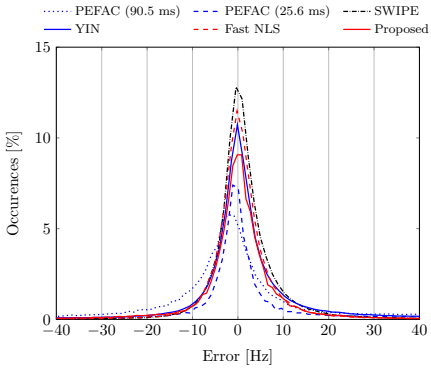


Fig. A.9: Pitch estimation error distributions of different algorithms for the Keele pitch database with oracle voicing detector in -5 dB white Gaussian noise

respectively, by the proposed. In fact, harmonic orders are estimated to three in the surrounding frames by both the fast NLS and the proposed. By using Bayesian tracking for the pitches and harmonic orders, smoother estimates of the pitches and harmonic orders are obtained. In conclusion, the proposed Bayesian pitch tracking algorithm obtains smooth estimates of the pitch and harmonic orders, and good voicing detection results by exploiting the past information.

The second experiment tests the performance of the proposed algorithm on musical instrument sounds (flute) from MIS database, decreasing from note B5 to C5. The spectrogram of the clean signals and the pitch estimates from fast NLS and the proposed algorithm are shown in Fig. A.5. The music signal is downsampled to 16 kHz. The SNR for Gaussian white noise is set to -5 dB. The pitch ω_0 is constrained to the range $2\pi\,[100\ 1500]\,/f_s$. The other parameters are set to the same as for Fig. C.1. As can be seen, the proposed algorithm not only has smoother estimates of the pitch than fast NLS but also better voicing detection results.

7.4 Experimental results on the Keele pitch database

In this subsection, the experimental results of different pitch estimation algorithms, using the Keele database, in white Gaussian noise, colored noise and reverberated conditions are presented.

First, we test the performance of the proposed algorithm on the Keele pitch database with white Gaussian noise. TER, GER and MAE in different SNRs for PEFAC, SWIPE, YIN, CREPE, fast NLS and the proposed algorithm are shown in Fig. A.6, Fig. A.7 and Fig. A.8, respectively. The error distributions of PEFAC, SWIPE, YIN, Fast NLS and the proposed algorithm with oracle voicing detector in -5 dB white Gaussian noise are shown in Fig. A.9.

Table A.1: Total error ratio in colored noise

SNR		-5.00	0.00	5.00	10.00
PEFAC (90.5 ms)	Babble	0.42	0.38	0.34	0.29
	Factory	0.34	0.30	0.27	0.25
PEFAC (25.6 ms)	Babble	0.41	0.35	0.29	0.24
	Factory	0.30	0.25	0.22	0.21
SWIPE	Babble	0.50	0.42	0.29	0.19
	Factory	0.52	0.49	0.40	0.28
CREPE	Babble	0.40	0.29	0.21	0.16
	Factory	0.39	0.28	0.20	0.15
YIN	Babble	0.50	0.43	0.32	0.22
	Factory	0.50	0.43	0.32	0.22
Prew-Fast NLS	Babble	0.35	0.27	0.18	0.12
	Factory	0.28	0.20	**0.14**	**0.11**
Prew-Proposed	Babble	**0.34**	**0.25**	**0.17**	**0.12**
	Factory	**0.28**	**0.20**	0.15	0.12

Table A.2: Gross error ratio in colored noise

SNR		-5.00	0.00	5.00	10.00
PEFAC (90.5 ms)	Babble	0.62	0.51	0.44	0.39
	Factory	0.56	0.47	0.41	0.38
PEFAC (25.6 ms)	Babble	0.72	0.65	0.60	0.57
	Factory	0.68	0.61	0.57	0.54
SWIPE	Babble	0.96	0.81	0.55	0.36
	Factory	1.00	0.94	0.76	0.54
CREPE	Babble	0.73	0.50	0.34	0.24
	Factory	0.75	0.53	0.36	0.26
YIN	Babble	0.95	0.83	0.61	0.42
	Factory	0.96	0.83	0.61	0.42
Prew-Fast NLS	Babble	0.57	0.41	0.30	0.24
	Factory	0.55	0.42	0.33	0.28
Prew-Proposed	Babble	**0.53**	**0.36**	**0.27**	**0.24**
	Factory	**0.51**	**0.37**	**0.29**	**0.25**

Table A.3: Mean absolute value [Hz] in colored noise with oracle voicing detector

SNR		-5.00	0.00	5.00	10.00
PEFAC (90.5 ms)	Babble	49.81	39.15	31.73	27.96
	Factory	36.20	31.24	27.97	26.69
PEFAC (25.6 ms)	Babble	81.49	72.65	65.71	60.54
	Factory	72.61	64.93	57.93	54.20
SWIPE	Babble	**31.73**	**17.94**	**10.95**	**8.04**
	Factory	43.91	27.02	16.02	10.51
CREPE	Babble	68.95	44.93	30.57	21.89
	Factory	79.00	52.41	34.51	24.70
YIN	Babble	56.25	39.05	23.86	14.96
	Factory	57.37	38.53	23.41	14.97
Prew-Fast NLS	Babble	64.81	45.79	31.45	23.79
	Factory	74.58	57.88	44.93	36.50
Prew-Proposed	Babble	33.33	17.91	12.22	10.81
	Factory	**19.32**	**13.20**	**11.23**	**10.48**

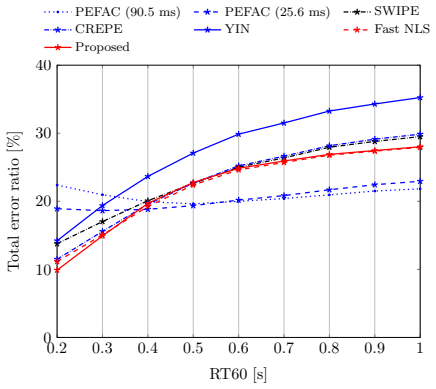


Fig. A.10: Total error ratio in different reverberation time for the Keele pitch database

For YIN, fast NLS and the proposed algorithm, the frame length is set to the same as the reference, i.e., 25.6 ms. Frame lengths 25.6 ms and 90.5 ms (default value) are used for PEFAC. The other parameters are set to the same as for Fig. C.1. Averages over 20 independent Monte Carlo experiments are used to compute TER, GER and MAE. The confidence intervals for them are not shown because they are not on the same scale as the mean values. For example, the 95% confidence intervals for GER and MAE estimates are on a scale of 0.1% and 0.1 Hz, respectively. As can be seen from Fig. A.6 and Fig. A.7, PEFAC has better performance in terms of both GER and TER than CREPE at -10 dB SNR. Moreover, using a longer frame length (90.5 ms) for PEFAC leads to a lower GER but a higher TER compared with a shorter frame length (25.6 ms). SWIPE and YIN have similar performance in terms of TER and GER.

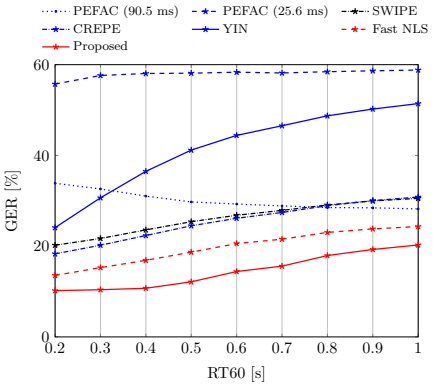


Fig. A.11: Gross error ratio in different reverberation time for the Keele pitch database

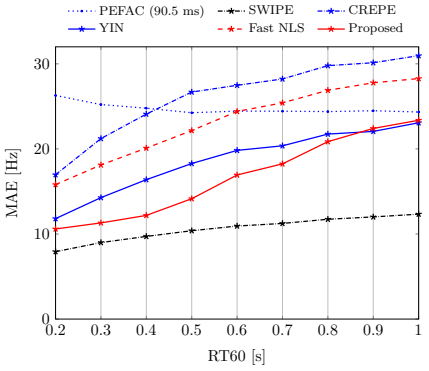


Fig. A.12: Mean absolute error in different reverberation time for the Keele pitch database with oracle voicing detector

The fast NLS method outperforms the PEFAC, SWIPE, YIN and CREPE. By imposing a smoothing prior on the pitches, harmonic orders and the voicing and using the harmonic model combined, the proposed algorithm achieves lower GER and TER than the fast NLS. As can be seen from Fig. A.8, when the oracle voicing detector is used, the SWIPE has the lowest MAE from 5 to 20 dB while the proposed algorithm achieves the best performance from -10 to 0 dB. From Fig. A.9, we can conclude that, for pitch estimation errors in the range $[-40, 40]$ Hz, the error distributions of SWIPE, PEFAC (25.6 ms), Fast NLS and the proposed algorithm in -5 dB white Gaussian noise are approximately symmetric around zero, while PEFAC (90.5 ms) tends to underestimate the pitch.

Second, the performance of the proposed algorithm with prewhitening is tested on the Keele pitch database in colored noise conditions, i.e., babble

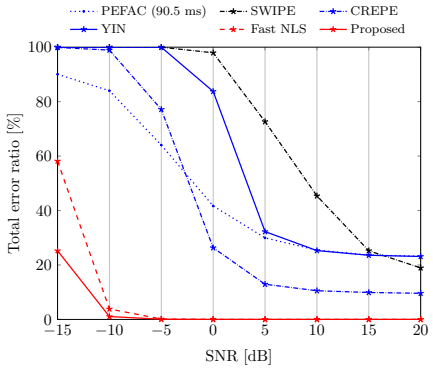


Fig. A.13: Total error ratio in different SNRs for the Parkinson's disease database in white Gaussian noise

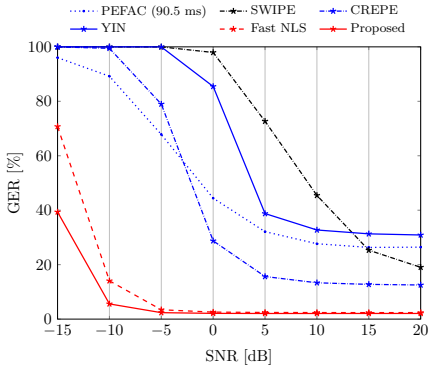


Fig. A.14: Gross error ratio in different SNRs for the Parkinson's disease database in white Gaussian noise

noise [4] and factory noise [5]. The time durations of these two files are both above 60 s. In each Monte Carlo trial, a randomly selected segment of the noise signals, according to the length of the speech signals, are scaled based on the desired SNR and added to the speech signals to simulate colored, noisy signals. The TER, GER and MAE results for Prew-proposed, Prew-fast NLS, PEFAC, Yin and SWIPE are shown in A.1, A.2 and A.3, respectively. The linear prediction order for the prewhitening is set to 30. The maximum allowed harmonic order for the proposed and fast NLS is set to 30. The other parameters are set to the same as for Fig. A.6. As can be seen from TABLE A.1 and A.2, PEFAC with 90.5 ms and 25.6 ms have a lower TER and GER than YIN and SWIPE in -5 and 0 SNR conditions. The Prew-Proposed and Prew-

[4]Crowd Talking 1 file in `https://www.soundjay.com/crowd-talking-1.html`
[5]Factory Floor Noise 2 file in `http://spib.linse.ufsc.br/noise.html`

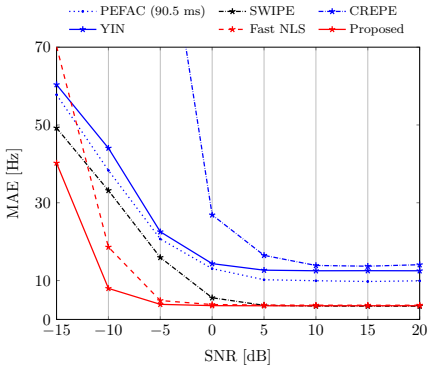


Fig. A.15: Mean absolute error in different SNRs for the Parkinson's disease database with oracle voicing detector in white Gaussian noise

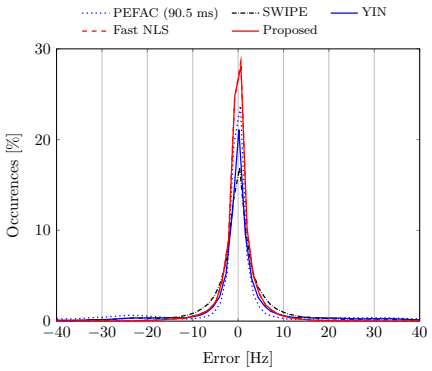


Fig. A.16: Pitch estimation error distributions of different algorithms for the Parkinson's disease database with oracle voicing detector in -5 dB white Gaussian noise

Fast NLS have lower voicing detection errors and Gross errors than YIN, PEFAC and SWIPE in both babble and factory noise conditions. Although similar performance in term of TER can be seen for Prew-Proposed and Prew-Fast NLS, the Prew-Proposed has a lower GER than Prew-Fast NLS. As can be seen from TABLE A.3, when the oracle voicing detector is used, the SWIPE achieves the lowest MAE in babble noise. The Prew-proposed has a comparable performance with the SWIPE in babble noise and has the best performance in factory noise.

Third, we investigate the effect of reverberation on the performance of different pitch estimation algorithms. Reverberation is the process of multi-path propagation and occurs when the speech or audio signals are recorded in an acoustically enclosed space. A commonly used metric to measure the reverberation is the reverberation time (RT60) [46]. The reverberated signals used

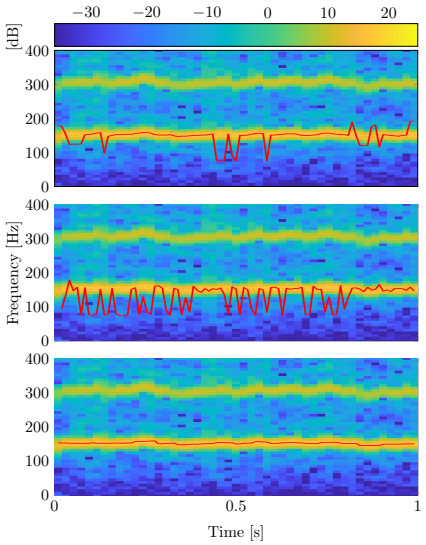


Fig. A.17: Pitch estimates from PEFAC (oracle), YIN (oracle), and the proposed for sustained /a/ sounds from a database of Parkinson's disease voices in 0 dB white Gaussian noise.

for testing are generated by filtering the signal by synthetic room impulse responses (RIRs) with RT60 varying from 0.2 to 1 s in 0.1 s step. The dimension of the room is set to $10 \times 6 \times 4$ m. The distance between the source and microphone is set to 1 m. The RIRs are generated using the image method [47] and implemented using the RIR Generator toolbox [48]. The position of the receiver is fixed while the position of the source is varied randomly from 60 degrees left of the receiver to 60 degrees right of the receiver for each Monte Carlo experiment. The TER, GER and MAE results on the Keele pitch database for the proposed, fast NLS, PEFAC, Yin and SWIPE are shown in Fig. A.10, Fig. A.11 and Fig. A.12, respectively, where the parameters are set to the same as for Fig. A.6. As can be seen from Fig. A.10, the PEFAC (90.5 ms) has the lowest voicing detection errors in more reverberated conditions (RT60 from 0.5 to 1 s) while the proposed algorithm has a better voicing detection performance in less reverberated conditions. The proposed and the fast NLS has similar performance in terms of TER. However, as can be seen from Fig. A.11, the proposed outperforms the PEFAC, SWIPE, CREPE, YIN and fast NLS in terms of GER. From Fig. A.12, we can conclude that SWIPE has the best performance while the proposed is the second best one in terms of MAE.

7.5 Experimental results on the Parkinson's disease database

In this subsection, the experimental results of different pitch estimation algorithms, using the Parkinson's disease database, in white Gaussian noise, is presented.

In the final experiment, the performance of the proposed algorithm is tested on sustained /a/ signals (voiced) from the Parkinson's disease database. The signals are downsampled to 16 kHz. TER, GER and MAE for different SNRs are shown in Fig. A.13, Fig. A.14 and Fig. A.15, respectively. The error distributions of PEFAC, SWIPE, YIN, FAST NLS and the proposed algorithm with oracle voicing detector in -5 dB white Gaussian noise are shown in Fig. A.16. The frame length is set to 80 ms for the fast NLS and proposed algorithms. The other parameters are set to the same as for Fig. A.6. Similar conclusions to Fig. A.6 and Fig. A.7 can be drawn from Fig. A.13 and Fig. A.14. The proposed algorithm has the best performance in terms of the TER and GER. Moreover, the proposed algorithm has similar performance as SWIPE in terms of MAE measure from 5 to 20 dB and presents the lowest MAE from -15 to 0 dB. As can be seen from Fig. A.16, for the Parkinson's disease database, the error distributions of PEFAC, SWIPE, Fast NLS and the proposed algorithm in -5 dB white Gaussian noise are all approximately symmetric around zero. The spectrogram of one of the sustained /a/ sounds from the Parkinson's disease database, pitch estimates of the PEFAC (oracle), YIN (oracle) and the proposed algorithm in 0 dB white Gaussian noise are shown in Fig. A.17. The oracle voicing detector from the ground truth (all voiced) is used for both PEFAC and YIN. As can be seen from Fig. A.17, the proposed algorithm outperforms the PEFAC (oracle) and YIN (oracle).

Based on the above experiments, PEFAC obtains a better pitch estimation and voicing detection performance than the neural network-based CREPE in low SNR scenarios. SWIPE offers good performance in terms of MAE in high SNRs. The proposed algorithm obtains superior performance in terms of GER, TER and MAE compared to PEFAC, SWIPE, YIN, CREPE, and the fast NLS in low SNR scenarios (under 5 dB) for the Keele pitch database and Parkinson's disease database. In high SNR scenarios (above 5 dB), the proposed algorithm has superior performance in terms of TER and GER, but not always the best performance in terms of MAE. In practice, choosing pitch estimation algorithm depends on the applications and needs.

8 Conclusions

In this paper, a fully Bayesian harmonic model-based pitch tracking algorithm is proposed. Using a parametric harmonic model, the proposed algorithm shows good robustness against noise. The non-stationary evolution of

the pitch, harmonic order and voicing state are modelled using first-order Markov chains. A fully Bayesian approach is applied for the noise variance and weights to avoid over-fitting. Using the hierarchical g-prior for the weights, the likelihood function can be easily evaluated using the fast NLS. The computational complexity of the recursive calculation of the predicted and posterior distributions is reduced by exploiting conditional independence between the pitch and harmonic order given the voicing indicators. Simulation results show that the proposed algorithm has good robustness against voicing state changes by carrying past information on pitch over the unvoiced/silent segments. The results of the pitch estimates and voicing detection for spoken sentences and sustained vowels are compared against ground truth estimates in the Keele and Parkinson's disease databases, showing that the proposed algorithm presents good pitch estimation and voicing detection accuracy even in very noisy conditions (e.g., -15 dB).

References

[1] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, April 2014.

[2] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010.

[3] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 2494–2498.

[4] D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003.

[5] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[6] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[7] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2042–2056, Oct 2013.

[8] K. Paliwal and K. Wójcicki, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Process. Lett.*, vol. 15, pp. 785–788, 2008.

[9] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2354–2367, 2016.

[10] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[12] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[13] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.

[14] G. Aneeja and B. Yegnanarayana, "Extraction of fundamental frequency from degraded speech using temporal envelopes at high SNR frequencies," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 829–838, 2017.

[15] B. Quinn and P. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, 1991.

[16] J. Sward, H. Li, and A. Jakobsson, "Off-grid fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 2, pp. 296–303, Feb. 2018.

[17] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[18] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "A kalman-based fundamental frequency estimation algorithm," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE Press, 2017, pp. 314–318.

[19] G. Zhang and S. Godsill, "Fundamental frequency estimation in speech signals with variable rate particle filters," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 5, pp. 890–900, 2016.

[20] C. Andrieu, M. Davy, and A. Doucet, "Efficient particle filtering for jump markov systems. application to time-varying autoregressions," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1762–1770, 2003.

[21] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004.

[22] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[23] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[24] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 2158–2168, 2014.

[25] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm using the deep stacking network for speech separation and pitch estimation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 6, pp. 1066–1078, 2016.

[26] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 2018, pp. 161–165.

[27] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[28] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[29] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, 2008.

[30] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[31] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *Proc. IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[32] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 438–449, 2002.

[33] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 176–183.

[34] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Process.*, vol. 109, pp. 236–247, 2015.

[35] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 225–238, 2014.

[36] A. Zellner, "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques*, 1986.

[37] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for bayesian variable selection," *J. Amer. Stat. Assoc.*, vol. 103, no. 481, pp. 410–423, 2008.

[38] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[39] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[40] K. Yoshii and M. Goto, "A nonparametric bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 3, pp. 717–730, 2012.

[41] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[42] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 6495–6499.

[43] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[44] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[45] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–2901, 2014.

[46] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[47] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[48] E. A. P. Habets, "Room impulse response generator," *Tech. Rep.*, 2010. [Online]. Available: https://github.com/ehabets/RIR-Generator

Paper B

A Kalman-based Fundamental Frequency Estimation Algorithm Using the Harmonic Model

Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen,
Max A. Little and Mads Græsbøll Christensen

The paper has been published in the
Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2017

Abstract

Fundamental frequency estimation is an important task in speech and audio analysis. Harmonic model-based methods typically have superior estimation accuracy. However, such methods usually assume that the fundamental frequency and amplitudes are stationary over a short time frame. In this paper, we propose a Kalman filter-based fundamental frequency estimation algorithm using the harmonic model, where the fundamental frequency and amplitudes can be truly nonstationary by modeling their time variations as first-order Markov chains. The Kalman observation equation is derived from the harmonic model and formulated as a compact nonlinear matrix form, which is further used to derive an extended Kalman filter. Detailed and continuous fundamental frequency and amplitude estimates for speech, the sustained vowel /a/ and solo musical tones with vibrato are demonstrated.

1 Introduction

Fundamental frequency can be described as the lowest rate for a periodic signal to repeat itself. Fundamental frequency information for voiced speech or audio signals has various applications, such as speech enhancement [1], voice disorder detection [2], automatic speech recognition [3] and music processing [4]. A very large number of fundamental frequency estimation algorithms have been proposed in the past, including those that could be broadly described as *non-parametric* and *parametric* methods. Here we will define non-parametric methods as those which are based on the autocorrelation function obtained within a specified time frame; examples include Yin [5] and RAPT [6]. These methods are computationally simple but they are prone to *subharmonic error* (that is, misidentifying multiples of the actual fundamental frequency, a.k.a. octave error). To reduce this subharmonic error problem, a recently devised method – the sawtooth waveform-inspired pitch estimator (SWIPE) [7], and variants – use the cross-correlation function against a sawtooth signal combined with frequency-domain information. By contrast, examples of parametric methods are *harmonic models* which use *nonlinear least squares* (NLS) model parameter estimation [8]. Under appropriate assumptions, such NLS estimators are optimal from a statistical perspective but are very computationally costly to run in practice. To lower this computational cost, recently a fast NLS has been proposed which exploits the matrix structure using a recursive matrix solver [9]. Most parametric harmonic models, as with non-parametric methods, assume signal stationarity at least over short time frames, but in practice this assumption is unrealistic. To account for the non-stationarity of voiced speech signals, a *harmonic chirp* model for voiced speech has been proposed, and the fundamental frequency and chirp rate parameters are obtained iteratively [10]. Another parametric model, the *adaptive*

quasi-harmonic model [11] has been proposed to attempt to capture time variation in both frequency and amplitude of voiced speech signals. Recently, *instantaneous* fundamental frequency estimation algorithms based on the harmonic model which use *non-linear recursive filters* have been proposed [12]. As the model parameter update results in a nonlinear state equation, classical extended (EKF), unscented and particle Kalman filters have been proposed to perform the parameter estimation in this time-varying model. Continuous variations in fundamental frequency are obtained. However, the size of the state space in [12] is $3K + 1$, where K is the harmonic order, leading to high computational effort.

In this paper, we propose a fixed order harmonic model to fit voiced speech and audio signals. A first order Markov chain is used to capture non-stationarity in fundamental frequency and amplitude. By exploiting linear relationships between the phases of different harmonics, the size of the state space is decreased to $K + 2$. The resulting nonlinear observation equation is formulated in compact matrix form, and an extended Kalman smoother is applied to track instantaneous fundamental frequency and amplitudes.

2 Harmonic model estimation

Consider the following general signal observation model

$$y_n = s_n + v_n, \tag{B.1}$$

where y_n is the observation signal and v_n denotes zero mean Gaussian noise with variance r_v, and n is the integer time index. We assume that the voiced speech or audio signal s_n is produced by a time-varying harmonic model, i.e.,

$$s_n = \sum_{k=1}^{K} A_{n,k} \cos(\theta_{n,k}), \tag{B.2}$$

$$\theta_{n,k} = k\omega_n n + \theta_{0,k}, \quad k = 1, \cdots, K, \tag{B.3}$$

where $A_{n,k}$ is the instantaneous amplitude of the k^{th} harmonic at time instant n, $\theta_{n,k}$ is the instantaneous phase, $\omega_n = 2\pi f_n / F_s$ is the instantaneous normalized digital radian frequency, $F_s = 1/T_s$ is the sampling rate, T_s is the sampling period, and $\theta_{0,k}$ is the initial phase, and K is the number of harmonics. Our objective is to estimate the fundamental frequency ω_n and amplitudes $A_{n,k}$, $1 \leq k \leq K$, simultaneously.

Assume that the fundamental frequency and amplitudes are time-invariant in a short time frame with a length N, and thus the time index n can be ignored, i.e. $\omega_n = \omega_0$ and $A_{n,k} = A_k$, $1 \leq k \leq K$. Combining (C.5),

(C.6) and (C.7), and using Euler's formula, we obtain

$$y_n = \sum_{k=1}^{K} \left(a_k z_n^k + a_k^* z_n^{-k} \right) + v_n, \tag{B.4}$$

where the superscript $*$ denotes complex conjugation, complex amplitude a_k is defined as $a_k = \frac{A_k}{2} e^{j\theta_{0,k}}$, and $z_n = e^{j\omega_0 n}$. Collecting N observation signals into a vector and writing (B.4) in matrix form yields

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{v}_n, \tag{B.5}$$

where $\mathbf{y}_n = [y_1, y_2, \cdots, y_N]^T$ and $\mathbf{v}_n$ is defined in the same form, $\mathbf{a} = [a_1, a_1^*, a_2, a_2^*, \cdots, a_K, a_K^*]^T$, and where $\mathbf{Z} = [\mathbf{z}(1), \mathbf{z}(-1), \mathbf{z}(2), \mathbf{z}(-2), \cdots, \mathbf{z}(K), \mathbf{z}(-K)]$ with $\mathbf{z}(k)$ defined as $\mathbf{z}(k) = [z_1^k, z_2^k, \cdots, z_N^k]^T$. With i.i.d. Gaussian assumptions on elements of vector $\mathbf{v}_n$ and fixed fundamental frequency ω_0, the maximum likelihood (ML) estimate of the complex amplitude vector $\mathbf{a}$ can be found using the normal equations, $\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{y}$ [8]. Replacing $\mathbf{a}$ in (B.5) with the ML estimate $\hat{\mathbf{a}}$, the ML estimator of the fundamental frequency can be formulated as the least squares problem

$$\begin{aligned} \hat{\omega}_0 =& \arg\min_{\omega_0} \left\| \mathbf{y} - \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{y} \right\|_2^2 \\ =& \arg\max_{\omega_0} \mathbf{y}^T\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{y}. \end{aligned} \tag{B.6}$$

where $\|\cdot\|_2^2$ is the squared 2-norm. The above NLS maximization problem is solved by a coarse grid search followed by a gradient ascent refinement process.

3 Proposed Kalman filter-based fundamental frequency estimation algorithm

We now proceed to consider the time-varying fundamental frequency and amplitude scenario. We first formulate the state and observation equations based on the time-varying harmonic model (C.6) and (C.7), and observation model (C.5), respectively. Then, the extended Kalman smoother framework is applied to solve the nonlinear observation equation problem.

3.1 State and observation equations

Assuming that the continuous phase can be written as $\Theta_{t,k} = k\Omega_t t + \Theta_{0,k}$, at sampling rate F_s, we the instantaneous frequency of the k^{th} harmonic is

$$\begin{aligned} k\omega_n = k\Omega_t T_s|_{t=nT_s} =& \frac{T_s \partial \Theta_{t,k}}{\partial t}|_{t=nT_s} \\ =& T_s \lim_{\Delta t \to 0} \frac{\Theta_{t,k} - \Theta_{t-\Delta t,k}}{\Delta_t}|_{t=nT_s} \\ \approx& T_s \frac{\Theta_{nT_s,k} - \Theta_{nT_s - T_s,k}}{T_s} \\ =& \theta_{n,k} - \theta_{n-1,k}, \end{aligned} \tag{B.7}$$

where Ω_t is the continuous radian frequency, $\omega_n = \Omega_{nT_s} T_s$ and $\theta_{n,k} = \Theta_{nT_s,k}$.

The approximation in (C.8) can be further verified from (C.7), i.e.,

$$\begin{aligned} \theta_{n,k} &= k\omega_n(n-1) + \theta_{0,k} + k\omega_n \\ &\approx k\omega_{n-1}(n-1) + \theta_{0,k} + k\omega_n, \\ &= \theta_{n-1,k} + k\omega_n, \quad k = 1, \cdots, K \end{aligned} \tag{B.8}$$

where, in the second step, we used the assumption that fundamental frequency is slowly changing relative to the timescale of the sampling rate, that is $\omega_n \approx \omega_{n-1}$. We collect the frequency, amplitudes and phase $\theta_{n-1,1} - \theta_{0,1}$ as a $(K+2) \times 1$ state vector

$$\mathbf{x}_n = [\omega_n, A_{n,1}, \cdots, A_{n,K}, \theta_{n-1,1} - \theta_{0,1}]^T. \tag{B.9}$$

From (B.8) and (C.11), we can further derive that the phases of different harmonics for $n \geq 1$ are related by

$$\begin{aligned} \theta_{n,k} =& \theta_{n-1,k} + k\omega_n \\ =& \theta_{0,k} + k\sum_{i=1}^{n} \omega_i \\ =& \theta_{0,k} + k(\theta_{n-1,1} - \theta_{0,1} + \omega_n) \\ =& kx_{n,1} + kx_{n,K+2} + \theta_{0,k}, \quad k = 1, \cdots, K, \end{aligned} \tag{B.10}$$

where $x_{n,i}$ denotes the i^{th} component of the vector $\mathbf{x}_n$. Substituting (C.11) and (C.9) into (C.6), the harmonic model can be re-formulated as

$$s_n = \sum_{k=1}^{K} x_{n,k+1} \cos(kx_{n,1} + kx_{n,K+2} + \theta_{0,k}). \tag{B.11}$$

We assume the frequency and amplitudes are changing in time according to a first order Markov chain random walk model

$$x_{n,k} = x_{n-1,k} + m_{n,k}, \quad k = 1, \cdots, K+1, \tag{B.12}$$

where $m_{n,k}$ are K zero mean, i.i.d. Gaussian processes. Moreover, based on the phase update (B.8) and definition (C.11), we have

$$\begin{aligned} x_{n,K+2} =& \theta_{n-1,1} - \theta_{0,1} \\ =& \theta_{n-2,1} + \omega_{n-1} - \theta_{0,1} \\ =& x_{n-1,K+2} + x_{n-1,1}. \end{aligned} \tag{B.13}$$

Based on (B.12) and (C.15), we can write the state equation in matrix form

$$\mathbf{x}_n = \mathbf{F}\mathbf{x}_{n-1} + \mathbf{\Gamma}\mathbf{m}_n, \tag{B.14}$$

where $\mathbf{F}$ is an $(K+2) \times (K+2)$ lower triangular Toeplitz matrix with first column $[1, 0, \cdots, 0, 1]^T$, $\mathbf{\Gamma}$ is a $(K+2) \times (K+1)$ Toeplitz matrix with first column $[1, 0, \cdots, 0]^T$ and the first row as $[1, 0, \cdots, 0]$, and the state noise vector is defined as $\mathbf{m}_n = [m_{n,1}, m_{n,2}, \cdots, m_{n,K+1}]^T$ with a covariance matrix $\mathbf{Q}_m$. Combining (C.5) and (B.11), we can write the observation equation in matrix form

$$y_n = (\mathbf{G}\mathbf{x}_n)^T \cos(\mathbf{B}\mathbf{x}_n + \boldsymbol{\theta}_0) + v_n, \tag{B.15}$$

where $\mathbf{G}$ is an $K \times (K+2)$ Toeplitz matrix with first column as a zero vector and first row as $[0, 1, 0, \cdots, 0]$, $\mathbf{B}$ is a $K \times (K+2)$ zero matrix except that the first and last columns are $[1, 2, \cdots, K]^T$, and $\boldsymbol{\theta}_0 = [\theta_{0,1}, \theta_{0,2}, \cdots, \theta_{0,K}]^T$.

Algorithm 2 Extended Kalman smoohter for fundamental frequency estimation

1: Initiate harmonic order K, state vector $\mathbf{x}_1$ and initial phase $\boldsymbol{\theta}_0$ with the NLS and Amp-LS algorithms
2: Choose initial state covariance $\mathbf{P}_{1|1}$, state noise covariance $\mathbf{Q_m}$ and background noise variance r_v
3: **Filtering step (forward, online):**
4: **for** $n = 2, 3, \cdots, N$ **do**
5: $\quad \mathbf{x}_{n|n-1} = \mathbf{F}\mathbf{x}_{n-1|n-1}$
6: $\quad \mathbf{P}_{n|n-1} = \mathbf{F}\mathbf{P}_{n-1|n-1}\mathbf{F}^T + \mathbf{\Gamma}\mathbf{Q_m}\mathbf{\Gamma}^T$
7: $\quad$ Calculate $\mathbf{H}_n$ based on (B.18)
8: $\quad \mathbf{K}_n = \mathbf{P}_{n|n-1}\mathbf{H}_n^T(\mathbf{H}_n\mathbf{P}_{n|n-1}\mathbf{H}_n^T + r_v)^{-1}$
9: $\quad$ Obtain $h(\hat{\mathbf{x}}_{n|n-1})$ based on (B.17)
10: $\quad \mathbf{x}_{n|n} = \mathbf{x}_{n|n-1} + \mathbf{K}_n(y_n - h(\hat{\mathbf{x}}_{n|n-1}))$
11: $\quad \mathbf{P}_{n|n} = \mathbf{P}_{n|n-1} - \mathbf{K}_n\mathbf{H}_n\mathbf{P}_{n|n-1}$
12: **end for**
13: **Smoothing step (backward, offline):**
14: **for** $n = N, N-1, \cdots, 2$ **do**
15: $\quad \mathbf{S}_{n-1} = \mathbf{P}_{n-1|n-1}\mathbf{F}^T\mathbf{P}_{n|n-1}^{-1}$
16: $\quad \mathbf{x}_{n-1|N} = \mathbf{x}_{n-1|n-1} + \mathbf{S}_{n-1}(\mathbf{x}_{n|N} - \mathbf{x}_{n|n-1})$
17: $\quad \mathbf{P}_{n-1|N} = \mathbf{P}_{n-1|n-1} + \mathbf{S}_{n-1}(\mathbf{P}_{n|N} - \mathbf{P}_{n|n-1})\mathbf{S}_{n-1}^T$
18: **end for**

3.2 Linearization via Taylor approximation

First, we linearise the nonlinear observation equation (C.12) using the first-order Taylor expansion about estimate $\mathbf{x}_n = \hat{\mathbf{x}}_{n|n-1}$

$$y_n \approx h(\hat{\mathbf{x}}_{n|n-1}) + \mathbf{H}_n(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}) + v_n, \tag{B.16}$$

$$h(\hat{\mathbf{x}}_{n|n-1}) = (\mathbf{G}\hat{\mathbf{x}}_{n|n-1})^T \cos(\mathbf{B}\hat{\mathbf{x}}_{n|n-1} + \boldsymbol{\theta}_0), \tag{B.17}$$

where $\mathbf{H}_n$ is a $1 \times (K+2)$ Jacobian matrix

$$\begin{aligned}
\mathbf{H}_n =& \frac{\partial (\mathbf{G}\mathbf{x}_n)^T \cos(\mathbf{B}\mathbf{x}_n + \boldsymbol{\theta}_0)}{\partial \mathbf{x}_n^T}\Big|_{\mathbf{x}_n = \hat{\mathbf{x}}_{n|n-1}} \\
=& [\cdots, \frac{\partial (\mathbf{G}\mathbf{x}_n)^T \cos(\mathbf{B}\mathbf{x}_n + \boldsymbol{\theta}_0)}{\partial x_{n,k}}, \cdots]\Big|_{\mathbf{x}_n = \hat{\mathbf{x}}_{n|n-1}} \\
=& [\cdots, \mathbf{i}_k^T \mathbf{G}^T \cos(\mathbf{B}\hat{\mathbf{x}}_{n|n-1} - \boldsymbol{\theta}_0) \\
& + (\mathbf{G}\hat{\mathbf{x}}_{n|n-1})^T (\sin(\mathbf{B}\hat{\mathbf{x}}_{n|n-1} + \boldsymbol{\theta}_0) \odot \mathbf{B}_{:,k}), \cdots] \\
=& \cos((\mathbf{B}\hat{\mathbf{x}}_{n|n-1} + \boldsymbol{\theta}_0)^T)\mathbf{G} \\
& - (\mathbf{G}\hat{\mathbf{x}}_{n|n-1})^T \mathrm{diag}(\sin(\mathbf{B}\hat{\mathbf{x}}_{n|n-1} + \boldsymbol{\theta}_0))\mathbf{B},
\end{aligned} \tag{B.18}$$

where $\mathbf{i}_k$ is a zero vector except that the k^{th} element is 1, $\odot$ denotes the element-wise product, $\mathbf{B}_{:,k}$ denotes the k^{th} column of the matrix $\mathbf{B}$, $\mathrm{diag}(\mathbf{z})$ denotes converting a column vector $\mathbf{z}$ to a diagonal matrix with the $(i,i)^{\text{th}}$ diagonal element set as the i^{th} element of $\mathbf{z}$.

3.3 Kalman-based fundamental frequency estimation

We use the extend Kalman filter (EKF) smoother to estimate the mean and covariance of the state vector $\mathbf{x}_n$. For completeness, the filtering and smoothing steps of the EKF are shown in Algorithm 2. For real-time application, the forward filtering step should be used without the backward smoothing step. Using only the forward filtering step leads to larger uncertainty over the parameter estimates [13]. This algorithm can be initialized with the NLS estimate and the complex amplitude estimator using least-squares (Amp-LS) [4]. For Kalman filter parameter tuning we refer the reader to [14] and [15].

4 Results

In this section, we test the performance of the proposed Kalman-based fundamental frequency tracking algorithm for real speech and music signals.

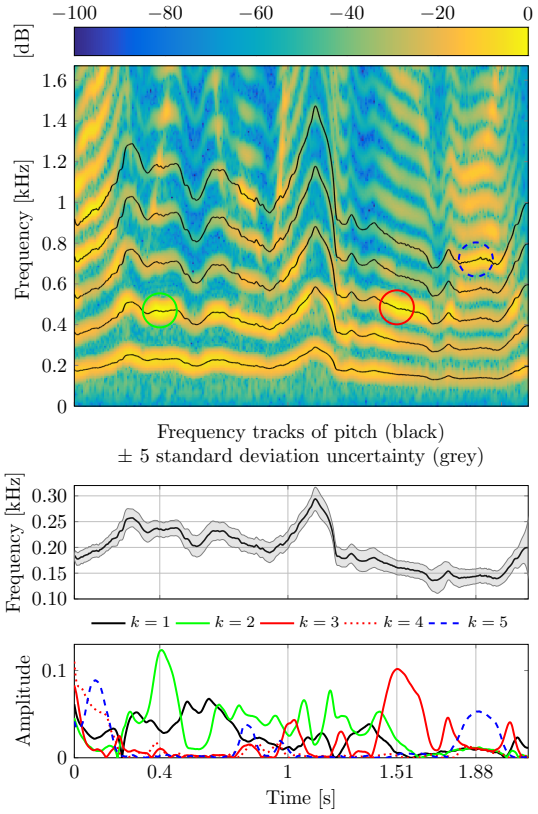


Fig. B.1: Fundamental frequency estimates of the speech signal "Why were you away a year, Roy?", number of harmonics $K = 5$.

4.1 Speech signal analysis

First, the proposed approach is tested on a speech signal of the spoken sentence "Why were you away a year, Roy?" uttered by a female speaker sampled at 8000 Hz. The spectrogram of the clean speech signals, fundamental frequency and amplitude estimates are shown in Fig. B.1, where $K = 5$, $r_v = 10^4$, the SNR for Gaussian white noise is set to 10 dB, $\mathbf{Q}_m$ and $\mathbf{P}_{1|1}$ are set to the identity matrices. As can be seen, the proposed algorithm generates continuous pitch estimates. Large amplitude estimates for harmonics $k = 2$, $k = 3$ and $k = 5$ are obtained in the high energy time-frequency area around 0.4 s, 1.51 s and 1.88 s. However, note that a clear delay in frequency estimate can be seen around 0.3 s (see the 4th and 5th harmonic tracks) due to the fixed harmonic order and r_v we used here. One approach to mitigating this delay is to re-initiate the algorithm with estimated harmonic order K and r_v based on a segmentation approach [16].

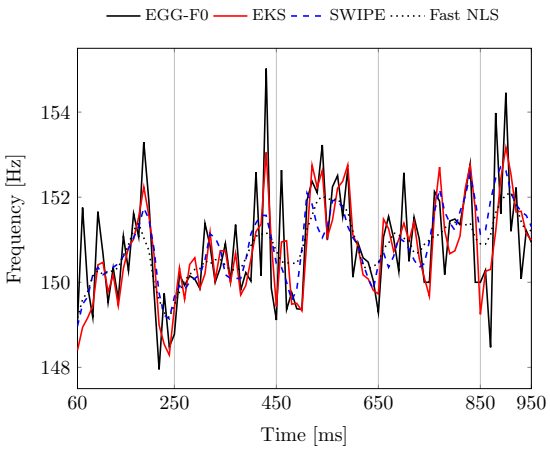


Fig. B.2: Fundamental frequency estimates of a sustained /a/ signal from a female patient with Parkinson's disease, number of harmonics $K = 4$, state noise variance $r_v = 10^5$.

Second, the performance of the proposed approach with different harmonic orders is compared with the classical extended Kalman filter (EKF) and SWIPE algorithms on a synthesized, sustained /a/ signal from a female with Parkinson's disease from a database of speech signals generated by a biophysical model of impaired voice production [17]. For this generated signal, the exact ground truth fundamental frequencies in 10 ms time frames are known. People with Parkinson's tend to exhibit increased vocal breathiness, tremor and roughness, and this presents a challenge for fundamental frequency estimation algorithms. The frequency estimates and the corresponding error measures of mean absolute error (MAE), mean relative error (MRE) and root mean squared errors (RMSE, see definitions in [17]) are obtained, where smaller values of error measures are better (see Fig. B.2 and Table B.1). The state noise variance r_v is set to 10^5 for the proposed EKS and traditional EKF. As can be seen from Fig. B.2, the proposed EKS with $K = 4$ achieves the closest approximation to the ground truth. Also, from Fig. B.2 and Table B.1 the performance of SWIPE and Fast NLS are similar and better than the traditional EKF. Furthermore, when $K = 4$, the performance of the proposed EKS is better than for other choices of K.

4.2 Music signal analysis

In this part, the sound of a musical instrument (flute) decreasing in frequency from note B5 to C5 from the University of Iowa Musical Instrument Samples [18] database is tested. The spectrogram of the signals and frequency estimates are shown in Fig. B.3, with number of harmonics $K = 2$, and the

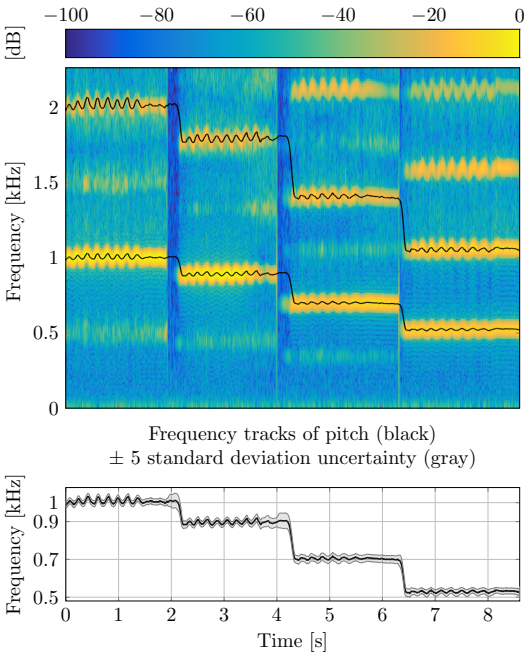


Fig. B.3: Fundamental frequency estimate of vibrato notes of a flute decreasing from B5 to C5 (eight notes of C-Major scale), number of harmonics $K = 2$.

SNR for Gaussian white noise is set to 30 dB, the sampling rate is 8000 Hz, and the other parameters have the same settings as in Fig. B.1. As can be seen, the proposed EKS can obtain a reasonably good estimate of fundamental frequency. Also, as expected, during a transition period from one note to another, the algorithm estimates large frequency uncertainty.

5 Conclusions

In this paper, we have proposed a fundamental frequency estimation algorithm based on a parametric harmonic model. Non-stationary temporal evolution of frequency and amplitude are modeled as first-order Markov chains. Compact nonlinear matrix forms of state and observation equations based are formulated, and an extended Kalman smoother for the problem is derived. The size of the state space is lowered by exploiting the linear relationships between the phases of different harmonics. Continuous fundamental frequency and amplitudes estimates for sustained vowels are compared to ground truth estimates from a biophysical model of impaired speech production, showing that this new algorithm outperforms existing algorithms in terms of accuracy.

Table B.1: Performance of the fundamental frequency estimation algorithms for a synthesized, sustained /a/ voice signals of 1 second duration generated from a biophysical model of impaired voice production. Performance results here are averaged over 90 frequency values from 60 to 950 ms in steps of 10 ms. MAE: mean absolute error, MRE: mean relative error, RMSE: root mean squared error.

Algorithms	MAE (Hz)	MRE (%)	RMSE (Hz)
EKF, $K = 4$	0.97	0.64	1.26
SWIPE	0.80	0.53	1.05
Fast NLS	0.80	0.53	1.05
EKS, $K = 1$	0.72	0.47	0.95
EKS, $K = 2$	0.71	0.47	0.92
EKS, $K = 3$	0.66	0.44	0.87
EKS, $K = 4$	0.63	0.42	0.86

References

[1] M. Krawczyk-Becker and T. Gerkmann, "Fundamental frequency informed speech enhancement in a flexible statistical framework," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 5, pp. 940–951, 2016.

[2] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012.

[3] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 2494–2498.

[4] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[5] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[7] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[8] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, 2008.

[9] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.

[10] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*. IEEE, 2014, pp. 1400–1404.

[11] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM–FM signal decomposition with application to speech analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 2, pp. 290–300, 2011.

[12] H. Hajimolahoseini, R. Amirfattahi, H. Soltanian-Zadeh, and S. Gazor, "Instantaneous fundamental frequency estimation of non-stationary periodic signals using non-linear recursive filters," *IET Signal Process.*, vol. 9, no. 2, pp. 143–153, 2015.

[13] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking a," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, 2012.

[14] R. Mehra, "On the identification of variances and adaptive kalman filtering," *IEEE Trans. Autom. Control*, vol. 15, no. 2, pp. 175–184, 1970.

[15] S. Formentin and S. Bittanti, "An insight into noise covariance estimation for kalman filter design," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 2358–2363, 2014.

[16] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2354–2367, 2016.

[17] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–2901, 2014.

[18] L. Fritts. [Online]. Available: http://theremin.music.uiowa.edu/index.html

Paper C

Instantaneous Bayesian Pitch Tracking in Colored Noise

Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen,
Max A. Little and Mads Græsbøll Christensen

The paper has been submitted to the
IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019

Abstract

Pitch information is one of the most important features of speech/audio signals. In noisy conditions, harmonic model-based pitch estimators present higher robustness than the more commonly used autocorrelation-based approaches. However, the pitch and amplitudes are assumed to be determinstic but unknown parameters and estimated frame by frame independently in traditional harmonic model-based pitch estimators. Moreover, noise is usually assumed to have a white Gaussian distribution, which is not realistic in real-life applications. In this paper, we propose instantaneous pitch tracking algorithms based on harmonic model and autoregressive model in a Bayesian tracking framework, which exploits the prior knowledge and the temporal information of the pitch and amplitudes. Modeling sound signals using the harmonic model, statistical optimal pitch estimates can be obtained, and the proposed pitch tracking algorithms have a high robustness against noise. Instead of assuming that the observation noise is white, the time-varying autoregressive model is used to model the noise in the proposed method. Experimental results on synthetic and real speech/audio signals show that the proposed pitch tracking algorithms are robust to colored noise and yields detailed and continuous pitch and amplitude estimates.

1 Introduction

The problem of estimating the fundamental frequency or pitch from noisy sound signals occurs in many applications, such as speech synthesis [1], voice disorder detection [2], and automatic speech recognition [3]. The fundamental frequency (or its inverse, the period) is a physical feature defined as the lowest frequency component of a periodic signal (e.g., the rate of vibration of the vocal folds for voiced speech signal). Pitch, on the other hand, is a perceptual feature related to the auditory sensation in terms of which sounds may be ordered on a musical scale [4]. The objective of this paper is to estimate the fundamental frequency. However, following [5], we use fundamental frequency and pitch interchangeably.

A widely used assumption in most pitch estimation methods is that the pitch is constant over a signal segment with fixed length (e.g., 15-40 ms for speech [6–8]). To ensure a smooth evolution of the pitch contour across segment boundaries, post-processing techniques, such as dynamical programming [9] or the Viterbi algorithm [10] are usually applied to remove outliers, often referred to as octave errors. Numerous segment-wise pitch estimation approaches have been proposed over the last fifty years. Segment-wise pitch estimation algorithms can be roughly categorized into two groups, i.e., non-parametric and parametric approaches. Non-parametric pitch estimation approaches such as YIN [11], Kaldi [3], SWIPE [12], and PEFAC [13] are based on various modifications of the autocorrelation method and are computa-

tionally cheap. In YIN, to obtain smooth pitch estimates, local estimates are obtained by considering neighbouring pitch estimates using median smoothing or dynamic programming techniques. In Kaldi, a cost function combining a modification of the autocorrelation function and a penalty-factor controlling how fast the pitch can change over time is defined. The pitch sequence is computed using the Viterbi algorithm. In SWIPE, both temporal and frequency interpolations are applied to obtain smooth and refined pitch estimates. In PEFAC, based on a cost function considering the rate of change of the pitches, the dynamical programming approach is applied to obtain smooth pitch estimates. Compared with non-parametric pitch estimation approaches, parametric approaches have a higher time-frequency resolution and robustness against noise [5]. A natural choice of parametric modeling of a periodic signal is the harmonic model. A nonlinear least squares (NLS) pitch estimator based on the maximum likelihood (ML) criterion has been proposed [14]. However, the naïve implementation of the NLS pitch estimator is computationally complex. A computational efficient algorithm based on the harmonic model was proposed in [15], where careful initialization of the parameters and refining based on the weighted least squares techniques were applied. This algorithm can be viewed as an approximated NLS estimator and works in moderate SNRs. More recently, the authors in [16] found that the linear parameters in the harmonic model can be computed recursively, and a fast NLS pitch estimation algorithm was derived. Considering the temporal continuity prior of the pitches, harmonic order and voicing, a Bayesian pitch tracking algorithm using the fast NLS algorithm has been proposed in [17].

For non-stationary sound signals (e.g., speech or vibrato musical sound), to obtain high temporal resolution of pitch estimates, we would like the segment length to be as small as possible. On the other hand, based on the Cramér-Rao Lower Bound [6] for harmonic model-based pitch estimator, the more pitch cycles (i.e., a larger segment length) we have, the higher frequency resolution of pitch estimates we can obtain. To resolve this contradiction, different segment lengths, dependent on the pitch values, are applied in SWIPE. Modifications to the standard harmonic model have been proposed, such as the harmonic chirp model [8, 18, 19], quasi-harmonic model [20, 21]. In the harmonic chirp model, instead of using a constant pitch value like the harmonic model, the pitch is considered to be either linearly increasing or decreasing over time. In the quasi-harmonic model, both the pitch and amplitudes are not constrained to be constant within a segment.

Despite the popularity of segment-wise pitch estimation methods, the pitch may present quickly even over very short segment. The ability to capture the small pitch fluctuations over time is important for voice disorder detection, such as Parkinson's disease detection [2]. Instantaneous pitch estimation algorithms have also been proposed [22–24]. Compared with the

segment-wise pitch estimation approaches, the pitch contours from instantaneous pitch estimation algorithms are smooth and no smoothing process is required [22], and a high time and frequency resolution pitch estimates can be obtained [23]. In [22], the pitch contour is obtained from the phase of each harmonic component, which is extracted using band-pass filters. In [23], Bayesian filtering based pitch tracking algorithms using the harmonic model are proposed, and a fine and continuous pitch contour is obtained. It is further shown in [24] that the dimension of the state space can be reduced by exploiting the linear relationships between the phases of different harmonics. In [25], sound signals are modelled as source-filter model with harmonic modeling of the excitation signals. The pitch contour is obtained by the Rao-Blackwellized variable rate particle filter. However, these instantaneous pitch estimation algorithms are based on a white Gaussian noise assumption, which is often not realistic in real life applications. Moreover, the extended Kalman filter method, used in [24] to deal with the nonlinear observation equation, may introduce errors and lead to states diverging over time.

In this paper, Bayesian filtering based instantaneous pitch tracking algorithms in colored noise based on the harmonic and autoregressive (AR) models are proposed. Using the harmonic model, a high robustness against noise can be obtained. The AR model is used to model the colored noise. Due to the nonlinearities of the observation and state models, the unscented Kalman filter and sequential Monte Carlo approaches are proposed to solve this problem. Fine and continuous pitch contours under very noisy and colored noise conditions is obtained using the proposed algorithms.

The rest of the paper is organized as follows. In Section 2, we briefly review general Bayesian tracking theory. In Section 3, we present the proposed observation and state evolution models, respectively. In Section 4 and 5, the unscented Kalman filter and sequential Monte Carlo methods are applied based on the proposed observation and state evolution models. Simulation results are given in Section 6.1, and the conclusions are given in Section 7.

Notation: Boldface symbols in lowercase and uppercase letters denote column vectors and matrices, respectively.

2 Bayesian tracking

In this section, we briefly review the Bayesian tracking framework. One general problem in engineering is to estimate the latent state sequence $\{\mathbf{x}_n\}, 1 \leq n \leq N$ from the noisy observation sequence $\{\mathbf{y}_n\}, 1 \leq n \leq N$ with the following observation equation

$$\mathbf{y}_n = h(\mathbf{x}_n, \mathbf{v}_n), \tag{C.1}$$

where $\mathbf{v}_n$ denotes the observation noise vector and $h(\cdot)$ is an arbitrary function, possibly nonlinear, mapping the state and noise vectors into observation vector. The state vector $\mathbf{x}_n$ is related to its previous state vector $\mathbf{x}_{n-1}$ by the state evolution equation

$$\mathbf{x}_n = f(\mathbf{x}_{n-1}, \mathbf{m}_n), \tag{C.2}$$

where $\mathbf{m}_n$ denotes the process noise. The notation $f(\cdot)$ is also an arbitrary function, possibly nonlinear, mapping the previous state and process noise vectors into the current state vector. In a Bayesian framework, the objective is to obtain the joint posterior density $p(\mathbf{X}_n|\mathbf{Y}_n)$ or the marginal posterior density $p(\mathbf{x}_n|\mathbf{Y}_n)$, where $\mathbf{X}_n$ denotes a collection of state vectors $\mathbf{X}_n = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ and $\mathbf{Y}_n$ is defined similarly to $\mathbf{X}_n$. Based on the Bayes' rule, a prediction and update recursion can be applied. In the prediction stage, the joint or marginal prediction pdfs can be obtained as

$$p(\mathbf{X}_n|\mathbf{Y}_{n-1}) = p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})p(\mathbf{x}_n|\mathbf{x}_{n-1}), \tag{C.3a}$$

$$p(\mathbf{x}_n|\mathbf{Y}_{n-1}) = \int p(\mathbf{x}_{n-1}|\mathbf{Y}_{n-1})p(\mathbf{x}_n|\mathbf{x}_{n-1})d\mathbf{x}_{n-1}, \tag{C.3b}$$

where $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ denotes the state transition pdf, which can be obtained using the state evolution equation (C.2). In the update stage, the target joint or marginal posterior pdfs can be expressed as

$$p(\mathbf{X}_n|\mathbf{Y}_n) = \frac{p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{X}_n|\mathbf{Y}_{n-1})}{p(\mathbf{y}_n|\mathbf{Y}_{n-1})}, \tag{C.4a}$$

$$p(\mathbf{x}_n|\mathbf{Y}_n) = \frac{p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{x}_n|\mathbf{Y}_{n-1})}{p(\mathbf{y}_n|\mathbf{Y}_{n-1})}, \tag{C.4b}$$

where $p(\mathbf{y}_n|\mathbf{Y}_{n-1})$ denotes the nomalization constant and $p(\mathbf{y}_n|\mathbf{x}_n)$ denotes the likelihood function, which can be obtained using the observation equation in (C.1). In most cases, closed-form solutions cannot be obtained for the prediction and update equations. However, there are at least two cases where closed-form solutions exist. In the first case, if the dynamical systems (C.1) and (C.2) are linear, and both the observation noise and process noise are Gaussian, (C.3b) and (C.4b) will reduce to the Kalman filtering algorithm. Another case is that when the state vector $\mathbf{x}_n$ is discrete and has a limited number of states, (C.3b) and (C.4b) reduces to the forward algorithm in forward-backward recursion for hidden Markov model (HMM) training. In other cases, the inference of the posterior pdfs $p(\mathbf{X}_n|\mathbf{Y}_n)$ or $p(\mathbf{x}_n|\mathbf{Y}_n)$ can be approximated using Monte Carlo approaches, such as particle filtering [26]. To use the Bayesian tracking framework for sample-by-sample pitch estimation, we define the mapping functions $h(\cdot)$ and $f(\cdot)$ based on the time-varying harmonic model and time-varying AR model in Section 3.

3 Signal models

Consider a general signal observation model, given by

$$y_n = s_n + v_n, \tag{C.5}$$

where y_n, s_n and v_n denote the observed, voiced speech/audio and additive noise signals, and n is the integer time index.

3.1 Time-varying harmonic model

We assume that the voiced speech/audio signal s_n can be modelled by a time-varying harmonic model, i.e.,

$$s_n = \sum_{k=1}^{K} A_{n,k}\cos(\theta_{n,k}), \tag{C.6}$$

$$\theta_{n,k} = k\omega_n n + \theta_{0,k}, \quad k = 1, \cdots, K, \tag{C.7}$$

where $A_{n,k}$ is the instantaneous amplitude of the k^{th} harmonic at time instant n, $\theta_{n,k}$ is the instantaneous phase, $\omega_n = 2\pi f_n / f_s$ is the instantaneous normalized digital radian frequency, f_s is the sampling rate, $\theta_{0,k}$ is the initial phase, and K is the number of harmonics. Following [24], at a typical sampling rate f_s (8 kHz or 16 kHz for speech), the instantaneous frequency of the k^{th} harmonic can be approximated by

$$k\omega_n \approx \ \theta_{n,k} - \theta_{n-1,k}. \tag{C.8}$$

Using (C.8), we can further obtain that the phases of different harmonics for $n \geq 1$ are related by

$$\theta_{n,k} = \theta_{0,k} + k((\theta_{n-1,1} - \theta_{0,1}) + \omega_n). \tag{C.9}$$

Substituting (C.9) into (C.6) and (C.7), the harmonic model can be reformulated as

$$s_n = \sum_{k=1}^{K} A_{n,k}\cos(k\omega_n + k(\theta_{n-1,1} - \theta_{0,1}) + \theta_{0,k}). \tag{C.10}$$

We collect the pitch, amplitudes and phase $\theta_{n-1,1} - \theta_{0,1}$ as the $(K+2) \times 1$ vector

$$\mathbf{x}_n^s = [\omega_n, A_{n,1}, \cdots, A_{n,K}, \theta_{n-1,1} - \theta_{0,1}]^T. \tag{C.11}$$

Combining (C.10) and (C.11), we can write the harmonic model in matrix form

$$s_n = (\mathbf{G}\mathbf{x}_n^s)^T \cos(\mathbf{B}\mathbf{x}_n^s + \boldsymbol{\theta}_0), \tag{C.12}$$

where $\mathbf{G}$ is a $K \times (K+2)$ Toeplitz matrix with first column as a zero vector and first row as $[0, 1, 0, \cdots, 0]$, $\mathbf{B}$ is a $K \times (K+2)$ zero matrix except that the first and last columns are $[1, 2, \cdots, K]^T$, and $\boldsymbol{\theta}_0 = [\theta_{0,1}, \theta_{0,2}, \cdots, \theta_{0,K}]^T$ denotes a collection of initial phases. In [24], the pitch ω_n is assumed to change in time according to a Gaussian random walk model. That is, the process noise for pitch is not constrained, and thus meaningless estimates of the pitch may be obtained (e.g., $\omega_n > \frac{f_s}{2}$). In fact, based on the applications, the pitch is usually constrained to a search range $[\omega_{\min}, \omega_{\max}]$, such as 70 to 400 Hz for speech. Another example is that when a coarse pitch estimate is obtained based on a grid search [14], a refined pitch estimate is required. The search range for the refined pitch can be even smaller, such as ± 10 Hz away from the coarse pitch estimate. To incorporate the prior pitch information in the model, we propose to use a constrained Gaussian random walk model for the pitch in this paper, i.e.,

$$\begin{aligned} &\omega_n = \omega_{n-1} + m^s_{n,1}, \\ &p(m^s_{n,1}|\omega_{n-1}) \propto \\ &\begin{cases} \mathcal{N}(m^s_{n,1}; 0, \sigma^2_\omega), & m^s_{n,1} \in [\omega_{\min} - \omega_{n-1}, \omega_{\max} - \omega_{n-1}] \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{C.13}$$

where $m^s_{n,1}$ is the process noise for the pitch with a constrained Gaussian pdf, and σ^2_ω is the variance. For the amplitudes, following [24], we assume they are changing in time according to Gaussian random walk models, i.e.,

$$A_{n,k} = A_{n-1,k} + m^s_{n,k+1}, \quad m^s_{n,k} \sim \mathcal{N}(0, \sigma^2_A), \tag{C.14}$$

where $\{m^s_{n,k}, 2 \leq k \leq K+1\}$ are i.i.d. zero mean Gaussian processes σ^2_A. It should be noted that the amplitude $A_{n,k}$ is not constrained to be positive value because of the identity $A\cos(x + \pi) = -A\cos(x)$. Based on the phase update (C.8), we have the recursive equation

$$\theta_{n-1,1} - \theta_{0,1} = \theta_{n-2,1} - \theta_{0,1} + \omega_{n-1}. \tag{C.15}$$

Based on (C.13), (C.14) and (C.15), we can write the state transition equation in matrix form

$$\mathbf{x}^s_n = \mathbf{F}\mathbf{x}^s_{n-1} + \boldsymbol{\Gamma}\mathbf{m}^s_n, \tag{C.16}$$

where $\mathbf{F}$ is a $(K+2) \times (K+2)$ lower triangular Toeplitz matrix with first column $[1, 0, \cdots, 0, 1]^T$, $\boldsymbol{\Gamma}$ is a $(K+2) \times (K+1)$ Toeplitz matrix with first column $[1, 0, \cdots, 0]^T$ and the first row as $[1, 0, \cdots, 0]$, and the noise vector is defined as $\mathbf{m}^s_n = [m^s_{n,1}, m^s_{n,2}, \cdots, m^s_{n,K+1}]^T$.

3.2 Time-varying AR model for noise

We assume the noise signal v_n can be expressed as a combination of white Gaussian noise and colored noise, i.e.,

$$v_n = v_n^u + v_n^c, v_n^u \sim \mathcal{N}(0, \sigma_u^2), \tag{C.17}$$

where v_n^u and v_n^c denote the white Gaussian noise with variance σ_u^2 and colored noise, respectively. The noise signal v_n^c is further modeled as a time-varying autoregressive (TVAR) model with order P, i.e.,

$$v_n^c = \sum_{p=1}^{P} a_{n,p} v_{n-p}^c + m_n^c, \quad m_n^c \sim \mathcal{N}(0, \sigma_c^2), \tag{C.18}$$

where $a_{n,p}, 1 \leq p \leq P$ are the TVAR coefficients, and m_n^c denotes the white Gaussian excitation with variance σ_c^2. We collect the colored noise signals as a $P \times 1$ vector, i.e.,

$$\mathbf{x}_n^c = \left[v_n^c, \cdots, v_{n-P+1}^c\right]^T. \tag{C.19}$$

Here, instead of using the TVAR coefficients $\mathbf{a}_n = [a_{n,1}, \cdots a_{n,P}]^T$ directly, we use the reflection coefficients $\mathbf{x}_n^a = g(\mathbf{a}_n)$, where $g(\cdot)$ is a nonlinear mapping function to transform the AR coefficients to the reflection coefficients. To obtain a stable AR model, the reflection coefficients are constrained in the range $[-1+\delta, 1-\delta]$ [27], where δ is a small positive number[1]. Using (C.19), the TVAR model (C.18) can be expressed into matrix form as

$$\mathbf{x}_n^c = \mathbf{f}(g^{-1}(\mathbf{x}_n^a))\mathbf{x}_{n-1}^c + \mathbf{c}m_n^c, \tag{C.20}$$

where

$$\mathbf{f}(g^{-1}(\mathbf{x}_n^a)) = \begin{bmatrix} a_{n,1} & \cdots & a_{n,P-1} & a_{n,P} \\ & \mathbf{I}_{P-1} & & \mathbf{0} \end{bmatrix}, \tag{C.21}$$

$$\mathbf{c} = \left[1, \ \mathbf{0}_{1\times(P-1)}\right]^T, \tag{C.22}$$

and $g^{-1}(\cdot)$ denotes the inverse function of $g(\cdot)$. To impose the stability constraint, similar to (C.13), we assume the reflection coefficients are changing in time according to constrained Gaussian random walk models, i.e.,

$$\begin{aligned} & x_{p,n}^a = x_{p,n-1}^a + m_{p,n}^a, \\ & p(m_{p,n}^a | x_{p,n-1}^a) \propto \\ & \begin{cases} \mathcal{N}(m_{p,n}^a; 0, \sigma_a^2), m_{p,n}^a \in \left[-1+\delta-x_{p,n-1}^a, 1-\delta-x_{p,n-1}^a\right] \\ 0, \qquad\qquad \text{otherwise}, \end{cases} \end{aligned} \tag{C.23}$$

[1] In this paper, we set $\delta = 10^{-4}$ for all the experiments in Section 6.1.

where $x^a_{p,n}$ denote the p^{th} element of the vector $\mathbf{x}^a_n$ and the process noises for the AR reflection coefficients are collected as $\mathbf{m}^a_n = \left[m^a_{1,n}, \cdots, m^a_{P,n}\right]^T$.

3.3 Noisy signal model

Collecting the vectors $\mathbf{x}^s_n$, $\mathbf{x}^c_n$ and $\mathbf{x}^a_n$, the state vector can be expressed as an $(K + 2P + 2) \times 1$ vector, i.e.,

$$\mathbf{x}_n = [\mathbf{x}^{s\,T}_n, \mathbf{x}^{c\,T}_n, \mathbf{x}^{a\,T}_n]^T. \tag{C.24}$$

Combing (C.16), (C.20), (C.23) and (C.24), the state equation can be expressed as

$$\begin{aligned} \mathbf{x}_n &= \text{diag}\left[\mathbf{F}, \mathbf{f}(g^{-1}(\mathbf{x}^a_n)), \mathbf{I}_P\right] \mathbf{x}_{n-1} + \mathbf{D}\mathbf{m}_n, \\ \mathbf{D} &= \text{diag}\left[\mathbf{\Gamma}, \mathbf{C}, \mathbf{I}_P\right], \mathbf{m}_n = \left[\mathbf{m}^{s\,T}_n, m^c_n, \mathbf{m}^{a\,T}_n\right]^T, \end{aligned} \tag{C.25}$$

where diag $[\cdot]$ denotes the block diagonal operator. The covariance matrix for the process noise vector $\mathbf{m}_n$ can be expressed as

$$\mathbf{Q}_m = \text{diag}\left[\sigma^2_\omega, \sigma^2_A \mathbf{I}_K, \sigma^2_c, \sigma^2_a \mathbf{I}_P\right].$$

Using (C.5), (C.12), (C.17) and (C.19), the observation equation can be expressed as

$$y_n = (\mathbf{G}\mathbf{x}^s_n)^T \cos(\mathbf{B}\mathbf{x}^s_n + \boldsymbol{\theta}_0) + \mathbf{C}^T\mathbf{x}^c_n + v^u_n. \tag{C.26}$$

As can be seen from (C.25) and (C.26), both the state evolution and observation equations are nonlinear. Moreover, the probability distributions of the process noises of the pitch and AR reflection coefficients are constrained Gaussian. Therefore, a traditional Kalman filter cannot be used. In [24], the nonlinear observation equation (C.12) is linearized using the first order Taylor expansion, and the pitch is estimated using the extended Kalman filter (EKF). However, this cannot simply be done here since it is difficult to calculate the Jacobian matrices for (C.20) (if not impossible). Moreover, using the EKF, the linear approximation of the system may introduce errors leading to states diverging over time. To deal with the nonlinearity problem, unscented Kalman filter (UKF) based on deterministic sampling technique is commonly applied [28]. It is shown in [29] that UKF produces more accurate results than the EKF. However, UKF has the limitation that it does not apply to general non-Gaussian distributions. A general approach to address the general nonlinearity and non-Gaussian problems is the particle filters (PFs), also known as sequential Monte Carlo (SMC) [30, 31]. In this paper, we propose a state constrained UKF pitch tracking algorithm in Section 4 and two SMC pitch tracking algorithms in 5 based on (C.25) and (C.26).

4 Unscented Kalman filter-based pitch tracking

The aforementioned nonlinearity problem can be mitigated by using an UKF, which uses a deterministic sampling approach. In the UKF framework, the state distribution is represented using a minimal set of carefully chosen sample points, known as the sigma points. When these points are propagated through a nonlinear system, using the unscented transform, the posterior mean and covariance can be captured accurately up to the third order (Taylor series expansion) for any nonlinearity [32].

4.1 Sigma points and unscented transform method

Next, we briefly review the definition of sigma points and the unscented transform method. Assuming that the random variable $\mathbf{x}$ with dimension L has a known mean $\widehat{\mathbf{x}}$ and covariance $\widehat{\mathbf{P}}$, the sigma points can be expressed as a matrix $\mathcal{X}$ of $2L+1$ sigma vectors with corresponding weighting vectors $\mathbf{w}^m$ and $\mathbf{w}^c$ defined as

$$\begin{aligned}
\mathcal{X} &= \left[\widehat{\mathbf{x}}, \widehat{\mathbf{x}} \pm \sqrt{(L+\lambda)\widehat{\mathbf{P}}}\right], \\
\mathbf{w}^m &= \left[\frac{\lambda}{(L+\lambda)}, \frac{1}{2(L+\lambda)}\mathbf{1}_{2L}^T\right]^T, \\
\mathbf{w}^c &= \left[\frac{\lambda}{(L+\lambda)} + (1-\alpha^2+\beta), \frac{1}{2(L+\lambda)}\mathbf{1}_{2L}^T\right]^T, \\
\lambda &= \alpha^2(L+\kappa) - L.
\end{aligned} \tag{C.27}$$

where α is usually set to a small positive value (e.g., 1e-3). κ is a scaling parameter usually set to 0, β is used to add prior knowledge of the distribution of $\mathbf{x}$ ($\beta = 2$ is optimal for Gaussian distribution), $\mathbf{1}_{2L}$ denotes an all ones column vector with dimension $2L$. Then, these sigma points are propagated through a nonlinear system $\mathbf{y} = f(\mathbf{x})$ to obtain the transformed sigma points $\mathcal{Y}$. The mean and covariance for the random variable $\mathbf{y}$ can be expressed as

$$\widehat{\mathbf{y}} = \mathcal{Y}\mathbf{w}^m, \;\; \widehat{\mathbf{P}}_y = (\mathcal{Y} - \widehat{\mathbf{y}}\mathbf{1})\mathrm{diag}(\mathbf{w}^c)(\mathcal{Y} - \widehat{\mathbf{y}}\mathbf{1})^T. \tag{C.28}$$

where $\mathbf{1}$ is an all ones row vector conformal to the column dimensions of $\mathcal{Y}$.

4.2 State constrains

Although the standard UKF algorithm is designed for dealing with nonlinear system, the state constraint issues are not considered. As we discussed before, the pitch should be constrained to a range based on the prior information, and the AR reflection coefficients should be constrained for model

stability. In this paper, a simple projection of the posterior mean is used, i.e.,

$$\widehat{\mathbf{x}}^{\mathrm{C}} = \mathcal{P}(\widehat{\mathbf{x}}). \tag{C.29}$$

where $\mathcal{P}(\cdot)$ denotes an element-wise projection function. Assume that the i^{th} element x_i in state vector $\mathbf{x}$ is subject to a box constraint $a_i \leq x_i \leq b_i$, then

$$x_i^{\mathrm{C}} = \mathcal{P}(x_i) = \begin{cases} x_i, & \text{for } a_i \leq x_i \leq b_i \\ a_i, & \text{for } x_i < a_i \\ b_i, & \text{for } b_i < x_i. \end{cases} \tag{C.30}$$

As can be seen from the state vector definition in (C.11) and (C.24), the pitch is contained in the 1^{st} element of $\mathbf{x}_n$ and the AR reflection coefficients are contained in the last P elements. The lower and upper boundaries for pitch and AR coefficients are set to $\omega_{\min}$ and $\omega_{\max}$, $-1+\delta$ and $1-\delta$, respectively. Moreover, as can be seen from (C.15), the state variable $\theta_{n-1,1} - \theta_{0,1}$ denotes the phase and it increases over time. To avoid overflow, it is wrapped to the range $[0, 2\pi]$. The proposed instantaneous pitch estimation algorithm using state constrained UKF is outlined in Algorithm 3.

Algorithm 3 Unscented Kalman Filter with constrained states

1: Initiate Harmonic order K, AR order P, initial phase $\boldsymbol{\theta}_0$, state mean $\hat{\mathbf{x}}_0$ and covariance matrix $\widehat{\mathbf{P}}_0$, and set $\hat{\mathbf{x}}_0^{\mathrm{C}} = \hat{\mathbf{x}}_0$
2: **for** $n = 1, 2, \cdots, \infty$ **do**
3: Construct augmented state vector and covariance matrix

$$\mathbf{x}'_n = [(\hat{\mathbf{x}}_{n-1}^{\mathrm{C}})^T, \mathbf{0}_{1\times(K+P+2)}, 0]^T,$$
$$\mathbf{P}'_n = \mathrm{diag}\left[\widehat{\mathbf{P}}_{n-1}, \mathbf{Q}_m, \sigma_u^2\right].$$

4: Calculate sigma points $\mathcal{X}'_{n-1}$ and and weighting vectors using (C.27) with $L = 2K + 3P + 5$.
5: Substituting $\mathcal{X}'_{n-1}$ to (C.25) to obtain transformed sigma points $\mathcal{X}_{n|n-1}$.
6: Calculate the mean $\hat{\mathbf{x}}_{n|n-1}$ and covariance $\widehat{\mathbf{P}}_{n|n-1}$ of the predicted state distribution of $\mathbf{x}_n$ by replacing $\hat{\mathbf{y}}$ with $\hat{\mathbf{x}}_{n|n-1}$, $\widehat{\mathbf{P}}_y$ with $\widehat{\mathbf{P}}_{n|n-1}$, and $\mathcal{Y}$ with $\mathcal{X}_{n|n-1}$ in (C.28).
7: Substituting $\mathcal{X}_{n|n-1}$ and the sub-matrix of $\mathcal{X}_{n|n-1}$ corresponding to the white noise v_n^u to (C.26) to obtain transformed sigma points $\mathcal{Y}_n$.
8: Calculate the mean $\hat{y}_n$ and covariance $\mathbf{P}_y$ of $\mathbf{y}_n$ by replacing $\hat{\mathbf{y}}$ with $\hat{y}_n$, $\widehat{\mathbf{P}}_y$ with $\mathbf{P}_y$ of $\mathbf{y}_n$ and $\mathcal{Y}$ with $\mathcal{Y}_n$ in (C.28).
9: Calculate the cross covariance matrix

$$\mathbf{P}_{xy} = (\mathcal{X}_{n|n-1} - \hat{\mathbf{x}}_{n|n-1}\mathbf{1})\times \mathrm{diag}(\mathbf{w}^c)(\mathcal{Y}_n - \hat{y}_n\mathbf{1})^T. \quad \text{(C.31)}$$

10: Calculate the Kalman gain and update the state statistics:

$$\mathbf{K}_n = \mathbf{P}_{xy}{\mathbf{P}_y}^{-1}, \quad \text{(C.32)}$$
$$\hat{\mathbf{x}}_n = \hat{\mathbf{x}}_{n|n-1} + \mathbf{K}_n(y_n - \hat{y}_n), \quad \text{(C.33)}$$
$$\widehat{\mathbf{P}}_n = \widehat{\mathbf{P}}_{n|n-1} - \mathbf{K}_n\mathbf{P}\mathbf{K}_n^T \quad \text{(C.34)}$$

11: Project $\hat{\mathbf{x}}_n$ to obtain the constrained state mean estimates $\hat{\mathbf{x}}_n^{\mathrm{C}}$ using (C.29).
12: Wrap the phase in $\hat{\mathbf{x}}_n^{\mathrm{C}}$ to the range $[0, 2\pi]$.
13: **end for**

5 Sequential Monte Carlo-based pitch tracking

In Section 4, state constrained unscented Kalman filter is used for dealing with the nonlinear mapping functions and constrained Gaussian process noise problems by using deterministic sampling techniques. As an alternative, the sequential Monte Carlo approaches [33] are proposed based on the

prediction equation (C.3a), and update equation (C.4a). Instead of imposing constraints to the states for the UKF, the SMC-based approaches can deal with non-Gaussian distributions naturally. However, they are more computational and storage complex than UKF-based approaches.

5.1 Pitch tracking using standard particle filter

Assume that the joint pdf $p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})$ cannot be obtained analytically, but it can be evaluated up to a constant number. Let a set of samples $\{\mathbf{X}_{n-1}^i\}_{i=1}^{N_s}$ with associated weights $\{w_{n-1}^i\}_{i=1}^{N_s}$ that characterize the joint posterior pdf $p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})$, where N_s denotes the number of particles and the weights are normalized such that $\sum_{i=1}^{N_s} w_{n-1}^i = 1$. Assume that weights are obtained using the importance sampling technique [10], i.e.,

$$w_{n-1}^i \propto \frac{p(\mathbf{X}_{n-1}^i|\mathbf{Y}_{n-1})}{q(\mathbf{X}_{n-1}^i|\mathbf{Y}_{n-1})}, \tag{C.35}$$

where $q(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})$ is often referred as an important density, which is easier to generate samples than $p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})$. Then, the joint pdf $p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})$ can be approximated as

$$p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1}) \approx \sum_{i=1}^{Ns} w_n^i \delta(\mathbf{X}_{n-1} - \mathbf{X}_{n-1}^i). \tag{C.36}$$

Combing (C.3a) and (C.4a), the joint posterior can be expressed as

$$\begin{aligned} p(\mathbf{X}_n|\mathbf{Y}_n) &= p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1}) \frac{p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{x}_n|\mathbf{x}_{n-1})}{p(\mathbf{y}_n|\mathbf{Y}_{n-1})} \\ &= p(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n). \end{aligned} \tag{C.37}$$

If the important density is constructed to have a factorized form similar to the target form (C.37), i.e.,

$$q(\mathbf{X}_n|\mathbf{Y}_n) = q(\mathbf{X}_{n-1}|\mathbf{Y}_{n-1})q(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n), \tag{C.38}$$

then the sample $\mathbf{X}_n^i$ can be obtained by augmenting the existing sample $\mathbf{X}_{n-1}^i$ with the new sample drawn from $\mathbf{x}_n^i \sim q(\mathbf{x}_n|\mathbf{x}_{n-1}^i, \mathbf{y}_n)$. Combining (C.35), (C.37) and (C.38), the weights update equation can be expressed as

$$w_n^i \propto w_{n-1}^i \frac{p(\mathbf{x}_n^i|\mathbf{x}_{n-1}^i)p(\mathbf{y}_n|\mathbf{x}_n^i)}{q(\mathbf{x}_n^i|\mathbf{x}_{n-1}^i, \mathbf{y}_n)}. \tag{C.39}$$

However, the above algorithm suffers from degeneracy problem since we are sampling from a high dimensional state space, i.e., the entire history

of the state sequence. To mitigate this problem, a resampling step is included to eliminate the particles with small weights [33]. As can be seen from (C.37) and (C.38), when $q(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)$, the important density is optimal. However, due to the nonlinearities and non-Gaussian properties of the state evolution and observation models (C.25) and (C.26), $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)$ cannot be evaluated analytically. Instead, in this paper, we only use the transition pdf from the state evolution equation (C.25), i.e. $q(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$ [30], and the weights update equation reduces to $w_n^i \propto w_{n-1}^i p(\mathbf{y}_n|\mathbf{x}_n^i)$. The standard particle filter algorithm for the proposed state transition and observation models (C.25) and (C.26) is outlined in Algorithm 4.

Algorithm 4 Particle filter

1: **for** $n = 1, 2, \cdots, \infty$ **do**
2: **for** $i = 1, 2, \cdots, Ns$ **do**
3: Substitute ${\omega_{n-1}}^i$ to (C.15) to obtain $(\theta_{n-1,1} - \theta_{0,1})^i$.
4: Draw pitch ${\omega_n}^i$ based on (C.13) using rejection sampling.
5: Draw amplitudes ${A_{n,k}}^i$, $1 \leq k \leq K$ based on (C.14).
6: Draw AR reflection coefficients $\mathbf{x}_n^{a\,i}$ based on (C.23) using rejection sampling.
7: Substitute $\mathbf{x}_n^{a\,i}$ to (C.21) and draw colored noise vector $\mathbf{x}_n^{a\,i}$ based on (C.20).
8: Evaluate the likelihood function $p(y_n|\mathbf{x}_n^i)$ using (C.26).
9: Compute weight w_n^i using $w_n^i \propto w_{n-1}^i p(y_n|\mathbf{x}_n^i)$.
10: Wrap the phase in $\widehat{\mathbf{x}}_n^i$ to the range $[0, 2\pi]$.
11: **end for**
12: Normalize the weights $w_n^i = \frac{w_n^i}{\sum_{i=1}^{Ns} w_n^i}$
13: Resampling algorithm
14: **end for**

5.2 Pitch estimation using Rao-Blackwellized particle filter

The inefficiency in sampling in a high-dimensional space is one of the major drawbacks of the standard particle filtering algorithm. However, in some cases, the dimension of the state space one need to sample from can be reduced when some state variables can be analytically marginalized out, conditioned on other stater variables. Assume that the state matrix $\mathbf{X}_n$ can be partitioned as $[\mathbf{Z}_n, \mathbf{\Phi}_n]$ and $\mathbf{Z}_n$ can be marginalized out analytically conditioned on $\mathbf{\Phi}_n$, where $\mathbf{Z}_n$ and $\mathbf{\Phi}_n$ are defined similarly to $\mathbf{X}_n$. We also assume that $\boldsymbol{\phi}_n$ and $\mathbf{z}_{n-1}$ are conditionally independent given $\boldsymbol{\phi}_{n-1}$, and $\mathbf{z}_n$ and $\boldsymbol{\phi}_{n-1}$

are conditionally independent given $\mathbf{z}_{n-1}$ and $\boldsymbol{\phi}_n$, i.e.,

$$\begin{aligned} p(\boldsymbol{\phi}_n|\mathbf{z}_{n-1}, \boldsymbol{\phi}_{n-1}) &= p(\boldsymbol{\phi}_n|\boldsymbol{\phi}_{n-1}), \\ p(\mathbf{z}_n|\mathbf{z}_{n-1}, \boldsymbol{\phi}_n, \boldsymbol{\phi}_{n-1}) &= p(\mathbf{z}_n|\mathbf{z}_{n-1}, \boldsymbol{\phi}_n). \end{aligned} \tag{C.40}$$

Our objective is to obtain a particle approximation of the marginal joint posterior distribution

$$p(\boldsymbol{\Phi}_n|\mathbf{Y}_n) = \int p(\mathbf{Z}_n, \boldsymbol{\Phi}_n|\mathbf{Y}_n) d\mathbf{Z}_n. \tag{C.41}$$

Substituting (C.3a), (C.4a) and (C.40) into (C.41), we can obtain

$$\begin{aligned} & p(\boldsymbol{\Phi}_n|\mathbf{Y}_n) \\ \propto & \int p(\mathbf{Z}_{n-1}, \boldsymbol{\Phi}_{n-1}|\mathbf{Y}_{n-1}) \times \\ & p(\mathbf{z}_n, \boldsymbol{\phi}_n|\mathbf{z}_{n-1}, \boldsymbol{\phi}_{n-1}) p(\mathbf{y}_n|\mathbf{z}_n, \boldsymbol{\phi}_n) d\mathbf{Z}_n \\ = & p(\boldsymbol{\Phi}_{n-1}|\mathbf{Y}_{n-1}) p(\boldsymbol{\phi}_n|\boldsymbol{\phi}_{n-1}) p(\mathbf{y}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n) \end{aligned} \tag{C.42}$$

where we used

$$p(\mathbf{y}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n) = \int p(\mathbf{y}_n|\mathbf{z}_n, \boldsymbol{\phi}_n) p(\mathbf{z}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n) d\mathbf{z}_n, \tag{C.43}$$

$$\begin{aligned} & p(\mathbf{z}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n) \\ = & \int p(\mathbf{z}_n|\mathbf{z}_{n-1}, \boldsymbol{\phi}_n) p(\mathbf{z}_{n-1}|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_{n-1}) d\mathbf{z}_{n-1}. \end{aligned} \tag{C.44}$$

If the integrations (C.43) and (C.44) can be evaluated analytically, similar to the derivations of particle filter (C.38) and (C.39), the sample $\boldsymbol{\Phi}_n^i$ can be obtained by augmenting the existing sample $\boldsymbol{\Phi}_{n-1}^i$ with the new sample drawn from $\boldsymbol{\phi}_n^i \sim q(\boldsymbol{\phi}_n^i|\boldsymbol{\phi}_{n-1}^i, \mathbf{y}_n)$. The weights update equation for the Rao-Blackwellized particle filter can be expressed as

$$w_n^i \propto w_{n-1}^i \frac{p(\boldsymbol{\phi}_n^i|\boldsymbol{\phi}_{n-1}^i) p(\mathbf{y}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n^i)}{q(\boldsymbol{\phi}_n^i|\boldsymbol{\phi}_{n-1}^i, \mathbf{y}_n)}. \tag{C.45}$$

Returning to the proposed signal models (C.25) and (C.26), $\mathbf{z}_n$ can be seen as a collection of amplitudes $A_{n,k}, 1 \leq k \leq K$ and colored noise vector $\mathbf{x}_n^c$, i.e., $\mathbf{z}_n = [A_{n,1}, \cdots, A_{n,K}, \mathbf{x}_n^{c\,T}]^T$. $\boldsymbol{\phi}_n$ can be seen as a collection of the pitch ω_n, phase $\theta_{n-1,1} - \theta_{0,1}$ and AR reflection coefficients $\mathbf{x}_n^a$, i.e., $\boldsymbol{\phi}_n = [\omega_n, \theta_{n-1,1} - \theta_{0,1}, \mathbf{x}_n^{a\,T}]^T$. Using the above definitions, the observation equation (C.26) can be expressed as

$$\begin{aligned} y_n =& (\mathbf{G}_1 \mathbf{z}_n)^T \cos(\mathbf{B}_1 \boldsymbol{\phi}_n + \boldsymbol{\theta}_0) + \mathbf{C}_1^T \mathbf{z}_n + v_n^u \\ =& (\mathbf{G}_1^T \cos(\mathbf{B}_1 \boldsymbol{\phi}_n + \boldsymbol{\theta}_0) + \mathbf{C}_1)^T \mathbf{z}_n + v_n^u, \end{aligned} \tag{C.46}$$

where $\mathbf{G}_1 = [\mathbf{I}_K, \mathbf{0}_{K\times P}]$, $\mathbf{B}$ is a $K \times (2+P)$ zero matrix except that the first and second columns are $[1, 2, \cdots, K]^T$, and $\mathbf{C}_1 = [\mathbf{0}_{1\times K},\ 1,\ \mathbf{0}_{1\times(P-1)}]^T$. The state evolution equation (C.25) for $\mathbf{z}_n$ can be expressed as

$$\mathbf{z}_n = \text{diag}\left[\mathbf{I}_K, \mathbf{f}(g^{-1}(\mathbf{F}_1\boldsymbol{\phi}_n)\right]\mathbf{z}_{n-1} + \mathbf{D}_1\mathbf{m}_n^{\text{linear}}, \tag{C.47}$$

where $\mathbf{F}_1 = [\mathbf{0}_{P\times 2}, \mathbf{I}_P]$, $\mathbf{m}_n^{\text{linear}} = [m_{n,2}^s, \cdots, m_{n,K+1}^s, m_n^c]$, $\mathbf{D}_1 = \text{diag}\,[\mathbf{I}_K, \mathbf{C}]$. The state evolution equations for elements of $\boldsymbol{\phi}_n$ are shown in (C.13), (C.15) and (C.23). As can be seen, given $\boldsymbol{\phi}_n$, the observation equation (C.46) and state evolution equation (C.47) are linear dynamical systems, and both the observation noise and process noise are Gaussian. Therefore, the integrations in (C.43) and (C.44) can be computed analytically using the one-step-ahead prediction of the Kalman filter algorithm and the prediction error decomposition. For completeness, the prediction, update and prediction error decomposition equations for the Kalman filter are given in the Appendix 8. The Rao-Blackwellized particle filter algorithm for the proposed model is shown in Algorithm 5.

Algorithm 5 Rao-Blackwellized particle filter

for $n = 1, 2, \cdots, \infty$ **do**
  for $i = 1, 2, \cdots, Ns$ **do**
    Substitute ${\omega_{n-1}}^i$ to (C.15) to obtain $(\theta_{n-1,1} - \theta_{0,1})^i$.
    Draw pitch ${\omega_n}^i$ based on (C.13) using rejection sampling.
    Draw AR reflection coefficients $\mathbf{x}_n^{a\,i}$ based on (C.23) using rejection sampling.
    Construct the vector $\boldsymbol{\phi}_n^i$, and substitute $\boldsymbol{\phi}_n^i$ to (C.46) and (C.47).
    Compute $p(\mathbf{z}_n^i|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n^i)$ using (C.47) and (C.52).
    Compute $p(\mathbf{y}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n^i)$ using (C.46) and (C.53).
    Compute $p(\mathbf{z}_n^i|\mathbf{Y}_n, \boldsymbol{\Phi}_n^i)$ using (C.46) and (C.54).
    Compute weight w_n^i using (C.45).
    Wrap the phase $(\theta_{n-1,1} - \theta_{0,1})^i$ to the range $[0, 2\pi]$.
  end for
  Normalize the weights $w_n^i = \frac{w_n^i}{\sum_{i=1}^{Ns} w_n^i}$
  Resampling algorithm
end for

6 Experimental results

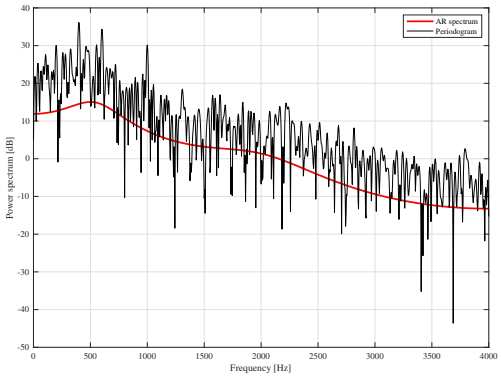


Fig. C.1: AR spectrum and the periodogram of the noisy signal.

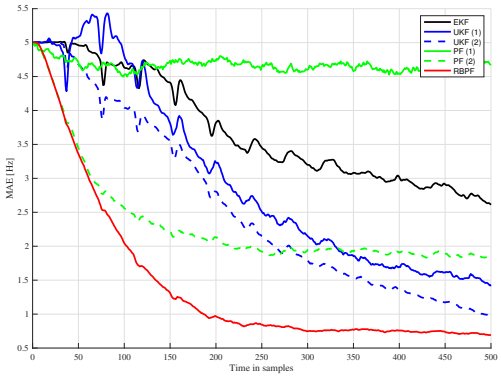


Fig. C.2: Pitch tracking curves for different algorithms in 0 dB colored noise.

6.1 Synthetic signal experiments

In this section, we test the performance of the proposed algorithms on synthetic signals.

First, the proposed approaches are tested on synthetic signals in pitch refining application. The noisy signal is generated based on the (C.25) and (C.26). To generate the synthetic periodic signals based on the time-varying harmonic model in section 3.1, we set the hyper-parameters $\sigma_\omega^2 = 0$ (pitch is constant), $\sigma_A^2 = 10^{-6}$. Using a constant pitch value, we are aiming to evaluate the performance of the proposed algorithms on pitch refining application, where a coarse pitch value is obtained by a grid-based approach [14, 34]. The harmonic order is set to $K = 5$, the initial amplitudes for all the harmonics are randomly generated based on the Gaussian distribution with variance 1, and the sampling frequency is 8000 Hz. The true pitch is set to 200 Hz. To generate the colored noise using the time-varying AR model in section 3.2,

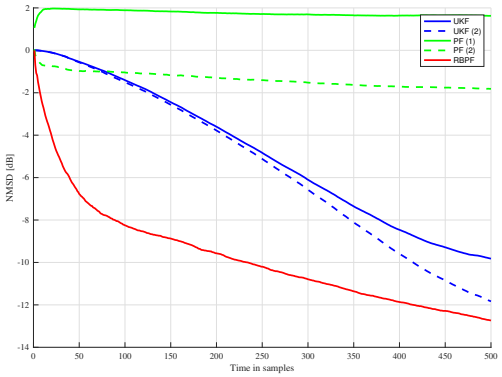


Fig. C.3: AR tracking curves for different algorithms in 0 dB colored noise.

we set the noise excitation variance $\sigma_c^2 = 1$, $\sigma_a^2 = 0$ (AR reflection coefficients are constant). The initial AR coefficients are obtained from 30 ms babble noise with AR model order $P = 4$ using the Levinson-Durbin algorithm [35]. The white noise variance is set to $\sigma_u^2 = 0$ (i.e., white Gaussian noise is not present). The SNR is set to 0 dB. The pitch tracking performance of the EKF [24], UKF, PF and RBPF is evaluated by the mean absolute error (MAE) [5]:

$$d_n^{\text{MAE}} = \left| \frac{(\hat{\omega}_n - \omega_n^{\text{true}})}{2\pi/f_s} \right|, \tag{C.48}$$

where ω_n^{true} and $\hat{\omega}_n$ denote the true and estimated pitch values. The AR tracking performance of UKF, PF and RBPF is evaluated by the normalized mean square deviation (NMSD), defined as

$$d_n^{\text{NMSD}} = \frac{\|\hat{\mathbf{x}}_n^a - \mathbf{x}_n^{a\,\text{true}}\|_2}{\|\mathbf{x}_n^{a\,\text{true}}\|_2}, \tag{C.49}$$

where $\|\cdot\|_2$ denotes the 2-norm, $\mathbf{x}_n^{a\,\text{true}}$ and $\hat{\mathbf{x}}_n^a$ are the true and estimated AR reflection vector, respectively. To initialize these algorithms, we assume a coarse estimate of the pitch at 205 Hz is obtained. The harmonic order and AR order are set to the same as the true model, i.e., 5 and 4, respectively. With an estimated pitch and white Gaussian assumption on the noise v_n, the amplitudes and initial phase are estimated using least squares method based on the first 240 samples (30 ms). The hyper-parameters are set to

EKF [24]: noise variance is set to 1, $\hat{\sigma}_\omega^2 = 5 \times 10^{-7}$, $\hat{\sigma}_A^2 = 10^{-6}$;

UKF (1): $\hat{\sigma}_c^2 = 0.1$, $\hat{\sigma}_A^2 = 10^{-6}$, $\hat{\sigma_a}^2 = 10^{-4}$, $\hat{\sigma}_u^2 = 10^{-4}$, $\hat{\sigma}_\omega^2 = 5 \times 10^{-7}$, $\alpha = 0.1$, $\beta = 2$, $\kappa = 0$;

UKF (2): $\hat{\sigma}_c^2 = 0.1$, $\hat{\sigma}_A^2 = 10^{-6}$, $\hat{\sigma_a}^2 = 10^{-4}$, $\hat{\sigma}_u^2 = 10^{-4}$, $\hat{\sigma}_\omega^2 = 10^{-7}$, $\alpha = 0.1$, $\beta = 2$, $\kappa = 0$;

PF (1): $\hat{\sigma}_c^2 = 0.1$, $\hat{\sigma}_A^2 = 10^{-6}$, $\hat{\sigma_a}^2 = 10^{-4}$, $\hat{\sigma}_u^2 = 10^{-4}$, $\hat{\sigma}_\omega^2 = 5 \times 10^{-7}$;

PF (2): $\hat{\sigma}_c^2 = 0.1$, $\hat{\sigma}_A^2 = 10^{-6}$, $\hat{\sigma_a}^2 = 10^{-4}$, $\hat{\sigma}_u^2 = 0.1$, $\hat{\sigma}_\omega^2 = 5 \times 10^{-7}$;
RBPF: $\hat{\sigma}_c^2 = 0.1$, $\hat{\sigma}_A^2 = 10^{-6}$, $\hat{\sigma_a}^2 = 10^{-4}$, $\hat{\sigma}_u^2 = 10^{-4}$, $\hat{\sigma}_\omega^2 = 5 \times 10^{-7}$.

The pitch is constrained to be in the range $[\omega_{\min}, \omega_{\max}] = 2\pi[-10 + 205, 10 + 205]/f_s$ for all the algorithms. The number of Monte Carlo experiments is set to 1000. The number of particles used for PF and RBPF are set to 200. The AR spectral model for generating the colored noise and the periodogram of the noisy observation from one Monte Carlo trial are shown in Figure C.1. The pitch tracking RMSE curves for EKF, UKF, PF and RBPF are shown in Figure C.2. The AR spectral distortion tracking curves for UKF, PF and RBPF are shown in Figure C.3. As can be seen from Figure C.2, the particle filtering algorithm with a larger white Gaussian noise variance $\hat{\sigma}_u^2$ (i.e., PF (2)) performs better than using a smaller value (i.e., PF (1)). By using AR modeling of the noise, the PF (2), UKF (1), UKF (2) and the RBPF achieve a faster convergence rate than the EKF algorithm. The RBPF algorithm has the fastest convergence rate. As can be seen from Figure C.3, the UKF algorithms have a better AR tracking performance than the PF algorithms. Moreover, the RBPF has the best performance.

Table C.1: MAE [Hz], MRE [%] and RMSE [Hz] in babble noise for "Why were you away a year, Roy?" from a male speaker

SNR		-10.00	-5.00	0.00	5.00	10.00
EKF	MAE	20.88	24.15	16.94	2.33	**0.75**
	MRE	16.22	18.82	13.63	1.95	**0.63**
	RMSE	23.85	27.05	19.13	3.13	**1.08**
UKF	MAE	21.82	26.76	29.11	17.52	4.31
	MRE	16.96	20.93	23.11	14.21	3.64
	RMSE	24.67	29.33	31.46	19.60	5.70
PF	MAE	22.17	14.71	6.99	2.70	1.82
	MRE	18.58	12.55	6.05	2.25	1.51
	RMSE	26.53	19.52	10.13	3.75	2.44
RBPF	MAE	**13.79**	**6.07**	**1.79**	**1.02**	0.89
	MRE	**11.98**	**5.49**	**1.55**	**0.84**	0.72
	RMSE	**18.09**	**9.34**	**2.69**	**1.34**	1.18

6.2 Results on audio and speech examples

In this subsection, we illustrate the performance of different pitch estimation algorithms using one speech and one audio sample.

First, the performance of the EKF and RBPF is tested on a speech signal of the spoken sentence "Why were you away a year, Roy?" uttered by a male speaker and sampled at $f_s = 8000$ Hz. The pitch is constrained to be in the range $[\omega_{\min}, \omega_{\max}] = 2\pi[70, 400]/f_s$, and $\hat{\sigma}_\omega^2 = 10^{-7}$, $\hat{\sigma}_A^2 = 10^{-5}$ and $\hat{\sigma}_a^2 = 10^{-5}$ (when applicable) are used for all the algorithms. The number of harmonics is set to $K = 5$. The noise variance for EKF is set to 1. The

Table C.2: MAE [Hz], MRE [%] and RMSE [Hz] in factory noise for "Why were you away a year, Roy?" from a male speaker

SNR		-10.00	-5.00	0.00	5.00	10.00
EKF	MAE	53.76	52.50	26.23	2.81	**0.56**
	MRE	42.16	41.33	21.87	2.62	**0.47**
	RMSE	56.00	54.50	31.47	5.33	**0.81**
UKF	MAE	48.69	46.50	41.16	17.76	3.73
	MRE	38.36	36.88	32.97	14.64	3.23
	RMSE	50.61	48.17	43.03	20.50	5.38
PF	MAE	**23.91**	18.05	8.68	3.09	1.69
	MRE	**19.33**	14.80	7.31	2.62	1.44
	RMSE	**27.45**	21.93	11.99	4.52	2.44
RBPF	MAE	28.73	**4.85**	**1.65**	**0.96**	0.86
	MRE	22.92	**4.24**	**1.46**	**0.79**	0.70
	RMSE	31.04	**7.08**	**2.43**	**1.25**	1.13

Table C.3: MAE [Hz], MRE [%] and RMSE [Hz] in babble noise for vibrato sound of a flute at note C5

SNR		-10.00	-5.00	0.00	5.00	10.00
EKF	MAE	41.08	2.95	1.22	0.65	0.34
	MRE	7.77	0.56	0.23	0.12	0.07
	RMSE	52.08	4.07	1.55	0.84	0.44
UKF	MAE	12.41	7.46	2.03	**0.56**	**0.29**
	MRE	2.34	1.41	0.38	**0.11**	**0.06**
	RMSE	13.87	9.18	2.91	**0.75**	**0.39**
PF	MAE	20.25	4.38	1.45	0.93	0.83
	MRE	3.83	0.83	0.27	0.18	0.12
	RMSE	24.70	6.25	1.90	1.23	1.07
RBPF	MAE	**5.91**	**1.54**	**0.79**	0.63	0.55
	MRE	**1.12**	**0.29**	**0.15**	0.12	0.10
	RMSE	**8.21**	**2.23**	**1.08**	0.83	0.70

Table C.4: MAE [Hz], MRE [%] and RMSE [Hz] in factory noise for vibrato sound of a flute at note C5

SNR		-10.00	-5.00	0.00	5.00	10.00
EKF	MAE	84.00	1.72	0.93	0.50	0.26
	MRE	15.90	0.33	0.18	0.09	0.05
	RMSE	113.65	2.10	1.13	0.61	0.32
UKF	MAE	10.49	5.00	1.17	**0.40**	**0.20**
	MRE	1.98	0.94	0.22	**0.08**	**0.04**
	RMSE	11.82	6.43	1.64	**0.53**	**0.26**
PF	MAE	8.00	1.61	1.23	0.81	0.80
	MRE	1.51	0.30	0.23	0.15	0.15
	RMSE	10.71	2.06	1.50	1.03	1.02
RBPF	MAE	**1.77**	**0.83**	**0.54**	0.46	0.46
	MRE	**0.34**	**0.16**	**0.10**	0.09	0.09
	RMSE	**2.24**	**1.06**	**0.69**	0.59	0.58

Table C.5: MAE [Hz] for Parkinson database in babble noise

SNR	-10.00	-5.00	0.00	5.00	10.00
PEFAC	58.01	48.96	32.45	20.16	13.92
YIN	62.94	51.46	36.17	25.19	18.12
SWIPE	53.29	34.98	16.10	5.93	**3.49**
F0NLS	58.23	42.58	16.39	7.55	6.02
BF0NLS	56.85	34.32	7.57	4.41	4.53
EKF	14.25	10.81	6.41	6.62	7.35
UKF	13.38	10.15	6.50	6.21	6.27
PF	13.12	11.59	8.98	6.18	5.09
RBPF	**10.81**	**8.32**	**5.68**	**3.97**	3.66

Table C.6: MRE [%] for Parkinson database in babble noise

SNR	-10.00	-5.00	0.00	5.00	10.00
PEFAC	46.29	39.45	26.49	16.46	11.07
YIN	46.86	37.67	25.58	17.70	12.88
SWIPE	39.83	25.55	11.48	4.47	2.93
F0NLS	40.89	30.10	12.05	6.12	5.19
BF0NLS	39.97	24.60	6.35	4.28	4.46
EKF	10.48	8.13	5.10	5.25	5.74
UKF	9.85	7.64	5.03	4.71	4.68
PF	9.84	8.70	6.76	4.78	4.06
RBPF	**8.17**	**6.23**	**4.31**	**3.15**	**2.92**

Table C.7: RMSE [Hz] for Parkinson database in babble noise

SNR	-10.00	-5.00	0.00	5.00	10.00
PEFAC	77.13	69.70	54.63	40.79	31.75
YIN	79.04	68.10	53.51	44.03	36.94
SWIPE	72.13	54.71	33.44	16.34	**9.96**
F0NLS	66.66	55.16	32.67	21.46	18.35
BF0NLS	65.50	48.60	20.76	15.42	14.63
EKF	19.25	16.13	13.02	14.58	16.12
UKF	18.21	15.30	12.53	13.10	13.58
PF	18.43	17.20	15.44	12.93	11.92
RBPF	**15.69**	**13.98**	**12.40**	**10.90**	10.92

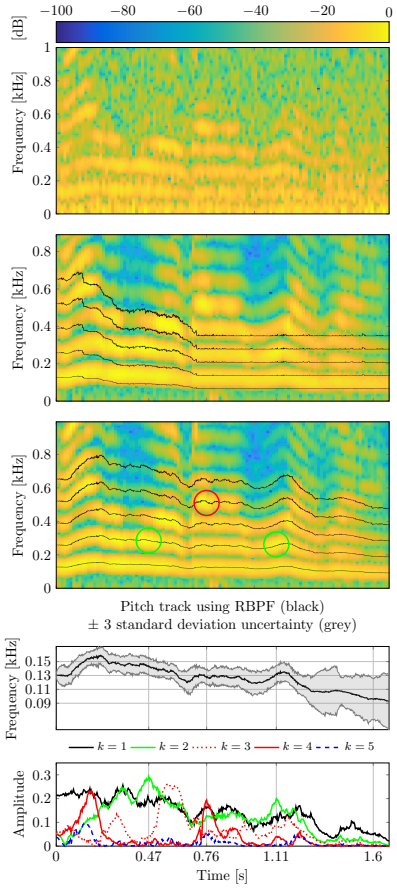


Fig. C.4: Noisy spectrogram under 0 dB factory noise, pitch track using EKF, pitch tracking using RBPF, pitch track and uncertainty using RBPF and tracks of amplitudes using RBPF (from top to bottom).

hyper-parameters for UKF and SMC based methods (i.e., PF and RBPF) are set to $\hat{\sigma}_u^2 = 1$, $\hat{\sigma}_c^2 = 0.01$, and $\hat{\sigma}_u^2 = 0.1$, $\hat{\sigma}_c^2 = 0.1$, respectively. The number of particles used for PF and RBPF are set to 500. The spectrogram of the signals, pitch and amplitude estimates in 0 dB factory noise are shown in Figure C.4. As can be seen, due to the white Gaussian assumption, the EKF pitch estimator breaks, while the proposed RBPF algorithm generates continuous and better pitch estimates. Large amplitude estimates for harmonics $k = 2$, $k = 4$ and $k = 2$ are obtained in the high energy time-frequency area around 0.47 s, 0.76 s and 1.11 s. The MAE, mean relative error (MRE) [36] and root mean square error (RMSE) [36] results for the EKF, UKF, PF, and RBPF pitch estimators in babble and factory noise scenarios with different SNRs

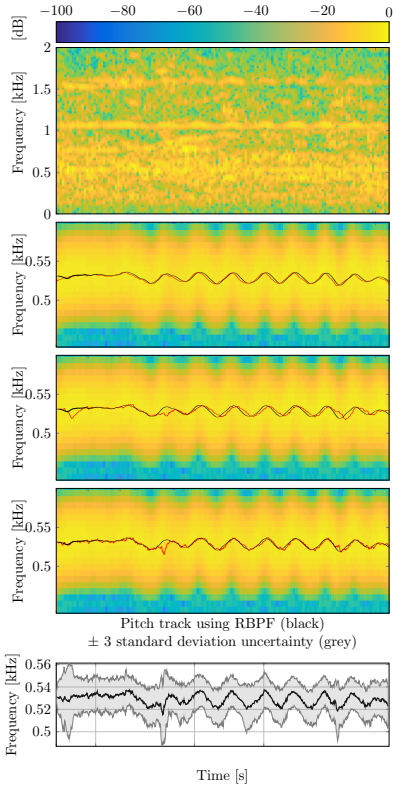


Fig. C.5: Noisy spectrogram under -5 dB babble noise, pitch track using SWIPE on clean signals (red line), pitch track using EKF on noisy signals (red line), pitch tracking using RBPF on noisy signals (red line), pitch track and uncertainty using RBPF (from top to bottom). Number of harmonic order $K = 5$. The black line on the spectrogram denotes the "ground truth" using EKF on clean signals.

are shown in TABLE C.1 and C.2, respectively. The ground truth pitch values are obtained using the EKF pitch estimator on clean speech. As can be seen from TABLE C.1 and C.2, the RBPF algorithm has the best performance from -5 to 5 dB, while the EKF algorithm has the best performance in 10 dB for both noise types. The PF algorithm achieves the best performance under -10 dB factory noise, and the RBPF algorithm has the best performance under -10 dB babble noise.

Next, we test the performance of the proposed algorithms on a vibrato sound of a musical instrument (flute) at C5 from the University of Iowa Musical Instrument Samples database. The pitch is constrained to be in the range $[\omega_{\min}, \omega_{\max}] = 2\pi[100, 1000]/f_s$ and $\hat{\sigma}_A^2 = 10^{-6}$ is used for all the algorithms. The other hyper-parameters are set to the same as the previous experiment. The spectrogram of the signals and pitch estimates for EKF and SWIPE on

clean signals, EKF and RBPF under -5 dB babble noise are shown in Figure C.5, where the black and red lines on the spectrogram denotes the ground truth and estimates, respectively. The ground truth pitch estimates are obtained using the EKF pitch estimator on clean signals. The pitch estimates of SWIPE in noisy conditions are not shown due to its poor performance. As can be seen, compared with the segment-wise pitch estimation algorithm SWIPE, a better time-frequency resolution and smoother pitch estimates are obtained by using the sample-wise pitch tracking approach EKF on clean signals. Moreover, both the EKF and RBPF algorithms generate good pitch estimates. However, a clear delay can be seen for the EKF algorithm. The MAE, MRE and RMSE results for the EKF, UKF, PF, and RBPF pitch estimators in babble and factory noise scenarios with different SNRs are shown in TABLE C.3 and C.4, respectively. As can be seen from TABLE C.3 and C.4, the RBPF algorithm has the best performance from -10 to 0 dB, while the UKF algorithm has the best performance from 5 to 10 dB.

6.3 Results on Parkinson disease database

In this subsection, the performance of the proposed algorithms is compared with the PEFAC, YIN, SWIPE, F0NLS, BF0NLS on sustained /a/ signals from the Parkinson's disease database in babble noise. The database contains 130 sustained /a/ phonations from patients with Parkinson's disease [36] at a sampling rate of 44.1 kHz. Each of the phonations is in one second length. The estimated "ground truth" pitches in 10 ms time frame increment are extracted from electroglottography (EGG). The signals are downsampled to 8000 Hz. The pitch is constrained to be in the range $[\omega_{\min}, \omega_{\max}] = 2\pi[50, 400]/f_s$, and $\hat{\sigma}_\omega^2 = 5 \times 10^{-8}$, $\hat{\sigma}_a^2 = 10^{-4}$ (when applicable) is used for all sample-wise algorithms. All the other hyper-parameters are set the same as Figure C.4. For the sample-wise pitch tracking methods, i.e., EKF, UKF, PF and RBPF algorithms, the pitch estimates are computed as the average values for every 10 ms. The MAE, MRE and RMSE for different SNRs are shown in TABLE C.5, C.6 and C.7, respectively. For initializing the EKF, UKF, PF, and RBPF, the initial pitch value is set to ± 5 Hz away from the ground truth value. As can be seen from TABLE C.5, C.6 and C.7, the EKF, UKF, PF and RBPF performs better than PEFAC, YIN, SWIPE, F0NLS, BF0NLS from -10 to 0 dB. RBPF algorithm has the overall best performance from -10 to 5 dB, while SWIPE has the best performance in 10 dB.

7 Conclusions

In this paper, a fully Bayesian harmonic model-based pitch tracking algorithm is proposed. Using a parametric harmonic model, the proposed algo-

rithm shows good robustness against noise. The non-stationary evolution of the pitch, harmonic order and voicing state are modelled using first-order Markov chains. A fully Bayesian approach is applied for the noise variance and weights to avoid over-fitting. Using the hierarchical g-prior for the weights, the likelihood function can be easily evaluated using the fast NLS. The computational complexity of the recursive calculation of the predicted and posterior distributions is reduced by exploiting conditional independence between the pitch and harmonic order given the voicing indicators. Simulation results show that the proposed algorithm has good robustness against voicing state changes by carrying past information on pitch over the unvoiced/silent segments. The results of the pitch estimates and voicing detection for spoken sentences and sustained vowels are compared against ground truth estimates in the Keele and Parkinson's disease databases, showing that the proposed algorithm presents good pitch estimation and voicing detection accuracy even in very noisy conditions (e.g., -15 dB).

8 Appendix

Assuming that the observation and the state evolution models conditioned on $\boldsymbol{\phi}_n$ are linear dynamical systems and can be written as

$$\mathbf{y}_n = \mathbf{H}(\boldsymbol{\phi}_n)\mathbf{z}_n + \mathbf{G}(\boldsymbol{\phi}_n)\mathbf{v}_n,\ \mathbf{v}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_v), \tag{C.50}$$

$$\mathbf{z}_n = \mathbf{T}(\boldsymbol{\phi}_n)\mathbf{z}_{n-1} + \mathbf{R}(\boldsymbol{\phi}_n)\mathbf{m}_n,\ \mathbf{m}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_m), \tag{C.51}$$

where $\mathbf{H}(\boldsymbol{\phi}_n), \mathbf{G}(\boldsymbol{\phi}_n), \mathbf{T}(\boldsymbol{\phi}_n)$ and $\mathbf{R}(\boldsymbol{\phi}_n)$ are model matrices dependent on $\boldsymbol{\phi}_n$. Given $\boldsymbol{\phi}_n$, the integrations in (C.43) and (C.44) can be computed analytically using the one-step-ahead prediction of the Kalman filter algorithm and the prediction error decomposition. Assuming that $p(\mathbf{z}_{n-1}|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_{n-1}) = \mathcal{N}(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1})$ and using (C.51), the analytical solution for (C.44) can be expressed as

$$p(\mathbf{z}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_{n|n-1}, \mathbf{P}_{n|n-1}), \tag{C.52}$$

where

$$\boldsymbol{\mu}_{n|n-1} = \mathbf{T}(\boldsymbol{\phi}_n)\boldsymbol{\mu}_{n-1},$$
$$\mathbf{P}_{n|n-1} = \mathbf{T}(\boldsymbol{\phi}_n)\mathbf{P}_{n-1}\mathbf{T}(\boldsymbol{\phi}_n)^T + \mathbf{R}(\boldsymbol{\phi}_n)\mathbf{Q}_m\mathbf{R}(\boldsymbol{\phi}_n)^T.$$

The equation (C.52) is known as the one-step-ahead prediction of the Kalman filter algorithm. Using the prediction error decomposition, the analytical solution for (C.43) can be expressed as

$$p(\mathbf{y}_n|\mathbf{Y}_{n-1}, \boldsymbol{\Phi}_n) = \mathcal{N}(\mathbf{y}_n|\bar{\mathbf{y}}_n, \mathbf{P}_n^{\mathbf{y}}), \tag{C.53}$$

where

$$\bar{\mathbf{y}}_n = \mathbf{H}(\boldsymbol{\phi}_n)\boldsymbol{\mu}_{n|n-1},$$
$$\mathbf{P}_n^{\mathbf{y}} = \mathbf{H}(\boldsymbol{\phi}_n)\mathbf{P}_{n|n-1}\mathbf{H}(\boldsymbol{\phi}_n)^T + \mathbf{G}(\boldsymbol{\phi}_n)\mathbf{Q}_v\mathbf{G}(\boldsymbol{\phi}_n)^T.$$

The conditional posterior $p(\mathbf{z}_n|\mathbf{Y}_n, \boldsymbol{\Phi}_n)$ can be evaluated using the update equation of the Kalman filter algorithm, i.e.,

$$p(\mathbf{z}_n|\mathbf{Y}_n, \boldsymbol{\Phi}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{P}_n), \tag{C.54}$$

where

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n|n-1} + \mathbf{K}(\boldsymbol{\Phi}_n)(\mathbf{y}_n - \bar{\mathbf{y}}_n),$$
$$\mathbf{P}_n = \mathbf{P}_{n|n-1} - \mathbf{K}(\boldsymbol{\Phi}_n)\mathbf{H}(\boldsymbol{\phi}_n)\mathbf{P}_{n|n-1},$$
$$\mathbf{K}(\boldsymbol{\Phi}_n) = \mathbf{P}_{n|n-1}\mathbf{H}(\boldsymbol{\phi}_n)^T(\mathbf{P}_n^{\mathbf{y}})^{-1}.$$

References

[1] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, April 2014.

[2] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010.

[3] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 2494–2498.

[4] A. S. Association *et al.*, "Acoustical terminology SI," *1-1960, American Standards Association*, 1960.

[5] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[6] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2042–2056, 2013.

[7] K. Paliwal and K. Wójcicki, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Process. Lett.*, vol. 15, pp. 785–788, 2008.

[8] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2354–2367, 2016.

[9] E. V. Denardo, *Dynamic programming: models and applications*. Courier Corporation, 2012.

[10] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[11] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[12] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[13] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[14] B. Quinn and P. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, 1991.

[15] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Process.*, vol. 80, no. 9, pp. 1937–1944, 2000.

[16] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.

[17] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust bayesian pitch tracking based on the harmonic model," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 11, pp. 1737–1751, Nov 2019.

[18] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Signals, Systems and Computers, 2014 48th Asilomar Conference on*. IEEE, 2014, pp. 1400–1404.

[19] T. L. Jensen, J. K. Nielsen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "A fast algorithm for maximum-likelihood estimation of harmonic chirp parameters," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5137–5152, 2017.

[20] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in *Proc. Interspeech*, 2008.

[21] ——, "Adaptive am–fm signal decomposition with application to speech analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 2, pp. 290–300, 2010.

[22] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1. IEEE, 1995, pp. 756–759.

[23] H. Hajimolahoseini, R. Amirfattahi, H. Soltanian-Zadeh, and S. Gazor, "Instantaneous fundamental frequency estimation of non-stationary periodic signals using non-linear recursive filters," *IET Signal Processing*, vol. 9, no. 2, pp. 143–153, 2015.

[24] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "A Kalman-based fundamental frequency estimation algorithm," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE Press, 2017, pp. 314–318.

[25] G. Zhang and S. Godsill, "Fundamental frequency estimation in speech signals with variable rate particle filters," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 5, pp. 890–900, 2016.

[26] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[27] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 438–449, 2002.

[28] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068. International Society for Optics and Photonics, 1997, pp. 182–193.

[29] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. A. Wan, "The unscented particle filter," in *Advances in neural information processing systems*, 2001, pp. 584–590.

[30] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *IEE proceedings F (radar and signal processing)*, vol. 140, no. 2. IET, 1993, pp. 107–113.

[31] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *Proc. IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[32] E. A. Wan and R. Van Der Merwe, "The unscented kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*. Ieee, 2000, pp. 153–158.

[33] A. Smith, *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[34] L. Shi, J. R. Jensen, J. K. Nielsen, and M. G. Christensen, "Multipitch estimation using block sparse bayesian learning and intra-block clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2018, pp. 666–670.

[35] J. Durbin, "The fitting of time-series models," *Revue de l'Institut International de Statistique*, pp. 233–244, 1960.

[36] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–2901, 2014.

Paper D

Multipitch Estimation Using Block Sparse Bayesian Learning and Intra-block Clustering

Liming Shi, Jesper Rindom Jensen, Jesper Kjær Nielsen and Mads Græsbøll Christensen

The paper has been published in the
Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2018

Abstract

Pitch estimation is an important task in speech and audio analysis. In this paper, we present a multi-pitch estimation algorithm based on block sparse Bayesian learning and intra-block clustering for speech analysis. A statistical hierarchical model is formulated based on a pitch dictionary with a fixed maximum number of harmonics for all the candidate pitches. Block sparse Bayesian learning is proposed for estimating the complex amplitudes. To deal with the problem of unknown harmonic orders and subharmonic errors, intra-block clustering structured sparsity prior is also introduced. The statistical update formulas are obtained by the variational Bayesian inference. Compared with the conventional group LASSO-type algorithms for multi-pitch estimation, experimental results indicate robustness against noise and improved estimation accuracy of the proposed method.

1 Introduction

Fundamental frequency (a.k.a., pitch) estimation has diverse applications in voice disorder detection [1], automatic music transcription [2], speech enhancement [3], etc. The pitch estimation algorithms can be broadly classified as *non-parametric* and *parametric* methods. The popular Yin [4] and RAPT [5] can be categorized as non-parametric methods since they are based on the *autocorrelation* function obtained within a specified time frame. These methods are computationally simple but they are sensitive to noise and prone to *subharmonic errors* (that is, misidentifying a rational number times the actual pitch). On the other hand, the pitch estimation methods using parametric model (e.g., *harmonic model*) are less commonly used but more robust to noise. In this model, both the pitch and complex amplitudes are assumed to be invariant during a short-time period (frame) (e.g., 20-40 ms for speech signals) [6]. Various kinds of estimators, such as the nonlinear least square estimator [6], have been proposed using the harmonic model or its variants.

When multiple speakers are present or multiple instruments are mixed in a music piece, the problem of multi-pitch estimation arises. In [7], different pitches were estimated by an iterative spectral subtraction process. That is, the estimated pitch from the most prominent sound was removed from the mixture signal repeatedly. The spectral smoothness principle was used to deal with the overlapping harmonics. A statistical harmonic model-based multi-pitch estimation algorithm was proposed in [8], where spectral smoothness was also imposed by modelling the spectral envelope of overtones as an autoregressive model. More recently, a multi-pitch estimation algorithm based on a pitch dictionary and group LASSO was proposed. A convex cost function, combining the advantages of l_2, sum of l_2, and l_1 norms, was designed, which was referred to as PEBS [9]. A total variation (TV) term

was further introduced to reduce the subharmonic errors (PEBS-TV). However, due to the difficulty of tuning the regularization parameters, an adaptive penalty estimator with self-regularization was proposed in [10], called PEBSI-Lite. The dictionary in this algorithm was initialized with pitch candidates estimated by frequency estimation methods (e.g., ESPRIT [11]). An iterative solution was obtained by the alternating direction method of multipliers (ADMM) [12]. Typically, these methods incorporate prior knowledge about the spectral smoothness, which can be exploited by the regularization techniques, or by the Bayesian framework with prior models on the unknown complex amplitude parameters.

In this paper, motivated by the work in Bayesian sparse signal recovery [13–15], a block sparse Bayesian learning-based multi-pitch estimation algorithm is proposed. By imposing the block sparse prior, the complex amplitudes of the active pitches in the dictionary can be recovered and thus also the corresponding pitches using block sparse Bayesian learning (BSBL) method. Moreover, to deal with an unknown number of harmonic orders and the subharmonic problem, intra-block cluster structured sparsity prior is introduced. By clustering the non-zero elements of the complex amplitudes within each block, the subharmonic errors can be reduced. Variational Bayesian inference is applied for obtaining statistical update formulas.

2 Fundamentals

We aim to fit the observed speech signals to an over-complete harmonic model with harmonic series including P candidate pitches and each pitch have up to $L_{\max}$ harmonics, i.e.,

$$y_n = \sum_{p=1}^{P} \sum_{l=1}^{L_{\max}} a_{p,l} e^{j\omega_p l n} + m_n, \tag{D.1}$$

where $a_{p,l}$ denotes the complex amplitude of the l^{th} harmonic of the p^{th} pitch in the dictionary, n is the time index, m_n is the complex Gaussian white noise, $\omega_p = 2\pi f_p / F_s$, f_p denotes the p^{th} pitch, and F_s is the sampling rate. Collecting N observed samples and writing (D.1) to a matrix form, we have

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{m}, \tag{D.2}$$

where $\mathbf{y} = [y_0, y_1, \cdots, y_{N-1}]^T$, the noise vector is given by $\mathbf{m} = [m_0, m_1, \cdots, m_{N-1}]^T$, the complex amplitude vector by $\mathbf{a} = \left[\mathbf{a}_1^T, \mathbf{a}_2^T, \cdots, \mathbf{a}_P^T\right]^T$, $\mathbf{a}_p = \left[a_{p,1}, a_{p,2}, \cdots, a_{p,L_{\max}}\right]^T$, $1 \le p \le P$, the dictionary $\mathbf{Z}$ is a $N \times PL_{\max}$ matrix denoted as $\mathbf{Z} = [\mathbf{Z}(\omega_1), \mathbf{Z}(\omega_2), \cdots, \mathbf{Z}(\omega_P)]$ and $\mathbf{Z}(\omega_p)$, for $1 \le p \le P$,

has a Vandermonde structure as follows:

$$\mathbf{Z}(\omega_p) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_p} & e^{j2\omega_p} & \vdots & e^{jL_{\max}\omega_p} \\ \vdots & \vdots & \vdots & \vdots \\ e^{j\omega_p(N-1)} & e^{j2\omega_p(N-1)} & \cdots & e^{jL_{\max}\omega_p(N-1)} \end{bmatrix}.$$

A key assumption in the over-complete harmonic model (D.2) is that the complex amplitude vector $\mathbf{a}$ is block sparse. However, when the actual number of harmonics of the p^{th} pitch candidate is less than $L_{\max}$, $\mathbf{a}_p$ also contains zeros. The sum of norms and L_1-norm regularization terms are introduced in [9] to impose both the block sparse and sparse priors for multi-pitch estimation. However, only using these two regularization terms may lead to subharmonic errors. For example, if the true pitch of an observed sinusoidal signal is 100 Hz and we have 50 Hz pitch in the dictionary, we may wrongly estimate the pitch as 50 Hz. This is because the observed signal can be fitted well with a block sparse complex amplitude vector estimate $\hat{\mathbf{a}}$ (e.g., $\hat{\mathbf{a}} = [\cdots, \mathbf{0}, \hat{\mathbf{a}}_p, \mathbf{0}, \cdots]^T$) and a sparse sub-block estimate $\hat{\mathbf{a}}_p$ that corresponds to the 50 Hz pitch (e.g., $\hat{\mathbf{a}}_p = [0, a_1, 0, a_2, 0 \cdots]^T$). To counter this problem, a total variation term is further added to the cost function to impose smoothness to the complex amplitudes.

3 Proposed block sparse Bayesian learning and intra-block clustering

As noted before, when subharmonic errors occur, the complex amplitude vector estimates of the subharmonics contain zeros. Instead of using the sparse and smoothness priors like the PEBS-TV, an alternative approach is to identify the complex amplitudes as cluster structured sparsity around the first several elements and up to the actual number of harmonics, which can be easily verified from the spectrogram of speech signals. In this paper, we impose both the block sparse prior and intra-block clustered structured sparse prior to the first several elements of each $\mathbf{a}_p$ for multipitch estimation. Block sparse prior is applied for estimating the complex amplitudes of the active pitches in the dictionary. Intra-block clustered structured sparse prior is exploited to counter the problem of unknown harmonic orders and subharmonic errors. In this section, we first formulate the problem using the hierarchical model and then give the update formulas using the variational Bayesian inference.

3.1 Hierarchical model

We proceed by assigning a circular, symmetric white complex Gaussian to the observed noises, i.e.,

$$p(\mathbf{m}|\gamma) = C\mathcal{N}(\mathbf{m}|\mathbf{0}, \gamma^{-1}\mathbf{I}_N), \tag{D.3}$$

where a complex Gaussian variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is defined as

$$C\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\pi^N|\boldsymbol{\Sigma}|}\exp\{-(\mathbf{x}-\boldsymbol{\mu})^H\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}. \tag{D.4}$$

A Gamma distribution is assigned to the precision γ of the complex Gaussian (conjugate prior), i.e.,

$$p(\gamma) \sim \Gamma(\gamma|c, d). \tag{D.5}$$

To motivate block sparsity and intra-block clustered sparsity for the complex amplitude vector $\mathbf{a}$, we first introduce a latent variable $\theta_{p,l}$ (the l^{th} element of the p^{th} block of $\boldsymbol{\theta}$) to indicate the zero/nonzero status of the corresponding complex amplitude coefficients $a_{p,l}$, i.e. $\mathbf{a} = \mathbf{u} \odot \boldsymbol{\theta}$, where $\odot$ denotes element-wise multiplication and

$$\begin{aligned} p(\mathbf{u}|\boldsymbol{\alpha}) &= C\mathcal{N}(\mathbf{u}|\mathbf{0}, \boldsymbol{\Lambda}^{-1}), \\ \boldsymbol{\Lambda} &= \text{diag}(\boldsymbol{\alpha}) \otimes \mathbf{I}_{L_{\max}}. \end{aligned} \tag{D.6}$$

The hyperparameter α_p (p^{th} element of $\boldsymbol{\alpha}$) is the precision of the p^{th} block, and when it is infinite, the p^{th} block will be zero [13]. A Gamma distribution is also assigned to the hyperparameter α_p as

$$p(\alpha_p) \sim \Gamma(\alpha_p|g, h). \tag{D.7}$$

Besides, the latent variable $\theta_{p,l}$ is drawn from Bernoulli distribution with success probability $\pi_{p,l}$, i.e.,

$$\theta_{p,l} \sim \text{Bernoulli}(\pi_{p,i}). \tag{D.8}$$

Three different patterns for clustered sparse recovery were introduced in [16, 17], i.e., $P0$: "strongly eliminate", when the two neighbours are both zeros; $P1$: "weakly eliminate", when one of the neighbour are zero; $P2$: "strongly plump", when both of the neighbours are non-zeros. However, in pitch estimation, non-zero clusters are formed around the first several elements of $\mathbf{a}_p$ of true pitches. Therefore, we propose to use the following four-pattern model for the latent variable $\theta_{p,l}, 1 \leq p \leq P, 1 < l < L_{\max}$, i.e., *P0*: "strongly elimination", when $\theta_{p,1} = 0$ (fundamental frequency is missing); *P1*: "mildly eliminate", when the two neighbours are both zeros and

$\theta_{p,1} = 1$; *P*2: "weakly eliminate", when one of the neighbour is zero and $\theta_{p,1} = 1$; *P*3: "strongly plump", when both of the neighbours are non-zeros and $\theta_{p,1} = 1$. According to these clustering patterns, the success probability for $1 < l < L_{\max}$ is chosen by

$$\pi_{p,l} = \begin{cases} \pi^0, \text{ if } P0 \\ \pi^1, \text{ if } P1 \\ \pi^2, \text{ if } P2 \\ \pi^3, \text{ if } P3 \end{cases}, \quad \pi^j \sim \text{Beta}(\pi^j | e^j, f^j),\ 0 \leq j \leq 3, \tag{D.9}$$

where π^j, for $0 \leq j \leq 3$ is drawn from the Beta distribution. Note that, the model for $l \in \{1, L_{\max}\}$ are not shown here for simplicity. However, we can follow the above definitions but use two patterns for $l = 1$ and three patterns for $l = L_{\max}$ because of their single neighbour characteristic. Using patterns *P*1, *P*2 and *P*3, we can expect that the non-zero elements within each block will be clustered together. Moreover, an all-zero cluster will be formed in the rear of the block since a large $L_{\max}$ is used. By introducing the pattern *P*0, a nonzero cluster around the first several elements of complex amplitude vector $\mathbf{a}_p$ is encouraged if the p^{th} pitch in the dictionary is active. We refer to the proposed algorithm as pitch estimation using block sparse Bayesian learning and intra-block clustering (PE-BSBL-Cluster). Note that if we set the latent variable $\theta_{p,l} = 1,\ 1 \leq p \leq P,\ l \leq l \leq L_{\max}$, the intra-block clustering scheme will be dropped and only block sparse prior will be applied, which we refer to as PE-BSBL.

3.2 Variational Bayesian inference

The exact joint posterior distribution can not be derived analytically. Instead, we resort to an approximation method, i.e., variational Bayesian inference [18]. For completeness, we give the update formulas in section 6. A detailed derivation and the results for $l \in \{1, L_{\max}\}$ is given in the technical report [19].

4 Results

We test the proposed PE-BSBL and PE-BSBL-Cluster in both synthetic and mixed speech signals scenarios [1]. All the modeling parameters are fixed as follows: $c = d = h = 10^{-6}$, $g = 1$, $(e^0, f^0) = (1, 10^6)$, $(e^1, f^1) = (1/L_{\max}, 1 - 1/L_{\max})$, $(e^2, f^2) = (1/L_{\max}, 1/L_{\max})$, $(e^3, f^3) = (1 - 1/L_{\max}, 1/L_{\max})$. The proposed algorithms are terminated if $\frac{\|\boldsymbol{\alpha}^i - \boldsymbol{\alpha}^{i-1}\|_2}{\|\boldsymbol{\alpha}^i\|_2} \leq 10^{-3}$ or 1000 iterations are

[1] An implementation of the proposed algorithms using MATLAB may be found in `https://tinyurl.com/y8orkosc`

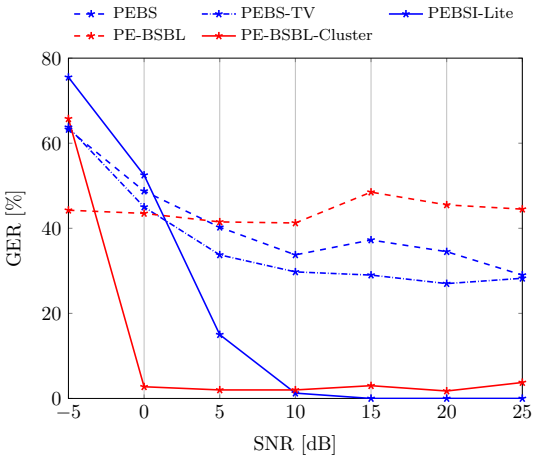


Fig. D.1: Gross error ratio for synthetic signal in different SNRs, Q=2, $f_1^0 = 160$ and $f_2^0 = 240$ Hz ($240 = \frac{3}{2} \times 160$) .

reached, where i denotes the iteration number. Pitch estimates are obtained by choosing the pitches that have the largest posterior energies defined as $\tilde{\boldsymbol{\mu}}_p^H \tilde{\boldsymbol{\mu}}_p + \mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}_p)$, where $\tilde{\boldsymbol{\mu}}_p$ and $\tilde{\boldsymbol{\Sigma}}_p$ denote the posterior mean and covariance of $\mathbf{a}_p$, and $\mathrm{Tr}(\cdot)$ is the trace operator. We compare the proposed algorithms with the PEBS [9], PEBS-TV [9] and PEBS-Lite [10]. For the PEBS and PEBS-TV, the regularization parameters are set to the same as in [10].

4.1 Synthetic signal analysis

The first experiment examines the performance for synthetic signals sampling of 8000 Hz, as shown in Fig. D.1. Two pitches with 160 and 240 Hz are used. The data length N is set to 240. Uniform grid ranging from 50 to 500 Hz with grid interval 2 Hz and $L_{\max} = 10$ is used for all the experiments. To simulate the off-grid effect, for each trial, the true pitches are drawn from the uniform distribution, i.e., $f_{0,q} \sim \mathrm{Unif}(f_q^0 - d/2, f_q^0 + d/2), 1 \leq q \leq Q$. The deviation d is the grid interval and f_q^0 denotes a pitch on the grid. The number of harmonics are uniformly drawn over the integer interval $[3, 10]$ in each simulation. The amplitude of each harmonic is set to unit magnitude and the phase is drawn uniformly on $[0, 2\pi)$ [10]. The performance is measured by the gross error rate (GER), defined by calculating the number of pitch estimates that is differed by more than a certain percentage from the ground truth [20, 21]. In this paper, we use 5% for all the experiments. The experimental results are obtained by the ensemble averages over 200 Monte Carlo simulations. As can be seen, the GERs of the PEBS, PEBS-TV and PEBSI-Lite, are lower than the PE-BSBL, especially in high SNRs. This is because that

the PE-BSBL only exploits the block sparse prior, and thus it is prone to subharmonic errors. Moreover, the PEBS-TV presents a better performance than the PEBS due to the TV term in the cost function. Furthermore, PEBSI-Lite obtains the lowest GER in high SNRs due to the built-in refining process and good performance of the ESPRIT in high SNRs. But its performance degenerates severely in low SNRs. By exploiting the block sparse and intra-block clustering structured priors together, the proposed PE-BSBL-Cluster achieves the lowest GER compared with PEBS, PEBS-TV and PE-BSBL in 0 to 25 dB SNRs. Although the PEBSI-Lite presents a slightly better performance than the proposed PE-BSBL-Cluster in high SNRs (10 to 25 dB), the proposed PE-BSBL-Cluster has a much lower GER in low SNRs (-5 to 5 dB). Thus, it is more robust to noise. Note that, high-resolution estimates for the proposed algorithm can be found by refining methods, such as gradient ascend method [6].

4.2 Mixed speech signal analysis

We also examine the performance of the PE-BSBL and PE-SBL-Cluster for a mixed speech signal of the spoken sentences "Why were you away a year?" from a female speaker and "Our lawyer will allow your rule." from a male speaker. The ground truth pitch estimates of each sentence are obtained by Yin in noise-free scenario. The sampling rate is 8000 Hz. The spectrogram of mixed speech (noise-free), pitch estimates of PEBSI-Lite and the proposed PE-BSBL-Cluster under 5 dB SNR are shown in Fig. D.2. On the spectrogram, the two black dotted lines (from top to bottom) denote the ground truth pitch estimates of the female and male sentences, respectively. The GER versus different SNRs, computed using 10 Monte-Carlo simulations, is shown in Fig. D.3. Analysis is performed every 30 ms with 50% overlap. As can be seen from Fig. D.2, the proposed PE-BSBL-Cluster has less estimation errors than the PEBSI-Lite. From the plots of both algorithms, the estimated pitch tracks of the male speaker can be clearly seen. However, it is easier to see the the estimated pitch track of the female speaker using the proposed PE-BSBL-Cluster than PEBSI-Lite. Similar conclusions to Fig. D.1 can be drawn from Fig. D.3. The proposed PE-BSBL-Cluster achieves the lowest GER in low SNRs (-5 to 10 dB) and has a comparable performance with the PEBSI-Lite in high SNRs (15-25 dB). Above all, due to the usage of the block sparse and clustering structured priors, compared with group-LASSO type algorithms, the proposed PE-BSBL-Cluster can deal with the problems of unknown harmonic orders and subharmonic errors, and presents a good performance even in low SNRs.

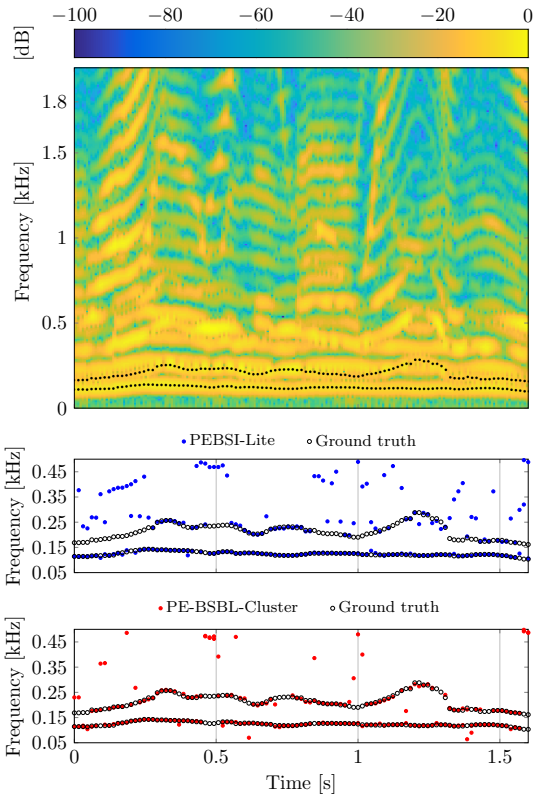


Fig. D.2: Pitch estimates of real mixed speech of the spoken sentences "Why were you away a year?" from a female and "Our lawyer will allow your rule." from a male speaker, $F_s = 8000$ Hz, SNR=5 dB.

5 Conclusion

A multi-pitch estimation algorithm using block sparse Bayesian learning and intra-block clustering has been proposed. Using a block sparse prior model, the complex amplitude vectors corresponding to the true pitches in the pitch dictionary can be recovered. Moreover, to deal with unknown number of harmonic orders and subharmonic errors, intra-block clustering structured sparsity are encouraged by imposing a clustering prior. Update equations are obtained by the variational Bayesian inference. Simulation results using both synthetic and real mixed speech show that the proposed PE-BSBL-Cluster has improved multipitch estimation accuracy in terms of GER and robustness against noise.

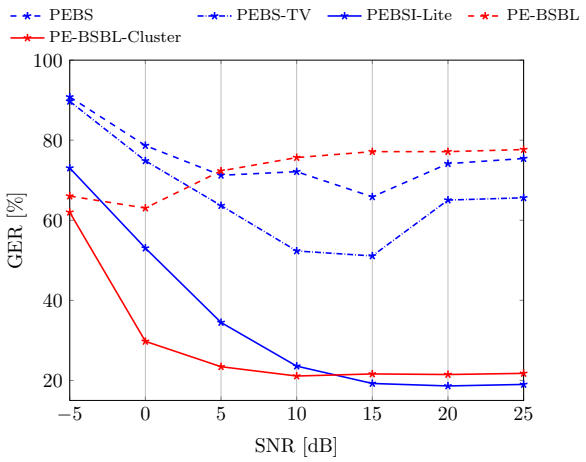


Fig. D.3: Gross error ratio for real mixed speech in different SNRs.

6 Appendix

The approximated posteriors are listed as follows:
(1) **the indicator variable** $\theta_{p,l}$, $1 \leq p \leq P$, $1 \leq l \leq L_{\max}$:

$$q(\theta_{p,l}) = \text{Bernoulli}(\tilde{\pi}_{p,l}), \tag{D.10}$$

where

$$\begin{aligned}&\tilde{\pi}_{p,l}\\ =&[1+\exp\{\langle\log(1-\pi_{p,l})\rangle - \langle\log(\pi_{p,l})\rangle + \langle\gamma\rangle[\langle u_{p,l}^* u_{p,l}\rangle \mathbf{z}_{p,l}^H \mathbf{z}_{p,l}\\ &-2\text{Re}(\langle u_{p,l}\rangle^* \mathbf{z}_{p,l}^H(\mathbf{y} - \sum_{(i,j)\neq(p,l)} \langle\theta_{i,j}\rangle\langle u_{i,j}\rangle \mathbf{z}_{i,j})]\}]^{-1},\end{aligned}$$

where $\langle\cdot\rangle$ denotes the expectation operator, $(\cdot)^*$ denotes the conjugate and $(\cdot)^H$ denotes conjugate transpose.
(2) **the complex amplitude u**:

$$q(\mathbf{u}) = \mathcal{CN}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{D.11}$$

where

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}} &= (\langle\boldsymbol{\Lambda}\rangle + \langle\gamma\rangle\langle\text{diag}(\boldsymbol{\theta})\mathbf{Z}^H\mathbf{Z}\text{diag}(\boldsymbol{\theta})\rangle)^{-1},\\ \tilde{\boldsymbol{\mu}} &= \langle\gamma\rangle\tilde{\boldsymbol{\Sigma}}\langle\text{diag}(\boldsymbol{\theta})\rangle\mathbf{Z}^H\mathbf{y},\end{aligned}$$

and $\langle\text{diag}(\boldsymbol{\theta})\mathbf{Z}^H\mathbf{Z}\text{diag}(\boldsymbol{\theta})\rangle = (\mathbf{Z}^H\mathbf{Z}) \odot (\langle\boldsymbol{\theta}\rangle\langle\boldsymbol{\theta}\rangle^T + \text{diag}(\langle\boldsymbol{\theta}\rangle \odot (1 - \langle\boldsymbol{\theta}\rangle)))$.

(3) **the noise precision** γ:

$$q(\gamma) = \Gamma(\gamma|\tilde{c}, \tilde{d}), \tag{D.12}$$

where

$$\begin{aligned}\tilde{c} =& c + N,\\ \tilde{d} =& d + \|\mathbf{y} - \mathbf{Z}(\langle\mathbf{u}\rangle \odot \langle\boldsymbol{\theta}\rangle)\|^2 + \mathrm{Tr}\{\mathbf{Z}^H\mathbf{Z}(\langle\mathbf{u}\mathbf{u}^H\rangle \odot (\langle\boldsymbol{\theta}\boldsymbol{\theta}^T\rangle)\\ & - (\langle\mathbf{u}\rangle \odot \langle\boldsymbol{\theta}\rangle)(\langle\mathbf{u}\rangle \odot \langle\boldsymbol{\theta}\rangle)^H)\}.\end{aligned}$$

(4) **the precision** α_p, $1 \le p \le P$ **of the complex amplitudes**:

$$q(\alpha_p) = \Gamma(\alpha_p|\tilde{g}_p, \tilde{h}_p), \tag{D.13}$$

where

$$\tilde{g}_p = g + L_{\max}, \qquad \tilde{h}_p = h + \langle\mathbf{u}_p^H\mathbf{u}_p\rangle.$$

(5) **the success probability** $\pi_{p,l}$, $1 \le p \le P$, $1 < l < L_{\max}$:

$$q(\pi_{p,l}^j) = \mathrm{Beta}(\pi_{p,l}^j|\tilde{e}_{p,l}^j, \tilde{f}_{p,l}^j), \tag{D.14}$$

where for $j \in \{0,1,2,3\}$,

$$\begin{aligned}\tilde{e}_{p,l}^j =& e^j + p(Pj)\langle\theta_{p,l}\rangle,\\ \tilde{f}_{p,l}^j =& f^j + p(Pj)(1 - \langle\theta_{p,l}\rangle),\end{aligned}$$

and

$$\begin{aligned}p(P0) =& 1 - \langle\theta_{p,1}\rangle,\\ p(P1) =& \langle\theta_{p,1}\rangle(1 - \langle\theta_{p,l-1}\rangle)(1 - \langle\theta_{p,l+1}\rangle),\\ p(P2) =& \langle\theta_{p,1}\rangle(\langle\theta_{p,l-1}\rangle(1 - \langle\theta_{p,l+1}\rangle) + \langle\theta_{p,l+1}\rangle(1 - \langle\theta_{p,l-1}\rangle)),\\ p(P3) =& \langle\theta_{p,1}\rangle\langle\theta_{p,l-1}\rangle\langle\theta_{p,l+1}\rangle.\end{aligned}$$

The Expectation of logarithm function can be calculated as

$$\langle\log\pi_{p,l}\rangle = \sum_{j=0}^{3} p(Pj)\langle\log\pi_{p,l}^j\rangle,$$

$$\langle\log(1 - \pi_{p,l})\rangle = \sum_{j=0}^{3} p(Pj)\langle\log(1 - \pi_{p,l}^j)\rangle.$$

References

[1] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010.

[2] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] M. Krawczyk-Becker and T. Gerkmann, "Fundamental frequency informed speech enhancement in a flexible statistical framework," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 5, pp. 940–951, 2016.

[4] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[6] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[7] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[8] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, 2010.

[9] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Process.*, vol. 109, pp. 236–247, 2015.

[10] F. Elvander, T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Process.*, vol. 127, pp. 56–70, 2016.

[11] P. Stoica, R. L. Moses, *et al.*, *Spectral analysis of signals*. Pearson Prentice Hall, Upper Saddle River, NJ, 2005, vol. 452.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. Jun, pp. 211–244, 2001.

[14] T. Park and G. Casella, "The bayesian lasso," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, 2008.

[15] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, 2008.

[16] L. Yu, H. Sun, J.-P. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Process.*, vol. 92, no. 1, pp. 259–269, 2012.

[17] L. Yu, C. Wei, J. Jia, and H. Sun, "Compressive sensing for cluster structured sparse signals: Variational bayes approach," *IET Signal Process.*, vol. 10, no. 7, pp. 770–779, 2016.

[18] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[19] L. Shi, J. R. Jensen, J. K. Nielsen, and M. G. Christensen, "Multipitch estimation using block sparse bayesian learning and intra-block clustering," *Tech. Rep.*, 2018. [Online]. Available: https://tinyurl.com/y8orkosc

[20] F. Flego and M. Omologo, "Robust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *Proc. European Signal Processing Conf.* IEEE, 2006, pp. 1–4.

[21] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Estimation of multiple pitches in stereophonic mixtures using a codebook-based approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 186–190.

Paper E

A Variational EM Method for Pole-zero Modeling of Speech with Mixed Block Sparse and Gaussian Excitation

Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been published in the
Proc. European Signal Processing Conference, 2017

Abstract

The modeling of speech can be used for speech synthesis and speech recognition. We present a speech analysis method based on pole-zero modeling of speech with mixed block sparse and Gaussian excitation. By using a pole-zero model, instead of the all-pole model, a better spectral fitting can be expected. Moreover, motivated by the block sparse glottal flow excitation during voiced speech and the white noise excitation for unvoiced speech, we model the excitation sequence as a combination of block sparse signals and white noise. A variational EM (VEM) method is proposed for estimating the posterior PDFs of the block sparse residuals and point estimates of modelling parameters within a sparse Bayesian learning framework. Compared to conventional pole-zero and all-pole based methods, experimental results show that the proposed method has lower spectral distortion and good performance in reconstructing of the block sparse excitation.

1 Introduction

The modeling of speech has important applications in speech analysis [1], speaker verification [2], speech synthesis [3], etc. Based on the source-filter model, speech is modelled as being produced by a pulse train or white noise for voiced or unvoiced speech, which is further filtered by the speech production filter (SPF) that consists of the vocal tract and lip radiation.

All-pole modeling with a least squares cost function performs well for white noise and low pitch excitation. However, for high pitch excitation, it leads to an all-pole filter with poles close to the unit circle, and the estimated spectrum has a sharper contour than desired [4, 5]. To obtain a robust linear prediction (LP), the Itakura-Saito error criterion [6], the all-pole modeling with a distortionless response at frequencies of harmonics [4], the regularized LP [7] and the short-time energy weighted LP [8] were proposed. Motivated by the compressive sensing research, a least 1-norm criterion is proposed for voiced speech analysis [9], where sparse priors on both the excitation signals and prediction coefficients are utilized. Fast methods and the stability of the 1-norm cost function for spectral envelope estimation are further investigated in [10, 11]. More recently, in [12], the excitation signal of speech is formulated as a combination of block sparse and white noise components to capture the block sparse or white noise excitation separately or simultaneously. An expectation-maximization (EM) algorithm is used to reconstruct the block sparse excitation within a sparse Bayesian learning (SBL) framework [13].

A problem with the all-pole model is that some sounds containing spectral zeros with voiced excitation, such as nasals, or laterals, are poorly estimated by an all-pole model but trivial with a pole-zero (PZ) model [14, 15]. The estimation of the coefficients of the pole-zero model can be obtained sepa-

rately [16], jointly [17] or iteratively [18]. A 2-norm minimization criterion with Gaussian residuals assumption is commonly used. Frequency domain fitting methods based on a similarity measure is also proposed. Motivated by the logarithmic scale perception of the human auditory system, the logarithmic magnitude function minimization criterion has been proposed [15, 19]. In [19], the nonlinear logarithm cost function is solved by transforming it into a weighted least squares problem. The Gauss-Newton and Quasi-Newton methods for solving it are further investigated in [15]. To consider both the voiced excitation and the PZ model, a speech analysis method based on the PZ model with sparse excitation in noisy conditions is presented [20]. A least 1-norm criterion is used for the coefficient estimation, and sparse deconvolution is applied for deriving sparse residuals.

In this paper, we propose a speech analysis method based on the PZ model with mixed excitation. Using the mixed excitation and PZ modeling together, we combine the advantages of non-sparse and sparse algorithms, and obtain a better fitting for both the excitation and SPF spectrum. Using the PZ model, instead of the all-pole model, a better spectral fitting can be expected. Moreover, we model both the voiced, the unvoiced excitation or a mixture of them by the mixed excitation. Additionally, block sparsity is imposed on the voiced excitation component, motivated by the quasi-periodic and temporal-correlated nature of the glottal excitation [12, 21]. The posterior probability density functions (PDFs) for the sparse excitation and hyperparameters, as well as point estimates of the PZ model parameters are obtained using the VEM method.

2 Signal models

Consider the following general speech observation model:

$$y\left(n\right) = s\left(n\right) + u(n), \tag{E.1}$$

where $y(n)$ is the observation signal and $u(n)$ denotes the noise. We assume that the clean speech signal $s(n)$ is produced by the PZ speech production model, i.e.,

$$s\left(n\right) = -\sum_{k=1}^{K} a_k s\left(n-k\right) + \sum_{l=0}^{L} b_l e\left(n-l\right) + m(n), \tag{E.2}$$

where a_k and b_l are the modeling coefficients of the PZ model with $b_0 = 1$, $e\left(n\right)$ is a sparse excitation corresponding to the voiced part and $m(n)$ is the non-sparse Gaussian excitation component corresponding to the unvoiced part. Assuming $s\left(n\right) = 0$ for $n \leq 0$ and considering one frame of speech

signals of N samples, (E.1) and (E.2) can be written in matrix forms as

$$\mathbf{y} = \mathbf{s} + \mathbf{u}, \tag{E.3}$$

$$\mathbf{As} = \mathbf{Be} + \mathbf{m}, \tag{E.4}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the $N \times N$ lower triangular Toeplitz matrices with $[1, a_1, a_2, \cdots, a_K, 0, \cdots, 0]$ and $[1, b_1, b_2, \cdots, b_L, 0, \cdots, 0]$ as the first columns, respectively. The block sparse residuals are defined as $\mathbf{e} = [e(1) \cdots e(N)]^T$, and $\mathbf{m}$, $\mathbf{s}$, $\mathbf{y}$ and $\mathbf{u}$ are defined similarly to $\mathbf{e}$. When $L = 0$, $\mathbf{B}$ reduces to the identity matrix and (E.4) becomes the all-pole model. Combining (E.3) and (E.4), the noisy observation can be written as

$$\mathbf{Ay} = \mathbf{Be} + \mathbf{m} + \mathbf{Au}. \tag{E.5}$$

In [20], we assumed that only the sparse excitation was present ($\mathbf{m} = \mathbf{0}$, but $\mathbf{u} \neq \mathbf{0}$). The sparse residuals and model parameters were estimated iteratively. The sparse residuals were obtained by solving

$$\min_{\mathbf{e}} \|\mathbf{e}\|_1^1 \quad \text{s.t.} \quad \left\|\mathbf{y} - \mathbf{A}^{-1}\mathbf{Be}\right\|_2^2 \leq C. \tag{E.6}$$

where C is a constant proportional to the variance of the noise. The model parameters was estimated using the l_1 norm of the residuals as the cost function (see [20] for details).

3 Proposed variational EM method

We now proceed to consider the noise-free scenario but with mixed excitation ($\mathbf{u} = \mathbf{0}$, but $\mathbf{m} \neq \mathbf{0}$). We consider the pole-zero model parameters $\mathbf{a} = [a_1, a_2, \cdots, a_K]^T$ and $\mathbf{b} = [b_1, b_2, \cdots, b_L]^T$ to be deterministic but unknown. Utilizing the SBL [13] methodology, we first express the hierarchical form of the model as

$$\begin{aligned} \mathbf{Ay} = \mathbf{Be} + \mathbf{m}, \qquad & \mathbf{m} \sim \mathcal{N}(\mathbf{0}, \gamma_m^{-1}\mathbf{I}_N), \\ \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_e^{-1}), \quad \gamma_m \sim \Gamma(c, d), \quad & \boldsymbol{\alpha} \sim \prod_{o=1}^{O} \Gamma(\alpha_o; e, f), \end{aligned} \tag{E.7}$$

where O is the number of blocks, $\mathbf{\Gamma}_e = \text{diag}(\boldsymbol{\alpha}) \otimes \mathbf{I}_D$, $\otimes$ is the Kronecker product, D is the block size, $N = DO$, $\mathcal{N}$ denotes the multivariate normal PDF and Γ is the Gamma PDF. The hyperparameter α_o is the precision of the o^{th} block, and when it is infinite, the o^{th} block will be zero. Note that it is trivial to extend the proposed method to any D. Moreover, when $D = 1$, each element in $\mathbf{e}$ is inferred independently. Here, block sparsity model is used to take the quasi-periodic and temporal-correlated nature of the voiced

excitation into account. The $\mathbf{m}$ is used for capturing the white noise excitation from unvoiced speech frame or a mixture of phonations.

Our objective is to obtain the posterior densities of $\mathbf{e}$, γ_m and $\boldsymbol{\alpha}$, and point estimates of the model parameters in $\mathbf{a}$ and $\mathbf{b}$. First, we write the complete likelihood, i.e.,

$$\begin{aligned} p(\mathbf{y}, \mathbf{e}, \boldsymbol{\alpha}, \gamma_m) &= p(\mathbf{y}|\mathbf{e}, \gamma_m) p(\mathbf{e}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\gamma_m) \\ &= \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)\mathcal{N}(\mathbf{e}|0, \boldsymbol{\Gamma}_e^{-1}) \\ &\quad \times \Gamma(\gamma_m; c, d) \prod_{o=1}^{O} \Gamma(\alpha_o; e, f), \end{aligned} \tag{E.8}$$

where we used $\mathcal{N}(\mathbf{y}|\mathbf{A}^{-1}\mathbf{B}\mathbf{e}, \gamma_m^{-1}(\mathbf{A}^T\mathbf{A})^{-1}) = \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)$ when $\det(\mathbf{A}) = 1$. Instead of finding the joint posterior density $p(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m|\mathbf{y})$, which is intractable, we adopt the variational approximation [22]. Assume that $p(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m|\mathbf{y})$ is approximated by the density $q(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m)$, which may be fully factorized as

$$q(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m) = q(\mathbf{e}) q(\gamma_m) \prod_{o=1}^{O} q(\alpha_0), \tag{E.9}$$

where the factors are found using an EM-like algorithm [22].

In the E-step of the VEM method, we fix the model parameters $\mathbf{a}$ and $\mathbf{b}$, and re-formulate the posterior PDFs estimation problem as maximizing the variational lower bound

$$\max_q \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{e}, \boldsymbol{\alpha}, \gamma_m)] + H[q], \tag{E.10}$$

where q is the shorthand for $q(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m)$, $H[q]$ is defined as $H[q] = -\mathbb{E}_q[\log(q)]$, and $\mathbb{E}_{q(x)}[f(x)]$ denotes the expectation of $f(x)$ w.r.t. the random variable x (i.e., $\mathbb{E}_{q(x)}[f(x)] = \int f(x)q(x)dx$). Substituting (E.8) and (E.9) into (E.10), and following the derivation from [22], we obtain

$$\begin{aligned} q(\mathbf{e}) &\propto e^{\mathbb{E}_{q(\boldsymbol{\alpha}, \gamma_m)}[\log \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)\mathcal{N}(\mathbf{e}|0, \mathrm{diag}(\boldsymbol{\alpha})^{-1} \otimes \mathbf{I}_D)]}, \\ q(\alpha_o) &\propto \Gamma(\alpha_o; e, f) e^{\mathbb{E}_{q(\mathbf{e})}[\log \mathcal{N}(\mathbf{e}|0, \mathrm{diag}(\boldsymbol{\alpha})^{-1} \otimes \mathbf{I}_D)]}, \\ q(\gamma_m) &\propto \Gamma(\gamma_m; c, d) e^{\mathbb{E}_{q(\mathbf{e})}[\log \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)]}. \end{aligned} \tag{E.11}$$

It is clearly seen that $q(\mathbf{e})$ in (E.11) is a Gaussian PDF, i.e.,

$$q(\mathbf{e}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{E.12}$$

where $\tilde{\boldsymbol{\Sigma}} = (\mathbb{E}[\gamma_m]\mathbf{B}^T\mathbf{B} + \mathbb{E}[\boldsymbol{\Gamma}_e])^{-1}$ and $\tilde{\boldsymbol{\mu}} = \mathbb{E}[\gamma_m]\tilde{\boldsymbol{\Sigma}}\mathbf{B}^T\mathbf{A}\mathbf{y}$. We also define the auto-correlation matrix $\tilde{\mathbf{R}} = \tilde{\boldsymbol{\Sigma}} + \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T$. The posterior PDF of α_o in (E.11) is a Gamma probability density, i.e.,

$$q(\alpha_o) = \Gamma(\tilde{e}_o, \tilde{f}_o), \tag{E.13}$$

where $\tilde{e}_o = e + D/2$, $\tilde{f}_o = f + \sum_{i=(o-1)D+1}^{oD} \tilde{\mathbf{R}}_{i,i}/2$ and $\tilde{\mathbf{R}}_{i,i}$ denotes the (i,i) element of $\tilde{\mathbf{R}}$. The expectation of the precision matrix is $\mathbb{E}[\boldsymbol{\Gamma}_e] = \text{diag}(\tilde{e}_1/\tilde{f}_1, \cdots, \tilde{e}_O/\tilde{f}_O) \otimes \mathbf{I}_D$. Similar to α_o, the posterior PDF of γ_m is

$$q(\gamma_m) = \Gamma(\tilde{c}, \tilde{d}), \tag{E.14}$$

where $\tilde{c} = c + N/2$, $\tilde{d} = d + (\text{tr}(\tilde{\boldsymbol{\Sigma}}\mathbf{B}^T\mathbf{B}) + \|\mathbf{A}\mathbf{y} - \mathbf{B}\tilde{\boldsymbol{\mu}}\|_2^2)/2$. The expectation of γ_m can be expressed as $\mathbb{E}[\gamma_m] = \tilde{c}/\tilde{d}$.

In the M-step, we maximize the lower bound (E.10) w.r.t. the modeling parameters $\mathbf{a}$ and $\mathbf{b}$, respectively. The optimization problems can be shown to be equivalent to $\min_{\mathbf{a}} \mathbb{E}_{q(\mathbf{e})}\|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2$ and $\min_{\mathbf{b}} \mathbb{E}_{q(\mathbf{e})}\|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2$, respectively. To obtain the estimate for $\mathbf{a}$, we first note that $\mathbf{A}\mathbf{y}$ can be expressed as $\mathbf{A}\mathbf{y} = \mathbf{C}\mathbf{a} + \mathbf{y}$, where $\mathbf{C}$ is a $N \times K$ Toeplitz matrix of the form

$$\mathbf{C} = \begin{bmatrix} 0 & \cdots & 0 \\ y(1) & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \vdots & & y(1) \\ \vdots & & \vdots \\ y(N-1) & \cdots & y(N-K) \end{bmatrix}_{N \times K}$$

Using this expression and $q(\mathbf{e})$ obtained in the E-step, the minimization problem w.r.t. $\mathbf{a}$ can be re-formulated as

$$\min_{\mathbf{a}} \mathbb{E}_{q(\mathbf{e})}\|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2 \iff \min_{\mathbf{a}} \|(\mathbf{B}\tilde{\boldsymbol{\mu}} - \mathbf{y}) - \mathbf{C}\mathbf{a}\|_2^2. \tag{E.15}$$

As can be seen, (E.15) is the standard least squares problem and has the analytical solution as

$$\mathbf{a} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T(\mathbf{B}\tilde{\boldsymbol{\mu}} - \mathbf{y}). \tag{E.16}$$

We can obtain the solution of $\mathbf{b}$, like $\mathbf{a}$, by setting the derivative of $\mathbb{E}_{q(\mathbf{e})}\|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2$ w.r.t. $\mathbf{b}$ to zero, i.e.,

$$\begin{aligned} \frac{\partial \mathbb{E}_{q(\mathbf{e})}\|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2}{\partial \mathbf{b}} &= \mathbb{E}_{q(\mathbf{e})}[2\mathbf{F}^T(\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e})] \\ &= \mathbf{0}_{L \times 1}. \end{aligned} \tag{E.17}$$

where $\mathbf{F}$ is an $N \times L$ lower triangular Toeplitz matrix of the form

$$\mathbf{F} = \begin{bmatrix} 0 & \cdots & 0 \\ e(1) & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \vdots & & e(1) \\ \vdots & & \vdots \\ e(N-1) & \cdots & e(N-L) \end{bmatrix}_{N \times L}$$

From (E.17), we obtain the estimate of $\mathbf{b}$, i.e.,

$$\mathbf{b} = (\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{F}])^{-1}(\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T]\mathbf{A}\mathbf{y} - \mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{e}]), \tag{E.18}$$

where $\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{F}]$ is an $L \times L$ symmetric matrix with the $(i,j)^{\text{th}}, j \geq i$ element given by $\sum_{k=1}^{N-j} \tilde{\mathbf{R}}_{k,k+j-i}$. The $\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T]$ can be obtained by simply replacing the stochastic variable $e(n), 1 \leq n \leq N-1$ in $\mathbf{F}^T$ with the mean estimate $\tilde{\mu}(n)$ (the n^{th} element in $\tilde{\boldsymbol{\mu}}$). The $\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{e}]$ is an $L \times 1$ vector with the l^{th} element given by $\sum_{k=1}^{N-l} \tilde{\mathbf{R}}_{k,k+l}$. Note that the estimation of $\mathbf{b}$ in (E.18) requires the knowledge of $\mathbf{a}$ and vice versa (see (E.16)). This coupling is solved by replacing them with their estimates from previous iteration. The algorithm is initialized with $\mathbf{a} = [1, 0, \cdots, 0_K]^T$, $\mathbf{b} = [1, 0, \cdots, 0_L]^T$, $\gamma_m = 10$ and $\alpha_o = 1, o = 1, \cdots, O$, and starts with the update of $\mathbf{e}$. We refer to the proposed variational expectation maximization pole-zero estimation algorithm as the VEM-PZ.

4 Results

In this section, we compare the performance of the proposed VEM-PZ, the two-stage least squares pole-zero (TS-LS-PZ) method [14], 2-norm linear prediction (2-norm LP) [1], 1-norm linear prediction (1-norm LP) [9] and expectation maximization based linear prediction (EM-LP) for mixed excitation [12] in both synthetic and real speech signals analysis scenarios.

4.1 Synthetic signal analysis

We first examine the performance of the VEM-PZ with different block size D and compares it with traditional algorithms using synthetic voiced consonant /n/, as shown in Fig. E.1 and Fig. E.2. The synthetic signals are generated by convolving an artificial glottal source waveform with a constructed filter. The first derivative of the glottal flow pulse is simulated with the Liljencrants-Fant (LF) waveform [23] with the modal phonation mode, whose parameter is taken from Gobl [24]. The voiced alveolar nasal consonant /n/ is synthesized

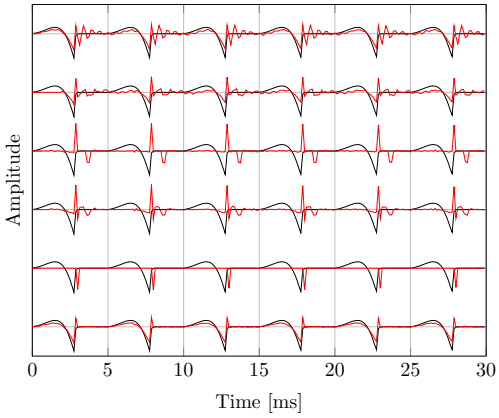


Fig. E.1: residuals estimate for synthetic /n/. The black line is the LF excitation. The red lines shown correspond to, from top to bottom, residuals of (1) 2-norm LP, (2) TS-LS-PZ, (3) 1-norm LP, (4) EM-LP, D=8, (5) VEM-PZ, D=1, (6) VEM-PZ, D=8.

Table E.1: The spectral distortion

F0	200	250	300	350	400
2-norm LP	1.79	**2.14**	2.12	2.53	2.13
TS-LS-PZ	2.41	4.77	1.88	1.46	2.86
1-norm LP	2.43	3.15	3.60	3.29	4.29
EM-LP	5.62	6.68	4.68	3.96	4.83
VEM-PZ, D=1	4.50	7.14	2.29	1.54	2.31
VEM-PZ, D=5	1.55	4.47	**0.69**	2.01	4.50
VEM-PZ, D=7	2.08	4.07	2.18	**1.41**	1.29
VEM-PZ, D=8	**0.77**	5.56	2.52	4.86	**0.53**

at 8 kHz sampling frequency with the constructed filter having two formant frequencies (bandwidths) of 257 Hz (32 Hz) and 1891 Hz (100 Hz) and one antiformant of 1223 Hz (52 Hz) [25]. N is set to 240 (30 ms), $e = 1$ and $c = d = f = 10^{-6}$ are used for all the experiments. The power ratio of the block sparse excitation **e** and Gaussian component **m** is set to 30 dB, and the pitch is set to 200 Hz. The K and L are set to 5 for the pole-zero modeling methods (i.e., VEM-PZ and TS-LS-PZ), but $K = 10$ is used for the all-pole modeling methods (i.e., 2-norm LP, 1-norm LP and EM-LP). In Fig. E.1, the means of the residuals of EM-LP and VEM-PZ are plotted. Note that the residuals of the TS-LS-PZ, 1-norm LP and EM-LP are prepended with zeros due to the covariance-based estimation methods. As can be seen in Fig. E.1, the residual estimate of the proposed VEM-PZ with $D = 8$ is closest to the true block

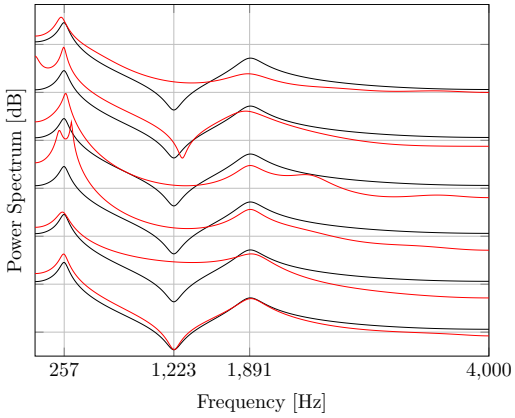


Fig. E.2: corresponding spectral estimates for synthetic /n/. The red lines have the same setting as Fig. 1.

sparse excitation. Moreover, when $D = 1$, the residuals of the VEM-PZ are the sparsest as expected. Although the EM-LP also produces block sparse residuals compared with the 1-norm LP, the proposed method achieves the best approximation to the true one due to the usage of the pole-zero modeling and the block-sparse motivated SBL method. The corresponding spectral estimates are shown in Fig. E.2. First, as can be seen, the VEM-PZ with $D = 1$ has a smooth power spectrum due to the sparse residuals in Fig. E.1. Second, the 1-norm LP tends to produce a better estimate of the formants than the 2-norm LP and TS-LS-PZ. Third, although the EM-LP has two peaks around the first formant, it performs well for second formant estimation. Finally, the proposed VEM-PZ with $D = 8$ has good performance for both formant, antiformant and bandwidth estimates because of the block sparse excitation and the pole-zero model.

Second, the spectral distortion is tested under different fundamental frequencies and block sizes. The measure is defined as the truncated power cepstral distance [26], i.e.,

$$d_{\text{ceps}} = \sum_{n=-S}^{S} (c_n - \hat{c}_n)^2, \tag{E.19}$$

where c_n and $\hat{c}_n$ are the true and estimated power cepstral coefficients, respectively. Cepstral coefficients are computed by first reflecting all the poles and zeros to the inside of the unit circle and then using the recursive relation in [25]. In our experiments, we set $S = 300$. The fundamental frequency rises from 200 to 400 Hz in steps of 50 Hz. The experimental results in TABLE E.1 are obtained by the ensemble averages over 500 Monte Carlo experiments. $D = 6$ is used for the EM-LP [12]. As can be seen from TABLE

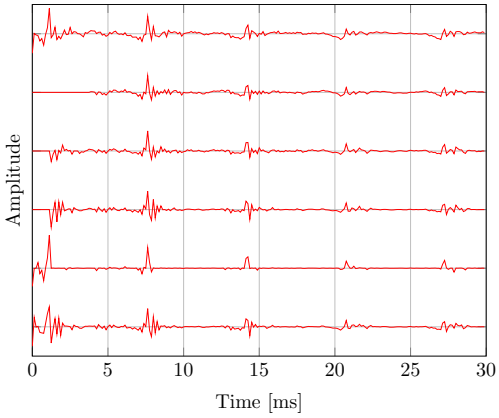


Fig. E.3: residuals estimate for /n/ in the word "manage". The red lines have the same setting as Fig. 1.

E.1, the 2-norm LP has a lower spectral distortion than the 1-norm LP, EM-LP and TS-LS-PZ (except for 300 and 350 Hz). The proposed VEM-PZ achieves the lowest spectral distortion for 200, 300, 350 and 400 Hz. However, note that the good performance of the VEM-PZ depends on a good choice of the block sizes for different fundamental frequencies, and there is a fluctuation when the frequency changes. This is because the length of correlated samples changes with the fundamental frequency. But, as can be seen from Fig. E.2 and from our experience, the VEM-PZ usually produces better formant, antiformant and bandwidth estimates than traditional ones.

4.2 Speech signal analysis

We examine the residuals and spectral estimate of the VEM-PZ for a nasal consonant /n/ in the word "manage" from the CMU Arctic database [27, 28], pronounced by an US female native speaker, sampled of 8000 Hz. The results are shown in Fig. E.3 and Fig. E.4. To improve the modeling flexibility, the K and L are set to 10 for the PZ methods (i.e., VEM-PZ and TS-LS-PZ), but $K = 10$ is still used for the all-pole methods (i.e., 2-norm LP, 1-norm LP and EM-LP). As can be seen from Fig. E.3, the residuals of the EM-LP and 1-norm LP are sparser than the 2-norm LP. The residuals of the proposed VEM-PZ with $D = 1$ are the sparsest. The proposed VEM-PZ with $D = 8$ is block sparse and is sparser than the TS-LS-PZ. From Fig. E.4, we can see that all the algorithms have formant estimates around 150 Hz. However, the TS-LS-PZ, 1-norm LP and VEM-PZ with $D = 1$ have very peaky behaviour. Also, the 2-norm LP produces bad bandwidth estimates around 2000 and 2900 Hz. Furthermore, compared to the EM-LP, the proposed VEM-PZ with D=8 has good antiformant estimates around 500 and 1500 Hz.

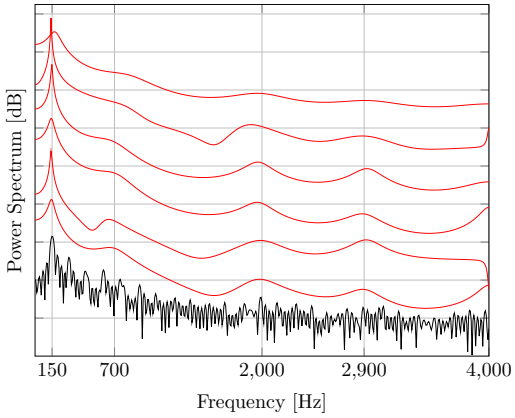


Fig. E.4: corresponding spectral estimates for /n/ in the word "manage". The red lines have the same setting as Fig. 1. The black line is the periodogram.

5 Conclusion

A variational expectation maximization pole-zero speech analysis method has been proposed. By using the pole-zero model, it can fit the spectral zeros of speech signals easily. Moreover, block sparse residuals are encouraged by applying the sparse Bayesian learning method. By iteratively updating parameters and statistics of residuals and hyperparameters, block sparse residuals can be obtained. The good performance has been verified by both synthetic and real speech experiments. The proposed method is promising for speech analysis applications. Next, further research into the formant, antiformant and bandwidth estimation accuracy, stability, and unknown sparse pattern should be conducted.

References

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[2] J. Pohjalainen, C. Hanilci, T. Kinnunen, and P. Alku, "Mixture linear prediction in speaker verification under vocal effort mismatch," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1516–1520, dec 2014.

[3] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[4] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 221–239, 2000.

[5] T. Drugman and Y. Stylianou, "Fast inter-harmonic reconstruction for spectral envelope estimation in high-pitched voices," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1418–1422, 2014.

[6] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, 1991.

[7] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 1, pp. 65–73, 2008.

[8] P. Alku, J. Pohjalainen, M. Vainio, A. M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction." *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1295–1313, Auguest 2013.

[9] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, jul 2012.

[10] D. Giacobello, M. G. Christensen, T. L. Jensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Stable 1-norm error minimization based linear predictors for speech modeling," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 5, pp. 912–922, may 2014.

[11] T. L. Jensen, D. Giacobello, T. V. Waterschoot, and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Commun.*, vol. 76, pp. 143–156, 2016.

[12] R. Giri and B. D. Rao, "Block sparse excitation based all-pole modeling of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014.

[13] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 2001, pp. 211– 244, 2001.

[14] P. Stoica and R. Moses, *Spectral Analysis of Signals*. New Jersey: Prentice Hall, Upper Saddle River, 2004.

[15] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 237–248, 2010.

[16] J. Durbin, "The fitting of time-series models," *Rev. Int'l Statistical Inst.*, vol. 28, no. 3, pp. 233–244, 1960.

[17] E. Levy, "Complex-curve fitting," *IRE Trans. Automat. Contr.*, vol. AC-4, no. 1, pp. 37–43, 1959.

[18] K. SteTiange, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 229–234, 1977.

[19] T. Kobayashi and S. Imai, "Design of IIR digital filters with arbitrary log magnitude function by WLS techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 247–252, 1990.

[20] L. Shi, J. R. Jensen, and M. G. Christensen, "Least 1-norm pole-zero modeling with sparse deconvolution for speech analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, to be published.

[21] P. Alku, "Glottal inverse filtering analysis of human voice production," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 36, 2011.

[22] C. M. Bishiop, *Pattern recognition and machine learning*, M. Jordan, J. Kleinberg, and B. Scho, Eds. Springer, 2006.

[23] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR4 (Speech, Music and Hearing, Royal Institude of Technology, Stochholm, Sweden)*, pp. 1–13, 1985.

[24] C. Gobl, "A preliminary study of acoustic voice quality correlates," *STL-QPSR4 (Speech, Music and Hearing, Royal Institude of Technology, Stochholm, Sweden)*, pp. 9– 22, 1989.

[25] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, Sep. 2012.

[26] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[27] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Speech Synthesis Workshop*. Pittsburgh, PA, 2004.

[28] CMU-ARCTIC Speech Synthesis Databases. [Online]. Available: http://festvox.org/cmu_arctic/index.html

Paper F

Least 1-norm Pole-zero Modeling with Sparse Deconvolution for Speech Analysis

Liming Shi, Jesper Rindom Jensen and Mads Græsbøll Christensen

The paper has been published in the
Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2017

Abstract

In this paper, we present a speech analysis method based on sparse pole-zero modeling of speech. Instead of using the all-pole model to approximate the speech production filter, a pole-zero model is used for the combined effect of the vocal tract; radiation at the lips and the glottal pulse shape. Moreover, to consider the spiky excitation form of the pulse train during voiced speech, the modeling parameters and sparse residuals are estimated in an iterative fashion using a least 1-norm pole-zero with sparse deconvolution algorithm. Compared with the conventional two-stage least squares pole-zero, linear prediction and sparse linear prediction methods, experimental results show that the proposed speech analysis method has lower spectral distortion, higher reconstruction SNR and sparser residuals.

1 Introduction

Speech modeling, as a fundamental speech analysis problem, has diverse applications in speech synthesis [1], speaker identification, speech recognition, etc. Based on the source-filter model of the speech production system, the speech production filter (SPF) is assumed to be time-invariant during a short-time period (frame) of approximately 20-40 ms, and excited by a pulse train or white noise for voiced or unvoiced speech.

Linear prediction (LP) with least squared error minimization criterion, based on an all-pole model, is commonly used for speech analysis [2]. The method performs well for white noise and small valued pitch harmonic excitations (aka residuals). However, for a large valued pitch, it tends to null out the input voiced speech harmonics and leads to an all-pole filter with poles close to the unit circle, and the estimated spectral envelope has a sharper contour than desired [3, 4]. Various improved schemes based on LP have been proposed, such as LP with the Itakura-Saito error criterion [5], all-pole modeling with a distortionless response at frequencies of harmonics [3] and the regularized LP [6]. More recently, motivated by the compressive sensing framework, sparse linear prediction based on the 1-norm criterion has been proposed for voiced speech analysis [7]. Unlike the conventional 2-norm method, sparse priors on the excitation signals and prediction coefficients are both utilized to offer an effective decoupling of the SPF and underlying sparse residuals. Moreover, the 1-norm method was shown to be robust against impulsive interference in all-zero plant identification [8, 9]. Fast methods and the stability of the 1-norm cost function for spectral envelope estimation are further investigated in [10, 11]. Another problem is that some sounds containing spectral zeros with voiced excitation, such as nasals, fricatives, or laterals, are poorly estimated by an all-pole model but trivial with a pole-zero model [12–14]. The estimation of the coefficients of the pole-

zero model can be obtained separately [15], jointly [16] or iteratively [12]. A model identification method is proposed for time-varying stochastic pole-zero model estimation [17]. A 2-norm minimization criterion with Gaussian residual assumption is usually used to obtain the parameter estimates in these methods. Motivated by the logarithmic scale perception of the human auditory system, the logarithmic magnitude function minimization criterion has also been proposed [14, 18]. Additionally, the performance of the all-pole method deteriorates severely in noisy conditions. Various noise robust approaches based on all-pole model have been proposed [19, 20].

In this paper, a speech analysis method based on sparse pole-zero modeling is presented. Using a pole-zero model for fitting the spectral envelope compared with the all-pole model, a better approximation can be obtained. The modeling coefficients and residuals are obtained in a iterative fashion. To consider the sparse priors of residuals, instead of conventional 2-norm minimization criterion, a least 1-norm criterion is used for the coefficient estimation. Moreover, sparse deconvolution is applied for deriving sparse residuals and denoising. The effectiveness of the proposed method for the spectral envelope estimation and signal reconstruction is verified using both synthetic signals and natural speech.

2 Fundamentals of the pole-zero estimation

The pole-zero speech production filter model is considered in this paper. A sample of speech is written in the following form:

$$\begin{aligned} s\left(n\right) =& -\sum_{k=1}^{K} a_k s\left(n-k\right) + \sum_{l=0}^{L} b_l e\left(n-l\right) \\ x\left(n\right) =& s\left(n\right) + m(n) \end{aligned} \tag{F.1}$$

where a_k and b_l are coefficients of the pole-zero model, $b_0 = 1$, $m(n)$ is Gaussian noise, and $e\left(n\right)$ is the residual.

When $L = 0$, (1) reduces to the all-pole model and the parameter estimation can be formulated as

$$\min_{\mathbf{a}} \left\|\mathbf{x} + \mathbf{X}\mathbf{a}\right\|_p^p + \gamma \left\|\mathbf{a}\right\|_q^q \tag{F.2}$$

where $\mathbf{x} = \left[x\left(N_1\right), x\left(N_1+1\right) \cdots x\left(N_2\right)\right]^T$, $\mathbf{a} = \left[a_1, a_2 \cdots a_K\right]^T$, $\left[\cdot\right]^T$ denotes matrix transpose, $\left\|\cdot\right\|_p$ is the p-norm, γ is the regularization parameter and

$$\mathbf{X} = \begin{bmatrix} x\left(N_1-1\right) & \cdots & x\left(N_1-K\right) \\ \vdots & & \vdots \\ x\left(N_2-1\right) & \cdots & x\left(N_2-K\right) \end{bmatrix}$$

N_1 and N_2 can be chosen in various ways. One way is setting $N_1 = 1$ and $N_2 = N + K$, which is the autocorrelation method. The covariance method is obtained by setting $N_1 = K + 1$ and $N_2 = N$ [21]. When the residual signal $e(n)$ is a Gaussian random variable, the standard 2-norm solution (i.e., $p = 2$ and $\gamma = 0$) of (1) is statistically equivalent to the maximum likelihood solution. However, the pulse train excitation for voiced speech can be better fitted as a super-Gaussian variable, known for its long-tail distribution. A sparse solution for the residuals can be obtained in principle by setting $0 \leq p \leq 1$. Moreover, prior knowledge of coefficients can be incorporated as a regularization term to improve the parameter estimation. For example, the generalized Gaussian or Laplacian distribution can be imposed to $\mathbf{a}$ by choosing $q = 2$ or $q = 1$, respectively. In [22], $p = 1$ and $\gamma = 0$ are used to obtain sparse residuals. [7] use $p = 1$ and $q = 1$ to encourage both sparse residuals and coefficients.

When $L > 0$, both poles and zeros present. The pole-zero model is known to be more effective than all-pole model, especially for fitting nasal sound [12–14, 18]. However, it inherently involves solving a nonlinear equation. A two-stage pole-zero method was proposed [15, 21]. In the first stage, a coarse estimate of $\widehat{\mathbf{e}} = \mathbf{x} + \mathbf{X}\widehat{\mathbf{a}}$ can be obtained by using (2) with a sufficiently high-order linear prediction K'. Then, replace $e(n)$ ($K' + 1 \leq n \leq N$) in (1) by $\hat{e}(n)$ determined in the first stage, and solve the following minimization problem:

$$\min_{\mathbf{z}} \left\| \mathbf{x}' + \mathbf{X}'\mathbf{z} \right\|_p^p + \gamma \left\| \mathbf{z} \right\|_q^q \tag{F.3}$$

where $\mathbf{x}' = [x(N_1'), x(N_1' + 1) \cdots x(N_2')]^T$, $\mathbf{X}' = [\bar{\mathbf{X}}, -\widehat{\mathbf{E}}]$ and

$$\bar{\mathbf{X}} = \begin{bmatrix} x(N_1' - 1) & \cdots & x(N_1' - K) \\ \vdots & & \vdots \\ x(N_2' - 1) & \cdots & x(N_2' - K) \end{bmatrix}$$

$$\widehat{\mathbf{E}} = \begin{bmatrix} \hat{e}(N_1' - 1) & \cdots & \hat{e}(N_1' - L) \\ \vdots & & \vdots \\ \hat{e}(N_2' - 1) & \cdots & \hat{e}(N_2' - L) \end{bmatrix}$$

and $\mathbf{z} = [a_1 \cdots a_K, b_1 \cdots b_L]^T$. N_1' and N_2' are usually set to $K' + L + 1$ and N, respectively. When we set p to 2 and γ to 0 in (2) and (3), the above formulation is the standard two-stage least square pole-zero (TS-LS-PZ) method [21].

Algorithm 6 SD-L1-PZ

Intiate γ, q, $p_1 = 2$ and $q_1 = 1$.
Initialization with the TS-L1-PZ:
  Solve (9) with a large K', and $\widehat{\mathbf{e}} = \mathbf{x} + \mathbf{X}\widehat{\mathbf{a}}$
  Coefficients estimation by (7) with ${N_1}' = K' + L + 1$
for $k = 1, \cdots$ **do**
  Calculate $\mathbf{A}^{-1}\mathbf{B}$
  Obtain refined sparse residual $\widehat{\mathbf{e}}_k$ using (5)
  Solve coefficients using (7) with ${N_1}' = \max(K, L) + 1$
  while poles or zeros are larger than 1 **do**
    Compute re-estimated coefficients using (8)
  end while
end for

3 Least 1-norm pole-zero modeling with sparse deconvolution (SD-L1-PZ)

There are problems with the above TS-LS-PZ method. First, to obtain an accurate estimate of the parameter vector $\mathbf{z}$, $K^{'}$ should be very large. As a result, a large segment of data $e(n)$ $(1 \leq n \leq K^{'})$ cannot be estimated, and the observations are not sufficiently used for estimating model parameters. Second, in presence of pulse train residuals, especially for a high valued pitch, the 2-norm based criterion is inappropriate for the estimation in both stages. Besides, this method does not consider the influence of the additive noise $m(n)$, which affects the reconstruction performance. To overcome these problems, we proceed to design a least 1-norm based pole-zero modeling with sparse deconvolution algorithm for speech analysis.

3.1 Finding a sparse residual

Instead of estimating the residuals using a sufficiently high-order all-pole model, a deconvolution method can be used [23]. First, (1) can be reformulated in a matrix form as follows

$$\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{e} + \mathbf{A}\mathbf{m} \tag{F.4}$$

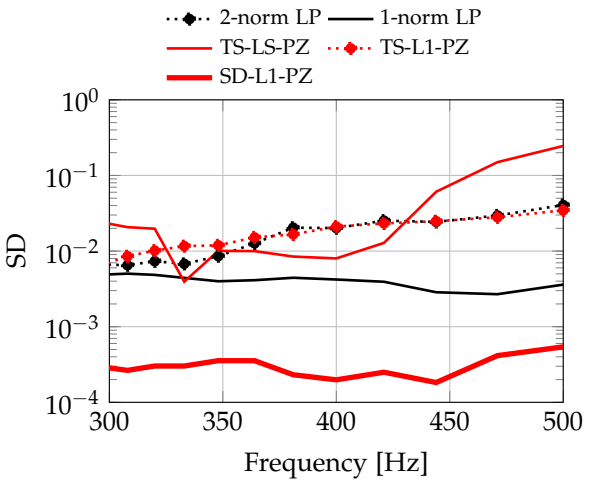


Fig. F.1: Spectral distortion for different excitation frequencies

where $\mathbf{e} = [e(1) \cdots e(N)]^T$, $\mathbf{m} = [m(1) \cdots m(N)]^T$, $\mathbf{A}$ and $\mathbf{B}$ have the form

$$\mathbf{A} = \begin{bmatrix} 1 & & & & & \\ a_1 & 1 & & & & \\ \vdots & \cdots & \ddots & & & \\ a_K & \cdots & a_1 & 1 & & \\ & \ddots & & & \ddots & \\ & & a_K & \cdots & a_1 & 1 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & & & & & \\ b_1 & 1 & & & & \\ \vdots & \cdots & \ddots & & & \\ b_L & \cdots & b_1 & 1 & & \\ & \ddots & & & \ddots & \\ & & b_L & \cdots & b_1 & 1 \end{bmatrix}$$

When $L = 0$, $\mathbf{B}$ reduces to the identity matrix and (4) becomes a standard AR estimation problem. The usage of both poles and zeros is to obtain a better fitting to sounds containing spectral zeros. With a known or estimated parameter vector $\mathbf{z}$, the estimation of residuals $\mathbf{e}$ can be formulated as

$$\min_{\mathbf{e}} \|\mathbf{e}\|_{q_1}^{q_1} \quad \text{s.t.} \quad \left\|\mathbf{x} - \mathbf{A}^{-1}\mathbf{B}\mathbf{e}\right\|_{p_1}^{p_1} \leq C \tag{F.5}$$

Assuming Gaussian noise and pulse train residuals, a reasonable choice is to set $p_1 = 2$ and $q_1 = 1$ in (5). A good approximation between observations

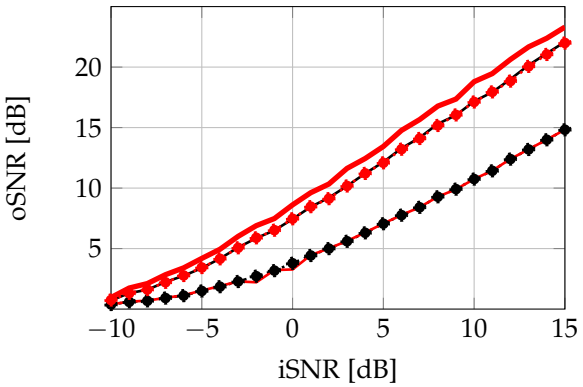


Fig. F.2: SNR of reconstructed signals over different input SNR (iSNR)

and modeling fitting is guaranteed by the 2-norm constraint. The 1-norm cost is to deemphasize the spiky residuals associated with pulse train excitation. Note that setting C larger yields an estimate with poorer data fitting but sparser residuals, and vice versa. When the variance of the noise is known, a good choice is to set $C = N\sigma_m^2$. The reconstruction of speech signal can be obtained as

$$\widehat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{B}\widehat{\mathbf{e}} \tag{F.6}$$

where $\widehat{\mathbf{e}}$ denotes the residual estimate obtained by (5).

3.2 Estimation of pole-zero modeling coefficients

With known or estimated residuals, we estimate the pole-zero modeling coefficients $\mathbf{z}$ using the second stage of the TS-LS-PZ algorithm but with $p = 1$ and ${N_1}' = \max(K, L) + 1$, instead of the conventional $p = 2$ and ${N_1}' = K' + L + 1$, i.e.,

$$\min_{\mathbf{z}} \left\|\mathbf{x}' + \mathbf{X}'\mathbf{z}\right\|_1^1 + \gamma \left\|\mathbf{z}\right\|_q^q \tag{F.7}$$

To account for the non-Gaussian distribution characteristics of the residuals, a 1-norm minimization criterion is again used here instead of conventional 2-norm. Also, since complete estimates of residuals are available, observations can be sufficiently used with a smaller ${N_1}'$. Furthermore, as noted before, prior knowledge of coefficients can be incorporated as regularization term. Especially, when $q = 1$, and high orders of K and L are used, it will lead to sparse pole-zero coefficient estimates (see [7] for details about sparse linear prediction). The estimation of residuals and pole-zero modeling coefficients are repeated until convergence. To guarantee the causality and stability of

the proposed method for both estimation and reconstruction, the parameter vector can be re-estimated using

$$\min_{\mathbf{z}} \left\|\mathbf{x}' + \mathbf{X}'\mathbf{z}\right\|_1^1 + \gamma \left\|\mathbf{z}\right\|_q^q \quad \text{s.t.} \quad \|\mathbf{a}\| \leq 1, \|\mathbf{b}\| \leq 1 \tag{F.8}$$

when the poles or zeros are outside of the unit circle.

Furthermore, for the initialization of this iteration procedure, we modify the first stage of the TS-LS-PZ to the 1-norm formulation [22]

$$\min_{\mathbf{a}} \|\mathbf{x} + \mathbf{X}\mathbf{a}\|_1^1 \tag{F.9}$$

In the second stage, (7) is used to replace the original cost function (3) with $p = 2$. Due to the 1-norm cost function, we refer this initialization approach to the two-stage least 1-norm pole-zero (TS-L1-PZ). We summarize the SD-L1-PZ in Algorithm 6.

4 Results

In this section, we test the performance of the proposed TS-L1-PZ and SD-L1-PZ in both synthetic and real speech signals analysis scenarios.

4.1 Synthetic signal analysis

Synthetic speech signals are generated by convolving an input excitation with a constructed filter to estimate the performance of the proposed method. The input excitation is a pulse train with the fundamental frequency between 300-500 Hz. The filter we used here has the following characteristics

$$H(z) = \frac{\sum_{i=1}^{4}(1 - \beta_i z^{-1})}{\sum_{j=1}^{5}(1 - \alpha_j z^{-1})} \tag{F.10}$$

where $\beta_1 = \beta_2^* = 0.5348 + 0.5529j$, $\beta_3 = \beta_4^* = -0.0263 + 0.7688j$, $\alpha_1 = \alpha_2^* = -0.5026 + 0.5976j$, $\alpha_3 = \alpha_4^* = 0.4449 + 0.7928j$, $\alpha_5 = 0.8602$. The SNR is set to 30 dB for additive Gaussian noise. As a measure for the accuracy of the estimated spectral envelope, the spectral distortion (SD) is defined as [24]

$$\mathrm{SD} = \frac{1}{\mathrm{S}} \sum_{s=1}^{\mathrm{S}} (\log \left|H(e^{j\omega_s})\right| - \log \left|\hat{H}(e^{j\omega_s})\right|)^2 \tag{F.11}$$

where $\hat{H}(e^{j\omega_s})$ denotes an estimate of the true envelope $H(e^{j\omega_s})$, S is the number of spectral samples. The experimental results are obtained by the ensemble averages over 2 s with 30 ms frame length. The SD curves for the SD-L1-PZ, TS-L1-PZ, TS-LS-PZ, 1-norm linear prediction (1-norm LP),

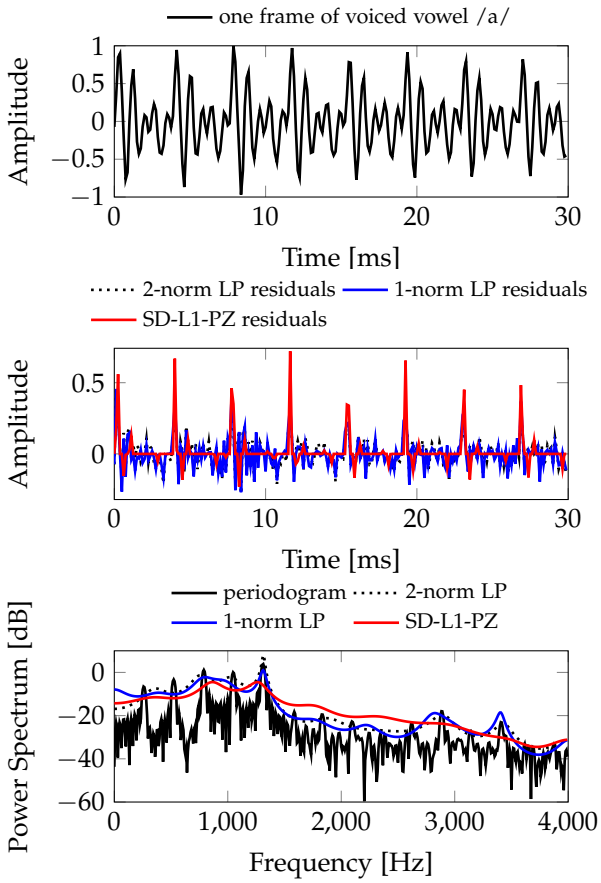


Fig. F.3: Residuals and spectral envelope estimates for the voiced vowel /a/

2-norm linear prediction (2-norm LP) with different excitation frequencies are shown in Fig. F.1, where 5 iterations are used for the SD-L1-PZ, $\gamma = 0$, $C = 0.01 \|\mathbf{x}\|_2^2$, K and L are set to 10. As can be seen, the SD of the 1-norm LP is lower than the 2-norm LP. Moreover, the plot of the TS-LS-PZ has more fluctuations than the TS-L1-PZ for different frequencies. Furthermore, by utilizing the 1-norm cost function and pole-zero modeling with sparse deconvolution together, the proposed SD-L1-PZ achieves the lowest spectral distortion compared with others.

Then, the reconstruction performance is tested in terms of the output SNR (oSNR). Synthetic speech signals are generated by convolving $e(n) = \delta(n-50) + 0.5\delta(n-80) - 0.3\delta(n-100)$ with a filter with transfer function

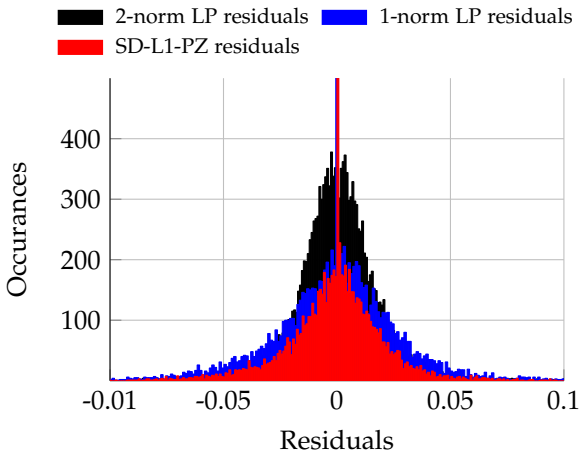


Fig. F.4: The histogram of residuals of the 2-norm LP, 1-norm LP and SD-L1-PZ

$H(z) = (1 + 0.8z^{-1})/(1 - 0.9z^{-1} + 0.81z^{-2})$ [23]. The oSNR is defined as

$$\text{oSNR} = E(\bar{x})^2 / E((\bar{x} - \hat{x})^2) \tag{F.12}$$

where $\bar{x}$ denotes the noise-free signal. The reconstructed signals $\hat{x}$ are obtained using (5) and (6) with $q_1 = 1$ for 1-norm based methods (i.e. the 1-norm LP, TS-L1-PZ and SD-L1-PZ), but with $q_1 = 2$ for the TS-LS-PZ. The experimental results are obtained by the ensemble averages over 100 Monte Carlo simulations. The oSNR curves for different algorithms are shown in Fig. F.2, where $\gamma = 0$, $C = \|\mathbf{m}\|_2^2$, $N = 300$, $K = 10$ and $L = 5$. As can be seen, the performance of the 2-norm and TS-LS-PZ, 1-norm LP and TS-LS-PZ are similar. The SD-L1-PZ presents a higher oSNR.

4.2 Speech signal analysis

This work also examines the performance of the SD-L1-PZ for a real voiced vowel /a/ sampling of 8000 Hz, as shown in Fig. F.3, where $\gamma = 0$, $C = 0.1\,\|\mathbf{x}\|_2^2$, $K = 20$, $L = 10$, the SNR for Gaussian white noise is set to 30 dB. As can be seen, the residuals of the SD-LS-PZ are sparser than both the 2-norm and 1-norm LP methods. Moreover, since we admit the existence of the pitch and harmonics, the spectral envelope estimate of the 1-norm LP and SD-L1-PZ is smoother than the conventional 2-norm LP, which tends to null out the harmonics [3]. In fact, due to the sparser residual estimates, the estimated spectral envelope of the SD-L1-PZ tends to be the smoothest one. Above all, due to the usage of the pole-zero model and 1-norm cost function, compared with all-pole model and 2-norm cost, the SD-L1-PZ presents sparser residuals and smoother spectral envelope estimation performance for voiced speech.

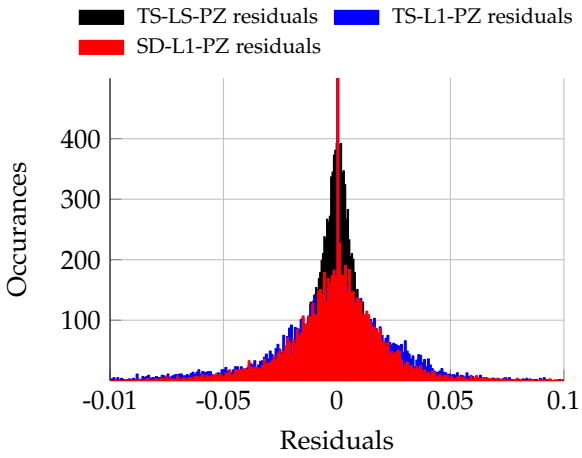


Fig. F.5: The histogram of residuals of the TS-LS-PZ, TS-L1-PZ and SD-L1-PZ

The residual estimates of the proposed approach are further tested on real speech signals "Why were you away a year, Roy?" uttered by a female speaker sampled at 8000 Hz. The histograms of the residuals for the 2-norm LP, 1-norm LP, TS-LS-PZ, TS-L1-PZ and SD-L1-PZ are shown in Fig. F.4 and Fig. F.5, where $\gamma = 0$, $C = \|\mathbf{m}\|_2^2$, $K = 10$, $L = 5$, and the SNR for Gaussian white noise is set to 20 dB. Analysis is performed every 30 ms without overlap. The residuals are obtained using (5) with $q_1 = 1$ for the 1-norm LP, TS-L1-PZ and SD-L1-PZ, but with $q_1 = 2$ for the TS-LS-PZ. As can be seen, the 1-norm-based approach, such as the 1-norm LP, or TS-L1-PZ and SD-L1-PZ, is thinner than the corresponding 2-norm method, which is the 2-norm LP or TS-LS-PZ, respectively. The SD-L1-PZ is the thinnest and highest among all the others, which means the residuals of the SD-L1-PZ are the sparsest.

5 Conclusion

A least 1-norm based pole-zero speech analysis method is proposed in this paper. By using the pole-zero model, it can fit the spectral zeros of speech signals easily than all-pole methods. Moreover, sparse residuals are encouraged by applying 1-norm criterion compared with the 2-norm methods. By iteratively updating parameters and residuals using the 1-norm cost and sparse deconvolution, robust coefficient estimates in noisy conditions can be obtained. Simulation results in both synthetic and real speech scenarios show that improved analysis performance in terms of lower spectral distortion, higher reconstruction SNR and sparser residuals can be obtained.

References

[1] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[3] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 221–239, 2000.

[4] T. Drugman and Y. Stylianou, "Fast inter-harmonic reconstruction for spectral envelope estimation in high-pitched voices," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1418–1422, 2014.

[5] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, 1991.

[6] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 1, pp. 65–73, 2008.

[7] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, jul 2012.

[8] T. Shao, Y. R. Zheng, and J. Benesty, "An affine projection sign algorithm robust against impulsive interferences," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 327–330, apr 2010.

[9] L. Shi, Y. Lin, and X. Xie, "Combination of affine projection sign algorithms for robust adaptive filtering in non-gaussian impulsive interference," *Electronics Lett.*, vol. 50, no. 6, pp. 466–467, 2014.

[10] D. Giacobello, M. G. Christensen, T. L. Jensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Stable 1-norm error minimization based linear predictors for speech modeling," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 5, pp. 912–922, may 2014.

[11] T. L. Jensen, D. Giacobello, T. V. Waterschoot, and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Commun.*, vol. 76, pp. 143–156, 2016.

[12] K. SteTiange, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 229–234, 1977.

[13] K. H. Song and K. U. Chong, "Pole-zero modeling of speech based on high-order pole model fitting and decomposition method," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1556–1565, 1983.

[14] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 237–248, 2010.

[15] J. Durbin, "The fitting of time-series models," *Rev. Int'l Statistical Inst.*, vol. 28, no. 3, pp. 233–244, 1960.

[16] E. Levy, "Complex-curve fitting," *IRE Trans. Automat. Contr.*, vol. AC-4, no. 1, pp. 37–43, 1959.

[17] Y. Miyanaga, N. Miki, and N. Nagai, "Adaptive identification of a time-varying ARMA speech model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 3, pp. 423–433, 1986.

[18] T. Kobayashi and S. Imai, "Design of IIR digital filters with arbitrary log magnitude function by WLS techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 247–252, 1990.

[19] C. Magi, J. Pohjalainen, T. Backstrom, and P. Alku, "Stabilised weighted linear prediction," *Speech Commun.*, vol. 51, no. 5, pp. 401–411, 2009.

[20] J. Pohjalainen, H. Kallasjoki, K. J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis in noisy conditions," *Proc. Interspeech*, pp. 1315–1318, 2009.

[21] P. Stoica and R. Moses, *Spectral Analysis of Signals*. New Jersey: Prentice Hall, Upper Saddle River, 2004.

[22] E. Denoël and J. P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1397–1403, 1985.

[23] I. Selesnick, "Sparse deconvolution (an MM algorithm)," Available: http://cnx.org/content/m44991/1.4/, pp. 1–17, 2012 [Online].

[24] H. Kameoka, N. Ono, and S. Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. 6, pp. 1507–1516, 2010.

Paper G

Automatic Quality Control and Enhancement for Voice-Based Remote Parkinson's Disease Detection

Amir Hossein Poorjam, Mathew Shaji Kavalekalam, Liming Shi, Yordan P. Raykov, Jesper Rindom Jensen, Max A. Little and Mads Græsbøll Christensen

The paper has been submitted to the
IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019

Abstract

The performance of voice-based Parkinson's disease (PD) detection systems degrades when there is an acoustic mismatch between training and operating conditions caused mainly by degradation in test signals. In this paper, we address this mismatch by considering three types of degradation commonly encountered in remote voice analysis, namely background noise, reverberation and nonlinear distortion, and investigate how these degradations influence the performance of a PD detection system. Given that the specific degradation is known, we explore the effectiveness of a variety of enhancement algorithms in compensating this mismatch and improving the PD detection accuracy. Then, we propose two approaches to automatically control the quality of recordings by identifying the presence and type of short-term and long-term degradations and protocol violations in voice signals. Finally, we experiment with using the proposed quality control methods to inform the choice of enhancement algorithm. Experimental results using the voice recordings of the mPower mobile PD data set under different degradation conditions show the effectiveness of the quality control approaches in selecting an appropriate enhancement method and, consequently, in improving the PD detection accuracy. This study is a step towards the development of a remote PD detection system capable of operating in unseen acoustic environments.

1 Introduction

Parkinson's disease (PD) is a neurodegenerative disorder which progressively makes the patients unable to control their movement normally and, consequently, decreases the patients' quality of life [1]. Since there is no cure for PD, it is necessary to develop tools to diagnose this disease in early stages in order to control its symptoms. Speech is known to reflect the PD symptoms since the majority of PD patients suffer from some forms of vocal disorder [2]. It has been demonstrated in [3] that early changes of clinical symptoms of PD are more reflected and pronounced in acoustic analysis of voice signals than in perceptual evaluation of voice by a therapist. This has motivated researchers to take advantage of advanced speech signal processing and machine learning algorithms to develop highly accurate and data-driven methods for detecting PD symptoms from voice signals [4–6]. Moreover, advances in smart phone technology provide new opportunities for remote monitoring of PD symptoms by bypassing the logistical and practical limitations of recording voice samples in controlled experimental conditions in clinics [5, 7]. However, there is a higher risk outside controlled lab conditions that participants may not adhere to the test protocols, which probe for specific symptoms, due to lack of training, misinterpretation of the test protocol or negligence. Moreover, voice signals in remote voice analysis might

be subject to a variety of degradations during recording or transmission. Processing the degraded recordings or those which do not comply with the assumptions of the test protocol can produce misleading, non-replicable and non-reproducible results [8] that could have significant ramifications for the patients' health. In addition, degradation of voice signals produces an acoustic mismatch between the training and operating conditions in automatic PD detection. A variety of techniques have been developed for compensating this type of mismatch in different speech-based applications [9–15] which can, in general, be categorized into four classes: (1) searching for robust features which parameterize speech regardless of degradations; (2) transforming a degraded signal to the acoustic condition of the training data using a signal enhancement algorithm[1]; (3) compensating the effects of degradation in the feature space by applying feature enhancement; and (4) transforming the parameters of the developed model to match the acoustic conditions of the degraded signal at operating time. However, to the best of the authors' knowledge, there is a lack of studies of the impact of acoustic mismatch and the effect of compensation on the performance of PD detection systems. Vasquez-Correa et al. proposed a pre-processing scheme by applying a generalized subspace speech enhancement technique to the voiced and unvoiced segments of a speech signal to address the PD detection in non-controlled noise conditions [16]. They showed that applying speech enhancement to the unvoiced segments leads to an improvement in detection accuracy while the enhancement of voiced segments degrades the performance. However, this study is limited in terms of degradation types as it only considered the additive noise. Moreover, they only evaluated the impact of an unsupervised enhancement method on PD detection performance, while the supervised algorithms have, in general, shown to reconstruct higher quality signals as they incorporate more prior information about the speech and noise.

Another open question which, to the authors' knowledge, has not been addressed is whether applying "appropriate" signal enhancement algorithms to the degraded signals will result in an improvement in PD detection performance. Answering this question, however, requires prior knowledge about the presence and type of degradation in voice signals, which can be achieved by controlling the quality of recordings prior to analysis. Quality control of the voice recordings is typically performed manually by human experts which is a very costly and time consuming task, and is often infeasible in online applications. In [17], the problem of quality control in remote speech data collection has been approached by identifying the potential outliers which are inconsistent, in terms of the quality and the context, with the majority of speech samples in a data set. Even though very effective in finding out-

[1]In this paper, by "signal enhancement", we refer to all algorithms intended to enhance the quality of degraded signals.

liers, it is not capable of detecting the type of degradation nor identifying short-term protocol violations in recordings. To identify the type of degradation in pathological voices, Poorjam et al. proposed two different parametric and non-parametric approaches to classify degradations commonly encountered in remote pathological voice analysis into four major types, namely background noise, reverberation, clipping and coding [18, 19]. However, the performance of these approaches is limited when new degradation types are introduced. Furthermore, the presence of outlier recordings, which do not contain relevant information for PD detection due to long-term protocol violations, is not considered in these methods and, therefore, there is no control over the class assignment for such recordings. To address the frame-level quality control in pathological voices, Badawy et al. proposed a general framework for detecting short-term protocol violations using a nonparametric switching autoregressive model [20]. In [21], a highly accurate approach for identifying short-term protocol violations in PD voice recordings has been proposed which fits an infinite hidden Markov model to the frames of the voice signals in the mel-frequency cepstral domain. However, these two approaches do not identify short-term degradations (e.g. the presence of an instantaneous background noise) in voice signals.

To overcome the explained limitations in the existing methods, we propose two approaches for controlling the quality of pathological voices at recording-level and frame-level in this paper. In the recording-level approach, separate statistical models are fitted to the clean voice signals and the signals corrupted by different degradation types. The likelihood of a new observation given each of the models is then used to determine its degree of adherence to each class of acoustic conditions. This gives us the flexibility not only to associate multiple classes to a voice signal corrupted by a combination of different degradations, but also to consider a recording as an outlier or a new degradation when it is rejected by all the models. In the frame-level approach, on the other hand, we extend the work in [21] to identify short-term protocol violations and degradations in voice signals at the same time. We show how the proposed quality control approaches can effectively inform the choice of signal enhancement methods and, consequently, improve the PD detection performance. The contribution of this paper is thus three-fold: (1) we investigate the impact of acoustic mismatch between training and operating conditions, due to degradation in test signals, on the PD detection performance; (2) to identify this mismatch, we propose two different approaches to automatically control the quality of pathological voices at frame- and recording-level; and (3) to efficiently reduce this mismatch, given that the specific degradation is known, we explore a variety of state-of-the-art enhancement algorithms and their effectiveness in improving the performance of a PD detection system. The rest of the paper is organized as follows. Section 2 explains the PD detection system that we have used for the

experiments throughout this paper. In Section 3, we investigate the impact of three major types of signal degradation commonly encountered in remote voice analysis, namely noise, reverberation and nonlinear distortion, on the performance of the PD detection system. Following that, in Section 4, we investigate on the influence of noise reduction and dereverberation algorithms on the performance of the PD detection system. In Section 5, we propose two different quality control approaches and investigate how these methods can improve the performance of PD detection. Finally, Section 6 summarizes the paper.

2 Parkinson's disease detection system

In this section, we describe the PD detection system we will use for further quality control and enhancement experiments. This approach, which was proposed in [22], fits Gaussian mixture models (GMMs) to the frames of the voice recordings of the PD patients and the healthy controls (HC) parametrized by perceptual linear predictive (PLP) coefficients [23]. The motivation for using PLP parametrization is that the perceptual features are more discriminative in PD detection than the conventional and clinically interpretable ones (such as standard deviation of fundamental frequency, jitter, shimmer, harmonic-to-noise ratio, glottal-to-noise exitation ratio, articulation rate, and frequencies of formants), particularly when the voice is more noisy, aperiodic, irregular and chaotic which typically happens in more advanced stages of PD [24–26].

Acoustic features of the PD patients' recordings and those of the healthy controls are modeled by GMMs with the likelihood function defined as:

$$p(\boldsymbol{x}_t|\lambda) = \sum_{c=1}^{C} b_c p(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{G.1}$$

where $\boldsymbol{x}_t$ is the feature vector at time frame t, b_c is the mixture weight of the c^{th} mixture component, C is the number of Gaussian mixtures, $p(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is a Gaussian probability density function where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean and covariance of the c^{th} mixture component, respectively. The parameters of the model, $\lambda = \{b_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$, are trained through the expectation-maximization algorithm [27].

Given $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$, a sequence of feature vectors, the goal in PD detection is to find the model which maximizes $p(\lambda_j|\boldsymbol{X})$, where $j \in \{\text{PD}, \text{HC}\}$. Using the Bayes' rule, independence assumption between frames, and assuming equal priors for the classes, the PD detection system

computes the log-likelihood ratio for an observation as:

$$\sigma(\boldsymbol{X}) = \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_{\text{PD}}) - \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_{\text{HC}}). \tag{G.2}$$

The final decision about the class assignment for an observation is made by setting a threshold over the obtained score.

Experimental Setup

In this study, we use the sustained vowel /a/ as the speech material for PD detection since they provide a simpler acoustic structure to characterize the glottal source and resonant structure of the vocal tract than running speech. Moreover, perceptual analysis of different vowels suggests that the best PD detection performance can be achieved when the sustained vowel phonation /a/ is parametrized by the PLP features [24]. We consider the mPower mobile Parkinson's disease (MMPD) data set [28] which consists of more than 65,000 samples of 10 second sustained vowel /a/ phonations recorded via smartphones by PD patients and healthy speakers of both genders from the US. The designed voice test protocol for this data set required the participants to hold the phone in a similar position to making a phone call, take a deep breath and utter a sustained vowel /a/ at a comfortable pitch and intensity for 10 seconds. A subset of 800 good-quality voice samples (400 PD patients and 400 healthy controls equally from both genders) have been selected from this data set. It should be noted that the health status in this data set is self-reported. To have more reliable samples, among participants who self-reported to have PD, we selected those who claimed that they have been diagnosed by a medical professional with PD and recorded their voice right before taking PD medications. For the healthy controls, we selected participants who self-reported being healthy, do not take PD medications, and claimed that they have not been diagnosed by a medical professional with PD. All speakers of this subset had an age range of 58 to 72. The mean $\pm$ standard deviation (STD) of the age of PD patients and healthy controls are 64 ± 4 and 66 ± 4, respectively. For all experiments in this paper, we downsampled the recordings from 44.1 kHz to 8 kHz since the enhancement algorithms used in this work are operating at 8 kHz. To extract the PLP features, voice signals are first segmented into frames of 30 ms with 10 ms overlap using a Hamming window. Then, 13 PLP coefficients are computed for each frame of a signal. To consider the dynamic changes between frames due to the deviations in articulation, a first- and a second-order orthogonal polynomials are fitted to the two feature vectors to the left and right of the current frame. These features, which are referred to as *delta* and *double-delta*, were appended to the feature vector to form a 39-dimensional vector per each frame. The number of mixture components for the GMMs were set to 32.

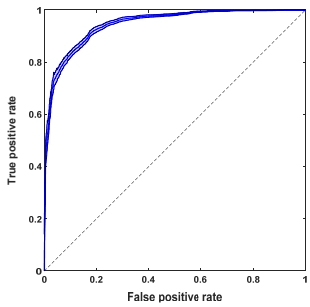


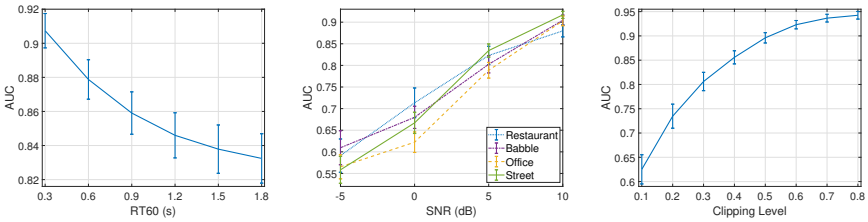
Fig. G.1: The ROC curve of the PD detection system, along with 95% confidence interval shaded in blue. The dashed line shows the chance level.

Results

To evaluate the performance of the PD detection system in a matched acoustic condition, we used 5-fold cross validation (CV) in which the recordings were randomly divided into 5 non-overlapping and equal sized subsets. The entire CV procedure was repeated 10 times to obtain the distribution of detection performance. Fig. G.1 shows the performance in terms of the receiver operating characteristic (ROC) curve, along with 95% confidence interval. In an ROC curve, the true positive rate is plotted against the false positive rate for different decision thresholds. The area under the curve (AUC) summarizes the ROC curve and represents the performance of a detection system by a single number between 0 and 1; the higher the performance, the closer the AUC value is to 1. Comparing with the commonly used classification accuracy, the AUC is a more preferred metric in this paper since it is a summary of the class overlap which sets a fundamental limit to the classification accuracy. The mean AUC for this PD detection system is 0.95.

3 Impact of signal degradation on PD detection

The PD detection system explained in the previous section gave a mean AUC of 0.95 in a matched acoustic condition. That is, when it was trained and tested using the clean recordings. However, as alluded to in the introduction, recordings collected remotely in an unsupervised manner are seldom clean as they are often degraded by different types of degradation. In this section we investigate the effect of 3 different commonly encountered degradations, namely reverberation, background noise and nonlinear distortion on the performance of the PD detection system. It should be noted that even though we tried to choose the most reliable samples from the MMPD data set, the labels might still not be 100% reliable as the diagnosis is self-reported. For this



(a) Reverberation **(b)** Additive Noise **(c)** Nonlinear Distortion

Fig. G.2: Performance of the PD detection system in acoustic mismatch conditions due to different degradations in test signals in terms of AUC, along with 95% confidence intervals.

reason, we are more interested in how the relative PD detection performance is influenced systematically under application of different experimental conditions.

3.1 Reverberation

Reverberation is a phenomenon that occurs when the signal of interest is captured in an acoustically enclosed space. Apart from the direct component, the microphone receives multiple delayed and attenuated versions of the signal, which is characterized by the room impulse response (RIR). A metric commonly used to measure the reverberation is the reverberation time (RT60) [29]. The presence of reverberation has shown to degrade the performance of speech-based applications such as speech and speaker recognition [30, 31]. In this section, we investigate the effect of reverberation on the PD detection performance. To this aim, we used 5-fold CV repeated 10 times to evaluate the performance. In each iteration, the model was trained using the clean recordings of the training subset, and evaluated on the recordings of the disjoint test subset which were filtered with synthetic room impulse responses of RT60 varying from 300 ms to 1.8 s in 300 ms steps measured at a fixed position in a room of dimension 10 m $\times$ 6 m $\times$ 4 m. The distance between source and microphone is set to 2m. The room impulse responses were generated using the image method [32] and implemented using the RIR Generator toolbox [33]. Fig. G.2a shows the impact of reverberation on the PD detection performance in terms of the mean AUC along with 95% confidence intervals. We can observe from the plot that the PD detection system exhibits lower performance in reverberant environments, as expected, and the amount of degradation is related to the RT60.

3.2 Background noise

Background noise is one of the most common types of degradation occurring during remote voice analysis. In this section we restrict ourselves to additive

background noise and investigate how this can influence the PD detection performance. To this aim, we performed the same CV procedure used for evaluating the impact of reverberation (explained in the previous section). In each iteration, the model was trained using the clean recordings of the training subset, and evaluated on the recordings of the test subset contaminated by an additive noise. The entire procedure was repeated for four different noise types, namely babble, restaurant, office and street noise[2] and different signal-to-noise ratios (SNRs) ranging from -5 dB to 10 dB in 5 dB steps. Fig. G.2b illustrates the impact of different noise types and different SNR conditions on the performance of the PD detection system in terms of the mean of AUC along with the 95% confidence intervals. We can observe a similar trends for all noise types that that the PD detection performance decreases as the noise level increases.

3.3 Clipping

In remote voice analysis, nonlinear distortion can manifest itself in speech signals in many different ways such as clipping, compression, packet loss and combinations thereof. Here, we consider clipping as an example of nonlinear distortion in signals which is caused when a signal fed as an input to a recording device exceeds the dynamic range of the device [34]. By defining the clipping level as a proportion of the unclipped peak absolute signal amplitude to which samples greater than this threshold are limited, we can investigate the impact of clipping on the PD detection performance. To this aim, the clean recordings of the test subset in each iteration of the CV were clipped with different clipping levels ranging from 0.1 to 0.8 in 0.1 steps. Fig. G.2c shows the performance as a function of clipping level. Similar to the other types of degradation, it can be observed that increasing the distortion level in voice signals decreases the PD detection performance.

4 Impact of noise reduction and dereverberation on PD detection

As seen in Section 3, the degradation introduced to the signals can lead to reduction in the performance of the PD detection system. Since there are practically an infinite number of possible types and combinations of nonlinear distortion that can be present in a signal, and since there is a lack of well-documented algorithms for dealing with most of the distortions (even in isolation), in this section, we only consider the degradations for which there

[2]The babble, restaurant and street noise files have been taken from https://www.soundjay.com/index.html and the office noise has been taken from https://freesound.org/people/DavidFrbr/sounds/327497

are well-documented and verified enhancement algorithms such as noise reduction and dereverberation and investigate the effects of these algorithms on the PD detection performance. To this end, from the 50 PD detection models developed and evaluated through 10 iterations of the 5-fold cross-validation procedure, as explained in Section (2), we selected one of the two models which showed the median performance and used it for further enhancement experiments in this section. We have used a total of 160 recordings for testing the algorithms used in this section. We will restrict ourselves to single channel enhancement algorithms. It should be noted that there exist a variety of objective and subjective metrics to measure the quality of the enhanced speech signal such as SNR, signal-to-distortion ratio [35], perceptual evaluation of speech quality [36] and short-time objective intelligibility [37]. However, since our main goal in this work is to study the influence of speech enhancement on the PD detection performance, we evaluate the effectiveness of the algorithms in terms of the AUC.

4.1 Dereverberation

Some of the popular classes of dereverberation techniques are the spectral enhancement methods [38], probabilistic model based methods [39, 40] and inverse filtering based methods [41, 42]. Spectral enhancement methods estimate the clean speech spectrogram by frequency domain filtering using the estimated late reverberation statistics. The probabilistic model based methods model the reverberation using an autoregressive (AR) process, and the clean speech spectral coefficients using a certain probability distribution function. The estimated parameters of the model are then used to perform dereverberation. Lastly, the inverse filtering methods use a blindly estimated room impulse response to design an equalization system. These methods, which are mainly developed for the running speech, assume that the signal at a particular time-frequency bin is uncorrelated with the signals at that same frequency bin for frames beyond a certain number [40]. However, this assumption is not valid for the sustained vowels which makes the dereverberation of the sustained vowels more challenging. Recently, deep neural network (DNN) based dereverberation algorithms have gained attention [43, 44] since they relax the assumption of uncorrelated neighboring time-frequency bins. The underlying principle of the DNN-based methods is to train a DNN to map the log-magnitude spectrum of the degraded speech to that of the desired speech.

In this section, we investigate the effectiveness of different dereverberation algorithms in improving the PD detection performance. For dereverberation experiments, we used three different algorithms: a probabilistic model based algorithm proposed in [40] (denoted as WPE-CGG, weighted prediction error with complex generalized Gaussian prior), an algorithm based on the inverse

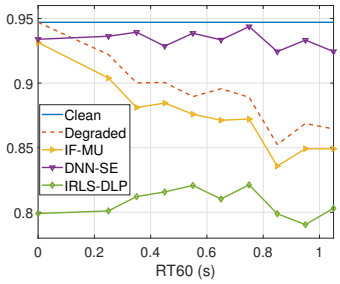


Fig. G.3: Impact of different dereverberation algorithms on the PD detection performance, in terms of AUC

filtering of the modulation transfer function [41] (denoted as IF-MU, inverse filtering with multiplicative update), and a DNN-based speech enhancement algorithm proposed in [43] (denoted as DNN-SE). It should be noted that the WPE-CGG and the IF-MU are unsupervised methods whereas DNN-SE is a supervised method. For the DNN-based algorithm, a feedforward neural network with 3 hidden layers of 1,600 neurons was used. To take into account the temporal dynamics, features of 11 consecutive frames (including the current frame, 5 frames to the left and 5 frames to the right over time) were provided to represent the input features of the current frames. To train the DNN model, we selected 640 clean recordings from the MMPD data set and filtered them with the synthetic room impulse responses of RT60 ranging from 200 ms to 1 s in steps of 100 ms using the implementation in [33] for a particular source and receiver position in a room of dimensions 10 m $\times$ 6 m $\times$ 4 m. For testing, the position of the receiver was fixed while the position of the source was varied randomly from 60 degrees left of the receiver to 60 degrees right of the receiver. Fig. G.3 shows the performance of the PD detection in terms of AUC for the different dereverberation algorithms. It can be observed from the figure that only DNN-SE is able to improve the PD detection performance while the other two methods degrade the performance. This is mainly due to two reasons: first, the DNN-SE is a supervised algorithm while the WPE-CGG and IF-MU are unsupervised; and second, the underlying assumption of the two unsupervised algorithms does not hold for the sustained vowels. We have also included the case of zero RT60 to investigate the impact of processing of the clean recordings by these dereverberation algorithms.

4.2 Noise reduction

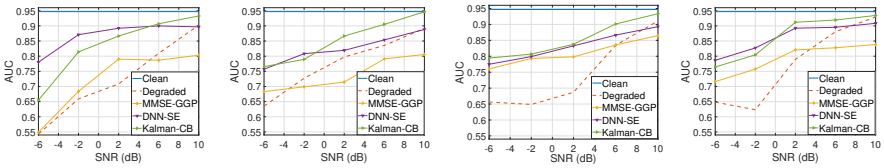
Methods for performing noise reduction can be broadly categorized into supervised and unsupervised methods. Unsupervised methods do not assume any prior knowledge about identity of the speaker or noise environment.

The supervised methods, on the other hand, make use of training data to train the models representing the signals of interest or the noise environment. Some of the popular classes of supervised speech enhancement methods include the codebook-based methods [45, 46], non-negative matrix factorization based methods [10, 47] and the DNN-based methods [48]. In the supervised method, the speech and noise statistics/parameters estimated using the training data are exploited within a filter to remove the noise components from the noisy observation. In this section, we used two supervised methods and one unsupervised method to investigate the effect of different noise reduction algorithms in reducing the acoustic mismatch between training and operating conditions.

The first supervised enhancement algorithm is based on the framework proposed in [49]. In this approach, a Kalman filter, which takes into account the voiced and unvoiced parts of speech [50], is used for enhancement. The filter parameters consist of the AR coefficients and excitation variance corresponding to speech and noise along with the pitch parameters (i.e. the fundamental frequency and the degree of voicing). Based on [49], the AR coefficients and excitation variance of the speech and noise are estimated using a codebook-based approach, and the pitch parameters are estimated from the noisy signal using a harmonic model based approach [51]. We refer to this method in the rest of this paper as the Kalman-CB. This algorithm has been selected because of its good performance in noise reduction in terms of quality and intelligibility based on both objective and subjective measures. The speech codebook was trained using 640 clean recordings selected from the MMPD data set (equally from both genders). To train the noise codebook, we used babble, restaurant, office and street noises to create four sub-codebooks. During the testing phase, all sub-codebooks, except the one corresponding to the target noise, were concatenated to form the final noise codebook. The size of the speech and noise codebooks were set to 8 and 12, respectively.

The second supervised enhancement method is the DNN-based algorithm proposed in [43]. This algorithm is the same as the one we used for dereverberation experiments, except it is trained using the noisy signals. This algorithm has been selected because, besides improvements in objective measures, it showed improvement in performance of automatic speech recognition in noisy environments. To train the DNN, we used the same 640 clean recording that we used for training the speech codebook in the Kalman-CB algorithm. The recordings were contaminated by three types of noise, namely babble, factory and F16 noises taken from NOISEX-92 database [52] under different SNR conditions selected randomly from the continuous interval [0,10] dB.

We used, as an unsupervised speech enhancement method, the algorithm proposed in [53] which is based on the minimum mean-square error (MMSE) estimation of discrete Fourier transform (DFT) coefficients of speech while



(a) Babble Noise **(b)** Restaurant Noise **(c)** Office Noise **(d)** Street Noise

Fig. G.4: Impact of different noise reduction algorithms on the PD detection performance, in terms of AUC, under different noise types and SNR conditions.

assuming a generalized gamma prior for the speech DFT coefficients. This method, denoted as MMSE-GGP, is a popular unsupervised algorithm which uses the MMSE-based tracker for noise power spectral density estimation.

Fig. G.4 shows the PD detection performance in terms of AUC for different noise types and SNR conditions. It can be observed from the figures that enhancing the degraded voice signals with the supervised methods in general improves the performance whereas the unsupervised method shows improvement only in the low SNR range and degrades the PD detection performance in higher SNR scenarios. The low performance of the unsupervised algorithm can be due to the fact that noise statistics in this case is estimated using a method proposed in [54] which has been designed for running speech rather than the sustained vowels. This observation is somewhat consistent with the statement in [16], which suggested that applying an unsupervised enhancement algorithm to the voiced segments results in a degradation in PD detection performance.

4.3 Joint noise reduction and dereverberation

In Sections 4.1 and 4.2, we showed the impact of noise reduction and dereverberation when one of these degradations was present in the signal. However, in some cases, the recordings may be degraded simultaneously by reverberation and background noise. There have been methods proposed for joint noise reduction and dereverberation with access to multiple channels [55, 56]. Since we have restricted ourselves to single channel enhancement methods, and motivated by the improvement in the PD detection performance as a result of using the DNN-SE algorithm for noise reduction and dereverberation, in this section, we investigate the effectiveness of this algorithm in performing joint noise reduction and dereverberation. In this case, the input to the DNN is the log-magnitude spectrum of the signal which is degraded by reverberation and background noise. For training the DNN model, the same 640 clean recordings that we used in the previous enhancement experiments were filtered with RIRs of different RT60s ranging from 400 ms to 1 s with 200

ms steps. Then, three types of noise, namely babble, factory and F16 noises (taken from NOISEX-92 database) were randomly added to the reverberant signals at different SNRs selected uniformly at random from the continuous interval [0,10] dB. Table G.1 summarizes the impact of joint noise reduction and dereverberation using the DNN-SE algorithm on the PD detection performance. In this table, we have also included the cases of infinite SNR and zero RT60 to investigate the effect of the enhancement system when the clean recordings or the ones degraded by only noise or reverberation were processed by this algorithm. It can be observed for the case of babble noise that the DNN-SE improves the PD detection performance in most of the cases when reverberation and background noise coexist and in the cases where only noise is present. However, in the case of only reverberation, the DNN-SE shows improvement only in the cases where RT60 is 400 ms and above. It should be noted that the babble noise used for training and testing were taken from two different noise databases. In the case of restaurant noise, improvement in PD detection performance is observed only in the low SNRs, namely -2 dB and -6 dB. The results of the restaurant noise is interesting in a sense that it shows how the DNN-SE algorithm can generalize for a noise type not seen during the training phase.

Table G.1: Impact of joint noise reduction and dereverberation using the DNN-SE algorithm on the PD detection performance. Bold numbers indicate the improvement in performance.

			Babble Noise: SNR (dB)						**Restaurant Noise: SNR (dB)**					
			-6	-2	2	6	10	inf	-6	-2	2	6	10	inf
RT60 (s)	0	Degraded	0.67	0.59	0.69	0.80	0.90	0.95	0.71	0.81	0.82	0.82	0.90	0.95
		DNN-SE	**0.80**	**0.89**	**0.89**	**0.89**	0.89	0.91	**0.77**	0.81	0.82	**0.83**	0.87	0.91
	0.2	Degraded	0.56	0.64	0.72	0.81	0.89	0.95	0.67	0.75	0.76	0.85	0.89	0.95
		DNN-SE	**0.82**	**0.89**	**0.87**	**0.89**	0.89	0.91	**0.74**	**0.79**	**0.79**	0.84	0.87	0.91
	0.4	Degraded	0.54	0.66	0.70	0.80	0.84	0.90	0.62	0.73	0.83	0.83	0.83	0.92
		DNN-SE	**0.78**	**0.84**	**0.85**	**0.89**	**0.86**	**0.91**	**0.73**	**0.77**	0.79	0.83	0.82	0.91
	0.6	Degraded	0.64	0.70	0.71	0.78	0.81	0.88	0.59	0.79	0.81	0.80	0.86	0.89
		DNN-SE	**0.75**	**0.83**	**0.85**	**0.85**	**0.88**	**0.89**	**0.69**	**0.81**	0.79	**0.81**	0.84	**0.91**
	0.8	Degraded	0.67	0.70	0.73	0.79	0.83	0.89	0.58	0.76	0.82	0.81	0.86	0.87
		DNN-SE	**0.81**	**0.83**	**0.86**	**0.87**	**0.88**	**0.91**	**0.75**	0.76	0.80	**0.84**	**0.87**	**0.90**
	1	Degraded	0.54	0.68	0.74	0.81	0.84	0.88	0.65	0.75	0.76	0.82	0.83	0.85
		DNN-SE	**0.80**	**0.81**	**0.86**	**0.86**	**0.88**	**0.90**	**0.76**	0.75	**0.78**	0.81	0.82	**0.90**

5 Automatic quality control in pathological voice recordings

We have shown in the previous section that, assuming the specific degradation is known, there exist algorithms to effectively transform a voice signal

from a degraded condition into the acoustic condition in which models are trained. Choosing the appropriate enhancement algorithm, however, requires prior knowledge about the presence and type of degradation in a voice signal. In this section, we introduce two approaches to automatically control the quality of recordings. The first approach detects, at recording level, the presence and type of degradation which has influenced the majority of frames of the signal. The second approach, on the other hand, detects short-term degradations and protocol violations in a signal.

5.1 Recording-level quality control

The major limitation of the classification-based approaches for identifying the type of degradation in a voice signal [18, 19] is that they do not consider the fact that a recording can be subject to an infinite number of possible combinations of degradations in real scenarios. This causes some problems when a signal is contaminated by a new type of degradation for which the classifier has not been trained. Moreover, there is no control in class assignment for a high-quality outlier which do not comply with the context of the data set.

To overcome these limitations, instead of using a multiclass classifier, we propose to use a set of parallel likelihood ratio detectors for the major types of degradations commonly encountered in remote voice analysis, each detecting a certain degradation type. This way, the likelihood ratio statistics of an observation given each of the models can be translated to the degree of contribution of each degradation to the degraded observation. Moreover, completely new degradation types and high-quality outliers can be detected if all models reject those observations according to a pre-defined threshold.

In this approach, the task of each detector is to determine whether a feature vector of the time frame t of a voice signal, $\boldsymbol{x}_t$, was contaminated by the corresponding degradation, H_0, or not, H_1. The decision about the adherence of each frame of a given speech signal to the hypothesized degradation is then computed as:

$$\log p(\boldsymbol{x}_t|H_0) - \log p(\boldsymbol{x}_t|H_1) \begin{cases} \geq \omega, & \text{accept } H_0 \\ < \omega, & \text{reject } H_0, \end{cases} \tag{G.3}$$

where ω is a pre-defined threshold for detection, and $p(\boldsymbol{x}_t|H_0)$ and $p(\boldsymbol{x}_t|H_1)$ are respectively the likelihood of the hypotheses H_0 and H_1 given $\boldsymbol{x}_t$.

To model the characteristics of each hypothesized degradation, we propose to fit a GMM of the likelihood function defined in (G.1) to the frames of the recordings in the feature space. The motivation for using GMMs is that they are computationally efficient models that are capable of modeling sufficiently complex densities as a linear combination of simple Gaussians. Thus, the underlying acoustic classes of the signals might be modeled by individual Gaussian components. While the hypothesized degradation models

can be well characterized by using training voice signals contaminated by the corresponding degradation, it is very challenging to model the alternative hypothesis as it should represent the entire space of all possible negative examples expected during recognition. To model the alternative hypothesis, instead of using individual degradation-specific alternative models, we train a single degradation-independent GMM using a large number of clean, degraded and outlier voice signals. Since this background model is used as an alternative hypothesis model for all hypothesized degradations, it is referred to as a universal background model (UBM).

When the UBM is trained, a set of degradation-dependent GMMs for modeling clean, noisy, reverberant and distorted recordings, $\mathcal{D} = \{\lambda_d\}_{d=1}^{4}$, are derived by adapting the parameters of the UBM through a *maximum a posteriori* estimation and using the corresponding training data. Given the UBM, λ_{ubm}, and the d^{th} trained degradation model, λ_d, and assuming that the feature vectors are independent, the log-likelihood ratio for a test observation, $\boldsymbol{X}_{\text{ts}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$, is calculated as:

$$\sigma_d(\boldsymbol{X}_{\text{ts}}) = \frac{1}{T}\left(\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_d) - \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_{\text{ubm}}) \right). \tag{G.4}$$

The scaling factor in (G.4) is used to make the log-likelihood ratio independent of the signal duration and to compensate for the strong independence assumption for the feature vectors [57]. The decision for the test observation can be made by setting a threshold over the scores.

To parametrize the recordings, we propose to use mel-frequency cepstral coefficients (MFCCs) [58]. Because it has been demonstrated in [18, 59] that degradation in speech signals predictably modifies the distribution of the MFCCs by changing the covariance of the features and shifting the mean to different regions in feature space, and the amount of change is related to the degradation level.

Experimental Setup

For training the UBM, we randomly selected 8,000 recordings from the MMPD data set. To make the training data balanced over the subpopulations to avoid the model to be biased towards the dominant one, we randomly divided this subset into 5 equal partitions of 1,600 samples. The recordings of the first partition were randomly contaminated by six different types of noise namely babble, street, restaurant, office, white Gaussian and wind noises under different SNR conditions ranging from -10 dB to 20 dB in 2 dB steps. The recordings of the second partition were filtered by 46 real room impulse responses (RIRs) of the AIR database [60], measured with mock-up phone in different realistic indoor environments, to produce reverberant data. As an example of non-linearities in signals, the recordings of the third partition

were processed randomly by either clipping, coding or clipping followed by coding. The clipping level was set to 0.3, 0.5 and 0.7. We used 9.6 kbps and 16 kbps code-excited linear prediction (CELP) codecs [61]. To consider the combination of degradations in signals, the recordings of the forth partition were randomly filtered by 46 different real RIRs and added to the noises typically present in indoor environments, namely babble, restaurant and office noise at 0 dB, 5 dB and 10 dB. The recordings of the last partition were used without any processing. The last subset also contains some outliers which do not contain relevant information for PD detection.

For adaptation of the degradation-dependent models, a subset of 800 good-quality recordings of PD patients and healthy speakers of both genders were equally selected from the MMPD data set. From this subset, 200 recordings were corrupted by babble, restaurant, street and office noises under different SNR conditions ranging from -5 dB to 10 dB in 5 dB steps. Another subset of 200 recordings were selected to be filtered by 16 real RIRs from AIR database. A subset of 200 recordings were also chosen to represent nonlinear distortions in signals by processing them in a same way the UBM data were distorted. The remaining 200 recordings were kept unchanged to represent the clean samples.

Using a Hamming window, recordings were segmented into frames of 30 ms with 10 ms overlap. For each frame of a signal, 12 MFCCs together with the log energy are calculated along with *delta* and *double-delta* coefficients. They are concatenated to form a 39-dimensional feature vector.

Results

To evaluate the proposed approach in identifying degradations in data not observed during the training phase, we used 10-fold cross validation with 10 iterations. For each experiment, we extended the test subset by adding 20 good-quality outlier recordings, including irrelevant sounds for PD detection randomly selected from the MMPD data set, to show whether the detectors could reject such outliers. Moreover, as an example of combination of degradations in speech signals, 20 good-quality recordings were selected from the MMPD data set, contaminated by noise and reveberation in a similar way we did for the UBM data, and appended them to the test subset to investigate whether both the noise and reverberation detectors could identify these recordings.

Fig. G.5 shows the performance of the detectors in terms of AUC, along with 95% confidence intervals, as a function of the number of mixture components in GMMs. We can observe from the results that the degradations in voice signals are effectively identified when GMMs with 1024 mixtures are used. The lower performance for reverberation detection model is mainly due to misdetection of some of the recordings in which noise and reverber-

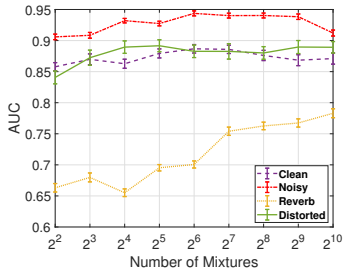


Fig. G.5: The performance of the proposed recording-level degradation detection in terms of AUC, along with 95% confidence intervals, as a function of number of mixture components.

ation coexist but the noise is more dominant than the reverberation. This can also be explained by considering the analysis of vowels in the presence of different degradations [18] which shows that MFCCs of the reverberant signals are, on average, positioned closer to the MFCCs of the clean signals, while noise and distortion (clipping) shift the MFCCs farther away from the position of clean MFCCs.

5.2 Frame-level quality control

While many types of degradation, such as reverberation and nonlinear distortions, typically influence the entire recording, additive noise can have a short-term impact on a signal. Moreover, the test protocol can be violated for a short period of time in a remotely collected voice signal. In recording-level degradation detection, we assumed that the majority segments of a voice signal are influenced by some types of degradation. Likewise, if a voice sample is an outlier, the majority segments of the signal are assumed to contain irrelevant information for PD detection. Even though beneficial in providing a global information about the quality of a signal, it does not say whether a degraded or an outlier signal still contains useful segments to be considered for PD detection. Identifying these segments facilitates making the most use of the available data.

In this paper, we consider additive noise as an example of a short-term degradation in a signal, and develop a framework which splits a voice signal into variable duration segments in an unsupervised manner by fitting an infinite hidden Markov model (iHMM) to the frames of the recordings in the MFCC domain. Then, the degraded segments and those that are associated with the protocol adherence or violation are identified by applying a multinomial naive Bayes classifier.

A HMM represents a probability distribution over sequences of observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$ by invoking a Markov chain of hidden state variables $s_{1:T} = (s_1, \ldots, s_t, \ldots, s_T)$ where each s_t is in one of the K possible states [62].

The likelihood of the observation $\boldsymbol{x}_t$ is modeled with a distribution of K mixture components as:

$$p(\boldsymbol{x}_t|s_{t-1}=i,\boldsymbol{\Theta})=\sum_{k=1}^{K}\pi_{i,k}p(\boldsymbol{x}_t|\boldsymbol{\theta}_k), \tag{G.5}$$

where $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$ are the time-independent emission parameters, $\pi_{ij}=p(s_t=j|s_{t-1}=i)$, $(i,j=1,2,\ldots,K)$, is the transition matrix of $K\times K$. We consider a HMM for clustering the frames of the signals in terms of different acoustic events. The prediction of the number of states required to cover all events such that we do not encounter unobserved events in the future is challenging. Moreover, it is reasonable to assume that as we observe more data, different types of protocol violations and acoustic events will appear and thus the inherent number of states will have to adapt accordingly. Here, we propose to use an infinite HMM to relax the assumption of a fixed K in (G.5), which is defined as:

$$\begin{aligned}
\boldsymbol{\beta} &\sim \text{GEM}(\gamma) \\
\boldsymbol{\pi}_k &\sim \text{DP}(\alpha,\boldsymbol{\beta}) && (k=1,2,\ldots,\infty) \\
\boldsymbol{\theta}_k &\sim H && (k=1,2,\ldots,\infty) \\
s_0 &= 1 \\
s_t|s_{t-1} &\sim \boldsymbol{\pi}_{s_{t-1}} && (t=1,2,\ldots,T) \\
\boldsymbol{x}_t|s_t &\sim f(\boldsymbol{\theta}_{s_t}) && (t=1,2,\ldots,T).
\end{aligned} \tag{G.6}$$

where $\boldsymbol{\pi}_k\sim\text{DP}(\alpha,\boldsymbol{\beta})$ are drawn from a Dirichlet process (DP) with a local concentration parameter $\alpha>0$, $\boldsymbol{\beta}$ is the stick-breaking representation for DPs which is drawn from Griffiths-Engen-McCloskey (GEM) distribution with a global concentration parameter $\gamma>0$ [63], each $\boldsymbol{\theta}_k$ is a sample drawn independently from the global base distribution over the component parameters of the HMM H, and f is the observation model for each state. The iHMM can possibly have countably infinite number of hidden states. Using the direct assignment Gibbs sampler, which marginalizes out the infinitely many transition parameters, we infer the posterior over the sequence of hidden states $\boldsymbol{\pi}$ and emission parameters $\boldsymbol{\Theta}$. In each iteration of the Gibbs sampling, we first re-sample the hidden states and then the base distribution parameters. For more details about the inference, we refer to [21].

Considering an iHMM as a clustering algorithm, segments of the voice recordings with similar characteristics are clustered together under the same state indicator values. To identify the segments of the signal that are sufficiently reliable for detecting PD voice symptoms, those that need enhancement before being used for PD detection, and those which do not contain relevant information for PD detection, we propose to use the multinomial naive Bayes classifier to map the state indicators $s_{1:T}$ to the labels

$y_{1:T} = (y_1, \ldots, y_t, \ldots, y_T)$, where $y_t = 1$ if $\boldsymbol{x}_t$ adheres to the protocol, $y_t = 2$ if it complies with the protocol but is degraded by additive noise, or $y_t = 3$ if it violates the protocol. In the multinomial naive Bayes, we assume that the samples in different classes have different multinomial distributions, and a feature vector for the t^{th} observation $\boldsymbol{\rho}_t = (\rho_{t,1}, \ldots, \rho_{t,K})$ is a histogram, with $\rho_{t,k}$ being the number of times state k is observed. The likelihood of the histogram of a new observation $\tilde{\boldsymbol{\rho}}$ is defined as:

$$P(\tilde{\boldsymbol{\rho}}|y_{1:T}, \tilde{y}, \boldsymbol{\rho}_{1:T}) = \frac{(\sum_{k=1}^{K} \rho_{t,k})!}{\prod_{k=1}^{K} \rho_{t,k}!} \prod_{k=1}^{K} p_{k,\tilde{y}}^{\rho_{t,k}}, \tag{G.7}$$

where $p_{k,\tilde{y}}$ is the probability of the k^{th} attribute being in class $\tilde{y} \in \{1, 2, 3\}$, which is trained using the training data. Using the Bayes rule and the prior class probability $P(\tilde{y})$, the class label for a new test observation is predicted as:

$$\hat{y} = \underset{y \in \{1,2,3\}}{\arg\max} \left(\log P(\tilde{y} = y) + \sum_{k=1}^{K} \tilde{\rho}_k \log(p_{k,y}) \right). \tag{G.8}$$

Experimental Setup

To evaluate the performance of the proposed method, a subset of 100 good-quality recordings (50 PD patients and 50 healthy controls equally from both genders) has been selected from the MMPD data set. From this subset, 50 recordings were selected and 60% of each signal were degraded by adding noise. We used babble, office, restaurant, street and wind noises, under different SNR conditions ranging from -5 dB to 10 dB in steps of 2.5 dB. In addition, 20 recordings from the MMPD data set containing several short- and long-term protocol violations were selected and added to the subset.

Using a Hamming window, recordings are segmented into frames of 30 ms with 10 ms overlap. For each frame of a signal, 12 MFCCs along with the log energy are calculated. The features of every five consecutive frames are averaged to smooth out the impact of articulation [59], and to prevent capturing very small changes in signal characteristics, which results in producing many uninterpretable states. Thus, each observation represents an averaged MFCCs of $\approx$100 ms of a signal. For the iHMM, we use the conjugate normal-gamma prior over the Gaussian state parameters, set the hyper-parameters α=$\gamma = 10$, and run the inference for 150 iterations.

Results

The top plot in Fig. G.6 shows a segment of 10 seconds duration selected from the data set. The segments of the signal which adhere to the test protocol and those that need enhancement are hand-labeled and shaded in green

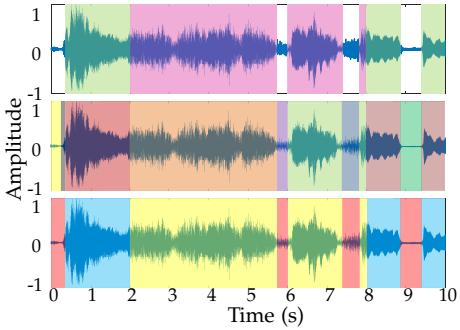


Fig. G.6: Illustrative results of applying the proposed frame-level degradation detection method to a 10-second segment of the voice recordings selected from the data set. In the top plot, the green shaded and pink shaded areas represent the segments of the signal which are hand-labeled as adhering to the protocol and those need enhancement, respectively. The middle plot shows the states, generated by the iHMM, in different colors. The bottom plot illustrates the result of applying a trained classifier to the state indicators to predict which segments adhere to (shaded in blue), which ones violate the protocol (shaded in red), and which ones need enhancement (shaded in yellow).

and pink, respectively. Fitting the iHMM to the data, 49 different states were discovered in this particular subset. The middle plot in Fig. G.6 illustrates the generated states in different colors. To evaluate the performance of the proposed approach for data not observed during the training phase (i.e. out of sample), we used 10-fold CV and repeated the procedure 10 times. The results, presented in Table G.2, indicate that the proposed method can automatically identify short-term degradation and protocol violations in pathological voices with a 0.1 second resolution and high accuracy.

Table G.2: The confusion matrix of the proposed frame-level quality control method. Results are in the form of mean±STD.

		Predicted		
		Adherence	Degraded	Violation
Actual	Adherence	$95\% \pm 1\%$	$3\% \pm 1\%$	$2\% \pm 1\%$
	Degraded	$10\% \pm 2\%$	$89\% \pm 2\%$	$0\% \pm 0\%$
	Violation	$5\% \pm 2\%$	$2\% \pm 1\%$	$93\% \pm 2\%$

5.3 Integrating quality control and enhancement algorithms

The proposed quality control approaches can be integrated with the enhancement algorithms for cleaning-up the remotely collected signals before they are being processed by a PD detection system. In this section, we evaluate how this integration can lead to improvement in PD detection accuracy.

The recording-level algorithm can be used in many different ways to provide information about the presence and type of degradation in a signal for

an automatic clean-up process. For example, one possible scenario could be to convert the parallel detectors to a multi-class classifier by calculating the maximum *a posteriori* probability for a new observation. Then, the enhancement algorithm for which the observation has the highest degradation class probability will be applied. Nevertheless, the advantage of the proposed method over the classification-based techniques is its capability to detect outlier recordings and those degraded by a new type of degradation. Thus, alternative approach could be to exploit the detectors to activate or bypass a set of enhancement blocks connected in series (e.g. noise reduction followed by dereverberation). This scenario not only allows enhancement of a signal degraded by more than one degradation, but also prevents outliers to be processed by the PD detection system. However, since there is no ground truth health status label for the outlier recordings, it is not possible to evaluate the performance of the PD detection system in the presence of outliers. For this reason, we considered a simple scenario in which the test subset only contains clean, noisy and reverberant recordings. Since there was no outlier in the test samples, the problem is simplified to a multi-class classification task. For the experiment, we used the same 160 test recordings we used for the enhancement experiments. From this subset, 60 recordings were randomly selected and corrupted by restaurant, office and street noises under different SNR conditions ranging from -5 dB to 7 dB in 4 dB steps. Another 60 randomly chosen recordings were filtered by 16 real RIRs from AIR database. The enhancement algorithm used in this experiment is the DNN-SE. The model for noise reduction was trained using the noisy recordings and the model used for dereverberation was trained using reverberant recordings. Table G.3 shows the PD detection performance in terms of AUC for four different scenarios: (1) when no enhancement is applied to the recordings, (2) when the recordings, regardless of the presence and type of degradation, were processed randomly by either of the enhancement algorithms, (3) when recordings were enhanced by the enhancement model selected based on the estimated degradation labels, and (4) when the degraded recordings were enhanced based on the ground truth degradation labels. Comparing the results of the first and the second rows with those of the third and the forth rows suggests that applying appropriate enhancement algorithms to the degraded signals leads to an improvement in PD detection performance, and the level of improvement is related to the accuracy of the degradation detection system.

In the next experiment, we investigate how the proposed frame-level quality control method can improve the performance of PD detection. To this aim, we randomly added babble, restaurant, office and street noises to all 160 test recordings at different SNRs ranging from -5 dB to 10 dB in 5 dB steps. However, for making a signal noisy, instead of adding a noise to the entire signal, we randomly corrupted 60% frames of the signal. The enhance-

Table G.3: Evaluation of the impact of applying the proposed Recording-level quality control in combination with DNN-SE on the PD detection performance.

Scenarios	AUC
No Enhancement	0.84
Enhancement based on Randomly Chosen Algorithm	0.86
Enhancement based on Predicted Labels	0.89
Enhancement based on Ground Truth Labels	0.90

ment algorithm used in this experiment is the Kalman-CB. In Table G.4, we compare the PD detection performance for four different scenarios: (1) when no enhancement is applied to the recordings, (2) when the entire signals are enhanced, (3) when the signals are enhanced based on the predicted labels, and (4) when the signals are enhanced based on the ground truth labels. For

Table G.4: Evaluation of the impact of applying the proposed frame-level quality control on the PD detection performance.

Scenarios	AUC
No Enhancement	0.86
Enhancement of Entire Recording	0.89
Enhancement based on Predicted Labels	0.92
Enhancement based on Ground Truth Labels	0.93

the last two scenarios, only the segments of the signals identified/labeled as degraded were enhanced. Moreover, we dropped the features of the frames identified as protocol violation. Comparing the result of second scenario with the last two scenarios, we can observe the superiority of integrating the proposed frame-level quality control and the enhancement algorithm in dealing with short-term degradation and protocol violations in recordings.

6 Conclusion

Additive noise, reverberation and nonlinear distortion are three types of degradation typically encountered during remote voice analysis which cause an acoustic mismatch between training and operation conditions. In this paper, we investigated the impact of these degradations on the performance of a PD detection system. Then, given that the specific degradation is known, we explored the effectiveness of a variety of the state-of-the-art enhancement algorithms in reducing this mismatch and, consequently, in improving the PD detection performance. We showed how applying appropriate enhancement algorithms can effectively improve the PD detection accuracy. To inform the choice of enhancement method, we proposed two quality control techniques operating at recording- and frame-level. The recording-level approach provides information about the presence and type of degradation in voice signals. The frame-level algorithm, on the other hand, identifies the short-term degradations and protocol violations in voice recordings. Experimental re-

sults showed the effectiveness of the quality control approaches in choosing appropriate signal enhancement algorithms which resulted in improvement in the PD detection accuracy.

This study has important implications that extend well beyond the PD detection system. It can be considered as a step towards the design of robust speech-based applications capable of operating in a variety of acoustic environments. For example, since the proposed quality control approaches are not limited to specific speech types, they can be used as a pre-processing step for many end-to-end speech-based systems, such as automatic speech recognition, to make them more robust against different acoustic conditions. They might also be utilized to automatically control the quality of recordings in large-scale speech data sets. Moreover, these approaches have the potential to be used for other sensor modalities to identify short- and long-term degradations and abnormalities which can help to choose an adequate action.

References

[1] L. S. Ishihara, A. Cheesbrough, C. Brayne, and A. Schrag, "Estimated life expectancy of Parkinson's patients compared with the UK population," *Journal of Neurol Neurosurg Psychiatry*, vol. 78, pp. 1304–1309, 2007.

[2] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease." *Behavioural Neurology*, vol. 11, no. 3, pp. 131–137, 1998.

[3] I. Eliasova, J. Mekyska, M. Kostalova, R. Marecek, Z. Smekal, and I. Rektorova, "Acoustic evaluation of short-term effects of repetitive transcranial magnetic stimulation on motor aspects of speech in Parkinson's disease," *Journal of Neural Transmission*, vol. 120, no. 4, pp. 597–605, 2013.

[4] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 1264–1271, 2012.

[5] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. R. Dorsey, S. Saria, and A. Terzis, "High frequency remote monitoring of Parkinson's disease via smartphone: platform overview and medication response detection," *arXiv preprint arXiv:1601.00960*, pp. 1–12, 2016.

[6] D. Gil and M. Johnson, "Diagnosing Parkinson by using artificial neural networks and support vector machines," *Global Journal of Computer Science and Technology*, pp. 63–71, 2009.

[7] J. Rusz, J. Hlavnička, T. Tykalová, M. Novotný, P. Dušek, K. Šonka, and E. Ružička, "Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1495–1507, 2018.

[8] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.

[9] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, 1995.

[10] M. Fakhry, A. H. Poorjam, and M. G. Christensen, "Speech enhancement by classification of noisy signals decomposed using NMF and Wiener filtering," in *Proc. European Signal Processing Conf.*, 2018.

[11] J. H. L. Hansen, A. Kumar, and P. Angkititrakul, "Environment mismatch compensation using average eigenspace-based methods for robust speech recognition," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 353–364, 2014.

[12] J. Alam, P. Kenny, G. Bhattacharya, and M. Kockmann, "Speaker verification under adverse conditions using i-vector adaptation and neural networks," in *Proc. Interspeech*, 2017, pp. 3732–3736.

[13] R. J. Mammone, Xiaoyu Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, p. 58, 1996.

[14] S. Nercessian, P. Torres-Carrasquillo, and G. Martinez-Montes, "Approaches for language identification in mismatched environments," in *IEEE Spoken Language Technology Workshop*, 2016, pp. 335–340.

[15] A. H. Poorjam, R. Saeidi, T. Kinnunen, and V. Hautamäki, "Incorporating uncertainty as a quality measure in i-vector based language recognition," in *Speaker and Language Recognition Workshop*, Bilbao, Spain, 2016, pp. 74–80.

[16] J. Vasquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londono, and E. Nöth, "Automatic detection of Parkinson's disease from continuous speech recorded in real-world conditions," in *Proc. Interspeech*, 2015, pp. 3–7.

[17] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "Quality control in remote speech data collection," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, 2019.

[18] A. H. Poorjam, J. R. Jensen, M. A. Little, and M. G. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in *Proc. Interspeech*, 2017, pp. 289–293.

[19] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A parametric approach for classification of distortions in pathological voices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 286–290.

[20] R. Badawy, Y. P. Raykov, L. J. W. Evers, B. R. Bloem, M. J. Faber, A. Zhan, K. Claes, and M. A. Little, "Automated quality control for sensor based symptom measurement performed outside the lab," *Sensors*, vol. 18, no. 4, 2018.

[21] A. H. Poorjam, Y. P. Raykov, R. Badawy, J. R. Jensen, M. G. Christensen, and M. A. Little, "Quality control of voice recordings in remote Parkinson's disease monitoring using the infinite hidden Markov model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019.

[22] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease," *Applied Soft Computing*, vol. 62, pp. 649–666, 2018.

[23] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[24] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Perceptual analysis of speech signals from people with Parkinson's disease," *Natural and Artificial Models in Computation and Biology - Lecture Notes in Computer Science*, vol. 7930, no. 1, pp. 201–211, 2013.

[25] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation," *Journal of Neural Transmission*, vol. 124, no. 3, pp. 303–334, 2017.

[26] J. Mekyska, Z. Smekal, Z. Galaz, Z. Mzourek, I. Rektorova, M. Faundez-Zanuy, and K. López-de Ipiña, "Perceptual features as markers of Parkinson's disease: the issue of clinical interpretability," in *Recent Advances in Nonlinear Speech Processing*, 2016, pp. 83–91.

[27] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[28] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, and A. D. Trister, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Scientific Data*, vol. 3, no. 160011, 2016.

[29] M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[30] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[31] P. Castellano, S. Sradharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 117–120.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, p. 943, 1979.

[33] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[34] J. Eaton and P. A. Naylor, "Detection of clipping in coded speech signals," in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[35] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Independent Component Analysis and Signal Separation*, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 552–559.

[36] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7-8, pp. 588–601, 2007.

[37] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.

[38] E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[39] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5172–5176.

[40] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[41] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 45–48.

[42] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, 2003.

[43] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[44] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 7, pp. 1236–1246, 2018.

[45] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[46] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 3, pp. 457–468, 2017.

[47] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[48] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[49] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 1, pp. 99–113, 2019.

[50] Z. Goh, K. C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 510–524, 1999.

[51] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.

[52] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[53] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[54] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[55] E. A. Habets, S. Gannot, I. Cohen, and P. C. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 8, pp. 1433–1451, 2008.

[56] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 4, pp. 680–693, 2016.

[57] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[58] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, 2nd ed. New York: IEEE Press, 2000.

[59] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018.

[60] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *International Conference on Digital Signal Processing*, 2009, pp. 1–5.

[61] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 10, pp. 937–940, 1985.

[62] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[63] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

Paper H

Online Parametric NMF for Speech Enhancement

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Liming Shi, Mads Græsbøll Christensen and Jesper B. Boldt

The paper has been published in the
Proc. European Signal Processing Conference, 2018

The layout has been revised.

Abstract

In this paper, we propose a speech enhancement method based on non-negative matrix factorization (NMF) techniques. NMF techniques allow us to approximate the power spectral density (PSD) of the noisy signal as a weighted linear combination of trained speech and noise basis vectors arranged as the columns of a matrix. In this work, we propose to use basis vectors that are parameterised by autoregressive (AR) coefficients. Parametric representation of the spectral basis is beneficial as it can encompass the signal characteristics like, e.g. the speech production model. It is observed that the parametric representation of basis vectors is beneficial while performing online speech enhancement in low delay scenarios.

1 Introduction

A healthy human auditory system is capable of focusing on desired signal from a target source while ignoring background noise in a complex noisy environment. In comparison to a healthy auditory system, the auditory system of a hearing impaired person lacks this ability, leading to degradation in speech intelligibility. In such scenarios, a hearing impaired person often relies on speech enhancement algorithms present in a hearing aid. However, the performance of the current hearing aid technology in this aspect is limited [1]. Speech enhancement algorithms that have been developed can be mainly categorised into supervised and unsupervised methods. Some of the existing unsupervised methods are spectral subtraction methods [2], statistical model based methods [3] and subspace based methods [4]. Supervised methods generally use some amount of training data to estimate the model parameters corresponding to speech and noise. The model parameters are subsequently used for enhancement. Examples of supervised enhancement methods include codebook based methods [5, 6], NMF methods [7–9], hidden Markov model based methods [10, 11].

In this paper, we propose a speech enhancement method based on non-negative matrix factorization (NMF) techniques. NMF for source separation and speech enhancement has been previously proposed [7, 8]. NMF techniques allow us to approximate the power spectrum or the magnitude spectrum of the noisy signal as a weighted linear combination of trained speech and noise basis vectors arranged as the columns of a matrix. Generally the basis vectors used in NMF based speech enhancement are not constrained by any parameters. Parameterisation of the basis vectors in the field of music processing has been previously done in [12]. In [12], harmonic combs parametrised by the fundamental frequency was used as the basis vectors. This parametrisation was found to efficiently represent the music signal in comparison to the non parametric counterpart.

In this work, we propose to use basis vectors that are parametrised by autoregressive (AR) coefficients. This parametrisation allows representation of power spectral density (PSD) using a small set of parameters. Parametrisation by AR coefficients is motivated by the source filter model of speech production. This model describes speech components as a combination of a sound source (excitation signal produced by the vocal chords) and an AR filter which models the vocal tract. In this work, we show that if we model the observed data in the time domain as a sum of AR processes, the maximisation of the likelihood corresponds to performing NMF of the observed data into a basis matrix and activation coefficients, using Itakura-Saito (IS) divergence as the optimisation criterion. The IS divergence has been extensively used in speech and music processing due to its similarity to perceptual distance. The basis matrix here consists of AR spectral envelopes parameterised by AR coefficients, and the activation coefficients can be physically interpreted as the excitation variance of the noise that excites the AR filter parametrised by the AR coefficients. A benefit of parametrically representing the spectral basis, is that, it can be represented by a small set of parameters, which means that fewer parameters have to be trained a priori for performing on-line speech enhancement.

The remainder of this paper is organised as follows. Section 2 explains the signal model and formulates the problem mathematically. Training of the speech and noise spectral bases is explained in Section 3. Section 4 explains the on-line estimation of the activation coefficients corresponding to the spectral bases followed by the enhancement procedure using the Wiener filter. Sections 5 and 6 give the experimental results and the conclusion respectively.

2 Mathematical formulation

This section explains the signal model and mathematically formulates the problem. The noisy signal is expressed as

$$x(n) = s(n) + w(n) \tag{H.1}$$

where $s(n)$ is the clean speech and $w(n)$ is the noise signal. The objective of a speech enhancement system is to obtain an estimate of the clean speech signal from the noisy signal. A very popular method for estimating the clean speech signal is by applying a Wiener filter onto the noisy signal. Wiener filtering requires the knowledge of the speech and noise statistics. Since there is no direct access to either speech or noise in practical scenarios, these statistics have to be estimated from the noisy observation. As the speech and noise properties change over time, these statistics are generally time varying. The majority of the speech processing algorithms consider these statistics to be

quasi-stationary. Thus, these statistics are assumed to be constant for short segments of time ($\approx$ 25 ms).

We now, explain the signal model used in the estimation of the statistics from a frame of noisy signal. It is assumed that a frame of noisy signal $\mathbf{x} = [x(0), \dots x(N-1)]^T$ can be represented as a sum of $U = U_s + U_w$ AR processes $\mathbf{c}_u$. This is mathematically written as

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{c}_u \;=\; \sum_{u=1}^{U_s} \mathbf{c}_u + \sum_{u=U_s+1}^{U} \mathbf{c}_u, \tag{H.2}$$

where the first U_s AR processes correspond to the speech signal and the remaining U_w AR processes correspond to the noise signal. Each of the AR process is expressed as a multivariate Gaussian [6] as shown below

$$\mathbf{c}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{Q}_u). \tag{H.3}$$

The gain normalised covariance matrix, $\mathbf{Q}_u$ can be asymptotically approximated as a circulant matrix which can be diagonalised using the Fourier transform as [13]

$$\mathbf{Q}_u = \mathbf{F}\mathbf{D}_u\mathbf{F}^H \tag{H.4}$$

where $\mathbf{F}$ is the DFT matrix defined as $[\mathbf{F}]_{k,n} = \frac{1}{\sqrt{N}} \exp(j2\pi nk/N)$, $n, k = 0 \dots N-1$ and

$$\mathbf{D}_u = (\mathbf{\Lambda}_u^H \mathbf{\Lambda}_u)^{-1}, \qquad \mathbf{\Lambda}_u = \text{diag}(\sqrt{N}\mathbf{F}^H \begin{bmatrix} \mathbf{a}_u \\ \mathbf{0} \end{bmatrix}) \tag{H.5}$$

where $\mathbf{a}_u = [1, a_u(1) \dots a_u(P)]^T$ represents the vector of AR coefficients corresponding to u^{th} basis vector and P is the AR order. The likelihood as a function of U excitation variances and AR spectral envelopes are expressed as

$$p(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{D}) \sim \mathcal{N}(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2 \mathbf{Q}_u) \tag{H.6}$$

where $\boldsymbol{\sigma}$ represents the excitation variances corresponding to the U AR processes and $\mathbf{D}$ represents AR spectral envelopes corresponding to the U AR processes. In this paper, we are interested in the maximum likelihood (ML) estimation of activation coefficients $\boldsymbol{\sigma}$ given the noisy signal $\mathbf{x}$. Since, we are performing supervised enhancement here, we assume that the spectral basis are trained a priori, which is explained in Section 3. Thus, in this work we only estimate the activation coefficients online while the basis vectors are assumed known. This is expressed mathematically as, To solve this, the log-

arithm of likelihood in (H.6) is written as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma},\ \mathbf{D}) = -\frac{N}{2}\ln 2\pi + \ln\Big|\sum_{u=1}^{U}\sigma_u^2\mathbf{F}\mathbf{D}_u\mathbf{F}^H\Big|^{-\frac{1}{2}} - \frac{1}{2}\mathbf{x}^T[\sum_{u=1}^{U}\sigma_u^2\mathbf{F}\mathbf{D}_u\mathbf{F}^H]^{-1}\mathbf{x}. \tag{H.7}$$

This is further simplified as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma},\ \mathbf{D}) = -\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\Big(\sum_{u=1}^{U}\sigma_u^2 d_u(k)\Big)^{-\frac{1}{2}} - \frac{1}{2}\mathbf{x}^T\mathbf{F}[\sum_{u=1}^{U}\sigma_u^2\mathbf{D}_u]^{-1}\mathbf{F}^H\mathbf{x} \tag{H.8}$$

where $d_u(k)$ represents the k^{th} diagonal element of $\mathbf{D}_u$ and number of frequency indices $K = N$. Further simplifying,

$$\ln p(\mathbf{x}|\boldsymbol{\sigma},\ \mathbf{D}) = -\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\Big(\sum_{u=1}^{U}\hat{\Phi}_u(k)\Big)^{-\frac{1}{2}} - \frac{1}{2}\sum_{k=1}^{K}\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} \tag{H.9}$$

where $\hat{\Phi}_u(k) = \sigma_u^2 d_u(k)$, $\Phi(k) = |X(k)|^2$ and $X(k) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}x(n)\exp(-j2\pi nk/N)$. Log-likelihood is then written as

$$\ln p(\mathbf{x}|\boldsymbol{\sigma},\mathbf{D}) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} + \ln\sum_{u=1}^{U}\hat{\Phi}_u(k)\right) \tag{H.10}$$

where

$$\sum_{u=1}^{U}\hat{\Phi}_u(k) = \sum_{u=1}^{U}\sigma_u^2 d_u(k) = \mathbf{d}_k\boldsymbol{\sigma} \tag{H.11}$$

where $\mathbf{d}_k = [d_1(k)\dots d_U(k)]$ and $\boldsymbol{\sigma} = [\sigma_1^2\dots\sigma_U^2]^T$. Thus maximising the likelihood is equivalent to minimising the IS divergence between $\boldsymbol{\phi} = [\Phi(1)\dots\Phi(K)]^T$ and $\mathbf{D}\boldsymbol{\sigma}$ subject to $\Phi(k) > 0\ \forall k$ where $\mathbf{D} = [\mathbf{d}_1^T\dots\mathbf{d}_K^T]^T$. In case we observe $V > 1$ frames, this corresponds to performing NMF of $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1\dots\boldsymbol{\phi}_v\dots\boldsymbol{\phi}_V]$ (where $\boldsymbol{\phi}_v = [\Phi_v(1)\dots\Phi_v(K)]^T$ contains the periodogram of the v^{th} frame) as

$$\boldsymbol{\Phi} \approx \underbrace{\begin{bmatrix} d_1(1) & \dots & d_U(1) \\ \vdots & \ddots & \vdots \\ d_1(K) & \dots & d_U(K) \end{bmatrix}}_{\mathbf{D}}\underbrace{\begin{bmatrix} \sigma_1^2(1) & \dots & \sigma_1^2(V) \\ \vdots & \ddots & \vdots \\ \sigma_U^2(1) & \dots & \sigma_U^2(V) \end{bmatrix}}_{\boldsymbol{\Sigma}}. \tag{H.12}$$

The first U_s columns of $\mathbf{D}$ corresponds to the spectral basis corresponding to the speech and the remaining U_w columns of $\mathbf{D}$ correspond to noise signal. The first U_s rows of $\mathbf{\Sigma}$ correspond to the activation coefficients for speech and the remaining U_w rows of $\mathbf{\Sigma}$ correspond to the activation coefficients corresponding to the noise signal, which leads to (H.12) being rewritten as,

$$\mathbf{\Phi} \approx [\mathbf{D}_s \, \mathbf{D}_w] \begin{bmatrix} \mathbf{\Sigma}_s \\ \mathbf{\Sigma}_w \end{bmatrix} = \mathbf{D}\mathbf{\Sigma}. \tag{H.13}$$

3 Training the Spectral Bases

This section explains the training of the basis vectors used for the construction of the basis matrix $\mathbf{D}$. In this work we use a parametric representation of the PSD [14] where the u^{th} spectral basis $\mathbf{d}_u = [d_u(1)...d_u(k)...d_u(K)]^T$ is represented as

$$d_u(k) = \frac{1}{\left|1 + \sum_{p=1}^{P} a_u(p)\exp(\frac{-j2\pi pk}{N})\right|^2}, \tag{H.14}$$

where $\{a_u(p)\}_{p=1}^{P}$ is the set of AR coefficients corresponding to the u^{th} basis vector. During the training stage, a speech and noise codebook is first computed using the generalised Lloyd algorithm [15] [16] [6]. The speech codebook and noise codebooks contain AR coefficients corresponding to the spectral envelopes of speech and noise. During the training process linear prediction coefficients (converted into line spectral frequency coefficients) are extracted from windowed frames, obtained from the training signal and passed as input to the vector quantiser. Once the speech codebook and noise codebooks are created, the spectral envelopes corresponding to the speech AR coefficients ($\{\mathbf{a}_u\}_{u=1}^{U_s}$) and noise AR coefficients ($\{\mathbf{a}_u\}_{u=U_s+1}^{U}$) are computed using (H.14), and arranged as columns of $\mathbf{D}$. The spectral envelopes generated here are gain normalised, so they do not include the excitation variance. Fig. H.1 shows a few examples of the trained speech and noise spectral envelopes.

4 Enhancement - Multiplicative Update

This section describes the estimation of speech and noise PSDs using the signal model explained in Section 2. Since we are interested in on-line processing of the noisy signal, we here assume that only a frame of noisy signal is available at particular time for enhancement. The method considered here assumes that

$$\boldsymbol{\phi} \approx \mathbf{D}\boldsymbol{\sigma} \tag{H.15}$$

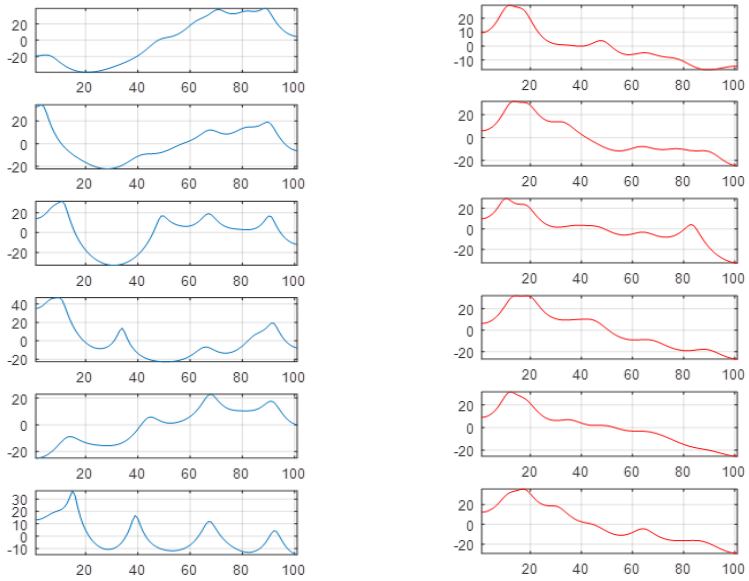


Fig. H.1: Figure showing a set of (a) trained speech spectral envelopes and (b) noise spectral envelopes.

where $\boldsymbol{\phi}$ is a $K \times 1$ vector containing the noisy PSD, $\mathbf{D}$ is $K \times U$ basis matrix and $\boldsymbol{\sigma}$ is $U \times 1$ vector containing the activation coefficients. The objective here, is to estimate $\boldsymbol{\sigma}$ given the noisy periodogram $\boldsymbol{\phi}$ and $\mathbf{D}$. As explained in Section 2, this is done by minimising the IS divergence as

$$\boldsymbol{\sigma}_{est} = [\boldsymbol{\sigma}_{s_{est}}^T \; \boldsymbol{\sigma}_{w_{est}}^T]^T = \underset{\boldsymbol{\sigma} \geq 0}{\arg\min} \quad d_{\text{IS}}(\boldsymbol{\phi} | \mathbf{D}\boldsymbol{\sigma}). \tag{H.16}$$

In this work, a multiplicative update (MU) method is used to estimate the activation coefficients which are calculated as [8, 17]

$$\boldsymbol{\sigma}_{est} \leftarrow \boldsymbol{\sigma}_{est} \frac{\mathbf{D}^T((\mathbf{D}\boldsymbol{\sigma}_{est})^{[-2]}.\boldsymbol{\phi})}{\mathbf{D}^T(\mathbf{D}\boldsymbol{\sigma}_{est})^{[-1]}}. \tag{H.17}$$

Once the gains are estimated, a Wiener filter can be constructed to extract the speech/noise components. The estimated clean speech PSD is obtained as $\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}}$ and the estimated noise PSD is obtained as $\mathbf{D}_w\boldsymbol{\sigma}_{w_{est}}$. The Wiener filter vector constructed to extract the speech component is denoted as

$$\mathbf{g}_{est} = \frac{\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}}}{\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}} + \mathbf{D}_w\boldsymbol{\sigma}_{w_{est}}}, \tag{H.18}$$

where the division is an element wise division.

5 Experiments

5.1 Implementation Details

This section explains the experiments that have been carried out to evaluate the proposed enhancement framework. The test signals used here consist of sentences taken from the GRID database [18]. The speech and noise PSD parameters are estimated (as explained in Section 4) for a segment of 25 ms with 50 percent overlap. The parameters used for the experiments are summarised in table H.1. For our experiments, we have used both a speaker-specific codebook and a general speech codebook. A speaker-specific codebook of 64 entries was trained using a training sample of 5 minutes of speech from the specific speaker of interest. A general speech codebook of 64 entries was trained from a training sample of approximately 150 minutes of speech from 30 different speakers. It should be noted that the sentences used for training the codebook were not included for testing. The proposed enhancement framework was tested on three different types of commonly encountered background noise: babble, restaurant and exhibition noise taken from the NOIZEUS database [19]. We have performed experiments for a noise specific codebook as well as general noise codebook. A noise-specific codebook of 8 entries was trained on the specific noise type of interest. For creating a general noise codebook, a noise codebook of 4 entries was trained for each noise type. While testing for a particular noise scenario, the noise codebook entries corresponding to that scenario are not used for the estimation of noise PSD. For example, while testing in the babble noise scenario, the noise codebook consists a total of 8 entries formed by concatenating the entries trained for restaurant and exhibition scenarios. After obtaining the speech and noise codebooks, the spectral basis matrix is constructed as explained in Section 3. The estimated PSD parameters are then used to create a Wiener filter for speech enhancement. Wiener filter is applied in the frequency domain and time-domain enhanced signal is synthesised using overlap-add.

5.2 Results

We have used the objective measures such as STOI and Segmental SNR to evaluate the proposed algorithm. We will denote the proposed parametric NMF as ParNMF. We have compared the performance of the proposed method to non parametric NMF where there is no parametrisation involved in the creation of the basis vectors. We will denote this method as NonParNMF. It should be noted that we have used the same training set for ParNMF and NonParNMF. We have also used the speech enhancement method proposed in [20] for comparison purposes, which we denote as MMSE-GGP. Traditionally, NMF methods for speech enhancement generally

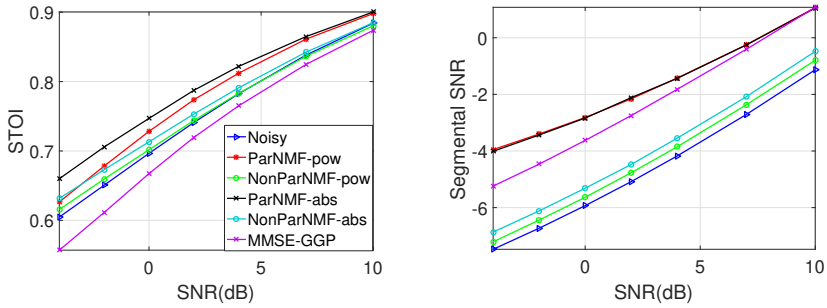
try to approximate the magnitude spectrum than the power spectrum. Even though, this is not theoretically well formulated, this has been observed to give better performance [21]. Thus, here we evaluated the performance of the proposed algorithm for both the cases, which we denote as ParNMF-abs while approximating the magnitude spectrum and ParNMF-pow while approximating the power spectrum. We do the same evaluation in the case of NonParNMF. Figures H.2-H.4 show these measures for different methods in different commonly encountered background noises while using a speaker specific codebook and a noise specific codebook. It can be seen that NMF based methods perform better than MMSE-GGP in terms of STOI. When comparing the ParNMF and NonParNMF, it is demonstrated that the former performs better in terms of STOI and Segmental SNR measures. We have also performed experiments when having an access to a general speech codebook and a general noise codebook. Figures H.5-H.7 shows the objective measures obtained for this case. It can be seen that performance in this case degrades in comparison to figures H.2-H.4 due to the mismatch in training and testing conditions. Even though there is a degradation in the performance, the proposed method is able to increase the STOI measure significantly over the conventional method.

6 Conclusion

In this paper, we have proposed an NMF based speech enhancement method where the basis vectors are parametrised using AR coefficients. Parametrisation of the spectral basis vectors helps in encompassing the signal characterestics. We have demonstrated, through objective measures, that the proposed parametric NMF based speech enhancement out performs its non-parametric counterpart in some of the typically encountered background noises.

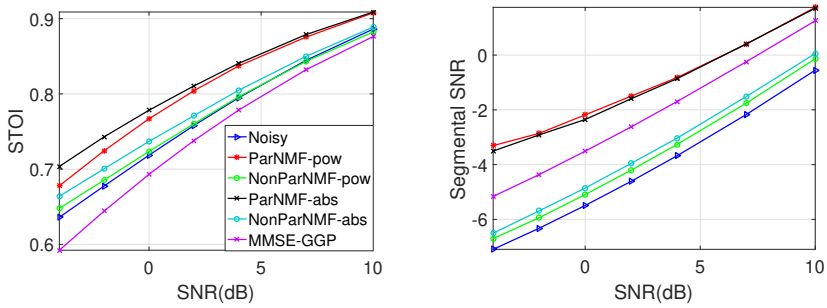
Table H.1: Parameters used for the experiments

Parameters	
sampling frequency	8000 Hz
Frame Size	200
Frame Overlap	50%
Speech AR order	14
Noise AR order	14
U_s	64
U_w	8
MU iterations	50



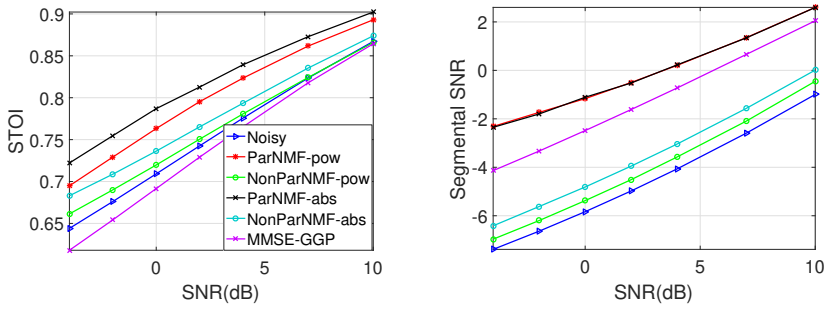
(a) STOI **(b)** Segmental SNR

Fig. H.2: Objective measures for babble noise when using speaker-specific codebook and a noise-specific codebook.



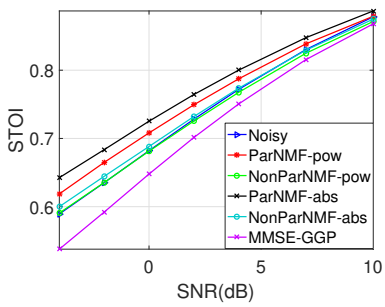
(a) STOI **(b)** Segmental SNR

Fig. H.3: Objective measures for restaurant noise when using speaker-specific codebook and a noise-specific codebook.

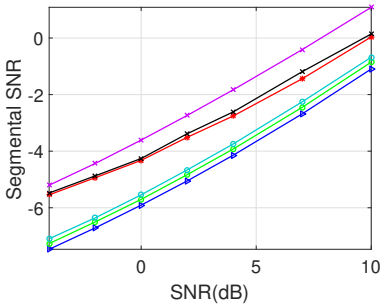


(a) STOI **(b)** Segmental SNR

Fig. H.4: Objective measures for exhibition noise when using speaker-specific codebook and a noise-specific codebook.

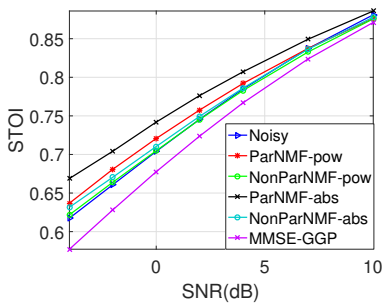


(a) STOI

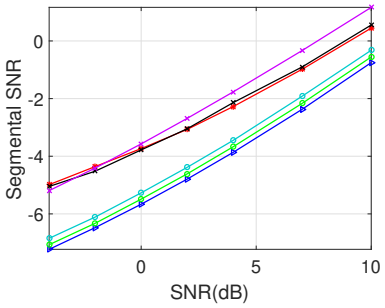


(b) Segmental SNR

Fig. H.5: Objective measures for babble noise when using general speech codebook and a general noise codebook.

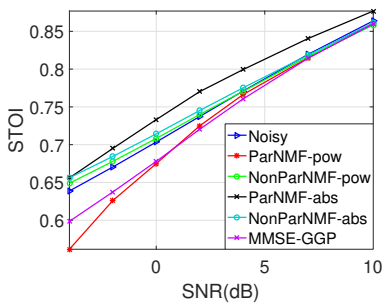


(a) STOI

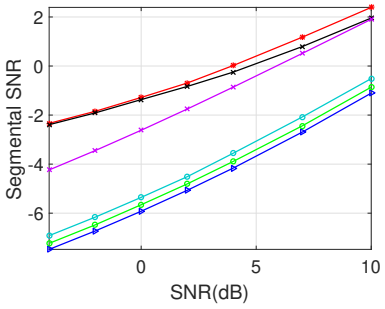


(b) Segmental SNR

Fig. H.6: Objective measures for restaurant noise when using general speech codebook and a general noise codebook.



(a) STOI



(b) Segmental SNR

Fig. H.7: Objective measures for exhibition noise when using general speech codebook and a general noise codebook.

References

[1] S. Kochkin, "10-year customer satisfaction trends in the US hearing instrument market," *Hearing Review*, vol. 9, no. 10, pp. 14–25, 2002.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, 1995.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[6] ——, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[8] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[9] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low snr conditions via noise estimation using sparse and low-rank nmf with kullback–leibler divergence," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, 2015.

[10] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, 1998.

[11] D. Y. Zhao and W. B. Kleijn, "Hmm-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 3, pp. 882–892, 2007.

[12] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proc. of the 13th International Conference on Digital Audio Effects (DAFx)*, 2010.

[13] R. M. Gray *et al.*, "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[14] P. Stoica, R. L. Moses, *et al.*, *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005, vol. 452.

[15] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[16] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2016, pp. 191–195.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," *IEEE 2015 Automatic Speech Recognition and Understanding Workshop*, 2015.

[19] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, 2007.

[20] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[21] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2015, pp. 266–270.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-531-4

AALBORG UNIVERSITY PRESS