New Stategies for Single-channel Speech Separation

Mowlaee Beikzadehmahalen, Pejman

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

New Strategies for Single-channel Speech Separation

Ph.D. Thesis

Pejman Mowlaee

Multimedia Information and Signal Processing
Department of Electronic Systems
Aalborg University
Niels Jernes Vej 12, 9220 Aalborg Ø, Denmark

New Strategies for Single-channel Speech Separation
Ph.D. Thesis

ISBN 978-87-92328-53-3
December 2010

Department of Electronic Systems
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg
Denmark

This thesis was written in LaTeX.

Dedicated to my lovely mother and father and my very kind brother.

Abstract

In many speech applications, the signal of interest is often corrupted by highly correlated noise sources. An extreme example is when several speakers are talking at the same time, a phenomenon called cock-tail party problem. Separating desired speaker signals from their mixture is one of the most challenging research topics in speech signal processing. The problem is called single-channel speech separation (SCSS) where the interfering signal is another speaker. Possible applications include speech coding, speech recognition, hearing aid and forensics where a high quality separation algorithm is required as a pre-processing stage to mitigate the effect of interfering signals.

In the introductory part of this thesis, we present the problem definition and give an overview of its different applications in real life. We then move on presenting the dominating previous SCSS methods and outline the problems they face. As our contribution, we present novel strategies to improve the separation performance in the form of proposing two SCSS systems, namely model-driven SCSS in sinusoidal domain and joint speech separation and speaker identification. We propose sinusoidal mixture estimator for speech separation. We generalize mask methods for speech separation from short-time Fourier transform to sinusoidal case. Experiments show that using sinusoidal masks improved the separation performance compared to the STFT counterpart. A separation system is proposed based on sinusoidal parameters composed of sinusoidal mixture estimator along with sinusoidal coders used as speaker models. To overcome the speaker dependency problem known as a common problem in model-driven SCSS methods, we present a joint closed loop speaker identification and speech separation considered as an attractive approach for speaker-independent SCSS. We also propose two contributions to identify speakers from single-channel speech mixture. We propose a new approach for speaker identification for single-channel speech mixture independent of the signal-to-signal ratio. We present a double-talk detection method to determine the single-talk/double-talk regions in a mixture. We also integrate a double-talk detector with a speaker identification module to improve the speaker identification accuracy. Finally, a joint speech separation and speaker identification system is proposed for separation challenge.

Resumé

I mange taleapplikationer er interessesignalet ofte korrumperet af yderst korrelerede støjkilder. I ekstreme tilfælde er der adskillige talere, som konverserer samtidigt. Dette fænomen kaldes cocktailparty problemet. At adskille det ønskede talesignale fra denne blanding er et af de mest udfordrende forskningsemner indenfor talesignalbehandling. Problemet kaldes for enkeltkanals taleseparation (på engelsk forkortet SCSS), hvor det interfererende signal er en anden taler. Mulige applikationer inkluderer talekodning, talegenkendelse, høreapparater og kriminaltekniske teknikker, hvor separationsalgoritmer af høj kvalitet kræves som præprocessor for at dæmpe effekten af interfererende signaler.

I den indledende del af denne afhandling definerer vi problemet, og giver et overblik over dets forskellige reelle applikationer, hvorefter vi præsenterer de hidtil dominerende SCSS metoder og skitserer deres problemestillinger. Som vores bidrag præsenterer vi nye strategier i form af to SCSS systemer til at forbedre separationsydelsen. Disse er modeldrevne SCSSer i sinusdomnet samt fælles taleseparation og taleridentifikation. Vi foreslår en blandingsestimator til estimation af sinusser til brug i taleseparering og generaliserer maskeringsmetoder til taleseparation fra korttids Fourier transformen (på engelsk forkortet STFT) til sinustilfældet. Eksperimenter viser, at når sinusmaskerne anvendes, forbedres separationsydelsen sammenlignet med STFT. Vi foreslår ydermere et separationssystem baseret på sinusparametre sammensat af en blandingsestimator til sinusser og sinuskodere. For at undgå taleafhængighedsproblemet, der er kendt som et typisk problem i modeldrevne SCSS metoder, præsenterer vi en fælles lukketsløjfe taleridentificering og taleseparering, hvilket anses for at være en attraktiv tilgang til en taleuafhængig SCSS. Derudover har vi to bidrag til at identificere talere i en enkeltkanals taleblanding. Vi foreslår en ny metode til at identificere talere, der kan anvendes på enkeltkanals taleblandinger uafhængigt af signal-støj-forholdet. Vi præsenterer en dobbelttale detektionsmetode til at kortlægge enkelttale/dobbelttale regioner i taleblandinger. Derudover integrerer vi en dobbelttaledetektor i et taleridentificeringsmodul for at forbedre taleridentificeringsprcisionen. Slutteligt foreslår vi et fælles taleseparations- og taleridentficeringssystem til separationsudfordringen.

List of Papers

The main body of this thesis consist of the following papers:

[1] P. Mowlaee, M. G. Christensen; S. H. Jensen, "Improved Single-channel Speech Separation Using Sinusoidal Modeling", *in Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 21–24, March, 2010.

[2] P. Mowlaee, M. G. Christensen; S. H. Jensen, "Sinusoidal Masks for Single-channel Speech Separation", *in Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4262–4266, March, 2010.

[3] P. Mowlaee, R. Saeidi, Z.-H. Tan, M. G. Christensen, P. Fränti, S. H. Jensen, "Joint Single-channel Speech Separation and Speaker Identification", *in Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4430–4433, March, 2010.

[4] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. G. Christensen, S. H. Jensen and P. Fränti, "Signal-to-signal Ratio Independent Speaker Identification for Co-channel Speech Signals", *in Proceedings of the 20th IEEE International Conference on Pattern Recognition (ICPR)*, pp. 4545-4568, August, 2010.

[5] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. G. Christensen, S. H. Jensen and P. Fränti, "Improving Monaural Speaker Identification by Double-Talk Detection", *in Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)* pp. 1069–1072, September, 2010.

[6] P. Mowlaee, M. G. Christensen, Z.-H. Tan, S. H. Jensen, "A MAP Criterion for Detecting the Number of Speakers at Frame Level in Model-based Single-Channel Speech Separation", *in Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers*, pp. 538–541, 2010.

[7] P. Mowlaee, M. G. Christensen, S. H. Jensen, "New Results on Single-Channel Speech Separation Using Sinusoidal Modeling", accepted for publication in *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 1265–1277, 2011.

[8] P. Mowlaee, R. Saeidi, M. G. Christensen, Z.-H. Tan, T. Kinnunen, P. Fränti, S. H. Jensen, "A Joint Approach for Single-Channel Speaker Identification and Speech Separation", submitted to *IEEE Trans. on Audio, Speech and Language Processing*, November, 2010.

Preface

This thesis is submitted to the Faculty of Engineering, Science and Medicine at Aalborg University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The work was carried out during the period from May 2009 through November 2010 at the Multimedia Information and Signal Processing (MISP) Group of the Department of Electronic Systems at Aalborg University. It was funded by the European Union Marie Curie SIGNAL Fellowship, contract no. MEST-CT-2005-021175.

I owe a great deal of gratitude to the people who have helped me on my journey to complete this dissertation. First, my sincere gratitude goes to my supervisor, Prof. Søren Holdt Jensen for giving me the opportunity of pursuing a Ph.D. degree and providing me with the freedom to pursue my interest and for his very valuable advices and supports on my decisions since I joined the MISP section. I would also like to extend my gratitude to my co-supervisor Mads Græsbøll Christensen, for his advice, encouragement and patience. I am so much grateful to him for very useful discussions we had on various topics and his deep understanding in signal and speech processing theorems. I have greatly benefited from our discussions on different topics. I am also so grateful to Zheng-Hua Tan for his supports and encouragement especially for his great contributions during ICASSP 2010.

This thesis is also the result of collaboration with other people who deserve an honorable mention here. I would like to extend my gratitude to Rahim Saeidi, Dr. Tomi Kinnunen and Prof. Pasi Fränti in University of Joensuu in Finland for their remarkable collaborations and invaluable comments on my papers we wrote together and also for making my stay in Finland a very pleasant experience. I would also like to acknowledge my colleagues at MISP, Aalborg University for their significant technical and social contributions to my life.

Last but not the least, I thank my family, my lovely mother and father and my very kind brother for their love and support for the years of my studying.

P. Mowlaee;
Aalborg, Denmark, September, 2010.

Contents

Introduction

Separating mixed signals has long been considered an important and fundamental issue with a wide variety of applications in telecommunications, audio and speech signal processing, and medical signal processing. Audio and speech separation systems find a variety of potential applications including automatic speech recognition (ASR) under adverse noise conditions, and multimedia or music analysis where signals are purposefully mixed from multiple sources.

This introductory overview is organized as follows. In Section 1, we first present two speech separation scenarios, namely multi-channel and single-channel. We define single-channel speech separation (SCSS) problem and its applications. We also compare the separation problem with the so-called speech enhancement problem. In this section, we also briefly present the differences between a machine and the human auditory system while separating a single-channel speech mixture. Finally, we present the masking effect and its role in separating a speech mixture. In Section 2, we provide a review of the previous methods used for SCSS. More specifically, we describe the methodology for each method, and its advantages and drawbacks. In Section 3, we describe the considerations required for a practical SCSS system. We present how a speaker identification module can be integrated to an SCSS module to solve SCSS problem when there is no knowledge of the identities of the underlying speakers in the mixture. In Section 4, we describe how different separation methods are being evaluated in the literature. We also emphasize the importance of two measurements: speech quality and speech intelligibility, known as the two most important issues when quantifying the separation performance. In Section 5, we address our own contributions where we propose two novel model-driven approaches for SCSS. Finally, in Section 6, we sum up the conclusions of this work. As an appendix to this introduction, Section 7 provides some conjectures on the future challenges that await the SCSS community and elaborates on how some of the topics discussed in this thesis could play a role in these challenges.

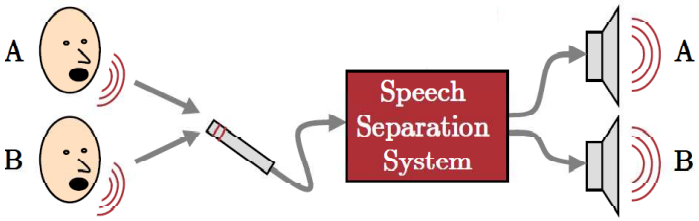

Fig. 1: Showing the system configuration for separating a single-channel recorded speech mixture composed of signals uttered by two speakers, namely speaker A and B. The goal here is to recover the underlying speakers signals according to their one observed mixture.

1 Background

1.1 Multi-Channel versus Single-Channel Speech Separation

In general, in terms of acoustical configuration and the number of microphones and speakers, source separation methods are categorized into the following two groups: over-determined where the number of microphones are more than or at least equal to the number of unknown speakers, and under-determined where the number of microphones are less than the number of unknowns.

Multi-channel speech separation scenario has been carefully studied and remarkable results have already been reported [2, 8]. They use spatial information for separating sources from mixtures either 1) by steering the main-lobe beampattern towards a desired source and placing deep nulls towards the interferers, or 2) by producing time-frequency masks to be applied on the given mixed signals to recover unknown sources [141].

In this thesis, we only focus on separating speech mixtures based on one single observation recorded by one microphone. Figure 1 shows the configuration of a single-channel speech separation problem. Assume that we have two speakers A and B, speaking simultaneously, and their mixture is recorded by a single microphone. The main objective of an ideal speech separation system is to recover the unknown speaker signals (here speakers A and B) accurately, according to only one observed mixture. As an everyday life experience, we all know that humans can easily and accurately separate mixed signals. However, such accurate separation can not be done easily by a machine.

In contrast to the multi-microphone separation scenario, no spatial phase information is provided to solve the separation problem in a single-channel scenario. Therefore, the problem results in an under-determined linear equation, which is too ill-conditioned to be solvable by applying the so-called methods since the mixing matrix is not invertible (equivalently, its condition number is infinity).

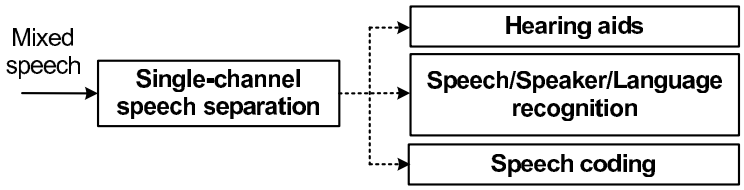

Fig. 2: Block diagram showing how a single-channel speech separation module can be used as a pre-processing stage to enhance the ultimate performance of a target application.

The problem is solved only by imposing *a priori* information, perhaps about the unknown speakers' models in the mixed signal [34, 52, 89, 107, 109, 111, 115, 116].

1.2 Applications for SCSS

In many practical applications like speech coding, speaker recognition, speech recognition and hearing aids, a separation system is often required since the input signals are rarely clean; they are mostly corrupted by some other undesired signals. Consequently, a high quality speech separation system is often required as a pre-processing stage in order to mitigate the effect of interfering signals on the overall performance of different applications. A high quality separation system would play an integral role in offering robustness to the target application. For example, the performance of a speech recognition system may significantly degrade in an adverse noise condition since it is trained for clean training-speech signals. Figure 2 shows a block diagram for an SCSS system, used as a pre-processor for target applications including speech coding, ASR, speaker identification and verification in adverse conditions. Another important application is the hearing aid. It has been well documented that hearing impaired listeners have great difficulty in perceiving speech corrupted with background noise [84]. Although human listeners are often able to attend to individual sources even in dense mixtures, this ability to separate sources from mixtures degrades significantly for hearing impaired listeners [141]. In fact, one of the primary complaints of hearing aid users relates to their poor separation performance in cocktail party-like situations. As a consequence, SCSS methods can potentially be used as tools to mitigate the background or interfering signals that corrupt the desired speaker signal.

It is remarkable to note that the SCSS problem can be viewed as an extreme case of the so-called speech enhancement problem. Therefore, it is intuitive to consider the connection between speech enhancement and separation. In the following section, we explain the differences between SCSS and speech enhancement.

1.3 Comparison to Speech Enhancement

Although there have been recent advances in many speech enhancement methods [37, 38, 46, 47, 59, 75, 120, 130–132], SCSS with high speech quality still remains a challenge for many applications, especially when the interfering signal is another competing speaker signal. Therefore, a reliable speech separation system is often considered as a pre-processor in order to improve the overall perceived speech quality of a certain application in an adverse noise condition.

In case of highly non-stationary noise, the conventional speech enhancement methods including spectral subtraction [14], Wiener filtering [36, 74, 82, 130] and subspace methods [46, 47, 59] are suitable for de-noising a given speech signal corrupted with noise. More advanced single-channel speech enhancement methods are: minimum mean square error (MMSE) amplitude estimator [38, 109, 120, 130, 131], and maximum *a posteriori* (MAP) amplitude estimator [49, 77]. These methods are mostly based on the second order statistics information of the noise signal to be estimated at some noise-only frames in the noisy signal. Such decision-directed methods cannot be used when the interfering signal is another speaker signal because of its fast varying statistics and that make it highly non-stationary [75]. More specifically, in case of a highly non-stationary noise where the noise PSD varies rapidly, it is not possible to apply the state-of-the-art noise estimation methods like minimum statistics [79, 81].

As important parts of speech enhancement methods, subspace-based methods are suggested in [46, 47, 59] where the goal is to find a projection by which the noisy speech signals are de-noised. However, for speech-speech scenario, finding such a projection is a difficult task since the bases of the underlying speakers in the mixture follow a rather similar probability density function (pdf) mostly super-Gaussian as reported in [49, 80]. This brings in the need of having a high quality separation system when speech enhancement is not sufficient for enhancing the desired signals.

1.4 Human Auditory System versus Machine

The human auditory system easily separates speech mixtures into the underlying speaker signals. It is still unknown what cues play a key role in the human ability to perceive different speakers in a speech mixture. However, scientists tried to mimic the procedure followed by the human auditory system for separating mixtures using a machine. In 1990, Bregman presented the auditory scene analysis (ASA) to describe this ability [15, 142] by suggesting strategies similar to the human auditory system.

In chronological order, the first attempts made for solving this problem date back many decades when Bregman [15] and Weintraub [146] used low level perceptual cues for time-frequency segmentation and for finding regions of presence

of one single source. Later, progress was made toward designing computational auditory scene analysis (CASA)-based systems, *machine listening* systems that aim to separate mixed signals by mimicking the same way as human listeners do in everyday life. Cooke and Ellis gave a detailed review of CASA grouping cues in [55]. In section 2.3, we will explain the CASA-based methods and their pros and cons in detail while separating speech mixtures.

1.5 Masking Effect on Separating Speech Mixtures

Auditory masking refers to the perceptual effect that, roughly speaking, a louder sound renders a weaker sound inaudible within a critical band [85]. In other words, it indicates that under certain conditions a stronger signal can mask a weaker one in some critical band (or blocking the target sound by acoustic interference [144]). The masking phenomenon has been studied carefully in psychoacoustics and is known as a highly important factor in the overall separation performance of the current SCSS systems at different signal-to-signal ratio (SSR) levels. The SSR is defined as the average ratio of the target speaker gain to the interfering signal gain. We define target signal as the desired signal in the speech mixture while the masker signal refers to the interfering speaker in the mixed signal. In single-channel speech separation, depending on the mixing level, each speaker can be a target or masker speaker. Therefore, the role of the target and the masker can interchange easily at different frequency bands. When two speech signals mix together, two types of maskings contribute to the separation performance [31]: 1) energetic masking, and 2) informational masking [29]. The listener's ability to hear a target speaker signal in the presence of other interfering sources is limited by both energetic and informational masking. Energetic masking is said to take place when some portions of one or both speakers are masked by another interfering speaker due to sharing their energy in the same critical frequency band. The energies of the sources most likely collide at a time-frequency cell [75, 85]. Energetic masking [132] makes the signal recovery of the speakers rather difficult. It occurs when the target signal overlaps with the interfering source at some time-frequency (T-F) cells. This makes some parts of the target speaker inaudible. Basically, energetic masking is a phenomenon where a weaker signal is dominated by a stronger speaker signal within a critical frequency band [42] rendering the speech signals inaudible at the periphery [29]. On the other hand, informational masking refers to the situation where the listener is unable to distinguish between target and interference when both are audible. In [132], informational masking was modeled as *errors in target segregation* in a speech separation system. Higher-level informational masking occurs when the target and masker signal are both audible, but the listener is unable to distinguish the elements of the target signal from a similar-sounding distracter.

It is very important to study how different separation methods perform with

the variations in SSR level in a mixed signal. This issue is arguable in case of the two types of masking discussed earlier in this section. According to the results reported in [5, 18, 54], the separation quality degrades as energetic masking takes place at some overlapping time-frequency cells.

2 Previous Methods

In general, previous SCSS methods can be categorized in the following groups: independent component analysis (ICA), non-negative matrix factorization (NMF), source-driven and model-driven methods. We now present the drawbacks and merits of each method.

2.1 ICA and ISA

ICA is a computational method for separating a multivariate signal into additive subcomponents assuming the mutual statistical independence of non-Gaussian source signals. It is a special case of blind source separation. In multi-channel processing, it is well known that under certain conditions, it is possible to completely separate the sources from their mixture using the ICA method [56]. The requirements are

- The mixing matrix must be full-rank.
- The number of observations should be larger than or at least equal to the number of unknown sources in the mixture.
- The independence assumption must be valid for the underlying sources in the mixture.
- Must have pre-knowledge regarding the number of sources in the observed mixture.

These factors serve as limiting restrictions for using ICA as an SCSS solution. In [58], as an extension of blind source separation (BSS) to SCSS, a maximum likelihood approach was used to propose a supervised ICA algorithm for separating single-channel speech mixtures. The method worked well on mixtures of speech and music sources, but led to poor performance for mixtures composed of speech signals, especially mixtures composed of same genders [58]. This is because the signal bases largely overlap in both time and frequency domains in a speech mixture. In [95], the authors combined ICA with binary time-frequency masking for separating speech mixtures by using two microphones. The method assumes no prior knowledge on the number of source signals in the mixture and separates up to six mixed speech signals under anechoic conditions. In [94, 96],

the same authors also extended their solution to separate speech signals from a convolutive mixture. They showed that their proposed method is applicable to separate signals in the presence of reverberation.

As another variant of ICA, independent subspace analysis (ISA) simulates an observation with the dimensionality increased from 1 to N. Performing ICA on the transformed signal yields a set of N independent bases, which must then be grouped together into subsets corresponding to the sources existing in the mixture. Davies and James showed that ISA is appropriate when the underlying sources have disjointed spectral supports, which guarantees that the bases are linearly independent [25]. ISA is, in fact, an extension of the ICA use of dynamic components to represent non-stationary signals. ISA extends ICA by identifying independent multi-component source subspaces of an input vector for separating over-complete convolutive mixtures [17]. ISA relaxes the usual ICA requirement on having at least as many sensors as sources. According to [17, 41], ISA is efficient when the underlying signals are stationary in pitch information, like in a drum signal. In ISA, the amount of information required to allow separation of sound sources varies from one signal to another. To overcome this indeterminacy and to improve the robustness of transcription, an extension of ISA was proposed in [40] to include sub-band processing.

2.2 NMF

NMF for Single-Channel Speech Separation

This method is based on decomposing a non-negative matrix representation of a mixed signal, i.e. the magnitude or power STFT, into the product of two low rank, non-negative matrices: $\mathbf{Y} = \mathbf{HW}$ where the columns of $\mathbf{H}$ are the basis vectors which together define the spectral structure of the time-frequency representation of the underlying signals in the mixture. Accordingly, the rows of $\mathbf{W}$, define the temporal structure and the weighings by which the underlying sources are active in the mixture, $\mathbf{Y}$.

The sparse NMF method separates a mixture by mapping a mixed feature vector onto the joint subspaces of the sources and then computes the parts which fall in each subspace [70]. In [69], an analysis was done to determine the circumstances under which only a unique NMF of a matrix exists. Several variants of NMF algorithms have been proposed for learning $\mathbf{H}$ and $\mathbf{W}$ from $\mathbf{Y}$, based on some temporal continuity constraints [138], or sparsity constraints [123, 138], to improve the separation quality. In [9], the authors suggested a sparse non-negative decomposition algorithm to separate audio mixtures recorded by a single microphone. The method generalized the Wiener filtering with locally stationary to non-Gaussian parametric source models. In [121], relying on Gaussian process priors, a general method was presented for employing

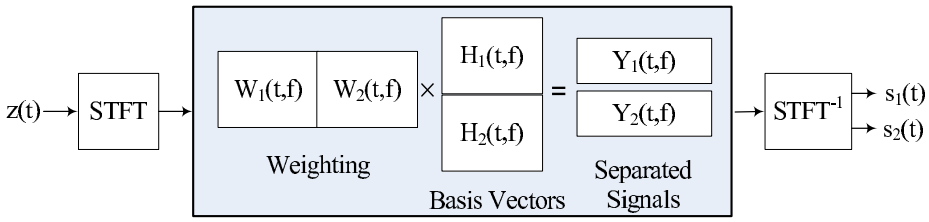

Fig. 3: Sparse factorization of a spectrum using non-negative matrix factorization [122].

the prior knowledge for nonnegative matrix factorization. It was shown that by selecting appropriate prior distributions, the proposed method in [121] achieved better results compared to conventional NMF.

The only requirement in implementing NMF is to specify the number of basis vectors to use. Similar to ISA methods in general, the NMF approach fails to perform well for speech mixtures where the underlying sources have a large amount of overlap. Fig. 3 shows how NMF factorizes a mixture spectrum composed of two speakers into the underlying two-source spectra. As shown in Fig. 3, the spectrum speakers' basis vectors, ($\mathbf{H}_1(t,f)$ and $\mathbf{H}_2(t,f)$), together with their corresponding weights, ($\mathbf{W}_1(t,f)$ and $\mathbf{W}_2(t,f)$), are found from the mixture and then used to recover their time-domain representation using an inverse Fourier transform.

NMF can be applied to SCSS in two ways: supervised and unsupervised. Unsupervised NMF directly decomposes the mixed signal into the underlying signals spectra without any knowledge of the sources. On the other hand, supervised NMF works by synthesizing sources from a learned set of bases of each source in the mixture. Rennie et al. [110] proposed a separation framework that combined ideas from sparse coding and NMF with the model-driven separation approach. In [63], a method was proposed to solve the complication of choosing an optimal number of bases in the training. In conventional NMF, phase information is thrown away and the spectrogram matrix of the mixed signal is factored into the sum of rank-one source spectrograms. In [92], the assumption of excluding phase from factorization and its consequences on separation was studied. More specifically, they proposed an improvement on NMF that followed the true distribution of the spectrogram points of the mixture more closely when the underlying source spectrograms were given *a priori* [92]. In practice, individual recordings of sources are unavailable. To mitigate this, [68] proposed an NMF algorithm for monaural blind source separation based on training the source models using mixed audio recordings.

Similarity between VQ and NMF

Both vector quantization (VQ) and NMF have been used for separating speech mixture in single-channel scenario. It is remarkable to note that although the two methods are different at first sight, the VQ approach to model-driven separation is very similar to the supervised NMF separation approach (model-driven SCSS methods are presented in the following subsections). Both methods employ source-specific constraints in their time-frequency representation as their learned spectral patterns based on clean training data. In the VQ approach to SCSS, the spectral patterns define a probability distribution on the underlying sources in the mixture. However, the NMF is basically a linear decomposition of mixture magnitude spectrum by describing it as a linear combination of the underlying sources' magnitude spectra. In this regard, NMF can be considered as a structured VQ- a multiple stage vector quantizer (MSVQ) [60] or a shape-gain vector quantizer (SGVQ) [118] widely used in speech coding literature.

One important concern while using NMF for separation is the model order of NMF bases. This concern also exists for other statistical models like VQ while being used for training speaker models in terms of the number of bits they use for quantization. It is well-known that employing a sufficiently large number of bases captures relatively complex source signals [147]. In this regard, the authors in [34] studied the effect of using different codebook sizes for short-time Fourier transform features and it was observed that even going towards large model orders does not necessarily achieve a high fidelity.

2.3 Source-driven Methods

CASA

In source-driven methods, the speech signals of individual speakers are extracted from the mixed signal without any *a priori* knowledge about the speakers [108]. CASA is known as the most representative method among the existing source-driven methods [54, 71, 99, 125, 139, 142]. A CASA-based method seeks discriminative features in the observed mixed signal to separate the underlying speech signals. CASA relies on extracting psychoacoustic cues from the given mixed signals. A CASA-based method works in two stages [15, 142]:

- Segmentation: The input mixed signal is decomposed into the time-frequency cells which are dominated by either the target or the masker.
- Grouping: The perceptual cues, namely, common onset/offset, harmonicity and periodicity found in segmentation stage, are grouped to find the specified regions where only one speaker is dominant.

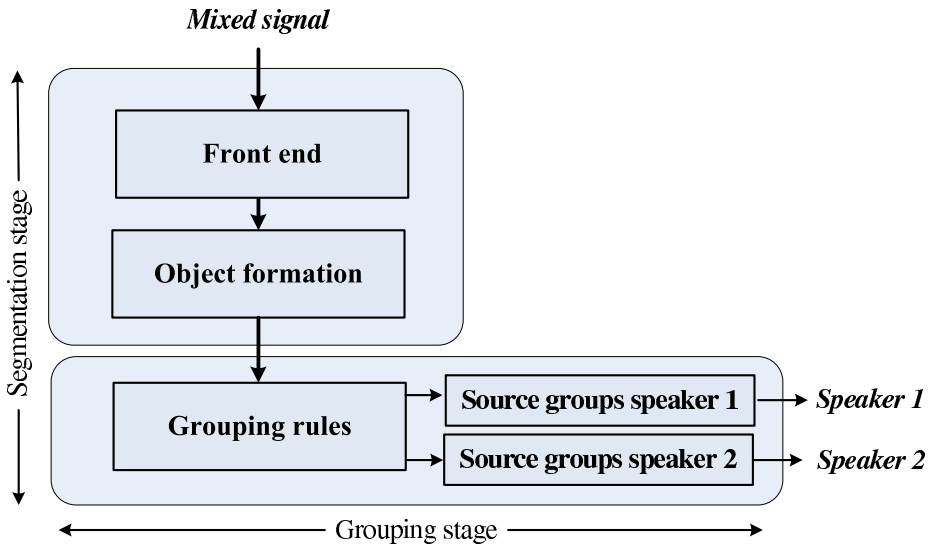

Fig. 4: Block diagram showing different steps in CASA-based method for separating speech mixture recorded by one microphone. The method relies on two steps: the segmentation stage followed by the grouping stage.

Fig. 4 shows the procedure followed in a CASA-based method for SCSS. See [5, 54, 71, 73, 125] for some examples of CASA or other source-driven methods.

Drawbacks

CASA-based methods have some limiting drawbacks as listed below:

- The grouping stage is based on periodicity; hence, it is applicable only to the voiced frames of speech [141, 142]. The unvoiced speech lacks harmonic structure and has weaker energy; therefore, it is more susceptible to interference [53]. However, in [143], an extended version of the CASA approach was proposed for separating speech mixture composed of both voiced and unvoiced sounds.
- The separation rules employed in CASA methods are experimental and heuristic. As a consequence, designing robust CASA-based methods working under different scenarios is often a difficult task. This is because the rules found for one scenario might not necessarily agree with other noise scenarios, and consequently, may lead to significant degradation in the overall separation performance.

- The most important cue used in CASA is the pitch information that is required to be estimated directly from the mixed signal. The CASA-based methods predominantly use the estimated pitch trajectories by applying a multi-pitch estimator. Therefore, the separation performance obtained by a CASA method is dominantly affected by the accuracy of the multi-pitch estimates of the underlying speakers obtained from the mixed signal. It is important to note that current multi-pitch trackers fail to assign pitch contours to the corresponding speakers when the two pitch contours cross each other, or they are unable to detect one of the pitch contours if one speaker's pitch frequency lies within a multiple integer of the other [18, 48, 61, 93, 151]. Their limited multi-pitch estimation accuracy leads to an inferior performance since the pitch estimation accuracy is relatively lost by large gross errors, especially at low SSRs as was reported in [18, 101].

- The output signals produced by the CASA-based methods often lack perceptual quality due to the severe cross-talk problem [54, 99, 108]. In general, applying masks inevitably cause cross-talk and artifacts in the separated signals, as was reported in [54, 89, 107]. The studies in [54, 107] measured the percentage of crosstalk of the CASA-based methods. Similarly, [107] reported the amount of the cross-talk suppression for several source-driven methods.

2.4 Classification Methods

Compared to the CASA methods, an alternative approach is to train a set of classifiers to decide whether or not each time-frequency cell in the mixed signal belongs to a particular underlying source in the mixture. In [124], such an approach was proposed to identify speech-dominated regions in a mixture composed of speech and non-stationary noise. The idea can be used for simple separation tasks. An alternative classifier was proposed in [148], treating the separation problem as a classification one by training a relevance vector machine (RVM). It was demonstrated in [148] that RVM mask performs better than the pitch tracking CASA approach. Although such classification-based source separation approach is suitable for simple denoising tasks, it is important to note that the classifiers are only applicable for separating mixtures composed of sources similar to those on which they were trained. Accordingly, the separation quality is poor if the underlying speaker signals are different from the training data. As a consequence, the method is unable to separate speech mixtures unless the training data consists of the same set of speakers as available in the mixture.

2.5 Model-Driven Methods

A model-driven SCSS method completely relies on *a priori* knowledge about the underlying speakers in the given mixture. The source-specific models are incorporated to learn the constraints on the feature parameters of individual speaker signals [33, 52, 71, 83, 149]. Model-driven methods are well-known machine learning methods like VQ, Gaussian mixture models (GMM) and hidden Markov models (HMM) that are used to learn the restrictive constraints for modeling the speakers in the mixture.

So far, many model-driven separation systems have been introduced. Early approaches to model-driven source separation focused on speech recognition in the presence of noise [44, 136]. The authors in [136] suggested the use of two hidden Markov models where both speech and noise sources are modeled using a HMM. From a historical viewpoint, the pioneer in the model-driven group is the work by Roweis [115, 116], called MAX-VQ. Roweis [115] presented the refiltering approach for SCSS, which estimated the masks spectra based on the VQ models of each of the speakers, independently trained on clean training data from each speaker. The variety of selected feature parameters together with different choices for speaker models lead to different model-driven techniques for SCSS methods. See [34, 52, 83, 89, 107, 109, 111, 115, 116, 123, 137] for some examples of model-driven speech separation methods.

Fig. 5 is an example showing how the magnitude STFT of a mixed signal can be represented by the entries of the two speaker models of the speakers in the mixture. Such speaker models are dictionaries represented in the form of vector quantizer codebooks trained to capture the range of the short-time spectral patterns of a particular voice. In the separation stage, at each time step, a finite search is performed over all the entries from the two codebooks to find the pair of entries (codevectors) that when combined, most likely match the observed spectrum. In the example shown in Fig. 5, the first and second speaker are represented by a codebook composed of three and four prototypes, respectively. Inferred codebook indices are shown for both sources at the right-bottom of Fig. 5, showing how a VQ-based approach can encode the possible parts of the feature space from the observed mixture.

Fig. 6 shows the block diagram of a typical model-driven SCSS approach based on source-specific models. We now explain the key parts of a model-driven SCSS system, namely, feature selection, mixture estimator, speaker modeling and reconstruction stage.

Feature Selection

In this stage, features are extracted from the signals. Common features already used for SCSS are: time waveform [58], log-spectrum [106, 116], discrete cosine transform (DCT) [7] and auto-regressive coefficients [12]. The conventionally

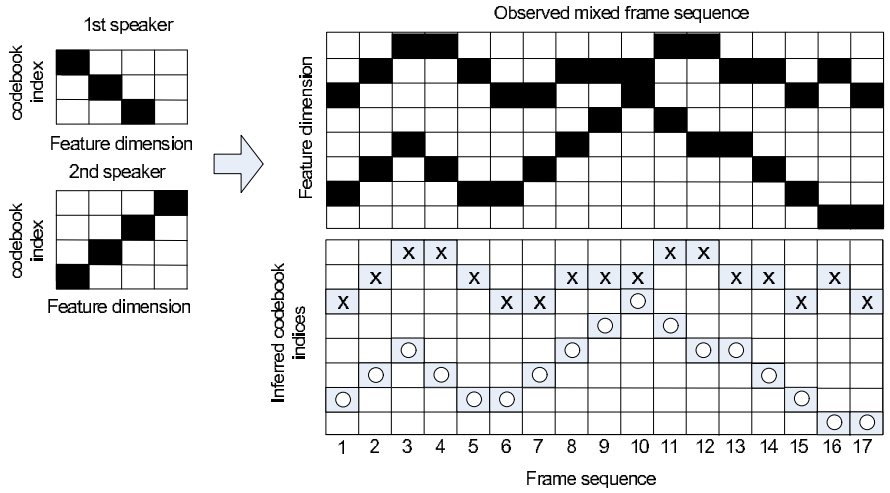

Fig. 5: An example showing how two speaker codebooks of the underlying sources (shown on the left) can define a given mixture [33].

used features in the CASA-based methods are divided into the following groups: directly in time domain [133, 150], gammatone frequency cepstral coefficients (GFCC) [71], and pitch information together with common onset and offset of the speech signals [54, 71, 73, 133]. Independent of the separation scheme used, the features used for SCSS are required to fulfill the following constraints:

- They are required to maintain a straightforward relationship between the features of the mixed signal and those in the underlying sources.
- In the discussion on the curse of dimensionality in [30], it was shown that if the goal is to optimize a function over a continuous product domain of a few dozen variables by exhaustively searching a discrete search space, we are faced with the problem of making millions of evaluations of the function. Hence, the dimensionality of feature parameters should be kept low to reduce computational complexity in the search process. More specifically, it is important to find features with sparse nature so that the signals can be represented with the minimum number of features without any noticeable reduction in the perceived signal quality.
- The selected feature type along with the statistical model determines the separation upper-bound performance of an SCSS system [36]. In [36], it was concluded that the ultimate quality of a model-based speech enhancement system is upper-bounded by the performance of the coder used. Similarly, we define the upper-bound performance of an SCSS method as the

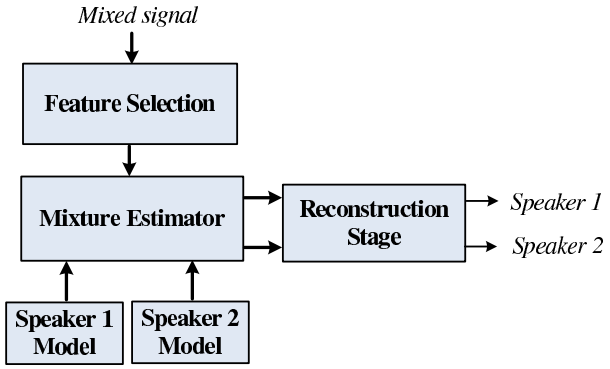

Fig. 6: Block diagram for a typical model-driven speech separation system composed of four stages: feature selection, speaker models, mixture estimator and re-synthesize stage.

highest achievable separation performance offered by a model-driven SCSS method. The upper-bound performance is obtained by the quantizer used when the mixture estimation in SCSS is ideal, i.e., without any error in the mixture estimation stage. In order to reach this upper-bound separation performance, the selected feature for SCSS is required to give a high quantization performance as reported in [34, 89]. Thus, both NMF and VQ-based SCSS methods require employing high fidelity source-specific models to achieve an acceptable high separation quality.

Speaker Models

Different machine learning methods have been employed for producing speaker models. The methods include: VQ [34, 71, 89, 107, 115], GMM [10, 106, 109, 111] and HMM [5, 52, 116, 149]. The speaker models are trained in the training stage and then used in the test stage to separate the speaker signals from their mixture. The trained models are used for estimating the best codewords in each of the speaker models, which when combined, best describe the observed mixture frame based on a certain optimization criterion. The estimated codevectors selected from the speaker models are then sent to the reconstruction stage.

Mixture Estimator

In a model-driven speech separation approach, joint best codevectors, each selected from one speaker model, need to be estimated, which when combined, best fit the observed mixture spectrum according to a given error criterion. This is accomplished by employing a mixture estimator that is arguably the key part of a model-driven SCSS system. Previous mixture estimators were carried out

using maximum likelihood [58], maximum *a posteriori* (MAP) [103, 104] and MMSE [106, 109] estimators.

In SCSS problem the relationship between the mixture and the underlying speech signals is given by

$$y(n) = x_1(n) + x_2(n), \quad n \in [0, N-1]. \tag{1}$$

Taking the Fourier transform from both sides in (1), we have

$$|Y(k)|\angle Y(k) = |X_1(k)|\angle X_1(k) + |X_2(k)|\angle X_2(k) + 2|X_1(k)||X_2(k)|\cos\theta(k), \tag{2}$$

where $k \in [1, K]$ indicates the kth frequency bin the spectrum representation, n is the time sample index, N is the window length in samples, and $\theta(k) = \phi_1(k) - \phi_2(k)$ is the angle between the speaker signals spectra. The goal in SCSS is to estimate both $\{x_1(n)\}_{n=0}^{N-1}$ and $\{x_2(n)\}_{n=0}^{N-1}$ based on the mixture at each frame.

Previous separation methods used either the *max-model* [115] or the *Algonquin model* [67] as their mixture estimator. The max-model suggests approximating the mixture logarithm power spectrum by the element-wise maximum of log-spectra of underlying speakers in the mixture. A high resolution signal reconstruction method was proposed in [67] highlighting the advantage of statistical model-driven separation method compared to other methods. The derived estimator was called *Algonquin*. In [52], both *Algonquin* and max-models were studied and compared. It was demonstrated in [52] that *Algonquin* performs slightly better than max-model.

Reconstruction Stage

According to the schematic block diagram shown in Fig. 6, the last part of a model-driven SCSS system is the reconstruction stage. The reconstruction stage works based on the indices of the speaker models found in the mixture estimation stage. Reconstruction of the separated signals can be done in two ways, 1) by employing an overlap and add procedure [52, 89, 149], and 2) by producing masks [71, 83]. We now present each method in detail.

The first group use overlap and add procedure for reconstructing the output separated signals. In this group, the joint optimal codevectors found in the mixture estimation stage are directly used to re-synthesize the separated speech signals for each speaker in the output stage. Then, an overlap and add procedure [98] is used to re-synthesize the separated signals after the correct states are found in the separation stage. Examples for the signal reconstruction-based SCSS methods are [52, 71, 83, 89, 149].

The second group reconstructing the separated signals use the mask methods targeted to produce masks to refilter the mixed signal in order to recover the

underlying speakers in the mixture. The mask methods produce masks based on the joint states selected from the speaker models of each speaker. These produced masks are then applied to the mixed signal to provide the separated signals. The masks to be applied are either binary [54, 71, 107, 115] or Wiener [10, 83, 106, 109].

Drawbacks of Model-Driven SCSS Methods

Model-driven methods have the following listed drawbacks:

- The computational complexity of MAX-VQ is high due to the factorial search. For instance, in [116], a separate HMM was applied for each speaker and a huge state space of 8,000 was required in order to carefully capture every possible transition between the codeword entries in each speaker codebook. Though using HMM enables the modeling of correlated speaker signals [36], it leads to a significantly more complex mixture estimation approach. In short, current model-driven methods are highly computationally complex and still lag far behind in being implemented in real-time.

- It is important to note that performance degradation in a model-driven SCSS can also come from the errors introduced in the mixture estimation stage that is targeted to find the two best codewords, which when combined, best describe the given mixture frames (see Fig. 6). Difficulty arises while mapping vectors of mixed signals onto states of speaker models. Such decoding errors result in wrong association of the codevectors with the max-model mixture estimator leading to the selection of poorly filtered signal vectors [71, 107]. Selecting incorrect states from the speaker models consequently degrades the perceptual quality of the separated signals.

- MAX-VQ tries to produce two masks based on the estimated VQ states. According to the results reported in [107] and [71, 89], using such masks provides re-synthesized signals often corrupted with undesirable cross-talk caused by another interfering speaker in the mixture. Therefore, the resulting signal-to-noise ratio (SNR) reported for the output signals are low [107, 116].

- Depending on having *a priori* knowledge of the identities of the underlying speakers in the mixture, a model-driven SCSS can be implemented in two ways: speaker-dependent or speaker-independent. Fig. 8 shows the two scenarios. It has been demonstrated that a speaker-dependent model-driven SCSS provides good separation performance [116]. Real practical scenarios, however, include separating speech mixtures when no *a priori* information is available for speakers in the mixture. Therefore, current

> SCSS methods suffer from the impractical assumption of *a priori* knowledge of speaker identities in the given mixture. In general, model-driven separation systems mostly suffer from the disadvantage of being specific to the sources present in the mixture. To overcome this difficulty, a speaker identification stage was integrated as a pre-processor to the single-channel separation module [52].

According to [127], the model-driven approach was expected to perform better than the pitch-based methods indicating that using only the pitch information showed limited discrimination for sequential grouping. This presents the idea that the integration of pitch and spectral envelope in [107] may not be the most efficient solution to recover both signals because multi-pitch estimation from a mixture at low SSRs is difficult [5, 18, 20, 54]. This motivates us to present pitch-independent separation methods in this thesis.

2.6 Binary Mask

Based on the masking phenomenon discussed earlier, the idea of binary mask was proposed in [54] where a T-F unit was assigned 1 if the target energy exceeded the interference energy and 0 otherwise. Binary masks have been widely used in CASA literature as output representations to label the origins of the mixed speech. In [71], binary masks were inferred according to the two speakers' VQ codebooks and then incorporated to produce binary masks to separate the signals in a CASA-based framework.

It is important to note that the term masking often referred to in binary masking literature is different from the one commonly used in psychoacoustics where it means blocking the target speaker signal by an acoustic interfering speaker. According to [144], in binary masking literature, a binary mask applies a pattern of binary gains onto the mixed signal in order to recover the original signals in the mixture. According to [144], such a binary pattern roughly reflects the formant structure and encodes the outline of energy variation in the time-frequency representation for the target speech relative to the interfering noise (here, another speaker).

As a limitation, the mask methods are not yet able to perform well especially at low SSR scenarios for recovering both target and masked speakers in the mixture. In [105], it was demonstrated that as SSR increased above 6 dB, the target speech signal was more easily recoverable since it masked the interference signal. The authors in [105] also reported that the recovered interference signals were not perceptually intelligible and consequently, the SCSS method in [105] was not able to recover the interference speech signal at high SSR levels.

Ideal Binary Mask

At each mixture frame, the ideal binary mask is defined as the mask produced by keeping all time-frequency cells where the target speaker dominates the interfering one and removing those where the target is masked by the interfering speaker. Theoretically speaking, an ideal binary mask gives the ceiling performance for all binary masks and is often considered as an important concept in studying the performance of the SCSS systems. More specifically, it is shown in [73] that the ideal binary mask leads to the optimal SNR performance when employing orthogonal T-F decomposition and rectangular windowing. The separation performance offered by the ideal binary mask is often used as the ground truth of target speech. In an ideal binary mask, it is assumed that we have *a priori* knowledge on the spectra of the underlying sources in the observed mixed signal [13, 16, 140, 144]. By considering this unrealistic assumption, it is possible to study the highest achievable separation performance for a binary mask approach. Fig. 7 shows the result of a simple setup showing how an ideal binary mask can separate the underlying signals from their mixture. In Fig. 7, time-domain representations are shown on the left panels while the time-frequency representations are shown on the right panels for original, mixture and separated speaker signals respectively (from top panel to bottom panel).

The ideal binary mask has also been suggested as the ultimate goal of a CASA-based separation method [16, 142]. It was shown that an ideal binary mask brings substantial intelligibility gains in speech separation [16] and speech perception [144]. In [91], authors studied the factors influencing the intelligibility of ideal binary-masked speech. The authors in [65] measured the speech intelligibility of ideal binary-masked noisy speech on a group of normal hearing individuals across different masker types as well as different mixing levels. The authors in [71] proposed a separation system that used the *a priori* masks to re-synthesize the separated speaker signals [71]. According to [21], using the ideal binary mask for ASR yields excellent recognition performance. More recent studies proposed strategies to estimate the ideal binary mask. In [13], the authors suggested a method for correcting errors in the estimated ideal binary mask by employing an HMM to model the error-free target binary mask and then using the Viterbi algorithm to find the most likely error-free binary mask for the target signal. Their results show that the proposed error correction approach reduces the noise energy as well as corrects some errors in the binary mask produced for the target signal.

2.7 Wiener Filtering

Wiener filtering is a classical speech enhancement method that relies on the MMSE estimation to restore the underlying clean signals [36, 82]. Further, the

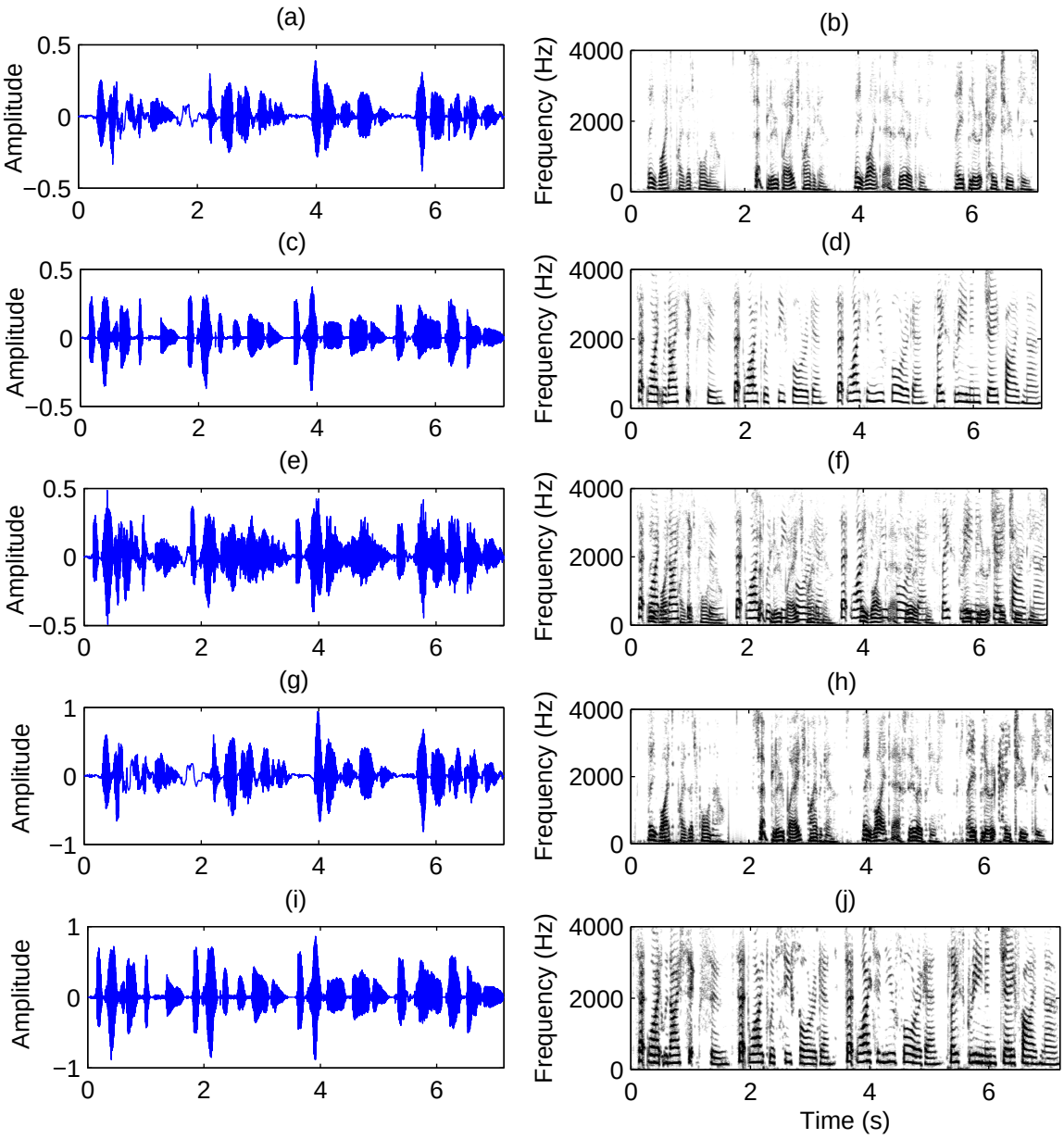

Fig. 7: An example showing how ideal binary mask can be used for separating single-channel speech mixture: the time signal for (a) speaker one original, (c) speaker two original, (e) mixture, (g) separated first speaker signal using ideal binary mask, (i) separated second speaker signal using ideal binary mask. The time-frequency representation for (b) speaker one original, (d) speaker two original, (f) mixture, (h) separated first speaker signal using ideal binary mask, (j) separated second speaker signal using ideal binary mask.

Wiener filtering has been used for the reconstruction stage of the SCSS [9, 11, 83]. In [83], the authors used the most likely codevectors selected from the speakers' codebooks to produce Wiener filters to segregate the speaker signals from the mixture.

The estimated source spectrum using the Wiener filtering is given by

$$|\hat{X}_1(k)| = \frac{S_{x_1}(k)}{S_{x_1}(k) + S_{x_2}(k)} |Y(k)| \tag{3}$$

$|Y(k)|$ and $|\hat{X}_1(k)|$ denote the amplitude spectrum for speech mixture and the first speaker signal, respectively, at the kth frequency bin speech, $S_z(k)$ denotes the power spectrum of $z \in \{y, x_1, x_2\}$, denoting speaker one, two and mixture, respectively. A similar expression is achieved for the estimated spectrum of the second speaker, $|\hat{X}_2(k)|$ by replacing $S_{x_1}(k)$ with $S_{x_2}(k)$. The classical Wiener filter commonly used in speech enhancement can be interpreted as a T-F mask (or equivalently, a T-F filter) where each T-F cell of the mask represents the ratio of the target speaker energy to the energy of the given mixture calculated in that specific T-F cell.

3 Considerations for Practical Separation Systems

In many model-driven SCSS methods, it is generally assumed that the speaker identities as well as the mixing level (SSR) are known *a priori*. Such simplifying assumptions are useful while studying a new SCSS method, but are very restrictive and impractical. In order to approach a more practical SCSS system, one needs to tackle these issues, i.e., estimate the speaker identities and the SSR level under which the underlying signals have been mixed together. Here, we first explain how to deal with a speaker-independent scenario where the speaker identities in the mixture are unknown. More specifically, we explain how to employ a speaker identification module as a pre-processor for SCSS to enable us to separate speech mixtures in a speaker-independent scenario (see Fig.8.b). After that we move on to the other practical issue called the gain estimation problem, indicating that SCSS methods require information regarding the mixing level of the underlying signals in the observed mixture.

3.1 Speaker Identification

Speaker identification is the task of recognizing the identity of a speaker based on an observed speech signal which can possibly be corrupted by noise or other interfering speakers [64]. A speaker identification or verification system helps to characterize the speakers in many applications, namely, telephone banking,

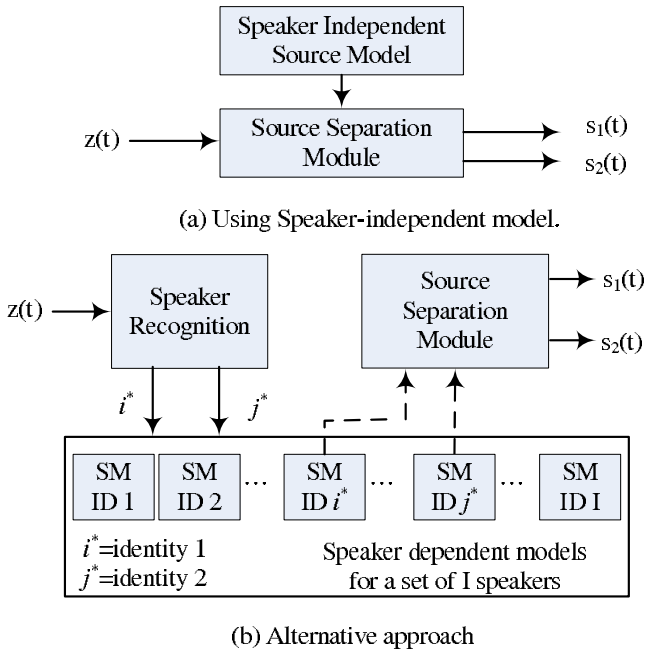

Fig. 8: (a) Using speaker-independent source model (SM), and (b) alternative approach: first estimating speaker identities (ID) and then choose the speaker model to change the separation problem from speaker-independent to speaker-dependent one.

voice dialing system, voice mail and security control. Fig. 9 shows the block diagram of a speaker identification system. As shown in Fig. 9, in a speaker recognition module, M speaker models work in parallel; the model achieving the maximum likelihood score will be selected as winner and identify the correct speaker identity. For recognizing an unknown speaker, the following steps are required to be taken in an orderly manner:

- Extracting features from the speakers.
- Forming speaker models for the extracted features.
- Decision taking based on some criterion like maximum likelihood or MAP to find the identity of the unknown speaker.

A GMM-based framework is often used in speaker recognition applications [113]. It is used as a reference for evaluating the effectiveness of a new algorithm [119]. The ML criterion is commonly used for training GMMs. In state-of-the-art systems, speaker-dependent GMMs are derived from a speaker-independent universal background model (UBM) by adapting the UBM components with

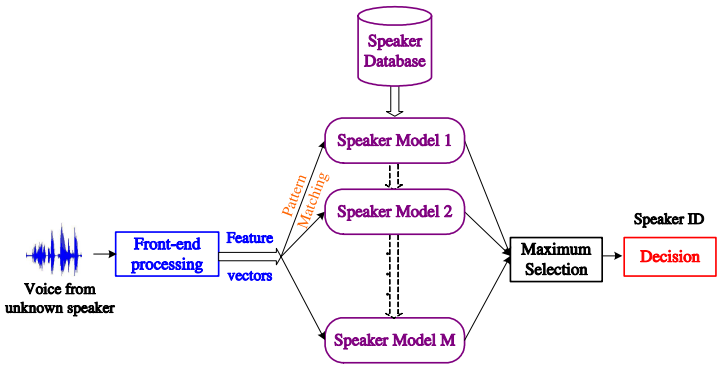

Fig. 9: Block diagram for a speaker recognition system.

maximum *a posteriori* (MAP) adaptation using the speakers personal training data [112]. This method constructs a natural association between the UBM and the speaker models. For each UBM Gaussian component there is a corresponding adapted component in the speakers GMM. In the verification phase, each test vector is scored against all UBM Gaussian components, and a small number of the best-scoring components in the corresponding speaker-dependent adapted GMM are chosen. The decision score is computed as the log-likelihood ratio (LLR) of the speaker GMM and the UBM scores.

Typical speaker identification systems consist of a short-term spectral feature extractor (front-end) and a pattern matching module (back-end). In traditional SID, the basic assumption is that only one target speaker exists in the given signal whereas in the *co-channel* SID, the task is to identify two speakers in a mixture. Distinct from the so-called *summed channel* speaker recognition task [135] where only one speaker is talking most of the time, in the co-channel SID problem, both speakers talk simultaneously, which makes the problem much more challenging. Although research on co-channel speaker identification has been going on for more than a decade [86], the problem remains largely unsolved.

Difficulties in Multi-talker Speaker Recognition

A major drawback of the current speaker recognition systems is their significant degradation in their performance under noisy conditions. This is mainly due to the mismatch existing between the likelihood calculation in the training and the test stages [113].

Several methods have been suggested for solving the problem of multi-talker speaker recognition in adverse conditions. For instance, [32] proposed using spectral subtraction to improve speaker recognition performance. However,

the method was suitable only for stationary noise scenarios. RASTA filtering [50] and cepstral mean normalization [43] have already been used to improve the speaker recognition performance. Recently, [128] proposed to improve the speaker recognition accuracy by employing CASA together with dynamic auditory features. They showed that using such advanced auditory features results in higher speaker identification accuracy compared to the conventional MFCC features [128]. Auditory features' uncertainties were used for robust speaker identification [126] showing that such features offer higher speaker identification accuracy compared to the conventionally used MFCC features. In [3], auxiliary features, namely pitch frequency and voicing state together with a Bayesian network were used showing a reduction in the influence of background noise as well as in transmission channel distortion.

Joint speech enhancement and speaker identification was proposed in [78]. They employed a Bayesian approach where the speech features were modeled using a mixture of Gaussians priors. A Gibbs sampler was used to estimate the speech source and the identity of the speaker.

Using Speaker Identification for SCSS

Interaction between a speaker identification module and a speech separation module can be implemented in two ways: 1) identification followed by separation, and 2) using separation module as a preprocessor for speaker identification. The two possible ways of integrating a speaker identification and speech separation module are shown in Fig. 10. The former approach is suitable for separating mixtures in a speaker-independent scenario as the SID module determines the most likely speakers in a speech mixture. This valuable information changes the speaker-independent separation problem to a speaker-dependent separation one, which is obviously more accurate. The method, however, requires a relatively accurate SID system since any possible error from the SID module propagates and consequently, results in performance degradation in the separation stage. In the latter approach, a separation system is employed as a pre-processing stage and helps to narrow down the speaker candidates in the mixture to be estimated by the SID module. However, due to imperfect separation, the SID performance might severely degrade. Further, the initial part of this configuration requires a separation to be implemented based on speaker-independent models that are known to show a relatively lower separation performance compared to the speaker-dependent one as was reported in [108]. If the separated signals are good enough for the SID module, then in the close loop, correct identities are given by the SID module and, consequently, the SCSS system changes into the speaker-dependent module based on the speaker models selected according to the estimated speaker identities. However, it is important to note that there is a high risk of choosing incorrect speaker models due to the possible errors in the

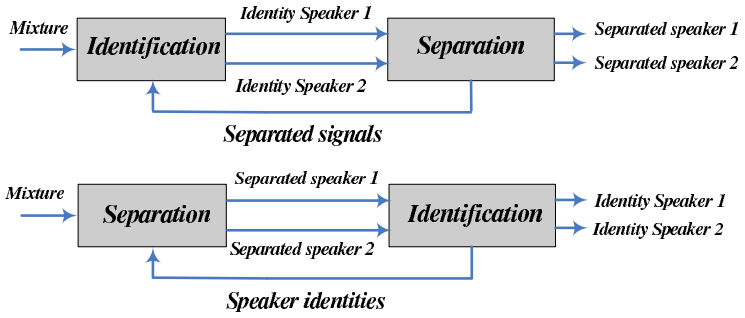

Fig. 10: Block diagram showing the two possible ways of integrating a speaker identification and speech separation module.

SID module, which consequently leads to a poor separation performance and poor SID estimates of the SID block. Based on these explanations, it is clear that the first module (see Fig.10(top)) is capable of providing a more reliable system.

Most of the current SCSS systems employ a model-based SID module, known as *Iroquois* [52], to identify the speakers in a mixed signal. Recognition accuracy as high as 98% on the speech separation corpus has been reported for *Iroquois* [22], which makes it a viable choice for use in SCSS systems [149]. In the *Iroquois* system, a short-list of the most likely speakers is produced based on the frames of the mixed signal that are dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* to find the SSR and the two speakers' identities with an exhaustive search on codebooks created for speech synthesis [52].

In [149], a single-channel source separation system was proposed based on the speaker adaptation principle that worked on the adaptation of a generic speech model to match each of the underlying speakers in the mixed signal. As another example, the system proposed in [71] identified the speakers based on GMM and employed a pitch-dependent method to re-synthesize the target speaker signal. In [52], the *Iroquois* system was used for determining which frames belonged to single-talk and mixture regions by choosing the most likely speakers based on the frames dominated by a single speaker.

3.2 Gain Estimation

Most of the previous model-driven SCSS methods [7, 67, 109, 116, 123] assume that the test speech files are recorded in conditions similar to that of the training phase. This unrealistic assumption highly limits the effectiveness of these techniques in real life since the observed speech files can be mixed at an energy ratio

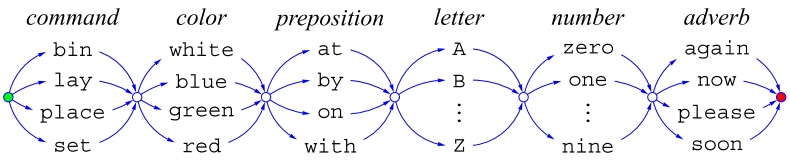

Fig. 11: Task grammar defined in the speech separation challenge [22]. Note that the letter W is excluded. An example sentence would be lay white with G six please.

different from that of the training speech files. The authors in [103] proposed a technique that estimated the gain associated with individual sources in the mixed signal. Their proposed gain-adapted MMSE SSR estimation approach estimated sources under different SSRs and worked by searching the whole space formed by the underlying speaker models, which is arguably too complicated.

4 Performance Evaluation for Separation Systems

Both objective and subjective measures can be used for evaluating the separation performance of an SCSS method. In the following, we explain the most important factors that play a role while assessing the separation performance of an SCSS system.

4.1 The Speech Separation Challenge

Database

For evaluating the separation performance of an SCSS method, it is common to use the comprehensive database GRID corpus provided in [23], which consists of simple sentences drawn from the grammar described in Fig. 11. The database consists of 34 speakers (18 male and 16 female speakers), each containing 500 utterances. The single-microphone recoded speech signals are sampled at a sampling rate of 25 kHz.

The database provided in [23] was targeted to measure the overall recognition performance of the separation systems submitted by participants for speech separation challenge. The clean training data were used for training speaker models. The test data was a mixture of the target and the masker speakers mixed at six SSR levels of {-9,-6,-3,0,3,6} dB and a clean signal for the target speaker in the mixture. For each six-test sets of two-talker signals, 600 utterances were provided among which 200 were for the same gender, 179 for different gender, and 221 for the same talker where the target and masker signals are from the same speaker.

The Task in Speech Separation Challenge

The recognition task is to identify the letter and number in the sentence spoken by the talker who said "white". The challenge contains two-talker speech recognition task. The main task in the speech separation and recognition challenge is to recognize the target speaker speech in the presence of another, masking speaker using a single microphone [22, 52]. The clean signal for the target speaker is available and is used to report the speech recognition baseline.

Participants in the Speech Separation Challenge

Table 1 lists all participants in the speech separation challenge in [22]. For each method, a short description is also given.

Table 1: List of all participants in the 2006 Speech Separation Challenge.

Excerpt	Descriptions
Human	N/A
Deshmukh [27]	Phase opponency
Every [39]	Pitch tracking
Runquiang [117]	CASA
Schmidt [123]	Source models, NMF
Virtanen [137]	Source models, factorial HMM (FHMM)
Ming [83]	HMM & Wiener filtering
Barker [5]	Combination of CASA & speech fragment decoder [6]
Kristjansson [52]	Grammatical constrains, source models, FHMM
Srinivasan [125]	CASA
Weiss [149]	Eigen voice models [149], source models, FHMM

4.2 Speech Quality versus Speech Intelligibility

It is still an open issue to design an SCSS approach that can improve both the quality and intelligibility of the separated signals. The problem refers back to the question on what is the optimal criterion for solving the SCSS problem. Equivalently, the fundamental question is how one can enhance both speech quality and intelligibility. As stated earlier, [73] showed that an ideal binary mask is optimal in terms of SNR but under some constraints. However, it is clear that SNR measure (or in general a l_2-norm optimality criterion) may not reflect the exact statistical optimality. In addition, an ideal binary mask is only achieved under crucial *a priori* assumption of having knowledge of the underlying speakers' spectra. Without such information, the estimated binary mask may have errors due to incorrectly changing zeros to one or vice-versa, which will

lead to considerable speech distortion in the separated output signal. Previous studies have shown that an ideal binary mask largely contributes to enhancing the intelligibility, but not the quality of the separated signals [142]. In fact, binary masks degrade the perceived speech quality of the separated signals. On the other hand, many speech enhancement methods [37, 38, 46, 47, 59, 75, 120, 130–132] have been proposed for enhancing the quality of the noise corrupted signals. As a consequence, it is a reasonable question to ask in which terms an SCSS method should be assessed and how well such a measure correlates with the subjective listening experiments.

4.3 Objective and Subjective Measures

Table 2 shows a list of previous measures used for quality assessment of speech enhancement, speech coding and single-channel speech separation applications. The table also shows the correlation coefficient of each measure with the subjective listening results [75]. Most of the previous methods reported either SNR or, at some points, segmental SNR (SSNR) measurements. As an example of the limitation in the SNR measurement, according to the results reported in [102], it was observed that the SNR results do not necessarily reflect the merits of HMM over VQ, confirming that SNR is not the most appropriate criterion for evaluating SCSS performance. This is in line with the conclusion in [72] that the quality of separated speech is not directly related to its SNR.

The SNR-based measures are simple measures and have been widely used due to their simplicity. They have only indirect relation to the perceived signal quality. The weighted spectral slope (WSS) is mainly related to differences in formant locations and provides a reliable measurement in assessing the signal reconstruction performance. Both LLR [97] and Itakura-Saito (IS) [97] have been widely used to assess the performance of speech coding methods. As objective measurements, speaker identification measures together with word error rate for speech recognition were both used in [52] to quantify the overall performances of different separation methods when being used for speech recognition task in the speech separation challenge [22]. The authors in [72] suggested combining CASA with objective quality assessment of speech. They showed that the proposed approach achieved substantially good subjective perceived speech quality of separated speech. Finally, [4] developed a standard for measuring the speech intelligibility for noisy speech mixtures.

In recent studies [28, 152], it has been observed that SNR-based measures are somehow illusive while evaluating speech enhancement methods. Moreover, in [49, 152] the use of SNR measures was discouraged as it correlates poorly with subjective rates and is very sensitive to both the experimental conditions and also because of its very common artifacts in BSS-like fractional delays between the signals to evaluate. As a result, the SNR-based measures (l_2-norm mea-

Table 2: List of quality measures used for speech enhancement, speech coding, and speech separation. The table shows the correlation coefficient of each measure with the subjective listening results [75].

Measure	Category
SNR & SSNR [26, 45, 97](ρ=0.31)	Speech separation and enhancement
CEP [45](ρ=0.56)	Cepstral distance measure
WSS [66] (ρ=0.53)	Speech enhancement and coding
LLR [97] (ρ=0.63)	Speech enhancement and coding
IS [57, 97](ρ=0.45)	Speech coding, separation and enhancement
PESQ [114] (ρ=0.65)	Speech enhancement
cSII [62](ρ=0.94)	Coherence SII best for additive-noise data
DAU [24](ρ=0.86)	Best for additive-noise data
Composite [152](ρ=0.85)	Speech separation and enhancement
Subjective tests [1]	Speech coding, separation and enhancement
Speech intelligibility [4]	Speech separation and enhancement

sures) are not the most appropriate criterion for evaluating SCSS performance. Attempts have been made to find some reliable and efficient metrics to evaluate the performance of a separation system [75].

Finding an effective measure to assess the crosstalk and separability of a given speech separation approach, has historically been considered a difficult task. The difficulty is mainly due to the differences in the testing methodology and lack of information on how the human brain assesses the speech signal. On the other hand, developing an objective speech quality measure that correlates highly with subjective speech measures has been viewed with much importance as subjective tests are generally expensive and time-consuming. In this aspect, a good objective speech quality measure would be a valuable assessment tool for developing new speech enhancement or separation algorithms, as the previous separation methods in a single-channel scenario have mostly reported their results in terms of SNR or word error rate (WER) measures.

5 Summary of Contributions

The main contributions of this work include proposing new approaches for solving the SCSS problem and suggesting the use of speaker identification module to solve speaker-dependency in model-driven SCSS systems. Additionally, we also present a solution to improve the speaker identification performance when applied to a speech mixture. Papers A through D deal with proposing new approaches in sinusoidal domain to improve the separation performance in SCSS. In these papers, we suggest the separation of the input mixed signal by apply-

ing a criterion based on sinusoidal parameters of the mixture along with those selected from the underlying speakers' codebooks (we use one sinusoidal coder for modeling each speaker signal in the mixture). Paper A proposes a sinusoidal mixture estimator based on unconstrained sinusoidal parameters estimated from the mixture and selected from the codebooks trained for the speakers in the mixture. Paper B derives sinusoidal binary mask and sinusoidal Wiener mask for solving the SCSS problem. The separation results obtained by the sinusoidal masks show improvement over the commonly used STFT-based masks like binary mask and Wiener filtering. Paper C presents a novel SCSS system where unconstrained sinusoidal parameters of the signals act as features, and sinusoidal coders act as speaker models suggesting the use of a sinusoidal mixture estimator as its mixture estimation stage. Paper D proposes a joint speaker identification and speech separation system in a closed loop, which successfully improves the separation quality without *a prior* knowledge of the identities of the speakers in the mixture. In Paper E, a novel speaker identification approach is proposed which works independent of the signal-to-signal ratio under which the underlying speakers are mixed together. In Paper F, we present a solution for double-talk detection based on a single-channel recorded speech mixture composed of two speaker signals. The proposed approach determines the number of speakers in a mixture composed of two speakers. Paper G proposes to combine a double-talk detection module as a pre-processing stage for speaker identification back-end. Finally, in Paper H, we present a full system for the speech separation challenge. The suggested system is composed of joint speaker identification, SSR estimation and speech separation, and is targeted at separating single-channel mixed signals into the underlying speakers. We will now go through the contributions of the individual papers that constitute the main body of this thesis.

Paper A

In this paper, a novel mixture estimator has been proposed and derived based on unconstrained sinusoidal parameters to improve the speech separation performance. The method is independent of pitch estimates and offers a new approach for single-channel speech separation, where pitch estimation is sometimes difficult because of the energetic masking that occurs in time-frequency cells in the mixture at different SSR levels. The proposed mixture estimator finds the optimal codevectors, one selected from each speaker codebook, which when combined, best describe the observed mixed frame. A variation of the sinusoidal coder proposed in [87] was used to model the underlying speakers in the mixture. Through several experiments it was observed that the proposed method achieved a higher score compared to the mask methods of MAX-VQ, Wiener filtering and the STFT VQ-based separation systems especially at low SSR levels. As the SSR increased, the proposed method asymptoted its separation upper bound perfor-

mance where it was assumed that the optimal indices were available *a priori*. It was also observed that the perceived speech quality of the proposed system was the highest. Finally, listening tests showed that the proposed method achieved lower cross-talk and was mostly preferred by the listeners.

Paper B

In this paper, we propose to use sinusoidal masks in lieu of the commonly used STFT-based mask methods. We generalize both the mask methods, binary and Wiener filtering in the STFT domain into the sinusoidal space. Instead of noise distortion, we suggest a trade-off between minimizing the speech distortion of the target signal and keeping the *cross-talk* of the other speaker below a given tolerance threshold. This is well-justified as in speech separation we are required to have no trace of *cross-talk* from the other speaker. We show that in a separation scenario, employing the new masks in sinusoidal space results in an improved separation performance compared to the STFT-based masks. According to our experiments, the masks defined in the sinusoidal domain, including both binary and Wiener masks, improve the separation quality as compared to the STFT masks. Finally, it was demonstrated that the proposed sinusoidal Wiener mask coincides with the so-called *Ephraim and Malah* noise suppression rule [37].

Paper C

In this paper, we present a model-driven separation system based on sinusoidal parameters. The proposed separation system employs sinusoidal coders as its speaker models for modeling each of the underlying speakers in the mixture. We suggest using a sinusoidal mixture estimator to find the optimal codevectors from each speaker, which when combined, best fit the observed mixed signal. It is shown that the proposed method results in an improved separation performance compared to source-driven and STFT-based methods in both speaker-dependent and speaker-independent scenarios. It was also shown that the proposed separation approach outperforms the source-driven and the fusion of both source and model-driven approaches proposed in [108] for separating single-channel mixtures. The proposed method cut down the computational complexity significantly by replacing the high dimension STFT features with sinusoidal parameters. Furthermore, in the mixture estimation stage, minimization is only calculated at the sinusoidal peaks of the mixture compared to the full-band calculation in other methods.

To assess the perceived speech quality of the separated output signals obtained by different methods, we conducted a subjective listening test using the multi-stimulus test with hidden reference and anchors (MUSHRA test) as described in ITU-R BS.1534-1 [1]. The MUSHRA test is a double blind test for the subjective assessment of intermediate quality level benefits obtained from

displaying all stimuli at the same time. This enables the subjects to carry out simultaneous comparison between the methods. Both objective and subjective results showed that the proposed method consistently outperformed other benchmark methods.

Paper D

In this paper, we propose to use the speaker-identification stage as a pre-processor for identifying the underlying speakers in the mixture. The speaker identification module is connected to the speech separation module in a closed loop. The proposed joint approach solves the SCSS problem when the precise source characteristics are not known *a priori*. We show that the proposed system successfully separates signals without *a priori* knowledge of the speaker identities. The proposed joint system achieves a higher separation quality compared to the case where speaker-independent source models were used, and achieves higher separation performance compared to the quality directly obtained from the mixed signal. From the experiments, it was observed that the results obtained by the suggested method are very close to those obtained by the speaker-dependent method where the correct speaker identities are known *a priori*.

Paper E

This paper introduces a novel speaker identification approach independent of the signal-to-signal ratio under which the underlying speakers in the mixture are mixed together in a single-channel scenario. The proposed method not only detects the speaker identities but also produces the SSR estimate as a bi-product. The proposed method uses a fusion of the adapted Gaussian mixture models and Kullback-Leibler divergence calculated between models. The experimental results show that the proposed speaker identification approach in this paper achieved an accuracy of 97% and 93% when the two target speakers enlisted as three and two most probable speakers respectively. The speaker identification results reported are compared to those reported by [52] showing that the proposed method achieves a close performance to the speaker identification results reported in [52].

Paper F

The problem of detecting the number of speakers for a particular segment occurs in many different speech applications. In single channel speech separation, for example, this information is used to simplify the separation process as the signal has to be treated differently depending on the number of speakers. In this paper, we propose a double-talk detection method in order to determine the single-talk/double-talk hypotheses in a mixed signal composed of two speakers.

We pose the double-talk detection problem as a model selection problem and derive a multiple hypotheses test for determining the number of speakers at a frame level in a mixed signal based on the underlying parametric speaker models, trained *a priori*. The experimental results indicate that the suggested method improves the quality of the separated signals in an SCSS scenario at different signal-to-signal ratio levels both for speaker-dependent and gender-dependent scenarios.

Paper G

We integrate a double-talk detector (DTD) with a speaker identification module for improving the speaker identification accuracy. This paper proposes a novel approach to improve single-channel speaker identification performance for a mixture composed of two speakers. The goal in this paper is to identify the identities of both speakers in the mixture. The proposed approach in this paper is to integrate a DTD as a pre-processing stage for speaker identification back-end. We demonstrate that including the DTD improves the speaker identification accuracy; the average recognition accuracy improves from 96.53% to 97.43%. Overall speaker identification performance is close to the results of the Iroquois system using computationally simple approach.

Paper H

In this paper, we present a joint speech separation and speaker identification system for the speech separation challenge. The system is composed of double-talk/single-talk detector, gain-estimation, speaker identification and sinusoidal mask. We show that the proposed method leads to an improved separation performance compared to base-line and other model-based separation systems in the STFT domain. We start from a situation where we have prior information of codebook index, speaker identities and SSR-level, and then, by relaxing these prior assumptions one-by-one, we demonstrate the efficiency of the complete system.

Contrary to previous studies mostly focused on speech recognition accuracy, we concentrate on reporting the signal quality of the separated signals. To this end, we report PESQ scores, objective speech intelligibility measure and cross-talk measure as objective measures, and MUSHRA test and speech intelligibility test as subjective measurements.

6 Conclusions

In this work, we have proposed several new approaches for solving the SCSS problem and improving speaker identification for identifying speakers in a single-

channel recorded speech mixture. We presented separation strategies that work independent of pitch estimates allowing freedom from performance degradation in source-driven methods due to the limited multi-pitch estimation accuracy. As a result, the methods proposed here prove attractive for SCSS or similar speech enhancement scenarios where pitch estimation methods show low accuracy due to energetic masking. In fact, the sinusoidal parameter estimation method employed here leads to a high frequency resolution at low frequencies, reflecting the pitch harmonic structure of each speaker signal and their mixture.

The methods proposed for single-channel speech separation were all in the sinusoidal domain. Replacing the high dimension STFT features with sinusoidal parameters cuts down the computational complexity of the separation stage significantly. We have also concentrated on the mixture estimator part in a model-driven SCSS method and derived a sinusoidal mixture estimator. The minimization is only calculated at the sinusoidal peaks of the mixture compared to the full-band calculation required in other methods, namely max-model and Wiener filtering methods. It was observed that using the proposed sinusoidal mixture estimator provides an improvement over other SCSS methods.

In the signal reconstruction stage, we generalized the STFT-based masks to sinusoidal masks, namely, sinusoidal binary mask and sinusoidal Wiener filter. We demonstrated that using the proposed sinusoidal masks lowers the undesirable cross-talk introduced by the interfering speaker while separating the target speaker. Using sinusoidal masks also provided higher separation performance compared to the conventional STFT masks.

To solve the speaker-dependency problem of the model-driven SCSS methods, we proposed to use a speaker-identification module before a speech separation module. It was demonstrated that the joint processor led to improvement in separation performance compared to a speaker-independent scenario and achieved a separation performance close to that of the speaker-dependent scenario.

We also proposed approaches for improving the speaker identification results on a speech mixture. To this end, we proposed a novel approach independent of the SSR level of the observed mixture. The proposed method achieved high speaker identification performance compared to the benchmark method. As another approach, we suggested combining a double-talk detection module to provide information on the number of speakers available at each frame extracted from the speech mixture. This information was sent to the speaker identification module which enabled it to perform more efficiently on those frames detected where only one speaker was active (single-talk frames). The results showed that the proposed method improved the speaker identification performance in different SSR levels and mixing scenarios.

We proposed a single-talk/double-talk detector to solve the problem of identifying the number of speakers in an observed speech mixture composed of two speakers. Finally, we presented a full-system composed of a single-talk/double-

talk detector, MMSE sinusoidal estimator, sinusoidal mask as its separation engine and joint speaker identification and SSR estimation. It was shown that the proposed full system achieved a high perceived speech quality as well as speech intelligibility compared to other benchmark methods that participated in the separation challenge [22].

7 Outlook

In the author's opinion, the SCSS problem is largely still an open issue and the methods proposed for solving it still require more improvements and studies. The methods discussed in this thesis for separating the mixed signals recorded by a single-channel still have some restrictions limiting their effectiveness in real scenarios. Some of these limiting assumptions are listed below:

- The training set used to train the speakers is noise-free and we assumed to have access to a large training material recorded from individual (non-mixed) speaker signals.
- The evaluation corpus consists of only digitally added mixtures which are synthetic. In practice, however, the underlying speakers in the mixture can form the mixed signal at any possible signal-to-signal ratio.
- For simplicity, it is often assumed that we only have two speakers in the mixture, however, the ultimate goal of a general separation problem is to separate a mixture composed of an unknown number of sources.
- At a higher level, the environmental or background noise effects as well as the reverberation problem are often neglected making the current separation methods likely to fail when being used under such practical scenarios.

The above mentioned restrictive assumptions are often made while trying to solve the SCSS problem. These limiting conditions are used for simplifying the separation problem, but are very impractical in a real separation scenario. In practice, each one of these issues and their effect on the overall separation performance should be carefully studied. Future work should address these issues by studying how to relax these simplifying yet restrictive and impractical presumptions.

In the following section, we give an overview of possible future ideas that can help to solve the SCSS problem and use it in relevant topics.

7.1 Generalizing SCSS Methods for Speech Enhancement

It is important to note that the methods proposed for SCSS in this thesis can also be generalized into speech enhancement in highly colored noise scenarios

including babble or harmonic noise [37, 38, 75, 120, 130–132]. In such scenarios, the noise corrupted signal includes fewer harmonics, which makes the separation task rather difficult. As a future work, the proposed methods are expected to be appropriately generalized to be applied to speech enhancement under highly colored noise. In this thesis, the proposed method offers an attractive candidate similar to the weighted codebook-mapping (WCBM) previously applied in [153], as an effective tool for speech enhancement. The WCBM in [153], however, was based on harmonic plus noise model (HNM) feature parameters that require voicing estimation and pitch. In contrast, the proposed method in this research works independently of the pitch estimates. It also benefits from the advantages offered by using unconstrained sinusoidal feature parameters as selected features for separation, by using sinusoidal coders as speaker models and finally, by using a sinusoidal mixture estimator.

7.2 Employ Dynamic Constraints for Solving SCSS

It is the author's opinion that using more constraints can lead to better results in a model-driven SCSS approach since the redundancy in features creates inaccuracies in speaker modeling and at mixture estimation stage. One possible idea is to include a kind of penalty function to capture the dynamic information of previous frames to preserve the continuity and spectral structure. Incorporating dynamic information into the separation problem builds a mechanism that enables us to deal with the possible errors in the mixture estimation stage while finding the most likely states of the speaker models.

7.3 Use SCSS with SID in a Closed Loop

In paper D, we demonstrated that a joint speaker identification and speech separation system overcomes the main drawback of previous source-specific methods, i.e., relying on *a priori* knowledge of the underlying speakers in the mixture. It is the author's opinion that a closed loop system comprising a speaker identification module followed by a speech separation module can be further developed to improve the overall separation performance. The idea we used in Paper D was to use the information from the separated output signals as feedback for correcting the speaker identities in a closed loop. Another idea is to employ an analysis-synthesis procedure in the separation module to correct the errors that occur in the mixture estimation stage. More specifically, employing information from signal reconstruction errors for the speakers in the output stage enables us to correct the errors in finding the optimal codevectors to be selected from speaker codebooks.

7.4 Minimize Distortion versus Cross-talk in Separated Signals

Despite the attractive appeal of using mask methods in speech enhancement, in a separation scenario, they introduce certain difficulties mainly attributed to the energetic masking where one speaker signal dominates the other [132]. The mask methods explicitly suggest filtering out one of the speakers as a jammer signal causing inferior performance while recovering the weaker signal [107, 142]. This decision taking in mask methods, in essence, results in producing musical noise and artifacts caused by the interfering signal. Similarly, in speech enhancement, using a mask method, in general, results in musical noise. Comparatively, in speech separation, the artifacts appear in the form of cross-talks, defined as parts of the interfering speaker still audible in the separated signals.

According to the above mentioned discussion, it is very crucial to analyze the potential factors that can influence the intelligibility while separating speech mixtures. More specifically, it is a very important future topic to analyze the distortions introduced by speech separation techniques. Similarly, in [76], it was hypothesized that if distortions introduced by a speech enhancement method are properly controlled, then large gains in intelligibility can be achieved in the enhanced signal.

References

[1] "Method for the subjective assessment of intermediate quality level of coding systems." 2003, ITU-R BS.1534-1.

[2] S.-I. Amari and J.-F. Cardoso, "Blind source separation-semiparametric statistical approach," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2692–2700, 1997.

[3] M. Arcinega, "speaker recognition, noisy environments, auxiliary information, bayesian networks, conditional models," Ph.D. dissertation, acult des sciences et techniques de l'ingnieur STI, Lausanne, Ecole polytechnique fdrale de Lausanne EPFL, 2006.

[4] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Commun.*, vol. 49, no. 5, pp. 402–417, 2007.

[5] J. Barker, M. Ning, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 94–111, 2010.

[6] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[7] T. Beierholm, B. D. Pedersen, and O. Winther, "Low complexity bayesian single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 2004, pp. 529–532.

[8] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *The Journal of Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[9] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 191–199, 2006.

[10] ——, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 191–199, 2006.

[11] L. Benaroya, L. Donagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2003, pp. 613–616.

[12] R. Blouet, G. Rapaport, and C. Fevotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008, pp. 37–40.

[13] J. Boldt, M. Pedersen, U. Kjems, M. Christensen, and S. Jensen, "Error-correction of binary masks using hidden markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010, pp. 4722 –4725.

[14] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[15] A. Bregman, *Auditory scene analysis.* Cambridge, MA: MIT Press, 1990.

[16] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[17] M. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *International Computer Music Conference*, 2000, pp. 1–5.

[18] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the em algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, pp. 728–731.

[19] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *Signal Processing, IEEE Transactions on*, vol. 58, no. 12, pp. 5969–5983, 2010.

[20] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation.* Synthesis Lectures on Speech and Audio Processing: Morgan and Claypool publishers, 2009.

[21] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[22] M. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[23] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[24] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[25] M. E. Davies and C. J. James, "Source separation using single channel ica," *Elsevier Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[26] J. Deller, J. Hansen, and J. Proakis, *Discrete-time Processing of Speech Signals*. IEEE Press, 2000.

[27] O. D. Deshmukh and C. Y. Espy-wilson, "Modified phase opponency based solution to the speech separation challenge," in *Proc. Interspeech*, 2006, pp. 101–104.

[28] L. Di Persia, D. Milone, H. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Elsevier Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.

[29] P. Divenyi, *Speech Separation by Humans and Machines*. Springer, 2004.

[30] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Math Challenges of the 21st Century*, 2000, pp. 1–33.

[31] S. B. Douglas, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2527–2538, 2001.

[32] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 121–124.

[33] D. P. W. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, chapter 4*, 2006, pp. 115–146.

[34] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, pp. 957–960, 2006.

[35] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, 1992.

[36] ——, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[37] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[38] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, 1995.

[39] M. Every and P. Jackson, "Enhancement of harmonic content of speech based on a dynamic programming pitch tracking algorithm," in *Proc. Interspeech*, 2006, pp. 85–88.

[40] D. Fitzgerald, E. Coyle, and B. Lawlor, "Sub-band independent subspace analysis for drum transcription," in *Digital Audio Effects Conference (DAFx)*, 2002, pp. 65–69.

[41] ——, "Independent subspace analysis using locally linear embedding," in *Digital Audio Effects Conference (DAFx)*, 2003, pp. 13–17.

[42] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, no. 1, pp. 47–65, 1940.

[43] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, 1981.

[44] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, 1996.

[45] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 2819–2822.

[46] P. Hansen and S. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, 2005.

[47] ——, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 1, pp. 1–20, 2007.

[48] B. Hanson and D. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1984, pp. 65–68.

[49] R. Hendriks and R. Martin, "MAP estimators for speech enhancement under normal and rayleigh inverse gaussian distributions," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 918–927, 2007.

[50] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[51] J. Hershey and P. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2007, pp. 317–320.

[52] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, 2010.

[53] G. Hu and B. Wang, "Segregation of unvoiced speech from nonspeech interference," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1306–1319, 2008.

[54] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[55] Y. Hu and P. Loizou, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, no. 3–4, pp. 141–177, 2001.

[56] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Journal of Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[57] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *IEICE Electronics Express*, vol. 53, no. 1, pp. 36–43, 1970.

[58] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *The Journal of Machine Learning Research*, vol. 4, no. 7-8, pp. 1365–1392, 2004.

[59] S. Jensen, P. Hansen, S. Hansen, and J. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, 1995.

[60] B.-H. Juang and J. Gray, A., "Multiple stage vector quantization for speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, May 1982, pp. 597 – 600.

[61] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 929–932.

[62] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2604–2604, 2004.

[63] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010, pp. 4206–4209.

[64] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[65] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[66] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982, pp. 1278–1281.

[67] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004, pp. 817–820.

[68] H. Laurberg, M. N. Schmidt, M. G. Christensen, and S. H. Jensen, "Structured non-negative matrix factorization with sparsity patterns," in *Asilomar Conference on Signals, Systems and Computers*, oct 2008.

[69] H. Laurberg, "Uniqueness of non-negative matrix factorization," in *IEEE 14th Workshop on Statistical Signal Processing*, 2007, pp. 44–48.

[70] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[71] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, 2010.

[72] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2014–2023, 2006.

[73] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[74] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[75] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[76] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 47–56, 2010.

[77] P. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, 2005.

[78] C. W. Maina and J. M. Walsh, "Joint speech enhancement and speaker identification using monte carlo methods," in *Proc. Interspeech*, 2009.

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

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

[81] ——, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Elsevier Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.

[82] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 28, no. 2, pp. 137–145, 1980.

[83] J. Ming, T. Hazen, and J. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67–76, 2010.

[84] B. C. J. Moore, *Cochlear Hearing Loss: Physiological, Psychological And Technical Issues*. UK: Wiley, Chichester, 2007.

[85] C. Moore, *An introduction to the psychology of hearing.* San Diego, CA: Academic Press, 2003.

[86] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Co-channel speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1995, pp. 828–831.

[87] P. Mowlaee and Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, 2008.

[88] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "FDMSM robust signal representation for speech mixtures and noise corrupted audio signals," *IEICE Electronics Express*, vol. 6, no. 15, pp. 1077–1083, 2009.

[89] ——, "Evaluating single-channel speech separation performance in transform-domain," *Journal of Zhejiang University - Science C*, vol. 11, pp. 160–174, 2010.

[90] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 34, no. 5, pp. 1124–1138, 1986.

[91] L. Ning and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[92] R. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2007, pp. 661–664.

[93] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, vol. 60, pp. 911–918, 1976.

[94] M. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *Independent Component Analysis and Blind Signal Separation*, vol. 3889, 2006, pp. 674–681.

[95] ——, "Overcomplete blind source separation by combining ica and binary time-frequency masking," in *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, 2005, pp. 15 –20.

[96] ——, "Two-microphone separation of speech mixtures," *IEEE Trans. Neural Netw.*, vol. 19, no. 3, pp. 475 –492, 2008.

[97] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *ObjectiveMeasures of Speech Quality*. New York, NY, USA: Prentice-Hall, 1988.

[98] T. Quatieri, *Discrete-time speech signal processing: principles and practice*. Upper Saddle River, NJ, USA: Prentice Hall Press, 2001.

[99] T. Quatieri and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 1, pp. 56–69, 1990.

[100] M. Radfar, R. Dansereau, and W. Y. Chan, "Monaural speech separation based on gain adapted minimum mean square error estimation," *Journal of Signal Processing Systems*, vol. 61, no. 1, pp. 21–37, 2010.

[101] M. Radfar, A. Sayadiyan, and R. Dansereau, "A new algorithm for two-talker pitch tracking in single channel paradigm," in *Proc. Int. Conf. Signal Processing*, 2006.

[102] M. H. Radfar, W. Y. Chan, R. M. Dansereau, and Wong, "Performance comparison of HMM and VQ based single channel speech separation," in *Proc. Interspeech*, 2009.

[103] M. H. Radfar and R. M. Dansereau, "Long-term gain estimation in model-based single channel speech separation," in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, 2007, pp. 143–146.

[104] ——, "Single channel speech separation using maximum a posteriori estimation," in *Proc. Interspeech*, 2007.

[105] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Speaker-independent model-based single channel speech separation," *Elsevier Neurocomputing*, vol. 72, no. 1–3, pp. 71–78, 2008.

[106] M. Radfar and R. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2299–2310, 2007.

[107] M. Radfar, R. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 1, p. 15, 2007.

[108] ——, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Speech Communication*, vol. 49, no. 6, pp. 464–476, 2007.

[109] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1766–1776, 2007.

[110] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008, pp. 1833–1836.

[111] M. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modeling for single-channel acoustic source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 2004, pp. 641–644.

[112] D. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Elsevier Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[113] 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.

[114] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Speech Communication*, vol. 2, pp. 749–752, 2001.

[115] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Interspeech*, 2003, pp. 1009–1012.

[116] ——, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 793–799.

[117] H. Runqiang, Z. Pei, G. Qin, Z. Zhiping, W. Hao, and W. Xihong, "CASA based speech separation for robust speech recognition," in *Proc. Interspeech*, 2006.

[118] M. Sabin and R. Gray, "Product code vector quantizers for waveform and voice coding," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 3, pp. 474 – 488, June 1984.

[119] R. Saeidi, H. Mohammadi, T. Ganchev, and R. Rodman, "Particle swarm optimization for sorted adapted gaussian mixture models," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 2, pp. 344–353, 2009.

[120] H. Sameti, H. Sheikhzadeh, L. Deng, and R. 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.

[121] M. N. Schmidt and H. Laurberg, "Non-negative matrix factorization with gaussian process priors," *Computational Intelligence and Neuroscience*, 2008.

[122] M. N. Schmidt, "Single-channel source separation using non-negative matrix factorization," Ph.D. dissertation, Technical University of Denmark, 2008.

[123] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Interspeech*, 2006, pp. 2614–2617.

[124] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[125] Y. Shao, S. Srinivasan, Z. Jin, and D. L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77–93, 2010.

[126] Y. Shao, S. Srinivasan, and D. L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2007, pp. 277–280.

[127] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 289–298, 2006.

[128] ——, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008, pp. 1589–1592.

[129] J. Spanier and K. B. Oldham, *An atlas of functions.* Bristol, PA, USA: Taylor & Francis/Hemisphere, 1987.

[130] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[131] ——, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[132] S. Srinivasan and B. Wang, "A model for multitalker speech perception," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3213–3224, 2008.

[133] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 7, pp. 2130–2140, 2007.

[134] C. Taal, R. 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 Processing*, 2010, pp. 4214–4218.

[135] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Elsevier Computer Speech and Language*, vol. 20, no. 3, pp. 128–158, 2006.

[136] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1990, pp. 845–848.

[137] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Proc. Interspeech*, 2006, pp. 89–92.

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

[139] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2000, pp. 765–768.

[140] D. L. Wang, *On ideal binary masks as the computational goal of auditory scene analysis in P. Divenyi, Editor, Speech Separation by Humans and Machines.* Boston, MA: Kluwer Academic, 2005.

[141] ——, "Time-frequency masking for speech separation and its potential for hearing aid design," *Journal on Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[142] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley-IEEE Press, 2006.

[143] D. L. Wang and G. Hu, "Unvoiced speech segregation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 2006, pp. 953–956.

[144] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[145] Z. X. Wang and D. R. Guo, *Special Functions.* world scientific, 1989.

[146] M. Weintraub, "A theory and computational model of auditory monaural sound separation (stream, speech enhancement, selective attention, pitch perception, noise cancellation)," Ph.D. dissertation, Stanford University, Stanford, CA, USA, 1985.

[147] R. J. Weiss, "Underdetermined source separation using speaker subspace models," Ph.D. dissertation, Department of Electrical Engineering, Columbia University, 2009.

[148] R. J. Weiss and D. P. W. Ellis, "Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, 2006, pp. 31–36.

[149] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[150] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2003, pp. 892–895.

[151] M. Wu, D. L. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 369–372.

[152] H. Yi and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[153] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1194–1203, 2007.

Paper A

Improved Single-channel Speech Separation Using Sinusoidal Modeling

Pejman Mowlaee, Mads Græsbøll Christensen, and
Søren Holdt Jensen

This paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 21–24, 2010.

The layout has been revised.

Abstract

We present a novel single-channel separation approach to improve the separation performance while recovering the signals from a mixture. The key idea in this research is to employ a mixture estimator based on unconstrained modified sinusoidal parameters. Compared to the mixmax (binary mask) and Wiener filter (softmask) approaches, the proposed approach works independently of pitch estimates. Furthermore, it is observed that it can achieve acceptable perceptual speech quality with less cross-talk at different signal-to-signal ratios while bringing down the complexity by replacing STFT with sinusoidal parameters. Improvements made by the proposed approach are demonstrated by employing PESQ as our objective measure and MUSHRA listening test as our subjective evaluation.

1 Introduction

Although there have been recent advances in many speech enhancement methods [1], single-channel speech separation (SCSS) systems with high quality are still of great importance and remain as an unsolved problem. Ideal separation systems are targeted to provide accurate estimations for both sources from their mixture. In this aspect having a high quality separation system would play an integral part offering robustness to many practical applications including speech recognition and speaker identification from mixtures of signals.

Previous single-channel speech separation systems are mainly divided into two categories: source driven [2], [3] and model-based methods [4], [5]. Most methods in either group are often required to estimate two masks at each frame and applying them to the given mixture to recover the unknown sources [3], [6–8]. The mask to be applied could be either binary (hard decision) [3], [6], [7] or soft mask [8] leading to MAX-VQ system (with log-max mixture approximation) [6], [7] and Wiener filter (soft masks) [8], respectively. Most of the previous separation systems led to rather satisfying performance for both sources mostly at signal-to-signal ratio (SSR) around 0 dB [4–8]. However, it is often expected that the SSR level vary from 0 dB since the underlying speakers in the mixture often mask each other as time evolves. As a consequence, the SSR level can vary in frames [1] making signal recovery of speakers difficult. One reason for this problem is the fact that usually at a frame level one speaker signal dominates the other and the energies of sources collide at a time-frequency cell. The mask-based methods explicitly suggest to filter out one of the speakers to recover the target speaker. This would degrade the performance of the signal recovery for the masked speaker. Further, using masks inevitably causes cross-talk and artifacts in the separated signals as reported in [3]. From these

aspects, there is a strong motivation in finding novel methods to recover both signals at different SSR levels. According to the results in [3] the Computationally Auditory Scene Analysis (CASA) often lacks enough perceptual quality due to severe cross-talk problems in the separated output signals. The separation performance of CASA-based methods are mainly determined by multi-pitch estimation accuracy. Further, according to the simulation results given in [9], the pitch estimation shows large gross errors especially at low SSR levels because of energetic masking. In this aspect, integration of pitch as proposed in [6] may not be the best solution at low SSR levels, since extracting pitch frequencies from a mixture is both challenging and difficult [9]. This, as a consequence, causes errors in mixture estimation stage which is targeted to find the pair of states of composite sources of the speakers that best fit the given mixture. These indices are then sent to the reconstruction stage, therefore any mixture estimation error would degrade the perceptual quality of the synthesized outputs. Compared to the mask-based methods, a model-based system is able to achieve a rather acceptable separation quality for known speakers at SSR of 0 dB. Model based systems are mostly based on statistical models including vector quantization (VQ) [4–6], Gaussian mixture models (GMM) [8] and Hidden markov models (HMM) [7]. As the most representative method of this group, the MAX-VQ separation system tries to produce two masks based on the estimated VQ states [2], [6], [7] and integrate them with the log-max approximation as its mixture estimation. According to the results reported in [2], [5], [6] using these estimated masks provides re-synthesis signals often corrupted with undesirable cross-talk effects. Furthermore, based on the analysis recently given in [10], we showed that log-max approximation in [6], [7] and Wiener filter [8] are both biased mixture estimators.

The main purpose of this paper is to propose a novel mixture estimator and apply it to modified unconstrained sinusoidal parameters. The separation result of the proposed method is compared with MAX-VQ [7], Wiener filter [8] and model-based VQ system by [4]. The paper is structured as follows: In the next section, we introduce modified unconstrained sinusoidal parameters to be employed as feature parameters. Parameter estimation is presented and followed by the proposed sinusoidal mixture estimator. We also explain the procedure to produce split-VQ speaker models composed of sinusoidal parameters to be used in our proposed method. In Section 3, we present the experimental results with PESQ as an objective measure and MUSHRA test as a subjective measure. Section 4 features the discussions and future work and Section 5 concludes on the work.

2 Proposed separation method

2.1 Sinusoidal model

Each speaker signals is denoted by $s_j(n)$ with $j \in [1, 2]$ and their mixture is shown by $z(n)$ with $n = 0, \ldots, N-1$ as the time sample index where N is the window length in samples. The sinusoidal model of speech in a fixed signal frame is

$$s(n) = \sum_{i=1}^{M} a_i \cos(2\pi f_i n + \phi_i) + e(n) \quad 0 \leq n \leq N-1, \tag{1}$$

where $e(n)$ is the sinusoidal modeling error assumed as an additive noise, M is model order and $i \in [1, M]$ is an index used to refer the ith sinusoidal component characterized by f_i, a_i, and ϕ_i as the frequency, amplitude, and phase, respectively. As a parametric feature vector we have $\mathbf{\Theta}$=$[\mathbf{a}, \mathbf{f}, \phi]$ of size $M \times 3$.

2.2 Sinusoidal Modeling and Parameter Estimation

We consider two modifications on unconstrained sinusoidal model developed in [11]. The modifications we made are described as follow; 1) the spectral coefficients are translated to Mel scale to take into account the logarithmic sensitivity of human auditory system, and 2) at each Mel band, the spectral peak with the highest amplitude is selected. By employing these two foundations as our sinusoidal parameter estimation rule, we find one peak per band and end up with three $M \times 1$ vectors of amplitude, frequency and phase for each speaker signal or their mixture. We define $\mathbf{v}_i = [1 \quad e^{j2\pi f_i} \quad \ldots \quad e^{j2\pi f_i (N-1)}]^T$ with $i \in [1, M]$ as the sinusoidal frequency vector of dimension $N \times 1$ and f_i is the selected peak at the ith band. All estimated sinusoidal frequency vectors for each speaker signal are represented in a matrix format as

$$\mathbf{V} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_M]^T \quad i \in [1, M] \quad , \tag{2}$$

where $\mathbf{V}$ is an $M \times N$ Vandermonde matrix whose rows are $\mathbf{v}_i$. Then signal representation in terms of sinusoids is an $N \times 1$ vector given by $\hat{\mathbf{s}} = \mathbf{V}^T \mathbf{a}$ where $\mathbf{a} = [a_1 \quad \ldots \quad a_M]^T$ and $\hat{\mathbf{s}}$ the reconstructed signal by the sinusoidal peaks in (2). Defining the complex amplitude for each sinusoid as $a_i = A_i e^{j\phi_i}$, the objective of the parameter estimation stage is to find peaks characterized by an amplitude frequency pair given by $A_i = |S_i(f_i)|$ and $f_i = \arg\max_{f \in \mathcal{F}_i} \log |S_i(f)|$, respectively where $\mathcal{F}_i$ denotes a set composed of all the frequencies within the frequency spectrum in the ith subband denoted by $S_i(f)$.

2.3 Sinusoidal Mixture Estimator

According to previous Section, we model the mixed signal as $\mathbf{z} = \mathbf{V}_z^T \mathbf{a}_z$ where $\mathbf{V}_z$ is a Vandermonde matrix composed of M frequency vectors of $N \times 1$ as $\mathbf{v}_{z,i} = [1 \quad e^{j\omega_{z,i}} \quad \dots e^{j\omega_{z,i}(N-1)}]^T$ related to $\{\omega_{z,i}\}$ as the set of sinusoidal frequency peaks retained for the mixture at the ith band. We derive mixture estimator based on unconstrained sinusoidal parameters of the underlying speakers and their mixture. The key idea is to project the mixture to its sinusoidal subspace spanned by the columns of $\mathbf{\Theta}_z$ and attempt to find a cost function to be minimized in mixture estimation stage. Based on the model in (1), for each speaker the power spectrum at the ith band is

$$P(e^{j\omega}) = \sigma_i^2 + A_i^2 \left[\delta(\omega - \omega_i) + \delta(\omega + \omega_i)\right] \quad , \tag{3}$$

where we can replace ω_i with underlying speakers signals frequency sets given by $\{\omega_{1,i}\}$ and $\{\omega_{2,i}\}$ or the mixture denoted by $\{\omega_{z,i}\}$ to define the related power spectrum. A cost function is defined as the squared error between the power spectra of the given and estimated mixture to be sampled only at sinusoidal peaks defined by set $\{\omega_{z,i}\}$. Sampling at sinusoidal frequencies of the mixed signal $\{\omega_{z,i}\}$ is not necessarily synchronous with $\{\omega_{1,i}\}$ and $\{\omega_{2,i}\}$, bringing the requirement of using an appropriate window denoted by $W(e^{j\omega})$ to reduce the spectral leakage. The expected value for the periodogram for each signal spectrum is given by $E\{\hat{P}(e^{j\omega})\} = P(e^{j\omega}) * W(e^{j\omega})$ where $E\{\cdot\}$ denotes expectation operator. The expected value for the mixture approximation error at the ith band is

$$\begin{aligned} E\{\epsilon_i(e^{j\omega})\} &= E\{\hat{P}_z(e^{j\omega}) - \hat{P}_1(e^{j\omega}) - \hat{P}_2(e^{j\omega})\} \qquad (4) \\ &= \sigma_{\epsilon,i}^2 + A_{z,i}^2[W(e^{j(\omega-\omega_{z,i})}) + W(e^{j(\omega+\omega_{z,i})})] \\ &\quad - \sum_{k=1}^{2} A_{k,i}^2[W(e^{j(\omega-\omega_{k,i})}) + W(e^{j(\omega+\omega_{k,i})})], \qquad (5) \end{aligned}$$

where we define $\sigma_{\epsilon,i}^2 = \sigma_{z,i}^2 - \sigma_{1,i}^2 - \sigma_{2,i}^2$ as the variance of the error. The key idea is to sample the expected mixture estimation error in (5) at sinusoidal frequencies of the mixture per ith band defined by set $\{\omega_{z,i}\}$. Replacing ω by $\omega_{z,i}$ in (5) we get

$$\epsilon_i = A_{z,i}^2 - A_{1,i}^2 W(e^{j(\omega_{z,i}-\omega_{1,i})}) - A_{2,i}^2 W(e^{j(\omega_{z,i}-\omega_{2,i})}), \tag{6}$$

which addresses the mixture approximation error defined between the original and estimated spectra at the ith subband. $A_{1,i}$,$A_{2,i}$ and $A_{z,i}$ indicate the first, second and the mixture sinusoidal amplitude selected at the ith band. According to (1), the mixture approximation error energy converges to zero when the

underlying speaker spectra are highly harmonic. Then the mixture estimation error energy termed as d at a given frames is $d = \sum_{i=1}^{M} |\epsilon_i|^2$. Finally, the sinusoidal mixture estimation is accomplished by searching for the optimal states of the composite sources denoted by $\{q^*, t^*\}$ obtained by solving the following minimization problem at each frame

$$\{q^*, t^*\} = \arg\min_{q,t} d_{q,t} \quad , \tag{7}$$

where q, t can be any possible state in the speaker models and $d_{q,t}$ is a 2D cost function defined based on the mixture approximation error in (6). At each frame, by in-place minimization of $d_{q,t}$ in (7), we achieve two states of the speaker models that when combined best fit the mixture. The selected codebook indices are then sent to a weighted overlap-add (OLA) to reconstruct two separated signals.

2.4 Split-VQ Speaker Codebooks

Recently, we reported improvements by applying perceptually weighted subband on the short-time Fourier transform (STFT) features especially at low SSR [5]. It was observed that the selected feature type along with the statistical model determine the upper bound of separation performance. Therefore, to achieve the upper bound separation quality, the selected feature for SCSS is required to perform a high quantization performance which is in agreement with the results reported in [4], [5], [12]. This is in accordance with the conclusion in [13] stating that the ultimate quality of the model-based speech enhancement system is upper bounded by the performance of the coder used. In this respect it was shown in [12] that by applying the split-VQ codebooks on sinusoidal parameters, it is possible to achieve a better quantization performance compared to the conventionally used STFT features. Due to this, we use split-VQ codebooks on sinusoidal amplitude and frequencies of the underlying signals as our speaker codebooks. Sinusoidal parameters from the training dataset of each speaker in the mixture are extracted and results in matrices whose entries are comprised of two distinctive parts; amplitude and frequency each of dimension $1 \times M$. Similar to [12], we apply two different distance measures to produce codebooks of amplitude and frequency. For the amplitude part we apply $d_a(\mathbf{a}, \hat{\mathbf{a}}) = \frac{1}{\|\mathbf{a}\|} \sum_{i=1}^{M} (a_i - \hat{a}_i)^2$ where $d_a(\cdot)$ denotes the distance measure applied to the amplitude part, M is the number of sinusoids used, and $\hat{}$ denotes the coded parameters. Let M_a be the codebook size for the amplitude part of our split-VQ codebook. After establishing M_a amplitude reference vectors, we select the most appropriate frequency vectors for each amplitude codeword. Another VQ of a lower size is performed on the frequency candidates for amplitude codeword. A VQ with frequency codebook size of $1, 2$ or 4 bits was

found as an appropriate choice [12]. To produce frequency codevectors, we apply a distance measure defined between the frequency part of the trained data matrix (defined by $\mathbf{V}$ in (2)) and their related codevectors denoted by $\hat{\mathbf{V}}$ as $d_w(\mathbf{V}, \hat{\mathbf{V}}) = \sum_{i=1}^{M} w_i(\mathbf{v}_i - \hat{\mathbf{v}}_i)^2$ where $d_w(\cdot)$ is defined as a weighted square error measure with $w_i = \frac{a_i}{\|\mathbf{a}\|}$ as the energy normalized amplitude vector used as a dynamic weighting to weight the Euclidean distance measure proportional to the sinusoidal amplitude at the peak frequencies indicated by $\mathbf{V}_i$. Concatenating the coded amplitude and frequency vectors denoted by $\hat{\mathbf{v}}$ and $\hat{\mathbf{a}}$, respectively, we achieve coded vectors in split-VQ of each speaker model.

3 Experimental Results

3.1 Dataset used and Separation Scenario

To evaluate the proposed separation algorithm, we selected four speakers including two male (speakers 9 and 19) and two female speakers (4 and 23) from the database [14]. Ten minutes of the speech signals of each speaker was used to produce split-VQ [12] and STFT codebooks similarly to [4], [5], [7], all with a codebook size of 2048 (for practical reasons 11 bits are used for amplitude and 3 bits for frequency part in split-VQ codebooks). As our separation scenario, we select two speaker signals, and mixed them together at a certain SSR ranging within $[-18, 18]$. The sampling frequency was decreased to 8 kHz from the original 25 kHz. A Hanning window of duration 32 ms is used with a frame rate of 8 ms. The benchmark methods used in our simulations are the mask based methods both binary mask (log-max) [6], [7] and Wiener filter (soft mask) [8]. Since most separation systems predominantly employ STFT or its logarithm as their signal representation [4], [6–8], we include the results obtained by the model-based VQ in [4], [5], [7].

3.2 Objective and Subjective Results

As a proof of concept, we evaluate the separation performance of the proposed method in a speaker dependent scenario. The core of the separation scenario is composed of two trained codebooks. Simulation results are conducted to assess the separation performance of the proposed method and compared them to those obtained by other separation methods. As our testing phase, fifteen pairs of utterances of each speaker (not used in the training set) were randomly selected to make mixtures. The separation results are quantified using PESQ [15]. The results for the separated signals were averaged at each SSR level over all pairs of test signals. Fig. 1 illustrates the separation results obtained by different methods for each speaker output. We also include the upper bound for the separation performance where it is assumed that the optimal indices are

known *a priori*. From Fig. 1 it is observed that the proposed method consistently achieves the highest PESQ score compared to the mask-based approaches of MAX-VQ in [2], [6], [7] and Wiener filter in [8]. The mask-based methods introduce significant mixture estimation error especially at low or high SSR levels. Further, compared to the STFT-based VQ system in [4], [5] denoted by STFT-CB in Fig. 1(a) and (b), the proposed separation approach outperforms the STFT upper bound performance [5]. From curves shown in Fig. 1 it is observed that the proposed mixture estimator asymptotically reaches to the upper bound performance achieved by the split-VQ codebooks [12] while there is on average a large gap between the separation upper bound and those obtained by the mask-based methods [6], [7] and the model-based VQ in [4], [5]. However, all methods exhibit their best performance as SSR increases for the target speaker. The test and the processed signals used in our MUSHRA test are presented on our webpage[1].

As our second experiment we set up a MUlti-Stimulus test with Hidden Reference and Anchors (MUSHRA) listening test as described in ITU-R BS.1534-1 [16] in order to assess the perceived speech quality of the separated signals. Eight listeners participated in the test (the authors not included) and the items used in our listening test are the separated signals produced by different methods at certain SSRs. Fig. 2 depicts the mean opinion score (MOS) obtained from different speech separation methods averaged over all listeners. The excerpts used are shown in Table 1. All of the played signals were monophonic sampled at 8 kHz of duration 2 sec. For each excerpt the listeners were asked to rank eight different separated signals relative to a known reference on a score from 0 to 100. The excerpts are composed of the hidden reference (denoted by HR), an anchor low-pass filtered at 2 kHz (denoted by Anchor 1). The remaining six excerpts are the separated signals defined in Table 1.

In our listening test, the separated signals produced by binary mask (MAX-VQ) in [6], [7] and the STFT-based VQ system [4], [5] were included. Two extreme cases of SSR level as 0 and -18 dB are included. It is observed that the proposed sinusoidal mixture estimator scores about twenty points higher on average than the mask-based method, and more than 25 points higher than STFT-based method. According to Fig. 2, no overlap exists between the proposed method and the benchmark methods. Hence, it can be concluded that the proposed method can achieve statistically significant improvement compared to other methods and consistently improves the performance of the synthesized speech for both target and interference separated signals. Compared to the mask-based approach, the proposed method shows improvements in the perceived signal quality. As indicated by the listening experiments, the separated output for the MAX-VQ method was found to suffer from severe crosstalk. Tests

[1] http://kom.aau.dk/∼pmb/IEEE_ICASSP.htm

Table A.1: Labels of the excerpts used in MUSHRA test.

Excerpt	Separation method and SSR scenario
BMssr0	Binary mask at SSR=0 dB
BMssr-18	Binary mask at SSR= −18 dB
SINssr0	Proposed method at SSR= 0 dB
SINssr-18	Proposed method at SSR= −18 dB
FFTssr0	STFT-based VQ at SSR=0 dB
FFTssr-18	STFT-based VQ at SSR= −18 dB

also revealed that the separation performance of the mask-based methods (especially at SSR=0 dB where their separation performance is often reported) do not necessarily produce the highest perceived quality for the separated signals. This is observed by comparing the MOS in Fig. 2 for the BMssr0 and SINssr0.

4 Discussion and Future work

The results obtained in our simulations are in agreement with [3] stating that the separation quality degrades as the energetic masking takes place at some overlapping time-frequency cells. The sinusoidal features used in this work lead to a high frequency resolution peak picking, reflecting the pitch harmonic structure of single speaker signals and their mixture. In this aspect, the idea is conceptually similar to the motivations behind the use of GF in CASA [2], [3]. By selecting the peak with the highest amplitude we simply exclude peaks mainly caused by windowing effect or modulation of low frequency components while still preserving high perceptual quality.

Comparing the upper bound separation performance in Fig. 1 confirms our recent findings in [5] stating that transforming full-band STFT features into perceptually weighted subbands can significantly provide improvements especially at low SSR levels. Correspondingly, the results in this paper show that by using split-VQ codebooks it is possible to achieve a higher separation upper bound compared to the conventionally used STFT features. The results presented here were in agreement with our recent findings in [5], [12], where the upper bound performance in SCSS was evaluated as the performance of the coder when the optimal codebook indices are known *a priori* (ideal separation).

In this paper, we only considered SCSS. Future work should consider the generalization of the proposed estimator for speech enhancement in non-stationary noise (babble or harmonic) where many researchers show growing interest in this field.

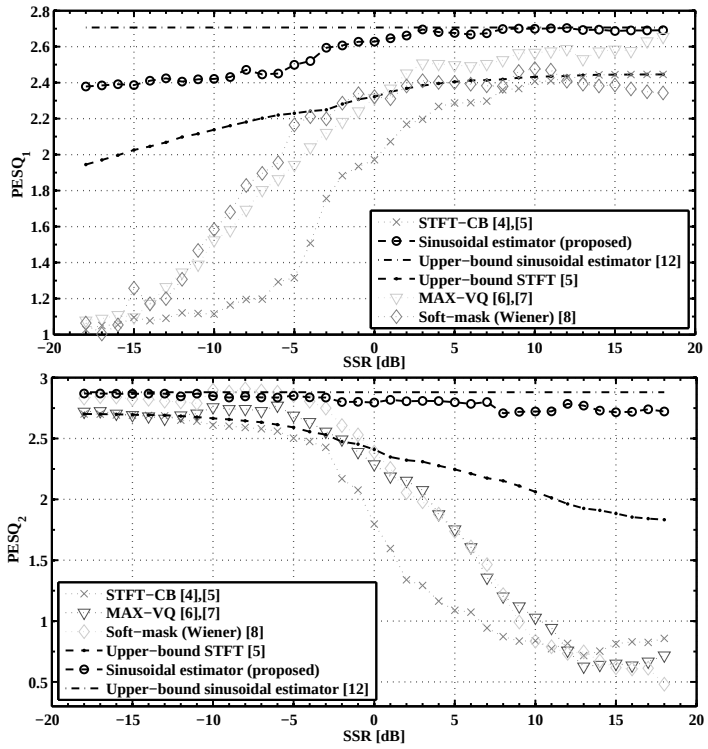

Fig. 1: Evaluation results for different separation methods in terms of PESQ for (a) speaker one (b) speaker two versus SSR.

5 Conclusion

In this paper, a novel mixture estimator has been proposed and derived based on a modified unconstrained sinusoidal parameters to improve the speech separation performance. The method is independent of pitch estimates and offer a new approach for single-channel speech separation, where pitch estimation is sometimes difficult because of energetic masking occurred at time-frequency cells in a mixture at different SSR. Through several experiments it was observed that the proposed method achieved a higher score compared to mask-based methods of MAX-VQ, Wiener filter and the STFT VQ-based separation system especially at low SSR levels. As SSR increases, the proposed method asymptote its separation upper bound performance where it is assumed that the optimal indices are *a priori* available. According to the MUSHRA listening test, it was observed that the perceived speech quality of the proposed system was the highest. Fur-

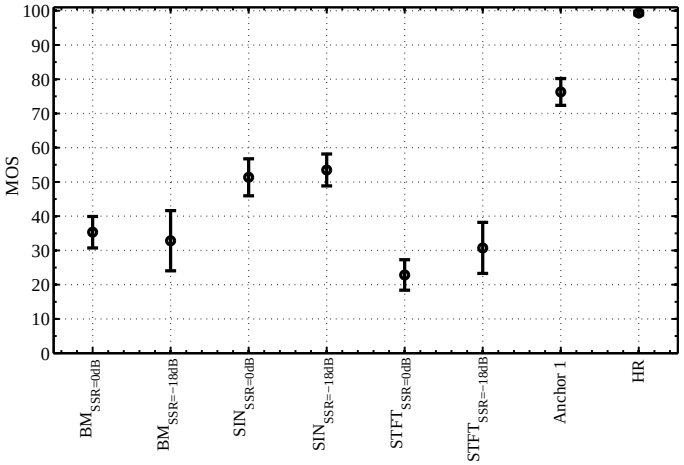

Fig. 2: MOS scores for different separation methods over all excerpts and all listeners. The error bars indicate the 95% confident intervals.

ther, compared to the benchmark methods, the proposed method achieved lower cross-talk and was mostly preferred by the listeners.

References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[2] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[3] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Neural Networks, IEEE Transactions on*, vol. 15, no. 5, pp. 1135–1150, Sept. 2004.

[4] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 957–960, May 2006.

[5] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single-channel separation performance in transform domain," *jzus*, vol. 11, no. 3, Jan. 2010.

[6] M. Radfar, R. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 84 186, March. 2007.

[7] S. Roweis, "Factorial models and refiltering for speech separation and denoising," *European Conference on Speech Communication and Technology*, pp. 1009–1012, 2003.

[8] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[9] M. Radfar, A. Sayadiyan, and R. Dansereau, "A new algorithm for two-talker pitch tracking in single channel paradigm," in *Proceedings of International Conference on Signal Processing ICSP*, Nov. 2006.

[10] P. Mowlaee, A. Sayadiyan, and M. Sheikhan, "Optimum mixture estimator for single-channel speech separation," *IEEE International Symposium on Telecommunications (IST)*, pp. 543–547, Aug. 2008.

[11] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[12] P. Mowlaee and Sayadiyan, "Model-based monaural sound separation by split-vq of sinusoidal parameters," in *European Signal Processing Conference EUSIPCO*, Aug. 2008.

[13] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.

[16] "Method for the subjective assessment of intermediate quality level of coding systems." 2003, iTU-R BS.1534-1.

Paper B

Sinusoidal Masks for Single Channel Speech Separation

Pejman Mowlaee, Mads Græsbøll Christensen, and
Søren Holdt Jensen

This paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4262–4266, 2010.

The layout has been revised.

Abstract

In this paper we present a new approach for binary and soft masks used in single-channel speech separation. We present a novel approach called the sinusoidal mask (binary mask and Wiener filter) in a sinusoidal space. Theoretical analysis is presented for the proposed method, and we show that the proposed method is able to minimize the target speech distortion while suppressing the crosstalk to a predetermined threshold. It is observed that compared to the STFT-based masks, the proposed sinusoidal masks improve the separation performance in terms of objective measures (SSNR and PESQ) and are mostly preferred by listeners.

1 Introduction

Speech signal processing in adverse environments has widely been studied during recent years. Many solutions have been proposed to improve the performance of speech enhancement systems under highly colored noise scenarios [1, 2]. In general, when the interfering noise is non-stationary, the overall performance of the enhanced signal is corrupted by undesired artifacts, speech distortion or residual noise in the background called musical noise [2]. In this regard, there is a crucial need to develop an efficient speech enhancement approach to minimize the residual noise while keeping the quality of the enhanced speech unchanged. We here focus on single-channel speech separation (SCSS).

Mask-based methods have predominantly been applied in many speech enhancement [2] and separation [3]. The key idea behind any mask method is to estimate two masks and apply them to the mixture spectrogram to recover the speaker signals. The mask-based methods are generally categorized into two groups: binary [3–5], and Wiener filter [6], [7]. Binary mask was applied as MAX-VQ [5]) which employs log-max mixture estimator [1] to find two binary masks extracted from the speaker codebooks and then apply them on the mixture. In [8] the MAX-VQ system was applied as a model-based where the codewords are provided by taking the mean value among the vectors trained by a clean speech dataset. According to [8], the separation stage leads to errors while estimating masks for the underlying speakers and the re-synthesis speech quality was reported relatively low because of crosstalk caused by the interfering signal [8].

The greatest asset of mask-based methods lies in its simplicity and the fact that all that is required, is an estimate of the masks time-frequency pattern. Although the use of a mask-based approach is often recommended in speech enhancement [2], it is not yet optimal for SCSS paradigm. The performance of the mask-based methods is influenced by the non-stationarity behavior of speech segments. It is of high interest to incorporate a model of non-stationary speech

into the binary mask or Wiener filtering frameworks. The main concern in mask-based method is attributed to the energetic masking occurring at frames where one speaker signal dominates the other. In such a case, the speaker signals energies collide at mixture time-frequency cells and make the signal recovery rather difficult. The mask-based methods explicitly suggest to filter out one of the speaker as a jammer signal which contradicts with the objective of an ideal separation system targeted to recover both signals.

In this paper, we present a new mask-based method for speech enhancement in general and in particular for SCSS. The proposed sinusoidal mask are constructed by using sinusoidal parameters extracted from the speaker models. It balances a tradeoff between the crosstalk suppression and the target speech distortion. Extensive simulation results are conducted to evaluate the speech separation performance for the proposed sinusoidal masks and compare them with those obtained by predominantly used STFT masks and VQ-based methods with STFT feature. The results show that the proposed masks could achieve a higher performance in terms of Perceptual Evaluation of Speech Quality (PESQ) as objective measure and are mostly preferred according to the informal listening experiments. The rest of the paper is organized as follows. In Section II the problem formulation for the mask-based SCSS is reviewed. The proposed method is presented in Section III. Section IV describes the separation algorithm. Section V presents the simulation results. Section VI concludes on the work.

2 Mask-based Speech Separation

We now briefly review the key idea behind mask-based methods for SCSS. The main objective here is to design two masks, either binary or Wiener filter based, to be applied to the mixture spectrogram. The filtered time-frequency representations are then used to recover the individual speaker signals. Note that the binary mask aims at retaining the dominant time-frequency cells in a mixture spectrogram. This is implemented by removing the interference-dominant units. Such masking approaches are mostly unable to recover both target and masked signals at the same time [4], [5]. On the other hand, the Wiener filter weights each time-frequency cell of the mixture spectrum by taking a soft-decision according the *a priori* SNR [2]. There are two deficiencies for STFT masks; 1) some portions of the weaker speaker signal (often of high importance) is relatively masked by the other speaker (causing speech distortion in target signal), and 2) in some parts of the recovered speech signal (target) some portion of the interfering speaker signal is still audible (called cross-talk). This is similar to musical noise in speech enhancement but introduces a more severe effect for the listeners. Furthermore, the Wiener filter in the STFT domain is not able to

recover both speaker signals with a high quality (especially when one of them is dominant). Hence, we aim at generalizing the STFT-based masks to sinusoidal space to improve the separation performance.

3 Proposed Sinusoidal Masks

In this section, we present the sinusoidal masks aimed at recovering the underlying speaker signals $\mathbf{s}_1$ and $\mathbf{s}_2$ according to the mixture $z = \alpha_1\mathbf{s}_1 + \alpha_2\mathbf{s}_2$ where α_1 and α_2 are the gains.

3.1 Sinusoidal Feature Parameters

According to the sinusoidal model of speech signals, each frame of the signal can be represented as a $N \times 1$ time vector as

$$\mathbf{s} = \mathbf{V}^T\mathbf{a} \quad , \tag{1}$$

where $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_M]^T$ is a Vandermonde matrix of $M \times N$ whose rows are $\mathbf{v}_k = [1 \quad e^{j\omega_k} \quad \dots \quad e^{j\omega_k(N-1)}]^T$ with $k \in [1, M]$ as the sinusoidal frequency vector of dimension $N \times 1$, ω_k indicates the frequency of the kth selected peak, N is the time window length in samples, M is the order, and $\mathbf{a} = [a_1 \dots a_M]^T$ is a $M \times 1$ complex sinusoidal amplitude vector whose components are defined as $a_k = A_k e^{j\phi_k}$. The sinusoidal model used here is [9]; however sinusoidal parameter estimation is a bit different which is described in [10], [11]. We simply select the peak of the highest amplitude per Mel scale band and is characterized with triple $M \times 1$ vectors of amplitude, frequency and phase of the selected peaks. The decision rule of taking the highest peak per band is similar to maximum approximation used as a minimum mean square error (MMSE) mixture estimator for log amplitude spectra [1].

3.2 Sinusoidal Binary Mask

We now consider the SCSS problem in a frame and define k as the frequency bin index. We incorporate the selected sinusoidal peaks within the bands to establish a sinusoidal binary mask defined as

$$H_1(\omega_k) = \begin{cases} 1 & \text{if} \quad A_{1,k} \geq A_{2,k} \\ 0 & \text{if} \quad A_{1,k} < A_{2,k} \end{cases} \quad , \tag{2}$$

where ω_k denotes the kth frequency component. The hard decision making in (2) can be summarized as an on-off keying (OOK) between two states C_1 the class of $A_1(k)$ and C_2 in favor of $A_2(k)$. A similar definition goes for $H_2(\omega_k)$ as complement of $H_1(\omega_k)$. The decision rule is similar to the ideal binary mask

which compares the gain ratio of each time-frequency cell to the 0 dB local SNR [3].

3.3 Sinusoidal Wiener Filter as a Constrained Optimization

Speech enhancement with negligible perceived distortion is of high interest. In order to achieve an ideal separation performance we need to satisfy two requirements [2]; 1) it is required to guarantee minimal speech distortion of the target signal, and 2) the separated signals are required to have no portions of the other speaker signal. Without loss of generality, we assume in the following that $\mathbf{s}_1$ is the target and the other speaker is the interfering signal. We aim to find the kth frequency bin of the sinusoidal gain function as $g_1(\omega_k)$ that solves a constrained minimization problem by keeping the cross-talk of the other speaker below a predefined threshold and minimizing the target speech distortion. We define $\varepsilon(\omega_k)$ as the separation error for the target signal in the kth frequency bin as

$$\varepsilon(\omega_k) \quad = \quad \underbrace{(g_1(\omega_k)-1)S_1(\omega_k)}_{\varepsilon_{s_1}(\omega_k)} + \underbrace{g_1(\omega_k)S_2(\omega_k)}_{\varepsilon_{s_2}(\omega_k)} \quad , \tag{3}$$

where $\varepsilon_{s_1}(\omega_k)$ is the speech distortion term for target speaker while $\varepsilon_{s_2}(\omega_k)$ is the the crosstalk term of the interfering speaker. We define $\mathbf{S}_i$ as a $N \times 1$ vector containing the spectral components of the ith underlying speakers defined as $\mathbf{S}_i = \mathcal{F}\{\mathbf{s}_i\} = [S_i(\omega_1) \cdots S_i(\omega_K)]^T$ where K is the number of frequency points used in calculating the DFT. The speech distortion energy for the target signal is calculated as $\varepsilon_{s_1}^2 = E\{\boldsymbol{\varepsilon}_{s_1}^H(\omega)\boldsymbol{\varepsilon}_{s_1}(\omega)\}$ and the cross-talk energy of the other speaker is $\varepsilon_{s_2}^2 = E\{\boldsymbol{\varepsilon}_{s_2}^H(\omega)\boldsymbol{\varepsilon}_{s_2}(\omega)\}$. We consider an optimization problem addressed as below

$$\min_{g_1,\mu} \quad \varepsilon_{s_1}^2 \quad \text{s.t.} \quad \varepsilon_{s_2}^2 \leq \delta \quad . \tag{4}$$

We define $\mathbf{G}_1$ as a $N \times N$ diagonal matrix with entries of $g_1(\omega_k)$ on its diagonal. The periodogram estimation of the PSD for the ith speaker is denoted by $\mathbf{P}_{s_i s_i}$ defined as the Fourier transform of the autocorrelation function, $\mathbf{R}_{s_i s_i}$ which is Toeplitz. Then, the power spectrum components are the diagonal elements of $\mathcal{F}^H R_{s_i s_i} \mathcal{F}$ where $\mathcal{F}$ is the N-point Fourier transform matrix [2] and $\mathbf{P}_{s_i s_i} = \text{diag}(P_{s_i s_i}(\omega_1), \cdots, P_{s_i s_i}(\omega_K))$. By using the Lagrangian multiplier method, we are required to solve the following constrained optimization in sinusoidal domain as

$$\mathbf{L} = (\mathbf{G}_1 - \mathbf{I})\mathbf{P}_1(\mathbf{G}_1 - \mathbf{I}) + \mu \mathbf{G}_1 \mathbf{P}_2 \mathbf{G}_1 \quad , \tag{5}$$

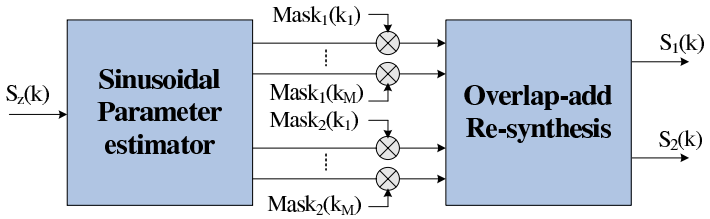

Fig. 1: SCSS system based on sinusoidal masks. The speaker signals are recovered by multiplying the mixture spectrum using a mask.

where $\mathbf{L}$ is a diagonal matrix whose (k,k)th element is given by the lagrangian of $\mathcal{L}(g_1(\omega_k),\mu)$ calculated at the kth frequency bin, and μ is the Lagrange multiplier as a parameter to trade off crosstalk suppression against speech distortion. Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{g}_1} = \mathbf{0}$ with $\mathbf{0}$ as a $N \times 1$ zero vector, we obtain

$$(g_1(\omega_k) - 1)P_{s_1s_1}(\omega_k) + \mu g_1(\omega_k)P_{s_2s_2}(\omega_k) = 0 \quad . \tag{6}$$

The kth component of the sinusoidal Wiener gain is

$$g_1(\omega_k) = \frac{P_{s_1s_1}(\omega_k)}{P_{s_1s_1}(\omega_k) + \mu P_{s_2s_2}(\omega_k)} \quad . \tag{7}$$

Since we have no access to speakers' PSD, we replace them by the squared spectral vectors in discrete frequency domain and we obtain

$$g_1(\omega_k) = \frac{\xi_k}{\xi_k + \mu} \quad , \tag{8}$$

where we define $\xi_k = \frac{P_{s_1s_1}(\omega_k)}{P_{s_2s_2}(\omega_k)}$ as the *a priori* SSR computed at sinusoidal frequency peaks. The idea is to make the noise imperceptible by a proper choice of μ. In this paper we assume that the SSR level is known *a priori* and we set $\mu = (\frac{\alpha_2}{\alpha_1})^2$ which is agreement with the relevant discussion in chapter 6 of [2] where μ was such chosen to minimize the speech distortion in speech dominated frames while reducing the residual noise in noise dominated frames. Replacing μ into (8) and taking square root from (8) we have

$$g_1(\omega_k) = \frac{\alpha_1 S_1(\omega_k)}{\sqrt{\alpha_1 S_1^2(\omega_k) + \alpha_2 S_2^2(\omega_k)}} \quad , \tag{9}$$

which is similar to parametric Wiener filter in [2] and we call it sinusoidal Wiener mask. The proposed masks: sinusoidal binary in (2) and Wiener mask in (8) are used to recover the signals.

4 Separation Algorithm

The key part of a separation algorithm is to find the optimal states of the source models of the speakers in the mixture. In this section we present the idea on how to find these states. These states refer to a codeword each composed of sinusoidal amplitude and frequency vectors denoted by $\mathbf{a} = [a_1 \dots a_M]^T$ and $\mathbf{f} = [f_1 \dots f_M]^T$, respectively. These codewords found by mixture estimation will be used to produce the sinusoidal masks. The codebooks are designed by using split-VQ of the sinusoidal parameters [10]. In the separation stage, two estimators are used. As our first estimator we use the optimum mixture estimation in [12]. In [12], it was demonstrated that under the uniformity assumption of mixture phase, the optimal estimator for SCSS, $S_{z,\mathrm{opt}}(k)$ in the MMSE sense is

$$S_{z,\mathrm{opt}}(\omega_k) = (S_1(\omega_k) + S_2(\omega_k)) \frac{E(\gamma_k)}{\pi} \quad , \tag{10}$$

with $\gamma_k = \frac{4S_1(\omega_k)S_2(\omega_k)}{S_1^2(\omega_k)+S_2^2(\omega_k)}$ and $E(\cdot)$ is the complete Elliptic integral of the second kind given by

$$E(\gamma_k) = \pi \left[1 - \frac{\gamma_k^2}{4} - \left(\frac{1 \times 3}{2 \times 4} \right)^2 \left(\frac{\gamma_k^4}{3} \right) - \dots \right] \quad . \tag{11}$$

As our second method, according to the optimum mixture estimator in (10) we replace the sinusoidal masks in (9) and we obtain

$$S_{z,m}(\omega_k) = \frac{1}{\pi} S_w(\omega_k)(g_1(\omega_k) + g_2(\omega_k))E(\gamma_k), \tag{12}$$

where $S_{z,m}(\omega_k)$ denotes the mask-based estimated mixture at the kth frequency bin and we define $S_w(\omega_k) = \sqrt{S_1^2(\omega_k) + S_2^2(\omega_k)}$ as the Wiener filter mixture estimation. To include SSR levels other than 0 dB in (12), we can consider the gain values α_1 and α_2. The sinusoidal mixture estimation is accomplished by searching for the optimal states of the composite sources by minimizing $\sum_{k=1}^{M} |S_z(\omega_k) - \hat{S}_z(\omega_k)|^2$, where $\hat{S}_z(\omega_k)$ can be replaced by either $S_{z,\mathrm{opt}}(\omega_k)$ in (10) or $S_{z,m}(\omega_k)$ in (12). The solution of this minimization problem gives two states in the split-VQ codebooks to be used produce the masks in (9). To re-synthesize the separated outputs the mixture phase ϕ_z is used. Using the sinusoidal binary mask the kth frequency bin of the refiltered spectrum is

$$S_i(\omega_k) = S_z(\omega_k) g_i(\omega_k) \quad i \in \{1, 2\} \quad , \tag{13}$$

where $S_z(\omega_k)$ is the mixture power spectrum, $S_i(\omega_k)$ is the recovered spectrum for the ith speaker signal and $g_i(\omega_k)$ is either sinusoidal binary mask or sinusoidal Wiener mask given by (2) and (8), respectively. By using IDFT along with the

mixture phase we get recovered time signals of each speaker. Fig. 1 shows the block diagram describing the SCSS based on the sinusoidal mask. In Fig. 1, $\{k_1, \ldots, k_M\}$ indicate the frequency bins of sinusoidal peaks defined in Section II and M is the sinusoidal model order.

According to the suppression rule of Ephraim and Malah in [13], the proposed sinusoidal Wiener filter can be expressed as

$$g_1(\omega_k) = \sqrt{\frac{\xi_k}{\xi_k + 1} \frac{S_1^2(\omega_k) + S_2^2(\omega_k)}{S_z^2(\omega_k)}} = \sqrt{\frac{\xi_k}{\xi_k + 1} \left(\frac{1 + \nu_k}{\zeta_k} \right)} \tag{14}$$

Similar to [13], we define $\zeta_k = \frac{S_z^2(\omega_k)}{S_2^2(\omega_k)}$ as the *a posteriori* SSR and $\nu_k = \zeta_k \frac{\xi_k}{\xi_k + 1}$ as the instantaneous SSR. Then the proposed mask given by (14) is similar to Ephraim and Malah suppression rule already given in [13].

5 Simulation Results

To assess the separation performance, we use the comprehensive database in [14] consisting of 34 speakers each containing 500 utterances. The sampling rate is decreased to 8 kHz from the original 25 kHz. Ten minutes of the speech signals of each speaker was used to produce the split-VQ and STFT codebooks with a codebook size of 2048. Twenty utterances are chosen from speakers 9 and 23 as test signals to evaluate the separation algorithms in a speaker-dependent scenario. The mixed signal is generated by adding the signals at different SSR. The separation performance for each method is reported in terms of PESQ [15] and segmental SNR (SSNR) [2]. The methods included in our simulations are the sinusoidal binary mask, sinusoidal Wiener filter and their STFT counterparts. As our benchmark methods, we applied algorithms similar to [4], [6], [16]. We also include the upper-bound performance for both STFT [17] and split-VQ [10] determining the highest performance obtainable by using the same source model if no mixture estimation error occurs. We used window length of 32 ms along with a frame shift of 8 ms. The codebook size for STFT and split-VQ was 2048. The number of sinusoidals used in our simulations is 50 and the number of DFT points in the STFT-based methods is 1024.

Fig. 2 shows the averaged PESQ scores for the separated signals obtained from their mixture[1]. From Fig. 2(a) it is observed that the optimal mixture estimator in sinusoidal space given in (10) is very close to the sinusoidal masks approximation in (12). Furthermore, by increasing the SSR level, both curves asymptotically attain the same performance, which is determined by the split-VQ upper-bound quantization performance [10]. Fig. 2(a) illustrates the PESQ

[1]The mixed and separated signals of different methods are downloadable from webpage at http://kom.aau.dk/∼pmb/IEEE_ICASSP2.htm

curves obtained by masking approaches in STFT and sinusoidal domain. It is observed that applying the sinusoidal Wiener mask to the mixture improves, the separation performance compared to the STFT-based methods. This is also validated by listening to the separated signals at different SSRs for different genders. The improvement introduced by sinusoidal wiener filter over the STFT-based mask is rather significant at low SSR. Fig. 2(b) shows the separation performance for the second speaker signal in terms of SSNR in dB, and it is observed that employing the proposed sinusoidal masks, either binary or Wiener filter, cause improvements in the separation performance compared to the STFT binary mask [3–5] or Wiener masks in [3], [6]. This improvement is significant at low SSRs. The curves in Fig. 2(b) show that the performance obtained by the optimal mixture estimator in (10) is very close to the operational upper-bound determined by the STFT VQ. It is also observed that the proposed sinusoidal masks outperform the results obtained by both optimal estimator in STFT domain and the STFT-based masks. The listening tests revealed that the re-synthesized signal quality is significantly improved compared to those obtained by the STFT based methods. From curves shown in Fig. 2(b) we observe that by employing the optimal estimator in (10) we reach to the operational upper-bound performance (where we assumed that the correct indices are known *a priori*). The results presented here are in accordance with our recent results in [17] where we showed that the model-based speech separation in transform domain results in improvements over the mask-based methods especially at low SSR.

According to the listening results, as SSR level decreases the STFT-based masks mostly lead to inferior performance. In contrast, the proposed sinusoidal masks achieve a superior performance and introduce significant improvement especially at low SSRs. This could be explained by the fact that the proposed sinusoidal mask minimizes the mixture estimation error at sinusoidal peaks of the mixture making a tradeoff between less crosstalk and small speech distortion. The proposed masks retain the highest peaks per bands and exclude other peaks mostly caused by main-lobe windowing or low-frequency modulation effect. This strategy would exclude those peaks vulnerable to be masked by the other speaker signal. Therefore the method is expected to result in lower crosstalk compared to the STFT masks.

6 Conclusion

In this paper, we proposed a new sinusoidal version for both binary mask and Wiener filter and compared their performance with their STFT counterparts. It was observed that the proposed sinusoidal masks could result in a significant improvement in the re-synthesized speech quality for both the recovered signals. We presented a framework to minimize the signal distortion while keeping the

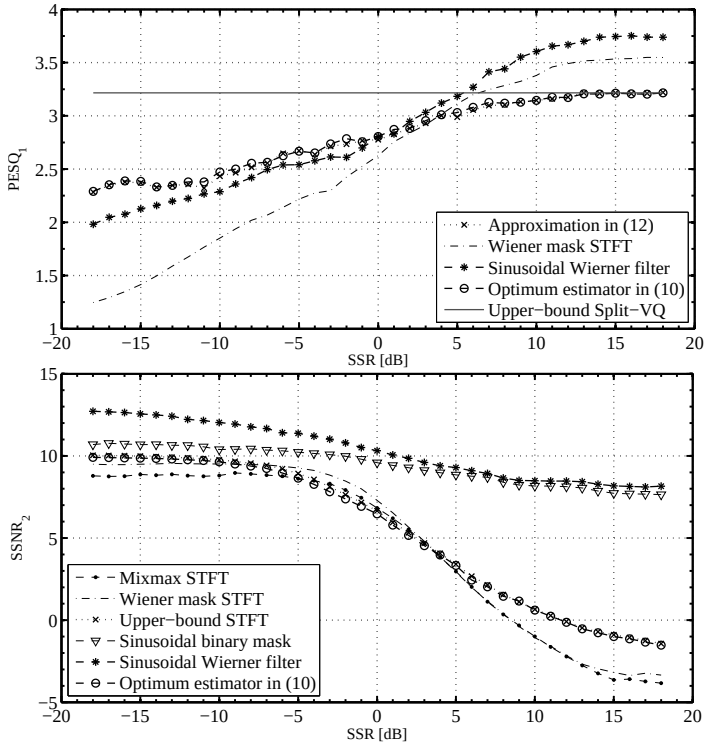

Fig. 2: Comparing the PESQ scores of the sinusoidal masks with the STFT masks and VQ-STFT versus SSR$\in [-18, 18]$.

crosstalk below a predefined threshold. It was demonstrated that by the proposed approach, it is possible to reach the optimal performance for SCSS in a MMSE sense. From the simulation results, It was observed that, compared to the STFT masks, sinusoidal masks improved the separation performance in terms of SSNR and PESQ and were mostly preferred by informal listening tests.

We focused on speech separation scenario. As a future work, it is highly desirable to evaluate the proposed masks in other noisy environments including babble noise, car noise and other noise types. It is expected that the proposed method results in improvements compared to the STFT masks.

References

[1] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 10, no. 6, pp. 341–351, Sep. 2002.

[2] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[3] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[4] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 84–186, March. 2007.

[5] S. Roweis, "Factorial models and refiltering for speech separation and denoising," *European Conference on Speech Communication and Technology*, pp. 1009–1012, 2003.

[6] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[7] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[8] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[9] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[10] P. Mowlaee and Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *European Signal Processing Conference*, Aug. 2008.

[11] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "FDMSM robust signal representation for speech mixtures and noise corrupted audio signals," *IEICE Electronics Express*, vol. 6, no. 15, pp. 1077–1083, 2009.

[12] P. Mowlaee, A. Sayadiyan, and M. Sheikhan, "Optimum mixture estimator for single-channel speech separation," *IEEE International Symposium on Telecommunications (IST)*, pp. 543–547, Aug. 2008.

[13] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[14] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.

[16] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 957–960, May 2006.

[17] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single-channel separation performance in transform domain," *jzus*, vol. 11, no. 3, Jan. 2010.

Paper C

New Results on Single-Channel Speech Separation Using Sinusoidal Modeling

Pejman Mowlaee, Mads Græsbøll Christensen,
and Søren Holdt Jensen

This paper has been published in
IEEE Transactions on Audio, Speech, and Language Processing,
vol. 19, no. 8, pp. 1265–1277, 2011.

The layout has been revised.

Abstract

We present new results on single-channel speech separation and suggest a new separation approach to improve the speech quality of separated signals from an observed mixture. The key idea is to derive a mixture estimator based on sinusoidal parameters. The proposed estimator is aimed at finding sinusoidal parameters in the form of codevectors from vector quantization (VQ) codebooks pre-trained for speakers that, when combined, best fit the observed mixed signal. The selected codevectors are then used to reconstruct the recovered signals for the speakers in the mixture. Compared to the log-max mixture estimator used in binary masks and the Wiener filtering approach, it is observed that the proposed method achieves an acceptable perceptual speech quality with less cross-talk at different signal-to-signal ratios. Moreover, the method is independent of pitch estimates and reduces the computational complexity of the separation by replacing the short-time Fourier transform (STFT) feature vectors of high dimensionality with sinusoidal feature vectors. We report separation results for the proposed method and compare them with respect to other benchmark methods. The improvements made by applying the proposed method over other methods are confirmed by employing perceptual evaluation of speech quality (PESQ) as an objective measure and a MUSHRA listening test as a subjective evaluation for both speaker-dependent and gender-dependent scenarios.

1 Introduction

There are many speech and audio applications where the signal of interest is corrupted by highly correlated noise sources. Separating such signals from their mixture has often been considered as one of the most challenging research topics in the area of speech enhancement. An extreme case of speech enhancement, single-channel speech separation (SCSS), is often considered as one of the most difficult scenarios where a speaker signal is corrupted with other interfering speaker signals. Although there have been recent advances in speech enhancement methods [1–10], SCSS with high speech quality still remains as a challenge. High quality separation systems could play an integral role in offering robustness in many practical applications including speech coding, speech recognition, speaker recognition in adverse mixture scenarios, and hearing aids [11].

The main objective for an ideal speech separation system is to recover the unknown speaker signals accurately, based on their observed mixed signal recorded by one microphone. The SCSS problem is ill-conditioned since the mixing matrix is non-invertible. The problem is in principle solvable by imposing *a priori* information e.g. about the speaker models [12–19]. Previous state-of-the-art SCSS systems can be divided into two groups; (i) source-driven or computation-

ally auditory scene analysis (CASA)-based method [20–25], and (ii) model-based method [12–19].

The main objective in the first group is to produce the binary masks required to separate the unknown speaker signals from their mixture. The methods predominantly use estimated pitch trajectories by applying a multi-pitch estimator. According to the results reported in [22, 26, 27], the separation quality degrades as energetic masking takes place at some overlapping time-frequency cells. Therefore, the overall separation performance is limited by the accuracy of the multi-pitch estimator especially when the relative amplitude levels of the signals differ substantially (the signal-to-signal ratio (SSR) gets either low or high). At these SSR levels, the pitch estimation accuracy is relatively lost by large gross errors [26, 28]. In addition, according to [20], the CASA-based methods are mostly able to segregate the voiced frames of the mixture and often lack perceptual quality due to a severe cross-talk problem.

The second group, model-based separation systems is based on statistical models including VQ [15–18], Gaussian mixture models (GMM) [13, 19, 29, 30], and Hidden Markov models (HMM) [12, 14, 27]. In [14], a separate HMM was applied for each speaker and a huge state space of 8,000 was required in order to carefully capture every possible signal transition state. Though using HMMs enables the modeling of correlated speaker signals, according to [31], it leads to a significantly more complex mixture estimation approach. MAX-VQ attempts to find two masks based on the estimated VQ codewords. According to the results reported in [17, 22, 23, 32], using such masks inevitably causes cross-talk and artifacts in the re-synthesized signals.

From a synthesis viewpoint, the methods in the second group are divided into two classes: overlap-add procedure and mask methods. The masks to be applied are either the binary [15, 16, 22, 23] or the Wiener filter masks [13, 29, 30, 33] leading to the separation approaches of the log-max estimator [12–15, 34] and the Wiener filtering [13, 29], respectively. Despite the attractive appeal of using masks in speech enhancement or separation, they have problems in dealing with the energetic masking [2]. These methods suggest filtering out one of the speakers as a jammer signal thereby causing inferior performance while recovering the masked speaker signal [16, 20].

In model-based methods, difficulties arise while mapping vectors of mixed signals onto states of speaker models resulting in wrong association of the codevectors with the log-max estimator leading to the selection of poorly filtered signal vectors [16, 23]. Selecting incorrect states from the speaker models could degrade the perceptual quality of the separated signals. According to [35], the model-based approach was expected to perform better than the pitch-based methods indicating that using only the pitch information shows limited discrimination for sequential grouping. This brings forward the idea that integration of pitch and spectral envelope in [16] may not be the most efficient solution to recover both

signals because accurate multi-pitch estimation from a mixture at low SSRs is still a problem [22, 26, 27, 36].

It is important to note that most of the previous separation systems achieve a rather acceptable separation quality for the underlying sources in the mixture by assuming speaker signals to have nearly the same long-term energy level, i.e., when SSR level is around 0 dB. In practice, however, a nonzero SSR level is expected since at each frame, one speaker signal often dominates others and the energies of the sources most likely collide [1, 37], a phenomenon called energetic masking [2] that makes the signal recovery of the speakers rather difficult. Therefore, studying novel methods to improve the separation quality at different SSRs is very important.

In this paper, we present new results for SCSS by proposing a mixture estimator based on sinusoidal parameters provided by codebooks for underlying speakers in the observed speech mixture. We consider a speech mixture composed of two speakers. The proposed model-based separation method aims to find optimal sinusoidal codevectors, one from each speaker model, that when combined best describe the observed mixture segment. The speaker models pretrained for speakers are VQ codebooks composed of sinusoidal amplitude and frequency vectors. In this paper, we focus on speaker-dependent scenario and then we relax this assumption by using gender-dependent codebooks as an intermediate scenario. Through extensive simulations and subjective evaluations, we assess the separation performance of the proposed method at different SSR levels. The separation results show that the performance of the proposed method outperforms those obtained by other previous SCSS methods.

The rest of the paper is structured as follows: In the next section, we review previous sinusoidal methods for separation. In Section III, we introduce modified unconstrained sinusoidal parameters to be employed as feature parameters. The parameter estimation procedure is presented and followed by the proposed sinusoidal mixture estimator. In Section IV, we present the experimental results to compare the separation performance of the proposed method with that of other methods. Section V presents subjective evaluations and results of our MUSHRA test to assess the perceived quality obtained by different methods. Section VI features the discussions and Section VII concludes the work.

2 Proposed Separation Method

We will now proceed to describe the proposed separation approach using sinusoidal modeling. Fig. 1 shows the block diagram of the proposed separation approach. The system is composed of the following blocks: sinusoidal parameter estimation, two trained speaker models, sinusoidal mixture estimator and overlap-add for signal reconstruction. In the following, we present our separation

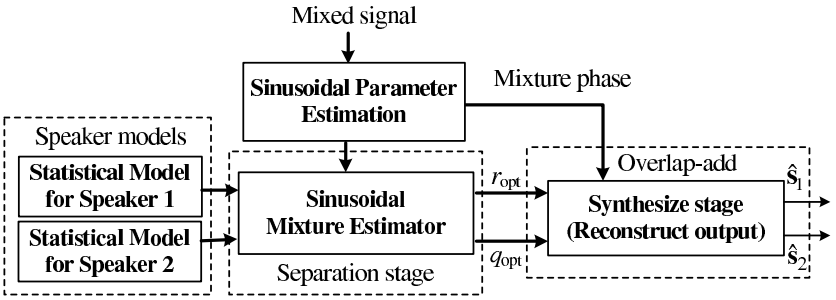

Fig. 1: Block diagram for the proposed speech separation method using sinusoidal modeling ($\hat{s}_1$ and $\hat{s}_2$ are separated signals).

approach.

2.1 Sinusoidal Modeling

Before presenting the sinusoidal modeling, we will introduce some basic notation. Assume that we have a mixed signal, $\{z(n)\}_{n=0}^{N-1} = \sum_{k=1}^{K} \{s_k(n)\}_{n=0}^{N-1}$ composed of K speakers where k is the speaker index and the kth speaker signal is denoted by $\{s_k(n)\}_{n=0}^{N-1}$ with $k \in [1, K]$, n is the time sample index and N is the window length in the samples. At each frame, we represent the kth speaker signal in additive noise $e_k(n)$ as

$$s_k(n) = \sum_{i=1}^{L} A_{k,i} \cos(n\omega_{k,i} + \phi_{k,i}) + e_k(n) \quad 0 \leq n \leq N-1 \tag{1}$$

where i is an index used to refer to the ith sinusoidal component characterized by the amplitude $A_{k,i}$, frequency $\omega_{k,i}$ and phase $\phi_{k,i}$, respectively. We define a parameter vector as $[\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\phi}]$ of size $L \times 3$ with $\boldsymbol{\alpha} = \{A_{k,i}\}_{i=1}^{L}$, $\boldsymbol{\omega} = \{\omega_{k,i}\}_{i=1}^{L}$ and $\boldsymbol{\phi} = \{\phi_{k,i}\}_{i=1}^{L}$ denoting the the k speaker's amplitude, frequency and phase vectors, respectively, and L being the sinusoidal model order. The signal model in (1) is also used for representing observed mixed signal, $z(n)$. For the sake of simplicity and tractability, here, we focus on separating speech mixture composed of two speakers, i.e. $K = 2$ and $k \in [1, 2]$.

2.2 Sinusoidal Parameter Estimation

We make two modifications to the unconstrained sinusoidal parameter estimator developed in [38] described as follows: i) the spectral coefficients are translated to the Mel scale to take into account the logarithmic sensitivity of the human auditory system, and ii) at each band the spectral peak with the highest amplitude is selected [39]. These changes allow us to select the most perceptually

relevant peak per band. The N-point Discrete Fourier Transform (DFT) vector for the ith frequency band of the kth speaker is represented by

$$\mathbf{v}_{k,i} = [1 \quad e^{j\omega_{k,i}} \quad \dots \quad e^{j\omega_{k,i}(N-1)}]^T \quad i \in [1, L], \tag{2}$$

where $\omega_{k,i}$ denotes the selected peak at the ith band for the kth speaker. We define

$$\mathbf{V}_k = [\mathbf{v}_{k,1} \quad \mathbf{v}^*_{k,1} \quad \mathbf{v}_{k,2} \quad \mathbf{v}^*_{k,2} \quad \dots \quad \mathbf{v}_{k,L} \quad \mathbf{v}^*_{k,L}]^T, \tag{3}$$

where $(\cdot)^*$ is the complex conjugate operator and $\mathbf{V}_k$ is a $2L \times N$ Vandermonde matrix whose rows are $\mathbf{v}_{k,i}$ defined in (2). The signal representation for the kth speaker in terms of sinusoids is given by an $N \times 1$ vector, $\hat{\mathbf{s}}_k = \mathbf{V}_k^T \mathbf{a}_k$, where

$$\mathbf{a} = [A_{k,1}e^{j\phi_{k,1}} \; A_{k,1}e^{-j\phi_{k,1}} \; \dots \; A_{k,L}e^{j\phi_{k,L}} \; A_{k,L}e^{-j\phi_{k,L}}]^T. \tag{4}$$

We define $S_k(\omega)$ as the complex spectrum for the kth speaker. The objective of the sinusoidal parameter estimation is to find peaks with the constraint [39]

$$\omega_{k,i} = \arg \max_{\omega \in \Omega_{k,i}} |S_k(\omega)| \quad , \quad \text{and} \quad a_{k,i} = S_k(\omega_{k,i}) \tag{5}$$

where $\Omega_{k,i}$ is a set composed of all continuous frequencies for the kth speaker within the ith band and $\arg\max(\cdot)$ returns the argument where $|S_k(\omega)|$ attains its maximum value.

2.3 Proposed Sinusoidal Mixture Estimator

In this section, we propose a mixture estimator based on the sinusoidal parametric vectors in our model-based separation approach shown in Fig. 1. Each speaker codebook is composed of a number of codevectors. The goal of a mixture estimator is to search the possible codevectors of the speaker models to find two optimal codevectors, one from each speaker model, such that when mixed, they satisfy a minimum estimation error criterion comparable to the mixed signal. These two best codevectors are denoted by $\{r_{\text{opt}}, q_{\text{opt}}\}$ in Fig. 1.

By applying the sinusoidal parameter estimator in (1) to the mixed signal, we obtain $\mathbf{z} = \mathbf{V}_z^T \mathbf{a}_z$ where $\mathbf{V}_z$ is a Vandermonde matrix composed of $2L$ frequency vectors of size $N \times 1$ as $\mathbf{v}_{z,i} = [1 \quad e^{j\omega_{z,i}} \quad \dots e^{j\omega_{z,i}(N-1)}]^T$ defined by $\{\omega_{z,i}\}_{i=1}^L$, which is the set of sinusoidal frequencies obtained for the mixture at the ith band. We define $\boldsymbol{\alpha}_z = \{A_{z,i}\}_{i=1}^L$, $\boldsymbol{\omega}_z = \{\omega_{z,i}\}_{i=1}^L$and $\boldsymbol{\phi}_z = \{\phi_{z,i}\}_{i=1}^L$ denoting, respectively, the amplitude, frequency, and phase of the ith component for the mixed signal. We derive a mixture estimator based on the sinusoidal parameters of the underlying speakers and their mixture. The key idea is to project the mixture onto its sinusoidal subspace spanned by the columns of the parametric vector $[\boldsymbol{\alpha}_z, \boldsymbol{\omega}_z, \boldsymbol{\phi}_z]$ and to find a cost function to be minimized in the mixture

estimation stage. Based on (1), we define $P_k(e^{j\omega})$ as the power spectrum for the kth speaker at the ith band as [40]

$$P_k(e^{j\omega}) = \sigma_{k,i}^2 + A_{k,i}^2 \left[\delta(\omega - \omega_{k,i}) + \delta(\omega + \omega_{k,i})\right], \tag{6}$$

where we assumed that $e_k(n)$ is white at each ith frequency band and $\sigma_{k,i}^2$ denotes its corresponding variance, $\{\omega_{k,i}\}_{i=1}^{L}$ is the frequency set for the peaks retained for the kth speaker signal. A similar definition holds for the mixed signal, and we define the mixture power spectrum as $P_z(e^{j\omega})$. The frequencies in $\{\omega_{z,i}\}_{i=1}^{L}$ are formed by applying (1) on the mixed signals. Considering an appropriate window denoted by $W(e^{j\omega})$ to reduce the spectral leakage, the expected value for the periodogram for each signal spectrum is $E\{\hat{P}_k(e^{j\omega})\} = P_k(e^{j\omega}) * W(e^{j\omega})$ where $\hat{P}_k(e^{j\omega})$ is the *periodogram* for the kth speaker, $E\{\cdot\}$ denotes the expectation operator and $*$ is the convolution operator. We define a cost function as the squared error between the power spectra of the mixed signal and its estimate to be sampled only at sinusoidal peaks given by $\{\omega_{z,i}\}_{i=1}^{L}$. The expected value for the mixture estimation error at the ith band is

$$\begin{aligned} E\{\epsilon_i(e^{j\omega})\} &= E\{\hat{P}_z(e^{j\omega}) - \hat{P}_1(e^{j\omega}) - \hat{P}_2(e^{j\omega})\} \qquad (7)\\ &= \sigma_{\epsilon,i}^2 + A_{z,i}^2[W(e^{j(\omega-\omega_{z,i})}) + W(e^{j(\omega+\omega_{z,i})})] \\ &\quad - \sum_{k=1}^{2} A_{k,i}^2[W(e^{j(\omega-\omega_{k,i})}) + W(e^{j(\omega+\omega_{k,i})})] \qquad (8) \end{aligned}$$

We define $\sigma_{\epsilon,i}^2 = \sigma_{z,i}^2 - \sigma_{1,i}^2 - \sigma_{2,i}^2$ as the variance of the error. The expected mixture estimation error in (8) is sampled at mixture sinusoidal frequencies per ith band defined by the set $\{\omega_{z,i}\}_{i=1}^{L}$. Replacing ω by $\omega_{z,i}$ in (8) and ignoring the negative part of the spectrum for real speech signals, we get

$$\epsilon_i = A_{z,i}^2 - A_{1,i}^2 W(e^{j(\omega_{z,i}-\omega_{1,i})}) - A_{2,i}^2 W(e^{j(\omega_{z,i}-\omega_{2,i})}) \tag{9}$$

where ϵ_i captures the mixture estimation error defined between the original and the estimated mixture spectra at the ith band. $A_{1,i}$, $A_{2,i}$ and $A_{z,i}$ are the sinusoidal amplitude selected at the ith band for the first, the second and the mixed signals, respectively. The mixture approximation error gets close to zero when the underlying speaker spectra are highly harmonic. The mixture estimation error termed as d at a given frame is $d = \sum_{i=1}^{L} |\epsilon_i|$. The distortion function in (9) only calculates the mixture estimation error at the sinusoidal peaks obtained from the mixture. The proposed mixture estimation is targeted to find the optimal indices by searching the possible codevectors in speaker one codebook ($\mathbb{C}_1$) and speaker two codebook, $\mathbb{C}_2$) by solving the following minimization problem at each frame [41]

$$\{r_{opt}, q_{opt}\} = \arg \min_{\mathbb{C}_1 \times \mathbb{C}_2} d(\{A_{z,i}, \hat{A}_{1,i}^r, \hat{A}_{2,i}^q, \hat{\mathbf{v}}_{1,i}^r, \hat{\mathbf{v}}_{2,i}^q\}_{i=1}^{L}) \tag{10}$$

where r and q are the codebook indices for speaker codebook one and two, respectively, and we define $\mathbb{C}_1 \times \mathbb{C}_2 = \{r \in \mathbb{C}_1\} \times \{q \in \mathbb{C}_2\}$ as the space formed by the union of the spaces defined by $\mathbb{C}_1$ and $\mathbb{C}_2$. In the minimization formula given by (3), $\{r_{opt}, q_{opt}\}$ addresses $\{\hat{A}_{1,i}^{r_{opt}}, \hat{\mathbf{v}}_{1,i}^{r_{opt}}, \hat{A}_{2,i}^{q_{opt}}, \hat{\mathbf{v}}_{2,i}^{q_{opt}}\}_{i=1}^{L}$ which are the optimal sinusoidal codevectors selected from codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$, and $d(\cdot)$ is the 2D cost function based on the mixture approximation error in (9). The set $\{r, q\} \in [1, M]$ can be any possible states in the speaker models with M as the codebook size. At each frame, by minimizing $d_{r,q}$ in (3), we obtain two codevectors of the speaker models, which when combined, satisfy the minimization criterion in (3). The selected codebook indices are then used to reconstruct the two separated signals by means of a weighted overlap-add (OLA) procedure as shown in Fig. 1.

It is important to note that, in this paper, we use a full search to consider all possible states during minimization of the distortion function in (3). However, it is also possible to apply some cyclic minimizer or expectation maximization (EM)-like algorithms as an approximation to solve the minimization problem more computationally efficient, which is generally sub-optimal.

Our goal here is to find the set of unknowns denoted as $\{\hat{A}_{1,i}, \hat{A}_{2,i}, \hat{\mathbf{v}}_{1,i}, \hat{\mathbf{v}}_{2,i}\}_{i=1}^{L}$ by solving the following minimization problem per band

$$\underset{\hat{A}_{1,i}^{r}, \hat{A}_{2,i}^{q}, \hat{\mathbf{v}}_{1,i}^{r}, \hat{\mathbf{v}}_{2,i}^{q}}{\arg\min} \sum_{i=1}^{L} \|(A_{z,i}\mathbf{v}_{z,i} - \hat{A}_{1,i}^{r}\hat{\mathbf{v}}_{1,i}^{r} - \hat{A}_{2,i}^{q}\hat{\mathbf{v}}_{2,i}^{q})\|_2^2 \tag{11}$$

where $\hat{A}_{1,i}^{r}$ and $\hat{\mathbf{v}}_{1,i}^{r}$ are referred to the rth codevector selected from codebook $\mathbb{C}_1$, $\hat{A}_{2,i}^{q}$ and $\hat{\mathbf{v}}_{2,i}^{q}$ are referred to the qth codevector selected from codebook $\mathbb{C}_2$. By taking the Fourier transformation of the expression in (11), we get the mixture estimation in (9).

Assume that the modeling error in (9) is a zero-mean white, i.i.d. (independent and identically distributed over observations) with Gaussian noise with constant variance $\sigma_i^2 \neq 0$ at each frequency band i. Using an l_2-norm and applying band decomposition, one can show that the log-likelihood of all bands is

$$P = K - \frac{1}{2}\sum_{i=1}^{L} \frac{\|\mathbf{z}_i - \boldsymbol{\mu}_{\hat{z}_i}\|_2^2}{\sigma_i^2}, \tag{12}$$

where $\boldsymbol{\mu}_{\hat{z}_i}$ is the estimated mixed signal formed by combining the selected codewords of the speakers for the ith band and $K = -\frac{L}{2}\log(2\pi) - \sum_{i=1}^{L}\log\sigma_i$. Minimizing the likelihood of all bands using the sinusoidal estimator approximates the exact likelihood of all bands in (5). The minimization results in two sinusoids (one for each speaker) per band.

2.4 Training Split-VQ codebooks on sinusoidal parameters

We use split-VQ codebooks composed of sinusoidal amplitude and frequency vectors as speaker models. Here, we briefly explain the split-VQ codebook generation used in our proposed separation method. The extracted sinusoidal parameters: amplitude and frequency, each of dimension L are entered to the training stage. Following [42], we apply different distance measures to produce codebooks of amplitude and frequency, respectively. For the amplitude part of the kth speaker, we apply distance measure

$$d_A = \sum_{i=1}^{L} \left(\frac{A_{k,i}}{\|\boldsymbol{\alpha}_k\|_2^2} - \frac{\hat{A}_{k,i}}{\|\hat{\boldsymbol{\alpha}}_k\|_2^2} \right)^2, \tag{13}$$

where $\|.\|_2^2$ is the l_2-norm and $\hat{\boldsymbol{\alpha}}_k = \{\hat{A}_{k,i}\}_{i=1}^{L}$ is the coded amplitude codevector, with $\hat{A}_{k,i}$ as the coded amplitude for the sinusoidal peak selected at the ith band for the kth speaker. Let M_A be the codebook size for the amplitude part of our split-VQ codebook. After establishing M_A amplitude codevectors, we select frequency vectors that are closest in terms of their related amplitude vectors. Another VQ of a lower size is performed on these frequency candidates for each amplitude codeword. To produce frequency codevectors for the kth speaker, we apply the following distance measure

$$d_w(\mathbf{V}_k, \hat{\mathbf{V}}_k) = \sum_{i=1}^{L} w_{k,i} \|\mathbf{v}_{k,i} - \hat{\mathbf{v}}_{k,i}\|_2^2, \tag{14}$$

where $w_{k,i} = \frac{A_{k,i}}{\|\boldsymbol{\alpha}_k\|_2^2}$ is the energy normalized amplitude vector used for dynamic weighting of the Euclidean distance measure to make it proportional to the sinusoidal amplitude at the peak frequencies.

3 Experimental Results

3.1 Separation Scenario and Database

As a proof of concept, we evaluate the performance of the proposed method in SCSS and compare it with other benchmark methods. In our implementations we first focus on speaker-dependent scenario. Then, we relax this assumption by using gender-dependent codebooks as an intermediate scenario. The SSR is defined as the averaged ratio of the target speaker gain to the gain of the interfering signal. In our experiments, we swept the SSR level within the range [-18,18] dB. Then, the separation results are averaged at each SSR level over all pairs of test signals and quantified using PESQ [43] as objective measure and

MUSHRA [44] listening test as subjective evaluation. As benchmark methods, the separation result of the proposed method is compared with other conventionally used methods: MAX-VQ [15, 23, 32], the Wiener filtering [13, 30], and STFT-VQ [17, 18]. We also compare the separation results of the proposed method with those obtained by HMS [24, 25] and fusion method [16].

To evaluate the proposed separation algorithm, we used the database provided for SCSS in [45] consisting of 34 speakers each uttering 500 sentences. For our speaker-dependent scenario, we selected four speakers including two male (speakers 9 and 19) and two female speakers (4 and 23) from the database. We used 10 minutes of speech signals from each of the four speakers to train the speaker models. The sampling frequency was decreased from the original 25 kHz to 8 kHz. We analyzed the performance of the proposed mixture estimator for many mixture pairs to find the best values of these parameters. According to our results, throughout all experiments presented here, we used 50 sinusoidal peaks and a von Hann window of duration 32 ms with a frame-shift of 8 ms. For practical reasons, throughout the simulations presented here, the desired frequency range was set to [60,3850] Hz at a sampling frequency of 8 kHz.

For practical reasons and according to findings reported in [42], we have opted for 11 bits for amplitude and 3 bits for frequency. For a fair comparison and consistent with the results in [17], the same codebook size was chosen for the STFT codebooks. In the experiments, we assumed that the double-talk regions in the mixture are known *a priori*. We only focus on separating the mixed regions to report the performance of different mixture estimators, which are arguably also the most difficult part. We also assumed *a priori* knowledge of speaker identities and SSR level in the observed speech mixture.

3.2 Ideal Separation Scenario

To assess the performance of our proposed mixture estimator, we consider the ideal separation scenario as was done in [32]. In an ideal separation scenario, we assume that we have access to the original underlying speakers, and from their spectral vectors, we find the optimal codevectors based on their corresponding trained speaker codebook. We select two utterances of one male and one female and add them together at SSR= 0 dB to form a mixture. Fig. 2 depicts how the proposed mixture estimator works by minimizing the error at the sinusoidal peaks estimated from the mixture. The sinusoidal peaks in magnitude spectrum are shown for the original and estimated mixture in Fig. 2(a), as well as for each of the underlying single speaker signals in Fig. 2(b) and (c). From the mixture estimation error shown in Fig. 2(d), it is observed that the estimation error is reasonably low especially at sinusoidal frequencies of the mixture, explaining the high accuracy of the proposed mixture estimator.

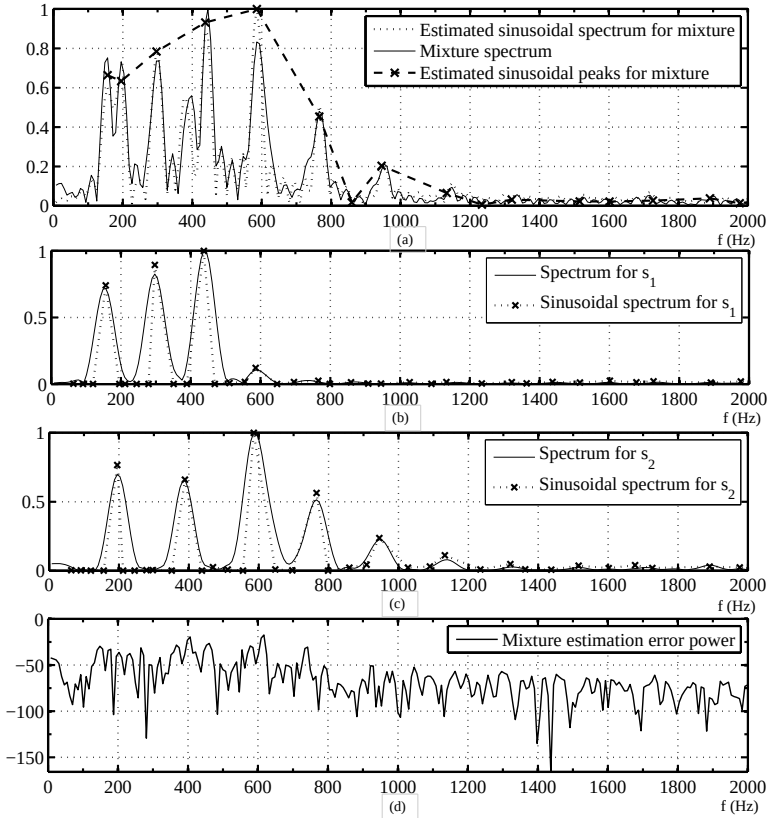

Fig. 2: Showing the magnitude spectrum for: (a) the original and estimated mixture, (b) speaker one, (c) speaker two and (d) mixture estimation error power in dB.

3.3 Evaluating Performance for Speaker-Dependent Case

We report the separation performance of the proposed method and compare it with respect to other benchmark methods. First, we consider speaker-dependent scenario where we assume that we have *a priori* knowledge of speaker identities. To this end, we randomly selected ten sentences from the test data of each speaker in order to forming the speech mixtures. The training and test sets were disjoint. Fig. 3 shows the PESQ scores of different separation methods versus SSR. To carefully assess the gap between methods, we also included the upper-bound for the separation performance achieved by the STFT [17] and split-VQ on the sinusoidals in [42]. The performance of the proposed method was compared to previous speaker-dependent methods. The methods we included in our simulations are binary mask, the Wiener filtering, and the STFT-based VQ methods. Each curve depicted in Fig. 3 is labeled with the related reference. Several results are inferred from Fig. 4: 1) according to the curves, the proposed method achieves a higher PESQ score compared to MAX-VQ and the Wiener filtering especially at low SSR levels, 2) it is observed that the proposed method achieves about 1 point improvement in PESQ score over the mask methods. The inferior performance of the mask methods can be further explained by the energetic masking effect of the dominant speaker at time-frequency cells [2, 27, 45]. The mixture estimation error observed in the mask methods is due to the fact that they originally filter out the competing speaker to recover a target signal and consequently lead to decoding errors while mapping vectors of the mixed signal into the codevectors in the codebooks of the underlying speaker in the mixture. Hence, using a log-max mixture estimator in a mask approach could result in the selection of wrong codevectors from the speaker models, and consequently, it leads to poorly filtered separated signals as reported in [23], 3) according to Fig. 3(a) and (b), the proposed method outperforms the STFT-based approach and its upper-bound separation performance. The significant degradation in performance caused by the STFT codebook-based method (denoted by STFT-CB), as compared to the proposed approach can be observed from the gap between the PESQ curves shown in Fig. 3(a) and (b). This agrees with the recent results reported in [17] stating that compared to mask methods, performing subband transformation on the STFT features could result in improvements in the perceived speech quality of the separated signals especially at low SSR levels, and 4) according to the curves shown in Fig. 3, the proposed method asymptotically reaches the upper-bound performance achieved by the split-VQ codebooks in [42].

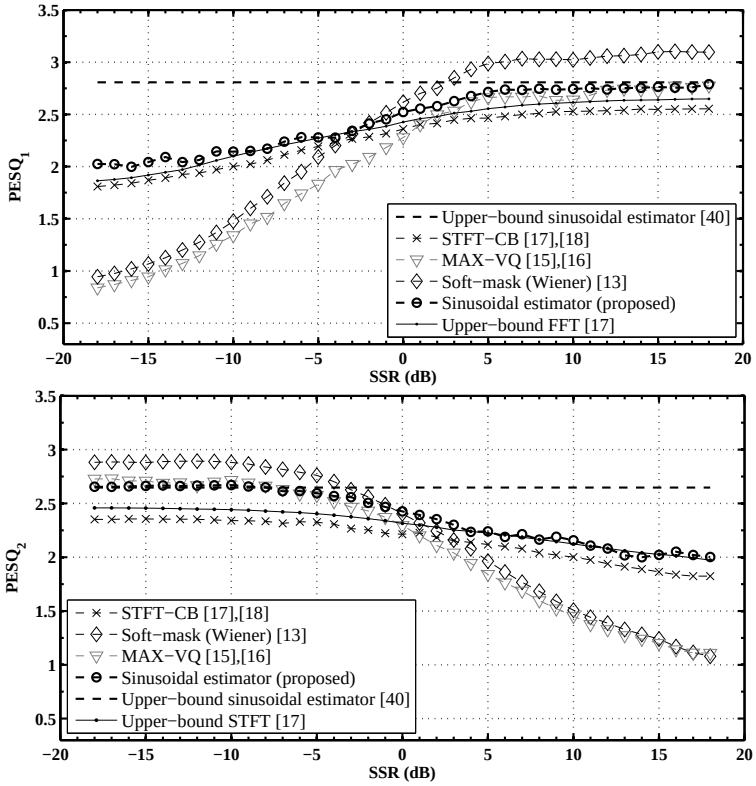

Fig. 3: Showing the separation results for speaker-dependent scenario for different methods in terms of PESQ score versus SSR levels for (a) speaker one and (b) speaker two [41].

Table C.1: Speaker labels used for training the gender-dependent models for male and female speakers.

Male	3	5	6	9	10	12	13	14	17	19
Female	4	7	8	11	15	16	21	22	23	24

3.4 Separation Results for gender-dependent Scenario

To relax the assumption of *a priori* knowledge of speaker identities, here, we study the separation results for gender-dependent scenario. As gender-dependent models, we selected 10 female and 10 male speakers each producing 35 s of speech signal. We trained a male speaker model using utterance from ten speakers and a female speaker model trained on ten female speakers. These two speaker models are gender-dependent considered as an intermediate scenario between speaker-dependent and speaker-independent. The speaker labels used for training our gender-dependent models are shown in Table F.1. To evaluate the separation performance we formed mixtures using fifteen utterances of speakers 29, 34 as female and 30, 32 as male speakers selected as our test speakers. The separation results were then averaged over the mixture pairs at different SSR levels and speakers. Fig. 4 illustrates the separation results obtained by different methods for gender-dependent scenario. Curves demonstrate the separation performance for each speaker in terms of SNR versus SSR. To assess the gap between different methods, we also included the upper-bound separation performance. From Fig. 4, it is concluded that compared to other methods, the proposed method shows a significant improvement for both speaker signals, especially at extreme SSR levels (both low and high).

It is important to note that the results shown in Figs. 3 and 4 can best be interpreted separately. According to the definition of SSR, high SSR means that speaker one is dominant in the mixture while a similar interpretation goes for the second speaker but for negative values of SSR. From Figs. 3 and 4, at high SSR levels, soft mask achieves a slightly higher PESQ score compared to our method. This can be explained because of the use of masks in soft-mask method which employs information directly from the mixed signal. Since at high SSR levels, target speaker (let speaker one) is more intelligible, then mask method achieves a higher PESQ score for this speaker compared to a model-based method since the latter employs no information directly coming from the mixture, but uses pre-trained speaker spectra for signal reconstruction. This observation can be further explained by noting the fundamental difference between mask and reconstruction-based methods while synthesizing the separated signals.

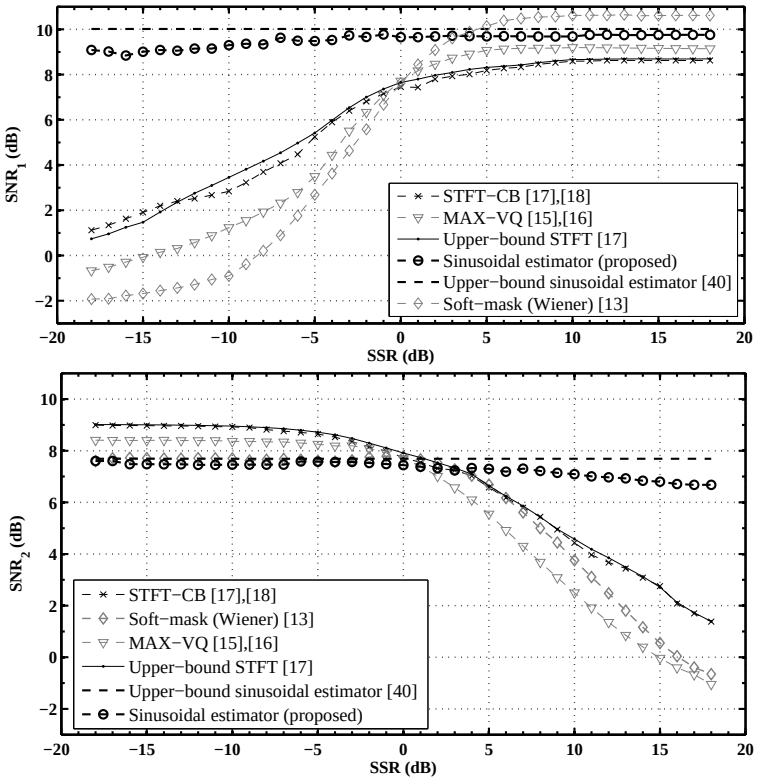

Fig. 4: Showing the separation results for gender-dependent scenario for different methods in terms of PESQ score versus SSR levels for (a) speaker one, and (b) speaker two.

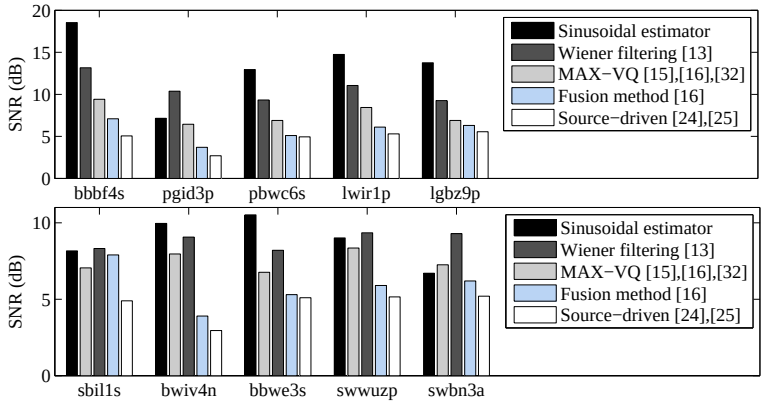

Fig. 5: Comparing the SNR results of the proposed method with MAX-VQ [15, 16, 32], Wiener filtering [13], STFT-VQ [18], source-driven [24, 25] and harmonic methods in [16] for speaker-dependent scenario.

3.5 Comparing the Separation Results with Harmonic Methods

We compare the separation performance of the proposed method in terms of SNR measure with source-driven in [24, 25] and fusion methods in [16] both based on pitch estimates of the underlying speakers in the observed speech mixture. These two methods serve as examples for source-driven and harmonic methods, respectively. To have a fair comparison, we used the same mixtures as described in Tables in [16] all formed at SSR=0 dB. Figures 5 and 6 show the SNR results measured in dB per mixture described on x-axis for speaker-dependent and speaker-independent scenarios, respectively. According to the results, it is observed that the proposed approach mostly achieves a higher score compared to source-driven and fusion methods in [16]. The improvement for speaker-independent scenario is lower but the proposed method still mostly outperforms other approaches including: Log-max, Wiener filtering, source-driven and fusion.

4 Subjective Evaluation

4.1 MUSHRA Test Setup

To assess the perceived speech quality of the separated output signals obtained by different methods, we conduct a subjective listening test by using the multi-stimulus test with hidden reference and anchors (MUSHRA test) as described

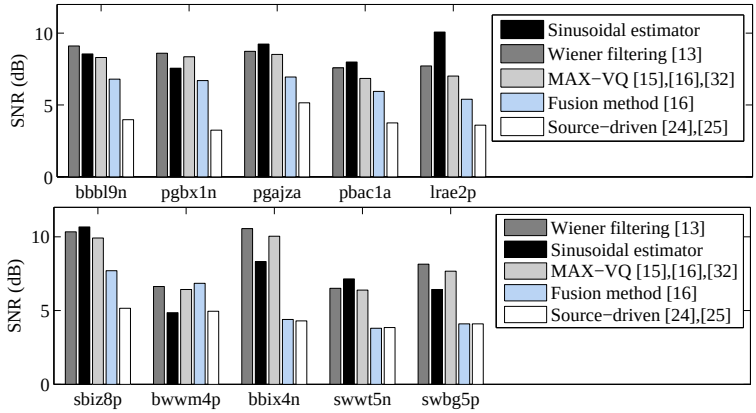

Fig. 6: Comparing the SNR results of the proposed method with MAX-VQ [15, 16, 32], Wiener filtering [13], STFT-VQ [18], source-driven [24, 25] and harmonic methods in [16] for speaker-independent scenario.

Table C.2: Labels of the methods used in MUSHRA test.

Excerpt	Separation method and SSR scenario
$\text{BMM}_{\text{SSR=0dB}}$	Binary mask at SSR=0dB
$\text{BMM}_{\text{SSR=-18dB}}$	Binary mask at SSR=-18dB
$\text{SIN}_{\text{SSR=0dB}}$	Proposed method at SSR=0dB
$\text{SIN}_{\text{SSR=-18dB}}$	Proposed method at SSR=-18dB
$\text{FFT}_{\text{SSR=0dB}}$	STFT-based VQ at SSR=0dB
$\text{FFT}_{\text{SSR=-18dB}}$	STFT-based VQ at SSR=-18dB

in ITU-R BS.1534-1 [44]. The MUSHRA test is a double blind test for the subjective assessment of intermediate quality level benefits obtained from displaying all stimuli at the same time. This enables the subjects to carry out simultaneous comparison between the methods directly. Seven untrained listeners participated in the test (the authors not included). The excerpts used in our listening test are shown in Table C.2, each indicating a separated signal at a specific SSR level. The experiments are conducted for both speaker-dependent and gender-dependent scenarios. Both MAX-VQ and STFT-based VQ methods were included as benchmarks for separation methods. All the played signals were monophonic of length 2 s and sampled at 8 kHz. Many more excerpts were used in our development phase, but the excerpts shown in Table C.2 are the ones that have been tested in our listening test. The excerpts consisted of the hidden reference denoted by HR and an anchor low-pass filtered at 2 kHz denoted by Anchor. The remaining six excerpts are the separated signals at different SSRs

shown in Table C.2. The hidden reference shows the known quality on the scale and is used to check the consistency of the responses of a subject during the listening test. A high score is expected at this point. The anchor point is included to enable comparisons between the different listening tests since it forms a simple but well-defined modification on the reference signal. Excerpts listed in Table C.2 were chosen and played for each subject. The listeners were asked to rank eight separated signals relative to a known reference on a scale of 0 to 100. By including different SSR levels, it is possible to assess any improvement observed in the synthesized speech quality of the proposed method compared to other methods. Further, the separation performance is evaluated for two SSRs.

4.2 Listening Test Results for Speaker-Dependent Scenario

We conducted the listening experiments on subjects in a silent room and a good sound quality audio, firewire interface, was used for digital to analog conversion. Moreover, we used a high quality headphone: AKG K240 MKII. The scores obtained from different methods were averaged over all listeners and excerpts. Fig. 7 depicts the mean opinion score (MOS) for the speaker-dependent scenario. Furthermore, the performance of individual excerpts can be observed by the numbers in the first row of Table C.3, which shows the results obtained by each clip[1]. For each entry, the first number is the averaged value over the scores obtained by seven listeners and the second number determines the confidence interval.

The results shown in Table C.3 are divided into two categories i.e. target and masker speakers. Odd columns show the results for the masked signal in the mixture explaining the low scores at SSR=-18 dB while the even columns show the results for the target speaker in the mixture. From Fig. 7, it is observed that the proposed method scores, on average, about 20 points higher than MAX-VQ, and more than 25 points higher than the STFT-based method. According to Fig. 7, no overlap exists between the confidence intervals of the proposed method and the other methods. Therefore, it can be concluded that the proposed method achieves statistically significant improvement by consistently enhancing the performance of the perceived speech quality for both target and interference separated signals especially at low SSRs. The proposed method achieves a slightly lower quality compared to those obtained by MAX-VQ and Wiener filtering. However, as indicated by the listening experiments, some of the separated outputs achieved by MAX-VQ were found suffering from severe crosstalk. Furthermore, listeners observed that in some cases the separated signals obtained by MAX-VQ were relatively poor compared to the reference signal. They observed that these methods suffer from the cross-talk phenomenon, mostly

[1]The mixed and separated wave files for different methods are downloadable from our webpage: http://kom.aau.dk/∼pmb/IEEE_Trans.htm

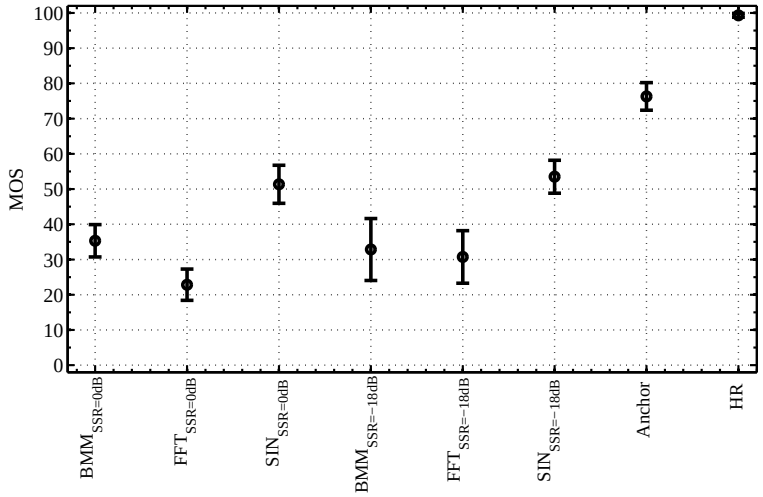

Fig. 7: Results of the MUSHRA listening test for the speaker-dependent scenario [41]. MOS scores for different separation methods over all excerpts and all listeners. Error bars indicate 95% confidence intervals.

while recovering masked signal, in which a portion of the other speaker signal exists in the separated output signal. This is mainly because a mask method applies a gain function to the mixed spectrum rather than finding a candidate from the codebooks. On the other hand, the proposed method produced artifacts in the separated signals often encountered in sinusoidal speech modeling especially in fricatives and sudden attacks [38]. However, the proposed method still outperforms the others by achieving, on average, 23 to 28 points, higher than the STFT-based method and 17 to 21 points higher than MAX-VQ. According to the listeners observations, the improvements brought about by the proposed method are perceived both as an increase in terms of speech signal quality and lower cross-talk. The tests also revealed that the separation performance of the mask methods (especially at 0 dB of SSR where their separation performance is often reported) does not necessarily produce the highest perceived quality for the separated signals. This can be observed by comparing the MOS results in Fig. 7 for $\text{BMM}_{\text{SSR}=0}$ and $\text{SIN}_{\text{SSR}=0}$ stating that the proposed method shows an advantage of 10 points in the resulting MOS compared to MAX-VQ. We also considered results shown in Table C.4 as the MOS results obtained by each listener averaged over eight clips defined in Table C.2. In gender-dependent scenario we only considered masked speaker output for subjective measurement at SSR=-18 dB while for speaker-dependent scenario we included both separated target speaker and masker speaker signals at SSR=-18 dB. By inspecting the

MOS results shown in Fig. 7 along with the results of the listening experiments in Table C.4, subjects often indicated that the signals related to the proposed method were close to the reference signal and showed a significant preference over other separated signals.

Table C.3: Results of the MUSHRA listening test for the speaker-dependent scenario. The MOS results obtained for each clip averaged over seven listeners are shown for different methods. For each case, the confidence interval is also included.

Excerpt\Clip	1	2	3	4	5	6	7	8
$BMM_{SSR=0dB}$	47.85±15.56	39.43±13.63	26.28±12.35	34.28±17.18	35.86±11.67	30.00±12.97	45.86±16.71	23.00±12.86
$BMM_{SSR=-18dB}$	3.85±2.68	61.57±15.31	2.43±1.47	62.86±13.04	1.71±0.65	57.71±24.03	2.00±0.69	70.57±18.31
$SIN_{SSR=0dB}$	81.14±13.39	50.86±22.82	51.00±14.87	47.43±15.69	41.43±14.08	36.57±7.16	47.14±16.46	55.28±12.54
$SIN_{SSR=-18dB}$	69.28±11.66	50.57±18.48	53.14±17.58	62.71±14.76	44.86±7.37	36.28±14.31	58.00±6.07	53.14±15.55
$FFT_{SSR=0dB}$	28.14±21.19	23.00±15.01	14.43±7.79	39.57±18.14	17.86±11.32	24.28±12.41	17.14±5.64	18.28±13.63
$FFT_{SSR=-18dB}$	6.85±3.75	51.00±10.75	4.71±2.51	60.86±16.61	4.28±2.60	48.28±9.99	2.86±1.35	67.00±13.63
HR	99.14±1.59	99.14±1.94	98.28±3.88	99.86±0.32	99.86±0.32	98.71±2.02	100.00±0.00	99.57±0.97
Anchor 2 kHz	76.14±13.45	63.86±18.85	78.57±8.84	68.86±18.09	82.14±5.84	77.14±11.89	80.28±10.19	83.28±5.93

Table C.4: Results of the MUSHRA listening test for the speaker-dependent scenario. The MOS results obtained by each listener averaged over eight clips are shown for different methods. For each case, the confidence interval is also included.

Excerpt\Listener	1	2	3	4	5	6	7
$BMM_{SSR=0dB}$	51.00±9.09	33.37±13.07	39.87±9.86	24.37±10.95	26.62±11.46	48.62±17.53	23.37±6.19
$BMM_{SSR=-18dB}$	39.12±31.37	33.87±26.06	40.12±32.88	19.12±16.12	31.87±24.85	43.12±34.77	22.62±18.23
$SIN_{SSR=0dB}$	66.12±10.71	47.50±16.92	58.25±15.82	47.00±21.93	34.87±10.48	55.12±16.81	50.62±12.31
$SIN_{SSR=-18dB}$	64.00±10.08	48.12±17.71	59.75±12.49	50.50±15.58	39.00±8.43	68.50±8.96	44.62±9.79
$FFT_{SSR=0dB}$	31.50±5.53	21.12±12.09	14.37±9.16	21.62±16.18	10.12±3.99	44.87±15.72	16.25±5.59
$FFT_{SSR=-18dB}$	34.12±26.97	32.12±23.87	34.37±28.56	25.37±17.48	31±20.40	31.75±26.07	27±20.09
HR	100.00±0.00	100.00±0.00	99.87±0.00	99.12±1.64	100±0.00	96.25±3.12	100±0.00
Anchor 2 kHz	72.00±7.26	73.12±7.16	96.00±2.74	66.25±14.21	79±5.94	68.75±18.02	80±2.39

4.3 Listening Test Results for Gender-dependent Scenario

Relaxing the *a priori* knowledge of speaker identities, we report the MOS results for the MUSHRA listening test in a gender-dependent scenario shown in Fig. 8. According to the results depicted in Fig. 8, since no overlap exists between the proposed method and the benchmark methods, it can be concluded that the proposed method can achieve statistically significant improvement compared to other methods and consistently enhances the performance of the synthesized speech quality for both target and interference separated signals. It is observed that at 0 dB of SSR the proposed method achieves greater improvement compared to other methods. From Fig. 8, it is observed that in extreme cases (low/high SSRs), the proposed method improves the perceived speech quality of the separated signals. The numbers in the first row of Table C.5 show the results obtained by each clip for the gender-dependent scenario. We also considered results shown in Table C.6 as the MOS results obtained by each listener

averaged over eight clips for gender-dependent scenario. It is observed that, for the gender-dependent scenario, the proposed method consistently outperforms the others in most of the cases.

By comparing the MOS results shown in Figs. 7 and 8 along with Tables C.3-C.6, we observe that the proposed method achieves a higher score both in speaker-dependent and gender-dependent scenarios. The MOS results for gender-dependent scenario are lower than those obtained for speaker-dependent scenarios. At low SSR levels, both mask and STFT-based methods show inferior performance especially in gender-dependent scenario. In contrast, the proposed method shows a shorter confidence interval both in speaker-dependent and gender-dependent scenario. By comparing the MOS results depicted in Figs. 7 and 8, it is observed that the relative difference between the methods in speaker-dependent and gender-dependent scenarios shows a remarkably similar pattern of overall performance.

Table C.5: Results of the MUSHRA listening test for gender-dependent scenario. The MOS results obtained for each clip averaged over seven listeners are shown for different methods. For each case, the confidence interval is also included.

Excerpt\Clip	1	2	3	4	5	6	7	8
$BMM_{SSR=0dB}$	32.57±8.99	29.71±8.56	29.14±18.10	18.86±10.03	24.86±7.25	25.43±9.78	22.86±12.00	28.14±11.79
$BMM_{SSR=-18dB}$	8.00±6.61	6.00±6.21	4.57±3.59	3.00±1.98	4.28±2.69	3.43±2.75	6.86±4.41	20.71±6.07
$SIN_{SSR=0dB}$	52.00±12.49	36.43±9.06	33.71±6.09	24.28±11.27	35.00±15.14	28.57±6.15	38.71±11.63	34.14±13.92
$SIN_{SSR=-18dB}$	35.57±13.85	33.86±13.05	32.00±5.61	17.00±6.09	41.00±12.33	28.71±7.36	38.14±10.83	33.86±10.16
$FFT_{SSR=0dB}$	34.86±8.11	28.86±13.07	28.14±9.37	16.57±3.24	32.00±13.96	27.57±9.77	29.71±10.14	19.71±11.76
$FFT_{SSR=-18dB}$	20.57±13.01	40.14±9.72	13.00±8.42	7.00±2.61	14.71±8.24	10.28±4.24	13.71±4.68	5.86±4.46
HR	97.28±3.98	97.28±3.98	97.28±3.48	97.71±3.48	97.71±4.92	97.00±3.49	99.14±1.94	99.14±1.94
Anchor 2 kHz	63.43±9.27	61.28±9.98	67.71±14.47	66.86±11.04	66.71±7.00	69.14±11.57	67.71±11.85	69.71±10.23

Table C.6: Results of the MUSHRA listening test for the gender-dependent scenario. The MOS results obtained by each listener averaged over eight clips are shown for different methods. For each case, the confidence interval is also included.

Excerpt\Listener	1	2	3	4	5	6	7
$BMM_{SSR=0dB}$	19.88±6.14	26.63±5.24	42.13±10.91	18.50±7.02	18.88±4.95	21.00±7.36	38.13±9.56
$BMM_{SSR=-18dB}$	4.75±5.89	4.50±3.39	14.38±6.03	4.25±2.49	4.38±5.72	4.38±3.55	13.13±6.59
$SIN_{SSR=0dB}$	36.13±6.94	32.63±10.11	47.38±11.93	19.75±9.52	31.38±8.62	33.38±7.99	46.88±9.56
$SIN_{SSR=-18dB}$	32.63±7.35	29.88±8.68	43.88±11.41	22.00±4.20	24.63±8.44	31.50±7.58	43.13±11.82
$FFT_{SSR=0dB}$	21.25±6.95	18.75±7.57	37.13±7.93	20.75±8.11	27.25±10.36	25.88±8.38	37.50±10.86
$FFT_{SSR=-18dB}$	11.37±7.68	10.38±5.82	26.38±12.39	11.63±8.77	13.00±7.12	12.88±11.27	24.00±12.73
HR	93.25±1.09	100.00±0	100.00±0	100.00±0	100.00±0	100.00±0	91.88±4.15
Anchor 2 kHz	71.50±3.49	53.13±3.70	85.88±6.67	62.75±3.15	52.13±3.77	70.00±5.07	70.63±3.26

5 Discussion

In previous separation methods based on either harmonic modeling [16, 24, 25, 32, 46–48] or CASA [20–22], the speech perceived quality for separated signals was directly determined by the accuracy of the multi-pitch estimator. However,

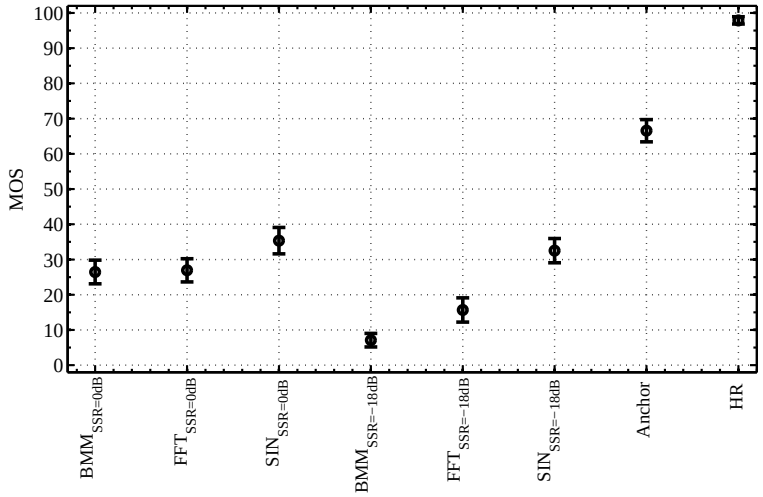

Fig. 8: Results of the MUSHRA listening test for the gender-dependent scenario. MOS scores for different separation methods over all excerpts and all listeners. Error bars indicate 95% confidence intervals.

due to energetic masking [2], the pitch detection accuracy of most of the existing pitch estimators, degrades significantly, especially at low SSRs [26, 28]. Hence, the proposed method offers an attractive candidate for SCSS or similar enhancement scenarios where pitch estimation with high accuracy is either rather erroneous [26, 28] or difficult because of the energetic masking [2, 27]. In addition, it was shown in [46] that a pitch-based method is not capable of attaining the same level of enhancement compared to a system based on sinusoidal frequencies. We confirmed this by comparing the separation performance of the proposed method with source-driven in [24, 25] and fusion methods in [16]. These motivate us to present a separation strategy independent of pitch estimates, in this paper. The sinusoidal parameter estimation taken in this work leads to a high frequency resolution at low frequencies, reflecting the pitch harmonic structure of each speaker signal and their mixture.

The proposed approach, like other well-known sinusoidal modeling methods in [46–48], has a major limitation in the failure to deal with unvoiced segments in a consistent manner. The methods in [46–48] were all suggested and limited by utterances composed of vocalic mixtures. Additionally, the sinusoidal modeling we used in this work is originally like the one described in [38] proposing that if we sample the spectrum of unvoiced speech with rate equal to 100 Hz, no perceivable degradation are observed in the synthesized speech signal at least from perceptually point of view. As a future work, it is possible to consider a more

complex modeling of speech and jointly estimating sinusoidal model parameters and voicing states of the two underlying signals.

The study in [46] reported the problems related to the frequency resolution of the discrete Fourier transform especially when two sinusoids related to different fundamental frequencies are arbitrary close to each other. As shown in [46, 47], the solution leads to singular ill-conditioned matrix as the frequency of one speaker close to the frequency of the other speaker and the problem is only solvable if two pitch frequencies and their integer multiples are not overlapping and are well separated; a condition which is often not met when two speech sources exist in the scene. This problem is equal to extracting two unknowns (two frequencies) from single observation (mixture frequency). In order to deal with such ambiguity, [46] suggested monitoring the spacing between neighboring frequencies and using a multi-frame interpolation procedure. However, in this work, we suggest testing all possible combinations of codevectors selected from underlying speakers' codebooks. This solution guarantees leading into the minimal error in the non-linear cost function. The work in [46, 48] only considered enhancing the target speech while current work addresses the more challenging problem of separating both speaker signals from their observed mixture. More specifically, in [46] the interference was suppressed while changing the interference speech to noise.

The present sinusoidal mixture estimator ignores the cross term components and phase differences which, in some situations, play a critical role and can change the position of peaks completely. This happens when the sinusoidal peaks of the underlying speakers get closer than 25 Hz. In such situations, the accuracy of the sinusoidal mixture estimator is limited but still finds the two states of the two speaker models (sinusoidal coders), which when combined, will best describe the mixture spectrum at certain frequencies (estimated from the mixture spectrum per bands).

The proposed technique uses pre-trained frequency codevectors based on peaks which makes the system more speaker-dependent. According to our simulations, the proposed method also led to good results for gender-dependent scenario which addresses an intermediate scenario. The more interesting speaker-independent scenario, most likely can be addressed by combining a speaker identification module with current separation system as reported in [49].

The present work considers the mixture scenario composed of two speaker signals. For mixtures with more than two speakers, it is possible to employ an EM-like algorithm in which for each speaker we update the signal parameters of one speaker at a time and then use these parameters in another searching scheme required for finding the optimal states of other two speakers' states. Separating mixtures of more than two speakers is an open problem and we have considered that as a potential future work.

The separation approach presented in this work neglected room reverber-

ation and echoes as well as background noise which exist in a real recording scenario. A deverberation approach [50, 51] together with a noise-suppression module can be integrated to each other, in order to mitigate the reverberation and background noise problem for achieving a robust speech separation system in a practical scenario. As an example, [51] proposed to suppresses noise components by spectral subtraction method, followed by a deverberation module applied to the noise-suppressed signal. In this way, it is possible to dereverberate the received echoic signal as well as to reduce background noise from the corrupted signal recorded by one microphone, and then apply our separation approach to the enhanced mixed signal.

By assuming *a priori* knowledge of double-talk regions in a given mixed signal, we apply the 2D search only to mixed frames to find the optimal states of the underlying speaker models (codebooks). For the single-talk regions, we simply re-synthesize the single-talk speaker signals according to the corresponding speaker codebooks. It should be noted that, the quantitative performance reported in our experiments are for the entire utterances.

The proposed approach cuts the computational cost in separation by substituting STFT feature vectors with sinusoidal peaks. We conducted simulations to quantify the computational complexity of the proposed method for ten 2-second mixtures. We observed that the STFT-VQ approach, used as our benchmark, took in average 26.71 s for separating each frame while the proposed one required 5.55 s. Hence, the proposed approach leads to approximately 5 times less computation time.

The upper-bound separation results presented here confirmed recent findings in [42], where it was demonstrated that by applying the split-VQ codebooks composed of sinusoidal parameters, it is possible to achieve a better quantization performance in terms of the re-synthesized speech quality compared to the conventionally used STFT or its logarithm as the selected feature vectors. This agrees with the conclusion in [31] stating that the ultimate quality of model-based speech enhancement system is upper-bounded by the performance of the coder used. Similarly in SCSS, the selected feature type along with the statistical model determines the separation upper-bound performance. Therefore, to achieve an acceptable separation upper-bound, the selected feature type for SCSS is required to perform a high quantization performance that is in agreement with the results reported in [17, 18, 42]. It was shown in [17] that by applying a subband perceptually weighted transform on the STFT vectors, it is possible to achieve improvements in the perceptual quality of the recovered signals especially at low SSRs. Similarly, in this work we observed that by changing STFT features with sinusoidal parameters, it is possible to achieve improvements in the separation performance.

We note that the method can also be generalized into speech enhancement in highly colored noise scenarios including babble or harmonic noise [1–7]. In

such scenarios, the mixed signal includes less harmonics which makes the separation task rather difficult. As a future work, the proposed method is expected to be appropriately applied to speech enhancement scenarios with highly colored noise. The proposed method in this paper offers an attractive candidate similar to the weighted codebook-mapping (WCBM) in [52], as an effective tool for speech enhancement. The WCBM in [52], however, was based on harmonic plus noise model (HNM) feature parameters that require voicing estimation and pitch. In contrast, the proposed method in this research is independent of pitch estimates and benefits from the advantages inherited from modified sinusoidal features, split-VQ codebooks and sinusoidal mixture estimator presented in this work.

6 Conclusions

In this paper, we presented new results on single-channel speech separation and also proposed a new method based on sinusoidal parameters. In our proposed method, we suggested to use a mixture estimator in the sinusoidal domain targeted to find the optimal sinusoidal codevectors selected from speaker codebooks that, when combined, best describe the observed mixed signal in each frame. The key idea in the proposed method is to separate the signals by mapping their mixture frames onto the joint subspaces of the sources and then compute the parts that fall in each subspace. We studied the performance of the proposed method and compared its results with those obtained by previous SCSS methods. Through extensive simulations, and by comparison to other methods, it was observed that the proposed method leads to rather good re-synthesized speech quality as well as lower undesirable cross-talk for both target and interference signals. It was also concluded that minimization at sinusoidal frequencies of the mixed signal, used in the proposed mixture estimator, makes significant improvement compared to both mask approach (log-max and Wiener filtering) and STFT-based VQ approaches. To assess the improvements made by the proposed method, we used PESQ as objective measure and MUSHRA listening tests as subjective evaluation for both speaker-dependent and gender-dependent scenarios. It was observed that the proposed method achieved a higher score compared to other separation methods. In addition, it was observed that by increasing the signal-to-signal ratio, the proposed method asymptotically reaches the upper-bound separation performance (ideal separation scenario). According to the MUSHRA listening tests, the perceived speech quality of the proposed method was the highest both in speaker-dependent and gender-dependent scenarios. Finally, compared to other methods, the proposed method achieved lower cross-talk and was mostly preferred by the listeners.

References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[2] S. Srinivasan and D. Wang, "A model for multitalker speech perception," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3213–3224, Nov. 2008.

[3] H. Sameti, H. Sheikhzadeh, L. Deng, and R. 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, Sept. 1998.

[4] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[5] ——, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[6] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[8] S. Jensen, P. Hansen, S. Hansen, and J. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[9] P. Hansen and S. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, Oct. 2005.

[10] ——, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 1, p. 24, March 2007.

[11] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Journal on Trends in Amplification*, vol. 12, no. 4, pp. 332–353, Dec. 2008.

[12] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[13] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[14] S. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 793–799.

[15] ——, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, 2003, pp. 1009–1012.

[16] M. Radfar, R. Dansereau, and A. Sayadiyan, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Elsevier speech communication*, vol. 49, no. 6, pp. 464 – 476, Jun. 2007.

[17] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single-channel speech separation performance in transform-domain," *Journal of Zhejiang University-SCIENCE C, Computers and Electronics*, vol. 11, no. 3, pp. 160–174, Jan. 2010.

[18] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, pp. 957–960, May 2006.

[19] M. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modeling for single-channel acoustic source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, May 2004, pp. 641–644.

[20] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[21] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77 – 93, 2010.

[22] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sept. 2004.

[23] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[24] B. Hanson and D. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1984, pp. 65–68.

[25] J. Naylor and S. Boll, "Techniques for suppression of an interfering talker in co-channel speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 205 – 208.

[26] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Apr. 1993, pp. 728–731.

[27] J. Barker, M. Ning, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 94 – 111, 2010.

[28] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 929–932.

[29] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[30] M. Radfar and R. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[31] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[32] M. Radfar, R. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 1, p. 15, March 2007.

[33] J. Ming, T. Hazen, and J. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67 – 76, Jan. 2010.

[34] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, Sept. 2002.

[35] Y. Shao and D. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 289–298, Jan. 2006.

[36] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. Synthesis Lectures on Speech and Audio Processing: Morgan and Claypool publishers, 2009.

[37] C. Moore, *An introduction to the psychology of hearing*. San Diego, CA: Academic Press, 2003.

[38] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[39] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "FDMSM robust signal representation for speech mixtures and noise corrupted audio signals," *IEICE Electronics Express*, vol. 6, no. 15, pp. 1077–1083, 2009.

[40] H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: John Wiley and Sons, 1996.

[41] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Improved single-channel speech separation using sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 21–24.

[42] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[43] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," iTU-T Rec. P.862, 2001.

[44] "Method for the subjective assessment of intermediate quality level of coding systems." iTU-R BS.1534-1, 2003.

[45] M. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[46] T. Quatieri and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 1, pp. 56 –69, Jan. 1990.

[47] F. Silva and L. Almeida, "Speech separation by means of stationary least-squares harmonic estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 809 –812 vol.2.

[48] T. Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," in *106th Audio Engineering Society Conv.*, 1999.

[49] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4430–4433.

[50] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[51] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, pp. 817–820.

[52] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1194–1203, May 2007.

Paper D

Joint Single-Channel Speech Separation and Speaker Identification

P. Mowlaee, R. Saeidi, Z. -H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen

This paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4430–4433, 2010.

The layout has been revised.

Abstract

In this paper, we propose a closed loop system to improve the performance of single-channel speech separation in a speaker independent scenario. The system is composed of two interconnected blocks: a separation block and a speaker identification block. The improvement is accomplished by incorporating the speaker identities found by the speaker identification block as additional information for the separation block, which converts the speaker-independent separation problem to a speaker-dependent one where the speaker codebooks are known. Simulation results show that the closed loop system enhances the quality of the separated output signals. To assess the improvements, the results are reported in terms of PESQ for both target and masked signals.

1 Introduction

A reliable and efficient speech separation system is desirable for many audio or speech applications where the best performance of the applications is only achieved when the signals are clean. Single-channel speech separation (SCSS) has been introduced as an ill-conditioned problem where we are required to estimate two unknown signals from one given mixture. Various statistical models have been adopted to solve the SCSS problem [1–3] where good results are only obtained under certain conditions like having *a priori* knowledge of speaker identities in the given mixture. This unrealistic but effective assumption is made to ease the SCSS problem. In practice, however, the speaker identities in the mixture are not known *a priori*. As a consequence, it is desired to design a model-based speaker independent SCSS system.

In [4], a system was proposed to capture speaker identities for enabling speaker dependent separation based on a Computationally Auditory Scene Analysis (CASA) framework on a given mixture of unknown speakers. Their system identifies the speakers identities based on Gaussian mixture models (GMM) and employed a pitch dependent method to re-synthesize the target speaker signal. In [5], the separation system used max approximation based on log-spectra of the underlying speaker signals along with Algonquin as the separation engine. In both systems reported in [4], [5], the speaker identification performance was successful almost in all trials, and system performance was assessed in terms of speech recognition word error rate. Most of previously proposed methods in speech separation literature were focused on speech recognition accuracy and not on re-synthesized speech quality of the separated speaker signals themselves.

In this paper, we consider a novel way of joining a speech separation system and a speaker identification system. The key idea is to put these systems (blocks) into a closed loop and send feedbacks from each system to another to

add more information to solve the SCSS problem. Each block could be viewed as a preprocessor for the other. The proposed approach benefits from the high separability of model-based speaker dependent separation methods and is also able to separate the speaker signals, without knowing their identities (i.e. generality). To assess the quality of the separated signals we report the separation results in terms of Perceptual Evaluation of Speech Quality (PESQ) [6] as our objective measure. We evaluate the performance of the system using a database consisting of 100 mixed speech signals with signal to signal ratio (SSR) ranging from -9 dB to +9 dB. The results show that the proposed approach significantly outperforms the technique that applies a single trained model for each gender (speaker independent case).

The paper is organized as follows. In Section 2 we explain the speech separation block and Section 3 presents the speaker identification block. In Section 4 we present the proposed joint speech separation speaker identification system. In Section 5 we present the simulation results and finally Section 6 concludes on the work.

2 Single-channel Separation System

The separation system transforms speaker signals from the DFT domain into modified sinusoidal features composed of amplitude, frequency and phase vectors. Each frame is modeled by using the sinusoidal model similar to [7] and we have $\mathbf{s} = \mathbf{V}^T\mathbf{a}$ where $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_M]^T$ is an $M \times N$ Vandermonde matrix whose rows are $\mathbf{v}_i$ defined as $\mathbf{v}_i = [1 \quad e^{j\omega_i} \quad \dots \quad e^{j\omega_i(N-1)}]^T$ with $i \in [1, M]$ as the sinusoidal frequency vector of dimension $N \times 1$, and ω_i is the frequency of the selected peak at the ith band, $\mathbf{s}$ is the time frame of the speaker signal by using the sinusoidal model, and $\mathbf{a} = [a_1 \dots a_M]^T$ is a $M \times 1$ complex sinusoidal amplitude vector found at a frame. The estimation process for obtaining the sinusoidal parameters is already described in [8]. We select only one peak with the largest amplitude per Mel scale subband.

In the core of the separation system we use a sinusoidal mixture estimator to find the most likely states of the composite sources of the underlying speakers. The performance of the proposed sinusoidal estimator is studied in [9]. In the following we explain how the separation system works. The mixed signal can be represented as $\mathbf{z} = \mathbf{V}_z^T\mathbf{a}_z$ where $\mathbf{V}_z$ is a Vandermonde matrix composed of M frequency vectors of $N \times 1$ with $\mathbf{v}_{z,i} = [1 \quad e^{j\omega_{z,i}} \quad \dots e^{j\omega_{z,i}(N-1)}]^T$ composed of set of sinusoidal frequency peaks for the mixture at the ith band denoted by $\omega_{z,i}$. The sinusoidal mixture estimator minimizes a cost function defined in the sinusoidal space in order to find the best indices from the speakers source models. We define the power spectrum for the selected sinusoid at at the ith

band as

$$P(e^{j\omega}) = \sigma_i^2 + A_i^2 \left[\delta(\omega - \omega_i) + \delta(\omega + \omega_i)\right] \quad , \tag{1}$$

where we define ω_i as the frequency of the selected sinusoid peak at the ith band which can be substituted with $\omega_{k,i}$ where $k = 1, 2$ indicates speaker one and two. Similarly to (1), we define $P_z(e^{j\omega_z})$ as the mixture power spectrum. We consider the squared error between the power spectra of the estimated and given mixture as our cost function for sinusoidal mixture estimator. This cost function is only sampled at sinusoidal peaks indicated by ω_z. The expected value for the periodogram for each signal spectrum is then given by $E\{\hat{P}(e^{j\omega})\} = P(e^{j\omega}) * W(e^{j\omega})$ with $E\{.\}$ as the expectation operator. At the ith band, the expected value of the mixture approximation error is given by

$$\begin{aligned} E\{\epsilon_i(e^{j\omega})\} &= E\{\hat{P}_z(e^{j\omega}) - \hat{P}_1(e^{j\omega}) - \hat{P}_2(e^{j\omega})\} \qquad (2)\\ &= \sigma_{\epsilon,i}^2 + A_{z,i}^2[W(e^{j(\omega-\omega_{z,i})}) + W(e^{j(\omega+\omega_{z,i})})] \\ &\quad - \sum_{k=1}^{2} A_{k,i}^2[W(e^{j(\omega-\omega_{k,i})}) + W(e^{j(\omega+\omega_{k,i})})], \qquad (3) \end{aligned}$$

with $\sigma_{\epsilon,i}^2 = \sigma_{z,i}^2 - \sigma_{1,i}^2 - \sigma_{2,i}^2$ as the variance of the error. We replace ω by $\omega_{z,i}$ to only sample this cost function at sinusoidal peaks of the mixture and we obtain

$$|\epsilon|_i^2 = \left| A_{z,i}^2 - A_{1,i}^2 W(e^{j(\omega_{z,i}-\omega_{1,i})}) - A_{2,i}^2 W(e^{j(\omega_{z,i}-\omega_{2,i})}) \right|^2 , \tag{4}$$

where $A_{1,i}$,$A_{2,i}$ and $A_{z,i}$ are the first, second and the mixture sinusoidal amplitude selected at the ith band. The mixture estimation error indicated by d is $d = \sum_{i=1}^{\hat{M}} \epsilon_i^2$. To find the most likely indices, it is required to search for among all possible states of the composite sources denoted by $\{q^*, t^*\}$. This index finding can be considered as the following minimization problem

$$\{q^*, t^*\} = \arg\min_{q,t} d_{q,t} \quad , \tag{5}$$

where q, t can be any possible state included in the composite source models and $d_{q,t}$ the cost function. At each frame, by solving this minimization problem we find two states of the speaker models (here split-VQ codebooks) that when combined best fit the given mixed signal. The source models used in our separation system are split-VQ on sinusoidal parameters as proposed in [8]. The source models are divided into amplitude and frequency part in a tree-like structure. Each entry of such source model is composed of a sinusoidal amplitude vector and several sinusoidal frequency vectors as its candidates. The selected codebook indices are then sent to a weighted overlap-add (OLA) block to reconstruct the separated signals.

3 Speaker Identification System

A GMM based framework is a common baseline system in speaker recognition applications [10]. Such a system is normally used as a reference when one needs to evaluate the effectiveness of novel algorithms or modeling approaches [11]. The GMM is a statistical approach for text-independent speaker recognition with a high computational load during the test phase. A popular method for training GMMs is based on the maximum-likelihood (ML) criterion, which has been shown to outperform several other existing techniques. In state-of-the-art systems, speaker-dependent GMMs are derived from a speaker-independent universal background model (UBM) by adapting the UBM components with maximum a posteriori (MAP) adaptation using speakers personal training data [12]. This method constructs a natural association between the UBM and the speaker models. For each UBM Gaussian component there is a corresponding adapted component in the speakers GMM. In the verification phase each test vector is scored against all UBM Gaussian components, and a small number of the best-scoring components in the corresponding speaker dependent adapted GMM are chosen. The decision score is computed as the Log Likelihood Ratio (LLR) of the speaker GMM and the UBM scores. Under the assumption of independent feature vectors, the log likelihood of a model λ for a sequence of T feature vectors, $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ is computed as follows

$$L(\lambda) = \frac{1}{T}\sum_{i=1}^{T}\log[p(\mathbf{x}_t|\lambda)] \quad , \tag{6}$$

where the mixture density used for the likelihood function is [12]

$$p(\mathbf{x}_t|\lambda) = \sum_{i=1}^{\hat{M}} w_i p_i(\mathbf{x}_t) \tag{7}$$

The density is a weighted linear combination of $\hat{M}$ unimodal Gaussian densities $p_i(\mathbf{x}_t)$, each parameterized by a $D \times 1$ mean vector $\boldsymbol{\mu}_i$ and a $D \times D$ covariance matrix Σ_i where D refers to the dimensionality of feature vector. Here $p_i(\mathbf{x}_t)$ is defined as

$$p_i(\mathbf{x}_t) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}}\exp\left[-\frac{(\mathbf{x}_t-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}_t-\mu_i)}{2}\right] \tag{8}$$

The mixture weights $\mathbf{w}_i$ further satisfy the constraint $\sum_{i=1}^{\hat{M}} w_i = 1$.

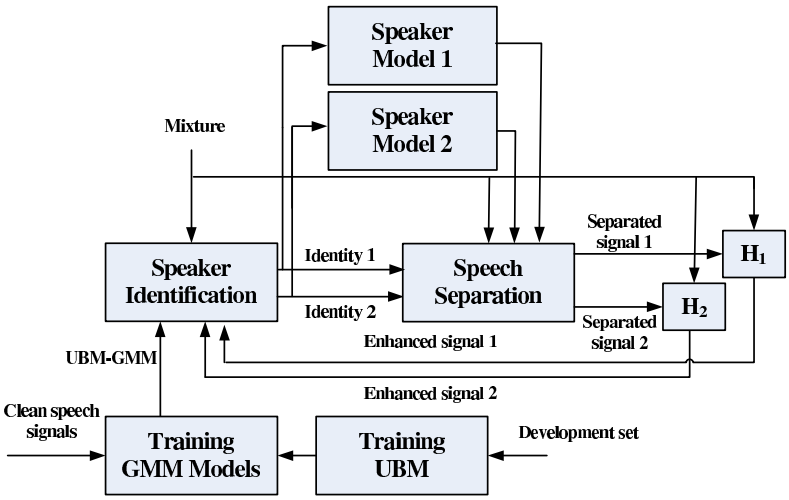

Fig. 1: Joint speaker identification and speech separation block diagram.

4 Joint Speaker Separation-Identification

Fig. 1 shows the block diagram of the proposed joint processor. The speech mixture is input to the speaker identification block where two identities for the underlying speakers are found. These speaker identities are then used to select the related speaker models and are applied in the speech separation block. The speech separation based on the speaker models, results in two separated signals. We then apply a filtering on each separated signal to enhance it. The filtering blocks are denoted by H_1 and H_2 shown in Fig. 1. The enhanced signals are fedback to the speaker identification block in order to achieve a more accurate speaker identity in one closed loop. The filtering method used here is similar to the binary mask approach with a hard decision but in sinusoidal domain. We call this filtering sinusoidal binary masks as proposed in [13]. Define k as the frequency bin index. At each frequency band we use the sinusoidal peaks of the underlying speaker signals to establish a sinusoidal binary mask defined as

$$H_1(\omega_k) = \begin{cases} 1 & \text{if} \quad A_{1,k} \geq A_{2,k} \\ 0 & \text{if} \quad A_{1,k} < A_{2,k} \end{cases} , \tag{9}$$

where ω_k denotes the kth frequency component. We also define $H_2(\omega_k) = 1 - H_1(\omega_k)$. Such binary masks in sinusoidal domain is very similar to the conventional binary masks in STFT or Gammatone filter bank in [14], except that here we compare the gain ratio of each band to the SSR level. Using the produced sinusoidal binary mask and applying it to the spectrogram of the given

mixture we then re-synthesize the separated output for each speaker as

$$\hat{s}_i(n) = F_D^{-1}\left(S_z(\omega)H_i(\omega)\angle F_D(z(n))\right) \quad i = 1, 2, \tag{10}$$

where F_D^{-1} denotes the inverse Fourier transform, $\angle F_D(z(n))$ is the phase of the Fourier transform of the mixture and $S_z(\omega)$ is the power spectral densities for the mixture, and $\hat{s}_i(n)$ is the recovered time signal for the ith speaker in the mixture. Since we have no access to the PSDs, we replace $S_z(\omega)$ with the approximation $|F_D(z(n))|^2$. These reconstructed speaker signals are then sent to the speaker identification block to get more accurate speaker identities.

In [4] SSR based speaker models are introduced for speaker identification purposes indicating that the number of GMM computations is a multiple of SSR level. Reference [5] introduces a more complex system that is composed of Expectation Maximization (EM) algorithm to find speaker combination in the mixed signal. Although the speech separation challenge could be considered as offline, here we consider much more on application potential. Our approach for speaker identification is not very accurate compared to those used in [4, 5], but its complexity is rather low. In GMM-UBM framework, only $\hat{M}$+C Gaussian mixtures evaluated per feature vector, where C is the number of top scoring Gaussians in UBM to be evaluated in speaker model.

5 Simulation Results

5.1 System Setup and Database

To evaluate the proposed separation algorithm in real world scenario, we used a comprehensive database provided by [15]. The database consists of 34 speakers each containing 500 utterances. The sampling rate is decreased to 8 kHz from the original 25 kHz. The mixed signal is generated by adding the signals according to the SSR level ranging in $[-9, 9]$ dB. The speaker signals to be mixed were selected from those used in the test setup of [15]. To put the performance of our proposed method into perspective, we report the separation performance for each method in terms of its PESQ score proposed by [6]. The proposed separation system is compared to a model-based system with correct speaker identities, speaker independent system and the speech quality obtained from the given mixture. The configurations for our separation setup are described as follows. We used window length of 32 msec along with a frame shift of 8 msec. The codebook size for STFT and Split-VQ was M=2048 and the sinusoidal model order was set to $M = 50$ in the Split-VQ. For speaker identification, a UBM with model order of $\hat{M} = 512$ was trained based on the two-talker development set [15]. Speaker GMMs are adapted accordingly based on 500 training data from individual talkers [12]. A 30 msec sliding Hamming window with a 15

msec shift was used to obtain a sequence of frames for extracting 12 dimensional mel-frequency cepstral coefficients (MFCCs), where 12 Δ-MFCCs and 12 Δ-Δ-MFCCs were concatenated to form a feature vector. Cepstral mean subtraction and RASTA filtering were also employed. Diagonal covariance matrices were employed and top-C scoring Gaussian mixtures were set to 5.

In the separation block, we first model the DFT spectral shape of each speaker using a split-VQ of modified sinusoidal model mixture. In order to show the superiority of multiple database approach over speaker-dependent separation modeling, we also fit a VQ to the training data of each gender for speaker-independent scenario. We quantify the degree of the separability by computing PESQ between the separated and original signals.

5.2 Results

We conducted evaluations for three scenarios: same gender (SG), different gender (DG) and same talker (ST). For each scenario we included five sentence of each speaker and combined them at nine SSR levels SSR=$\{-9, -6, -3, 0, 3, 6, 9\}$ dB. The PESQ curves obtained by the proposed method is shown in Fig. 2 along with the results obtained by speaker-dependent method with correct identities indicated by SD. For each speaker, the closed loop in the joint processor results in improvements compared to the signal quality of the related speaker in the mixture (shown with dotted line). Furthermore, as the SSR increases the PESQ scores of the proposed method reaches the separation upper-bound determined by the correct speaker dependent models. The gap at low SSR levels between the proposed method and correct speaker dependent approach is large for different gender scenario (see Fig. 2(b)). This can be caused by high masking of male female combination since their time-frequency pattern is quite different compared to the same talker or same gender. In the same gender scenario, it is observed that the gap between the PESQ scores of the proposed method and speaker dependent approach is small even at low SSR levels (see Fig. 2(a)). In the same talker scenario, the speaker identification could find the identities most accurately and this explains why two PESQ curves (shown in Fig. 2(c)) are very close to each other. Fig. 3(a),(b) shows the improvement made by using the proposed method and compares its performance with speaker dependent, speaker independent and speech quality obtained from the mixture. The curves are shown for different SSR levels for same talker scenario. It is observed that the results obtained by the proposed method, are very close to the one obtained by the speaker dependent method where the correct speaker identities are known *a priori*. In the separation block, we first model the DFT spectral space of each speaker using a split-VQ of modified sinusoidal model mixture. In order to show the superiority of multiple database approach over speaker-dependent separation modeling, we also fit a VQ to the training data for each gender. We quantify

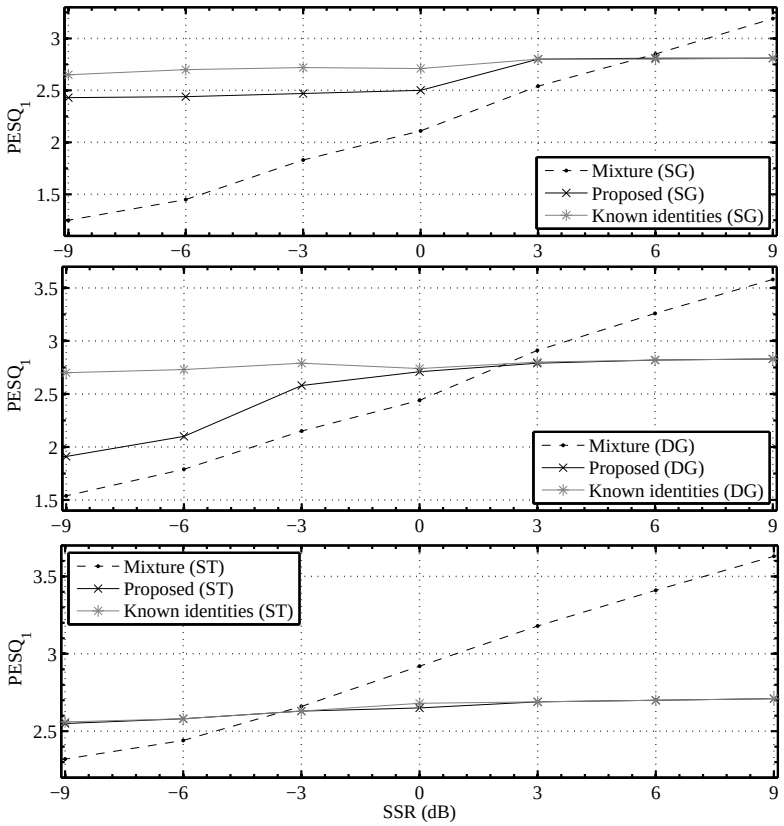

Fig. 2: Comparing the proposed approach with known identities, the mixture PESQ scores for each speaker at different SSRs for (a) same gender (SG), (b) different gender (DG), and (c) same talker (ST).

the degree of the separability by computing PESQ between the separated and original signals in the time domain. We also include the separation upper bound performance obtained by the correct speaker models where the optimal indices are *a priori* available.

6 Conclusion

In this paper, a novel approach has been proposed to combine speech separation with speaker identification in single-channel scenario. The system is implemented in a closed loop with the idea to improve the separation performance

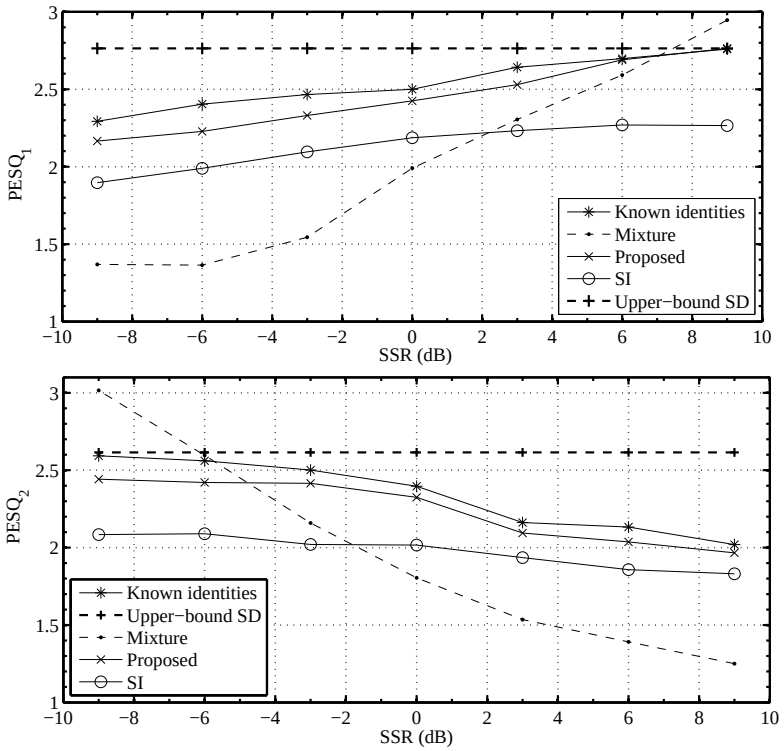

Fig. 3: Comparing the proposed approach for same gender (SG) scenario with correct speaker dependent (correct SD), speaker independent (SI) and mixture PESQ scores for (a) speaker one and (b) speaker two for different SSR levels.

by getting speaker identities from the mixture by using speaker identification. The separation result is fedback to speaker identification to achieve correct identities. The separation method is independent on pitch estimates and is based on sinusoidal feature parameters which have low feature dimension. Experimental results showed that the proposed method achieved a high score close to the separation performance with the correct identities and a higher performance compared to the speaker independent case. All methods asymptotically reached the operation upper bound performance determined by the VQ source models of the speaker in the mixture. As the SSR increases, the proposed method asymptote its separation upper bound performance, where it is assumed that the optimal indices are *a priori* available. According to the informal listening tests, it was observed that the perceived speech quality of the proposed system was improved after the identification stage in one closed loop.

References

[1] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 957–960, May 2006.

[2] S. Roweis, "Factorial models and refiltering for speech separation and denoising," *European Conference on Speech Communication and Technology*, pp. 1009–1012, 2003.

[3] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[4] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[5] J. Hershey, J. R. Steven, Rennie, P. A. Olsen, and K. T. T., "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[6] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.

[7] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[8] P. Mowlaee and Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *European Signal Processing Conference*, Aug. 2008.

[9] P. Mowlaee, M. Christensen, and S. Jensen, "Improved single-channel speech separation using sinusoidal modeling," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, accepted, 2010.

[10] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[11] R. Saeidi, H. Mohammadi, T. Ganchev, and R. Rodman, "Particle swarm optimization for sorted adapted gaussian mixture models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 344–353, Feb. 2009.

[12] D. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[13] P. Mowlaee, M. Christensen, and S. Jensen, "Sinusoidal masks for single channel speech separation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, accepted, 2010.

[14] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[15] M. Cooke, J. Hershey, and S. Rennie, "Elsevier computer speech and language," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

Paper E

Signal-to-Signal Ratio Independent Speaker Identification for Co-Channel Speech Signals

R. Saeidi, P. Mowlaee, T. Kinnunen, Z. -H. Tan,
M. G. Christensen, S. H. Jensen, and P. Fränti

This paper has been published in
Proceedings of the IEEE International Conference on Pattern Recognition, pp. 4545–4568, 2010.

The layout has been revised.

Abstract

In this paper, we consider speaker identification for the co-channel scenario in which speech mixture from speakers is recorded by one microphone only. The goal is to identify both of the speakers from their mixed signal. High recognition accuracies have already been reported when an accurately estimated signal-to-signal ratio (SSR) is available. In this paper, we approach the problem without estimating SSR. We show that a simple method based on fusion of adapted Gaussian mixture models and Kullback-Leibler divergence calculated between models, achieves an accuracy of 97% and 93% when the two target speakers enlisted as three and two most probable speakers, respectively.

1 Introduction

Speaker identification (SID) is the task of recognizing one's identity based on observed speech signal [1]. Typical speaker identification systems consist of short-term spectral feature extractor (front-end) and a pattern matching module (back-end). In traditional SID, the basic assumption is that only one target speaker exists in the given signal whereas in *co-channel* SID, the task is to identify two target speakers in one given mixture. Distinct from the so-called *summed channel* speaker recognition task [2], where only one speaker is talking most of the time, in the co-channel SID problem, both speakers talk simultaneously. Research on co-channel speaker identification has been done for more than one decade [3], yet the problem remains largely unsolved.

Most of the current *single-channel speech separation* (SCSS) systems use a model-based SID module, known as *Iroquois* [4] to identify the speakers in a mixed signal. The goal of an SCSS system is to estimate the unknown speaker signals according to their observed mixture. Interaction of the SID and speech separation modules can be managed in a closed loop to increase the overall performance [5]. Recognition accuracy as high as 98% has been reported for *Iroquois* in [6] which makes it as a first choice to be included in SCSS systems [7]. The database in [6] is provided for speech separation challenge and consists of 2 seconds of small vocabulary speech for 34 speakers. In the *Iroquois* system, a short list of the most likely speakers are produced based on the frames of the mixed signal that are dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* to find the signal-to-signal ratio (SSR) and two speakers identity with an exhaustive search on codebooks created for speech synthesis [4].

The SSR estimation in *Iroquois* system is based on finding the most likely combination of speakers codebooks to produce the current speech frame, where in text-independent case gets more challenging compared to the database in [6].

Although the SSR can be continuous and time-varying over a recording in realistic conditions, in database presented in [6] and in this study the discrete SSR levels of $\{-9, -6, -3, 0, 3, 6\}$ dB are considered. Furthermore, in real-time applications of SCSS and in forensic applications it is necessary to have a *fast* and *accurate* system to identify the underlying sources in mixed signal without SSR estimation required.

To this end, in this paper, we propose an SSR-independent SID module for co-channel speech. More specifically, we examine different frame-level likelihood scores and model level distances to solve the problem and propose a combination of the most successful ones to compare the accuracy with respect to *Iroquois*. Since the proposed system is SSR-independent and tuned on 8 kHz speech, it is believed that it could be an alternative approach for the SID in SCSS and useful for telephony data found, for instance, in forensic applications.

2 Speaker Recognition Approach

We use two main approaches for speaker recognition: frame-level log-likelihood calculation for a given mixed signal against a speaker GMM and between-models distance of a GMM model trained on mixed signal to speaker GMMs.

2.1 Frame-Level Likelihood Scores

From the frame-level likelihood estimation originally defined for the *Iroquois* system in [4, 8] and which aims at determining the frames where only one speaker exists, we derive three different scores defined at the end of this section. A maximum likelihood (ML) trained GMM has been used in [4]; however, maximum *a posteriori* (MAP) derived GMMs [9] are more accurate in speaker verification and we follow this latter approach. Let λ denote speaker GMM. The likelihood function is defined as,

$$\ell(\mathbf{x}) = p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x}). \tag{1}$$

The density is a weighted linear combination of M unimodal Gaussian densities $p_m(\mathbf{x})$, where $p_m(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m)$ and the mixture weights w_m further satisfy the constraints $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$. Speaker-dependent GMMs are adapted from universal background model (UBM) [9]. The UBM is a GMM trained on a pool of feature vectors extracted from as many speakers as possible to serve as *a priori* information for feature distribution. GMM means are the only parameters updated and weights and covariances are copied directly from UBM to GMMs.

2.2 Model Distance Scores

We define λ_{ig} as the SSR-dependent model for ith speaker at SSR level g. Another approach to measure similarity of a speech segment with a speaker model (λ_i) is to make a model from the test utterance with MAP adaptation (λ_e) and calculate the distance between λ_e and the speaker model. We use the *Kullback-Leibler divergence* (KLD) as a distance measure between the two probability distributions. Since this distance cannot be directly evaluated for GMMs, we use the upper bound of KLD which has successfully been applied to speaker verification [10]:

$$\begin{aligned}\mathrm{KLD}_i = \frac{1}{2}\sum_{g=1}^{G}\sum_{m=1}^{M} w_m(\boldsymbol{\mu}_{me}-\boldsymbol{\mu}_{mig})^T\boldsymbol{\Sigma}_m^{-1} \\ (\boldsymbol{\mu}_{me}-\boldsymbol{\mu}_{mig}).\end{aligned} \tag{2}$$

Here G ranges in a set of SSR levels, $\boldsymbol{\mu}_{me}$ is the mth mean vector in λ_e and $\boldsymbol{\mu}_{mig}$ is the mth mean vector in λ_{ig}, whereas w_m and $\boldsymbol{\Sigma}_m$ are the weights and the covariances of the UBM, respectively. An alternative approach to measure the distortion between GMMs is *approximate cross entropy* (ACE) [11]. As shown in [11], assuming infinite number of test utterance feature vectors, log-likelihood for a given λ_i equals to negative cross entropy between λ_e and λ_i. It can be approximated as follows:

$$\begin{aligned}\mathrm{ACE}_i = \sum_{g=1}^{G}\sum_{m=1}^{M} w_m \max_n \Bigg[\log w_n \\ -\frac{1}{2}(\boldsymbol{\mu}_{me}-\boldsymbol{\mu}_{nig})^T{\boldsymbol{\Sigma}_m}^{-1}(\boldsymbol{\mu}_{me}-\boldsymbol{\mu}_{nig}) \\ -\frac{1}{2}\log|\boldsymbol{\Sigma}_n| - \frac{D}{2}(1+\log 2\pi+\frac{1}{Tw_m+r})\Bigg],\end{aligned} \tag{3}$$

where T is the total number of frames for training λ_e, D is features dimension and r is a relevance factor that controls compromise between UBM statistics and adaptation data in GMM adaptation [9]. The value r =0 corresponds to barely standing on adaptation data.

2.3 Proposed Method

In this work, we train the UBM (λ_{UBM}) using digitally mixed speech signals at different SSR levels formed by different speakers. Moreover, we train each target speaker i, the set of gain-dependent models λ_{ig} that are adapted from the UBM based on ith speaker speech files corrupted by other speakers signal at SSR level

g. Using SSR-based speaker models, the system captures speaker-dependent information when it is contaminated by other speakers data. This is similar to the idea of having an SSR-based bias in GMM parameters in [4], however, it has the major difference that we build separate GMMs for each SSR level based on the UBM. It enables the system to function independent of the SSR level.

For a feature vector extracted from a speech segment at time instance t, and denoted by $\mathbf{x}_t$, frame level score for speaker i is defined as,

$$s_{it} = \frac{1}{G}\sum_{g=1}^{G} \log[p(\mathbf{x}_t|\lambda_{ig})] - \log[p(\mathbf{x}_t|\lambda_{UBM})], \tag{4}$$

We average over all SSR levels to be independent of the underlying SSR in the given signal and normalize all speakers scores at time instance t with the corresponding UBM score. To emphasize dominant speaker score in a frame, the score in (4) is further normalized by $s'_{it} = s_{it}/\sigma_t$, where σ_t is standard deviation of all speakers scores for the frame t. To sum up, we consider five different scores for a speaker:

NWF: *number of winning frames*, where speaker i is the most probable speaker in that frame, $NWF_i = \sum_t \varphi(s'_{it})$ where $\varphi(s'_{it}) = 1$ for $i = \arg\max_j s'_{jt}$ and 0 otherwise.

NCF: *number of confident frames* for speaker i where s'_{it} is above threshold α: $NCF_i = \sum_t \psi(s'_{it})$ where $\psi(s'_{it}) = 1$ for $s'_{it} > \alpha$ and 0 otherwise.

LL: *Log-likelihood* mean for which s'_{it} is above threshold α: $LL_i = (1/NCF_i)\sum_t \psi(s'_{it})s'_{it}$.

KLD: *Kullback-Leibler divergence* between λ_e and a set of models λ_{ig}, computed using (2).

ACE: *approximate cross entropy* between λ_e and a set of models λ_{ig}, computed using (3).

As it is common in speaker recognition, to enable using benefits from different recognizers, we considered the fusion of the scores. We used an approximate brute-force search to find the optimal weights for score fusion. It should be mentioned that we normalized (and reverted for KLD) the range of scores from different recognizers before fusion. A block diagram of proposed system is presented in Fig. 1.

3 Experimental Setup

We evaluate the proposed SID module using the *speech separation challenge* corpus provided in [6]. The corpus is composed of 34 speakers (18 male, 16 female),

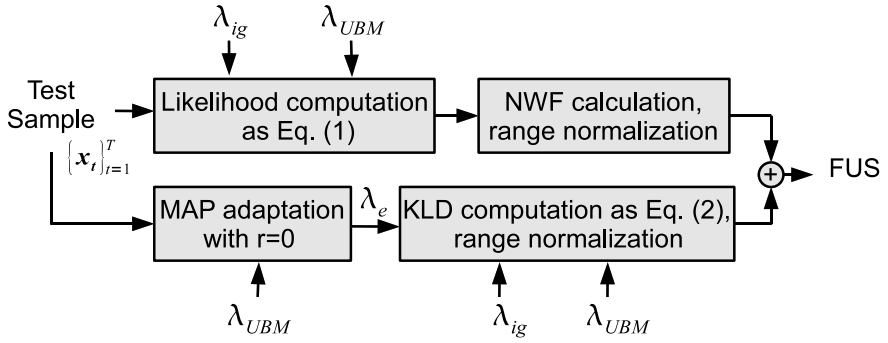

Fig. 1: Proposed SID module is a combination of frame level likelihood score and model level distance: FUS = 0.54NWF + 0.46KLD.

with a total number of 34,000 utterances, each following a command-like structure, and all having a unique grammatical structure. Each sentence is formed by different syntaxes of command, color, letter, number and code, for instance "bin white by A 3 please". The test data in the corpus is composed of 500 laboratory-quality signals for each of the 34 target speakers, as well as test set consisting of mixed signals at six signal-to-signal ratio levels of $\{-9, -6, -3, 0, 3, 6\}$ dB. For each of these six test sets for two-talker signal, 600 utterances are provided, from which 221 are for same talker (ST), 200 for same gender (SG), and 179 for different gender (DG). The utterances were originally sampled at 25 kHz with a duration of 2 second.

Since we are interested in telephone-quality speech bandwidth, we downsample the signals from 25 kHz to 8 kHz. We extract features from 30 msec frames multiplied by a Hamming window. A 27-channel mel-frequency filterbank is applied on discrete Fourier transform (DFT) spectrum to extract 12-dimensional mel-frequency cepstral coefficients (MFCCs), followed by appending Δ and Δ^2 coefficients, and using an energy-based voice activity detector (VAD) for extracting the feature vectors. We digitally add the signals with an average frame-level SSR to construct the UBM and the target speakers GMMs. For each of 34 speakers, 50 random files from each speaker were mixed at SSRs levels $\{-9, -6, -3, 0, 3, 6\}$ dB with 50 random files from other speakers which gives us about 180 hour of speech for training UBM. The number of Gaussians, M, is set to 2048.

Speakers SSR-dependent GMMs, λ_{ig} , trained by mixing 100 random files from each speaker with 100 random files from other speakers yielding about 1.8 hours data for each SSR. Relevance factor was set to 16 for training speaker models, λ_{ig}, where its value was set to 0 in training test model, λ_e, because of availability of only 2 seconds of data for adaptation. We set the threshold α to 1 in frame-level scores calculation. The accuracies defined here are to identify both of the speakers existing in mixed signal as the two most probable speakers.

4 Experimental Results

We first analyze the performance of speaker identification system using each of the 5 scores individually. The results shown in Table E.1 indicate that NWF and KLD have the best average performance compared to the other methods. To the best of our knowledge, SID accuracy for *Iroquois* is not reported without SSR estimation included. Compared to *LL* score, our proposed method, *NWF*, is more accurate. It is observed that, the number of frames above the confidence level, NCF is more important than their mean value, *LL*. On the other hand, the model based approach, *ACE*, works equally well as the frame-level method but it is more complex and has slightly worse accuracy than *KLD*.

Table E.1: Speaker Identification accuracy for different systems (percentage of utterances with both speakers in the 2-best list output). FUS is proposed system composed of $0.54NWF+0.46KLD$ and *IRO* stands for Iroquois

SSR (dB)	-9	-6	-3	0	3	6	Ave
NWF	81	90	94	95	92	88	90
NCF	75	88	93	94	92	86	89
LL	74	84	90	91	87	82	85
KLD	79	89	92	93	91	87	88
ACE	79	87	92	92	89	84	87
FUS	92	93	96	97	93	87	93
IRO [4]	96	98	98	99	99	98	98

Score fusion was then done by using two most successful methods: $\text{FUS}_i = 0.54\text{NWF}_i + 0.46\text{KLD}_i$. The fusion weights were optimized on development set consisting of 300 mixed signals for each SSR level. We found that, for the fusion system, in all of the experiments, one of the speakers in the mixed signal is *always* identified. The accuracy of the proposed system (FUS) for listing two target speakers in 3-best list is shown in Table E.2. This accuracy suggests to use proposed SID module as a concise "short-list" generator for the SSR estimation in *Iroquois* to reduce complexity. To understand the system performance better, we look for combinations of speakers that are identified in any given SSR. Surprisingly, in 68% of cases both speakers are correctly identified in the mixed signal at all SSR levels, and in 80% of experiments possibly only for one SSR we cannot identify both speakers but one of them. From the results, it is observed that mixed signals with different genders (DG) are more problematic than the same gender, which there are almost no significant difference in identification accuracy between males and females.

5 Conclusion

A new method for speaker identification in co-channel scenario was introduced based on the existing approaches in speaker verification and compared the accuracy to *Iroquois* approach. From the simulation results conducted on speech separation challenge database, we observed that the proposed simple SID module performs well in listing two target speakers as three most probable speakers without any requirement on the estimates of the SSR level. As a future work, since we already got satisfactory results on 8 KHz speech, we plan to examine the proposed algorithm on telephony quality spontaneous speech and more realistically when signals are not synthetically mixed.

Table E.2: Speaker Identification accuracy for proposed *FUS* system (percentage of utterances with both speakers in the 3-best list output) ST, Same Talker, SG, Same Gender and DG, Different Gender).

SSR	ST	SG	DG	Ave
-9 dB	100	93	83	92
-6 dB	100	97	94	97
-3 dB	100	100	98	99
0 dB	100	98	99	99
3 dB	100	97	93	97
6 dB	100	94	91	95
Ave	100	97	93	97

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Elsevier speech communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Elsevier Computer Speech and Language*, vol. 20, no. 3, pp. 128–158, 2006.

[3] D. Morgan, E. George, L. Lee, and S. Kay, "Co-channel speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1995, pp. 828–831.

[4] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[5] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4430–4433.

[6] M. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[7] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[8] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[9] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[10] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, May 2006.

[11] H. Aronowitz and D. Burshtein, "Efficient speaker recognition using approximated cross entropy (ACE)," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2033–2043, Sept. 2007.

Paper F

A MAP Criterion for Detecting the Number of Speakers at Frame Level in Model-based Single-Channel Speech Separation

P. Mowlaee, M. G. Christensen, Z. -H. Tan, and S. H. Jensen

This paper has been published in
44th Annual Asilomar Conference on. Signals, Systems, and Computers,
pp. 538–541, 2010.

The layout has been revised.

Abstract

The problem of detecting the number of speakers for a particular segment occurs in many different speech applications. In single channel speech separation, for example, this information is often used to simplify the separation process, as the signal has to be treated differently depending on the number of speakers. Inspired by the asymptotic maximum a posteriori *rule proposed for model selection, we pose the problem as a model selection problem. More specifically, we derive a multiple hypotheses test for determining the number of speakers at a frame level in an observed signal based on underlying parametric speaker models, trained a priori. The experimental results indicate that the suggested method improves the quality of the separated signals in a single-channel speech separation scenario at different signal-to-signal ratio levels.*

1 Introduction

An open problem in speech processing is the detection of the number of speakers present in a given segment of a signal. A special case of this problem is the classification of speech segments into what is often referred to as single-talk (one speaker), double-talk (speech mixture), and noise-only regions, with the resulting detector commonly referred to as a double-talk detector. Knowledge of such regions is useful since in many speech applications, it is required to process the underlying signals differently depending on the type. In this regard, a detector solving this problem can be effectively used as a pre-processor for improving the performance.

Double-talk detection has been used for a number of applications, two examples being acoustic echo cancellation and single-channel speech separation (SCSS). In acoustic echo cancellation, the double-talk detector is used to freeze the adaptation of an adaptive filter during double-talk regions (when both far-end and near-end speech is present) in order to avoid divergence of the adaptive filter, and, as a consequence, avoid the cancellation of the desired speech signal [1]. However, in SCSS, it is used to classify an observed speech mixture into single-talk, double-talk, and noise only regions, regions that have to be processed differently.

In the context of SCSS, a few separation methods implicitly detect double-talk regions in various contexts, e.g., [2–4]. In [2], a state-based hypothesis test was proposed in order to determine the reliability of each time-frequency cell in a given noise-corrupted speech signal. It was observed that the method led to a significant improvement in speech recognition performance in presence of other competing speaker signals. Similarly, in [4], a silence state was added to the speaker codebooks in order to deal with frames where only one speaker is

active.

A few participants in the speech separation challenge [5], made use of a model-based speaker identification (SID) module, called *Iroquois* [3] to identify speakers existing in the mixture. *Iroquois* works based on excluding silence and mixture segments from its parameter update procedure. Instead, it selects segments where only one speaker is dominated which are known as discriminating features for speaker recognition purpose. This decision-taking helped narrowing down what speakers are present in the mixture, hence, leading to an improvement in speaker recognition performance [3]. This required the calculation of speaker posteriors for different trained models of speakers present in the whole dataset (e.g. 34 speakers in [5]). *Iroquois* used a fixed threshold for calculating the uncertainty in speaker identification, and, as a consequence, could result in errors while determining which frame belongs to single-talk and double-talk regions.

Source-driven approaches, mostly known as computationally auditory scene analysis (CASA) [6], suggest to combine time-frequency segments of the mixed signal that are likely to arise from the same source and then concatenate them into a single stream. As a consequence, CASA-based methods implicitly detect the number of speakers in the speech mixture independently of *a priori* knowledge of any speaker model [6]. However, the methods predominantly use estimated pitch trajectories by applying a multi-pitch estimator. For the masked signal, as a consequence, the overall accuracy for CASA-based method is limited by the accuracy of the multi-pitch estimator.

To solve the problem of detecting the number of speakers in a speech mixture, we take a different approach. We integrate the *maximum a posteriori* (MAP) criterion proposed in [7] into SCSS to solve the model selection problem. We derive multiple hypothesis tests to determine double-talk/single-talk regions in segments of the mixed signal. We present the results of signal classification by applying the proposed method to speech mixtures composed of two speakers at different signal-to-signal ratio (SSR) levels. In addition, to put the idea into perspective, we demonstrate how using the proposed detector will affect the quality of the separated output signals. More specifically, by finding single-talk regions thanks to a double-talk detector, the remaining problem to be solved in SCSS is only to separate the mixture segments. For single-speaker frames, the observed signal is directly re-synthesized according to the corresponding speaker models.

The paper is structured as follows: In the next section, we introduce basic notation, definitions and the model-selection problem. In Section 3, we derive multiple-hypothesis rules required for detecting single-talk and double-talk regions in a segment of mixture. In Section 4, we present the experimental results with showing the accuracy of the proposed method. We also present the results showing the improvements achieved by employing the proposed double-talk detector in a SCSS scenario. Section 5 concludes on the work.

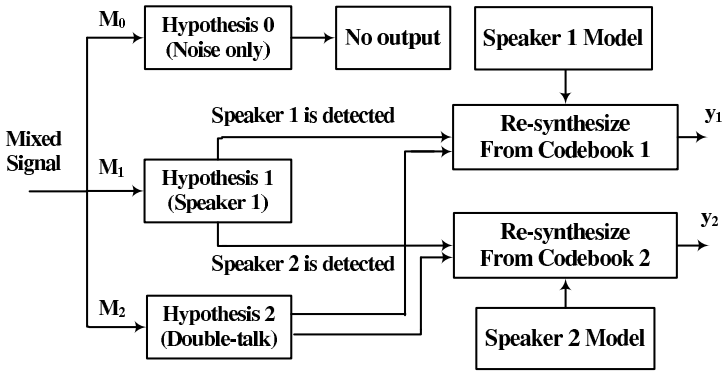

Fig. 1: The schematic block diagram for the proposed method for detecting the number of speakers in mixture and showing how it can be used in the SCSS problem. The decision lies in one of the following three models: $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ showing, noise-only, single-talk, and double-talk classes, respectively. The separated output signals are shown as $\mathbf{y}_1$ and $\mathbf{y}_2$ for speaker one and two, respectively.

2 Model Selection for Detecting the Number of Speakers

We will now proceed to introduce some basic notation and definitions. Consider a mixed signal with N samples $\mathbf{y} \in \mathbb{R}^N$ composed of up to J speaker signals as $\mathbf{y} = \sum_{j=1}^{J} \mathbf{s}(\boldsymbol{\psi}_j) + \mathbf{e}$, where the superscript T represents the matrix transpose, $j \in [1, J]$ the number of signals in the mixed signal, $\mathbf{s}(\boldsymbol{\psi}_j) \in \mathbb{R}^N$ the jth signal characterized by parameter vector $\boldsymbol{\psi}_j$ and $\mathbf{e} \in \mathbb{R}^N$ the noise signal incorporated in the model. For simplicity in the ensuing derivations and simulations, we focus on $J = 2$, that is, a mixture of two speakers and noise. As our signal model, we use sinusoidal modeling as described in [8]. More specifically, we model the jth speaker signal in the mixture as a parametric feature vector $\boldsymbol{\psi}_j$, composed of sinusoidal parameters: amplitude, frequency and phase vectors. We here use $K = 3$ candidate models each denoted by $\mathbf{M}_k$, for describing the mixed signal, $\mathbf{y}$, namely: $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ to indicate noise-only, single-talk, and double-talk, respectively. Each of these models is described by parameter vector $\boldsymbol{\theta}_k$ with L_k sinusoids. A block diagram of the proposed method for detecting the number of speakers in mixture is shown in Fig. 1. The proposed approach addresses the following problem: given the mixed signal, select the model which is the most likely. We consider three models for $\mathbf{y}$ as:

$\mathbf{M}_0$**:** $\mathbf{y} = \mathbf{e}$,
$\mathbf{M}_1$**:** $\mathbf{y} = \mathbf{s}(\boldsymbol{\psi}_j) + \mathbf{e} \quad \text{for} \quad j \in [1, 2]$,

$\mathbf{M}_2$**:** $\mathbf{y} = \mathbf{s}(\boldsymbol{\psi}_1) + \mathbf{s}(\boldsymbol{\psi}_\mathbf{2}) + \mathbf{e}$,

where $\mathbf{s}(\boldsymbol{\psi}_1) + \mathbf{s}(\boldsymbol{\psi}_2)$ represents an estimate for the mixed signal, and $\mathbf{s}(\boldsymbol{\psi}_j)$ with $j \in [1, 2]$ indicates the jth signal modeled by the parameter set $\boldsymbol{\psi}_j$.

Following the model selection approach in [7], we adopt a MAP criterion for multiple-hypothesis tests to determine double-talk/single-talk regions in segments of a mixed signal. To this end, we need to evaluate the posterior probabilities of $\mathbf{M}_k$ with $k \in Z_K = \{0, 1, 2\}$. The MAP estimate of the most likely hypothesis is denoted by $\hat{\mathbf{M}}_k$, and is obtained as

$$\hat{\mathbf{M}}_k = \arg \max_{\mathbf{M}_k : k \in Z_K} \left\{ \int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{M}_k) p(\boldsymbol{\theta}_k|\mathbf{M}_k) d\boldsymbol{\theta}_k \right\}. \tag{1}$$

The problem in (2) is a complicated nonlinear maximization problem due to the used models. As proposed in [7], instead of numerical integration for the evaluation of marginal density in (2), we employ the asymptotic MAP criterion, which under certain conditions can be shown to be

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k : k \in Z_K} \left\{ -\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathbf{M}_k) + p_c \right\}, \tag{2}$$

with p_c being the model-dependent penalty of the MAP criterion, $\hat{\boldsymbol{\theta}}_k$ an estimate of $\boldsymbol{\theta}_k$ for the kth model $\mathbf{M}_k$, and $-\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathbf{M}_k)$ the log-likelihood term obtained from an approximation of (2).

3 Multiple-hypothesis Algorithm

The problem is now to determine $-\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathbf{M}_k)$ for each of the three underlying candidate models $\mathbf{M}_k$ with $k \in Z_K = \{0, 1, 2\}$. Here, we use sinusoidal modeling in [8] to model the speaker signals in the mixture. Let $\mathbf{s}_i(\hat{\boldsymbol{\psi}}_j)$ be the jth speaker signal with $j \in [1, 2]$ for the ith frequency band modeled by the parametric vector $\hat{\boldsymbol{\psi}}_j$. Here we assume that the signal modeling error, $\mathbf{e}$ has a Gaussian distribution and the modeling error subband signal, $\mathbf{e}_i$ is white in each ith frequency band. Then from the subband decomposition and the independece assumption for all frequency bands, assuming that $\mathbf{e}_i$ is independent from one band to another, one can show that the likelihood function for all bands for each class $\mathbf{M}_k$ is given by

$$\begin{aligned} p(\mathbf{e}|\sigma^2) &= \prod_{i=1}^{Q} p(\mathbf{e}_i|\sigma_i^2) \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{i=1}^{Q} \sigma_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{Q} \frac{\mathbf{e}_i^T \mathbf{e}_i}{\sigma_i^2} \right), \end{aligned} \tag{3}$$

where Q is the total number of frequency bands, σ_i denotes the variance due to the modeling error signal in the ith band, $\mathbf{e}_i$.

For single speaker class, $\mathbf{M}_1$, the modeling error at the ith frequency band, is given by $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{s}_i(\hat{\boldsymbol{\psi}}_j)$. For the mixed class, $\mathbf{M}_2$, let us define the estimated error as $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{s}_i(\hat{\boldsymbol{\psi}}_1) - \mathbf{s}_i(\hat{\boldsymbol{\psi}}_2)$ as the noise estimated for the ith frequency band as a colored noise not fitted by $\mathbf{M}_2$. The MAP criterion [7] for sinusoids composed of unknown amplitudes and frequencies reduces to

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k \in Z_K} \left\{ \frac{N}{2} \sum_{i=1}^{Q} \ln \hat{\sigma}_i^2 + \frac{5L_k}{2} \ln N \right\}. \tag{4}$$

where we define $\hat{\sigma}_i^2 = \frac{1}{N} \hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_i$ as the estimated variance for the ith frequency band and we remind the reader that L_k is the number of sinusoids. In the mixture class $\mathbf{M}_2$, we require a mixture estimate to replace $\mathbf{s}(\hat{\boldsymbol{\psi}}_1) + \mathbf{s}(\hat{\boldsymbol{\psi}}_2)$ in order to find the best pair of $\{\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2\}$ from the speaker models of the underlying speakers. Here, we use the minimum mean square error (MMSE) estimator for the mixture magnitude spectrum in [9], in order to find the the joint best states in the speaker models which when combined best describe the magnitude spectrum for the observed mixture, $\mathbf{y}$.

We include the noise model, $\mathbf{M}_0$ as one of the examined models by setting $\mathbf{y} = \hat{\mathbf{e}}$ and setting the number of sinusoids equal to zero ($L_k = 0$). The estimated noise variance is given by $\hat{\sigma}_i^2 = \frac{1}{N} \mathbf{y}_i^T \mathbf{y}_i$.

Finally, using the estimated value for σ_i depending on each possible class of $\mathbf{M}_k$ with $k \in Z_K = \{0, 1, 2\}$, the best model, as a result, is the one which yields high log-likelihood and low model order, which is achieved in (6). The proposed method for detecting the number of speakers in the speech mixture can be summarized in the following three steps:

(1) Find the variance of noise, $\hat{\sigma}_i$ at each ith band.
(2) Compute the MAP criterion for each class: $\{\mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2\}$.
(3) Select the model with largest log-likelihood.

4 Simulation Results

4.1 System Setup and Database

To evaluate the proposed approach, we used the database in [5] with a sampling rate of 8 kHz. The speaker models are obtained by the split-VQ (vector quantization) [8] composed of sinusoidal amplitude and frequencies trained based on 10 minutes of speech signals for each speaker. For training the speaker models we used 2048 codevectors for amplitude and 8 codevectors for frequency part.

Table F.1: Speaker labels used for training the gender-dependent models for male and female speakers.

Male	3	5	6	9	10	12	13	14	17	19
Female	4	7	8	11	15	16	21	22	23	24

Throughout the experiments, a Hamming window of length 32 ms with frame-shift equal to 8 ms was used to segment the speech files both in the training and test phase. As our test data, we used the mixture of target and masker speakers in the test setup of [5] mixed at six SSR levels of {-9,-6,-3,0,3,6} dB. To relax the speaker-dependent assumption, we used gender-dependent models and we trained a male speaker model using utterance from ten speakers and a female speaker model trained on ten female speakers. The speaker labels used for training our gender-dependent models are shown in Table F.1.

4.2 Experiment 1: Detection Accuracy

Figure 2, shows the clean signal (prior to mixing) for speaker one and two together with their mixture. In Fig. 2, the detection results of the number of speakers in speech mixture are shown for gender-dependent scenario. The hypotheses for single-talk and double-talk regions are also shown as ground truth. It is observed that, the double-talk detector effectively finds the regions of the non-speech and mixture segments and determines at each frame that which speaker(s), if any, are active. Comparing with the ground-truth, it is observed that the accuracy of the proposed double-talk detector is high. In our experiments, the models $\mathbf{M}_k$ with $k \in Z_K$ are considered as either speaker-dependent or gender-dependent. It is important to note that, in the speaker-dependent scenario, the proposed method solves a four class problem, namely noise, speaker one, speaker two, and mixture classes. However, using gender-dependent speaker models, the proposed double-talk detector solves a three-class problem for same gender or same talker scenario, since the estimated error signal, given by single-talk classes, will be the same.

4.3 Experiment 2: Speech Separation

In another experiment, we aim to study the effectiveness of employing the proposed double-talk detector in a SCSS system. More specifically, as a proof of concept, we report the signal quality of the separated signals obtained by using a model-based separation system with and without double-talk detector proposed in this work. Figure 3 shows the perceptual evaluation of speech quality (PESQ) [10] scores averaged over 50 mixtures The results are reported for both speaker-dependent and gender-dependent scenarios. From the PESQ

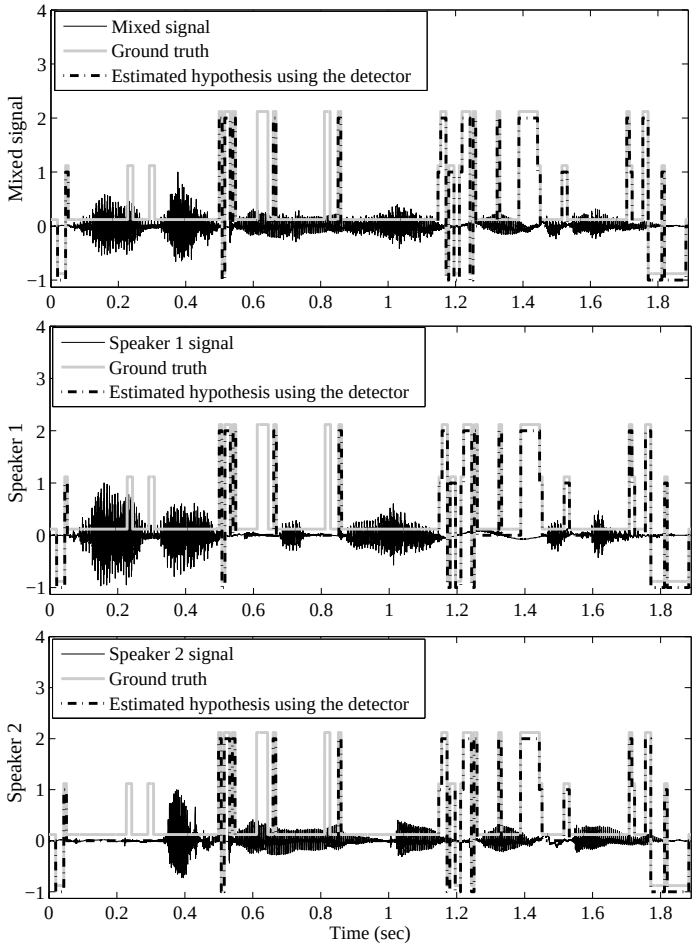

Fig. 2: Showing the performance for detecting the number of speakers in a mixture of a male and a female speaker mixed at 3 dB SSR. The mixed signal is composed of a male (speaker 12) uttering *"Lay white with e 8 again"* with female (speaker 11) uttering *"Set green with v 3 soon"*. Decisions are -1 for no speech, 1 for speaker one, 2 for speaker two and 0 for mixed signal regions.

curves shown in Fig. 3, it is observed that integrating double-talk detector into a model-based SCSS improves the speech quality of the re-synthesized signals. It is also observed that the PESQ scores obtained in the gender-dependent scenario were slightly lower than those obtained in speaker-dependent scenario.

However, as the SSR increases the performance of gender-dependent scenario aysmptotates to the one offered by speaker-dependent scenario. From informal listening test, it was observed that, the improvement obtained by employing the proposed detector is noticeable.

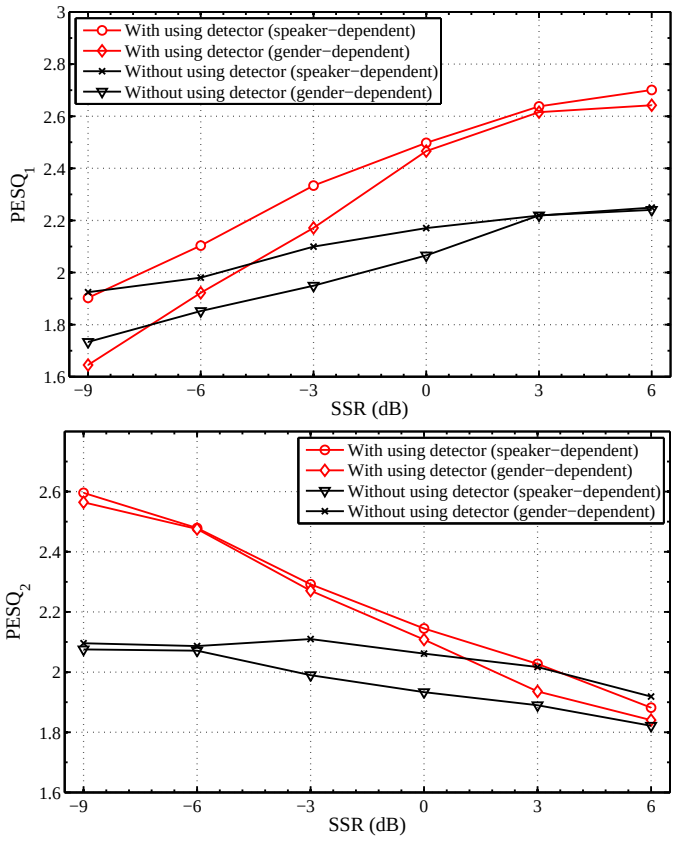

Fig. 3: Showing the PESQ scores obtained for speech separation in speaker-dependent and gender-dependent scenarios for two cases: with and without using the proposed method for detecting the number of speakers in a given speech mixture: (top panel) the PESQ scores for the first speaker and (bottom panel) for the second speaker in terms of the SSR level in decibels.

5 Conclusion

To conclude on our work, we have presented a solution to detecting the number of speakers in an observed segment of mixed speech signal. To solve the problem, we applied the MAP criterion already proposed for model selection and derived the multiple-hypothesis test algorithm to determine double-talk/single-talk regions for a particular segment in a given mixed signal in SCSS framework. We showed that, such information can be used to narrow down the separation problem only for mixed frames. Experiments showed that the proposed method successfully determines the single-talk and double-talk regions in both speaker-dependent and gender-dependent scenarios. The proposed detector approach also led to improvement in the signal quality of the separated signals compared to the scenario where no detector was used.

References

[1] J. Benesty, D. Morgan, and J. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168 –172, Mar. 2000.

[2] S. Srinivasan and D. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, March 2005, pp. 89–92.

[3] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Superhuman multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[4] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67 – 76, Jan. 2010.

[5] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, Jan. 2010.

[6] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77 – 93, Jan. 2010.

[7] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[8] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[9] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Sinusoidal masks for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2010, pp. 4262–4266.

[10] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.

Paper G

Improving Monaural Speaker Identification by Double-Talk Detection

R. Saeidi, P. Mowlaee, T. Kinnunen, Z. -H. Tan,
M. G. Christensen, S. H. Jensen, and P. Fränti

This paper has been published in
Proceedings of the 11th Annual Conference of the International Speech Communication Association INTERSPEECH-ICSLP, pp. 1069–1072, 2010.

The layout has been revised.

Abstract

This paper describes a novel approach to improve monoaural speaker identification where two speakers are present in a single-microphone recording. The goal is to identify both of the underlying speakers in the given mixture. The proposed approach is composed of a double-talk detector (DTD) as a pre-processor and speaker identification back-end. We demonstrate that including the double-talk detector improves the speaker identification accuracy. Experiments on GRID corpus show that including the DTD improves average recognition accuracy from 96.53% to 97.43%.

Index Terms: speaker identification, double-talk detection, single-channel, Gaussian mixture models.

1 Introduction

Speaker recognition systems have evolved to reach high accuracy on clean speech signals [1]. However, speaker recognition under adverse conditions remains a challenging problem. Depending on the noise type and the way that it affects the speech signal, the more complicated methods are required to handle speaker recognition task. One of the most challenging cases are speech signals mixed with other speech signals known as monaural speech. This happens in such applications as *single-channel speech separation* [2] where accurate speaker identification is crucial for the entire system. Here we consider the task of identifying both of the speakers' identity in a given speech mixture of two speakers. Current approaches for handling this task are combined with speech separation where we cannot say exactly there is a stand-alone speaker identification system for monaural speech [3]. We have recently independently proposed methods for both speaker identification (SID) [4] and speaker-dependent double-talk detection (DTD) [5] for speech signals mixture. Our proposed method [4] does not depend on speech separation but it works directly on monaural signal without any prior information about mixing scenario of the two speech signals. In this work we improve the SID accuracy by introducing external information of mixed frames and single-talk frames provided by enhanced version of proposed DTD module in [5]. A block diagram of the proposed system is shown in Fig. 1.

Majority of the current single-channel speech separation systems use *a priori* knowledge of speaker identities [6] which is both impractical and restrictive regarding real applications. A joint system composed of speaker identification and speech separation blocks was proposed in [7] for relaxing the need for *a priori* speaker identities. The proposed system [7] improved the overall perceived speech quality of the separated output signals compared to speaker-independent and the observed speech mixture. To make speaker identification system more efficient, in this work, we introduce gender-dependent DTD and apply it to

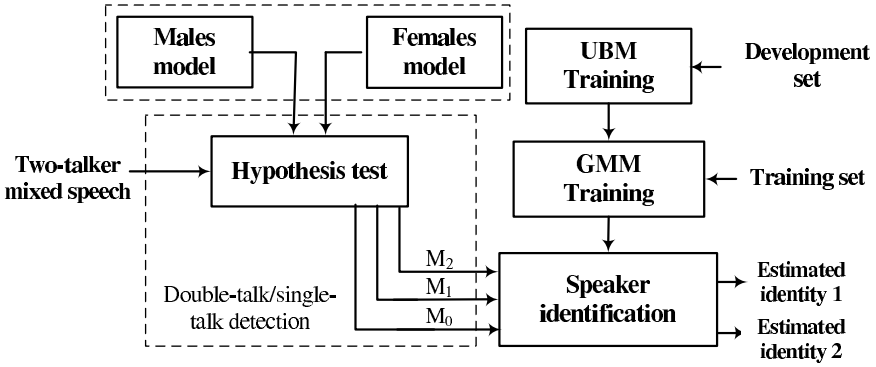

Fig. 1: *The proposed method is composed of a double-talk detector followed by SSR-independent speaker identification.*

monaural SID.

State-of-the-art single channel speech separation (SCSS) systems use a model-based SID module known as *Iroquois* [3] to identify the speakers in monaural speech. Recognition accuracies as high as 98% and 99% were reported on GRID corpus for *Iroquois* in [2, 8] for locating the target speakers in short-lists of top-2 and top-3 most probable speakers respectively. In the *Iroquois* system, a short-list of the most likely speakers is produced based on frames dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* for estimating both the signal-to-signal ratio (SSR) and the identities of the two speakers using exhaustive search on codebooks created for speech synthesis [3]. Based on the sizes of the short-list and code-books, this search can be time consuming. It is important to notice that if we wish to apply *Iroquois* system on a conversational mixed speech, it also requires a reliable speech separation system to produce meaningful results. Independent performance of our proposed method could be considered as a bonus in this situation. In view of this problem, the proposed system could also be used as a pre-processor for *Iroquois* system to reduce the search time.

The new contributions in this study are summarized as follows. We include a sophisticated MAP-based double talk detector (DTD) to our recent recognition system [4]. The double-talk detector was earlier introduced for monaural speech assuming known speaker identities [5]. In this paper, we adopt the method to monaural speaker identification, by using gender-dependent models to enable speaker-*independent* processing. The DTD module is utilized in the identification system so that the mixed-signal recognition score is enhanced by using "bonus" scores obtained from the more reliable single-talk regions of the mixed signal.

2 Double-Talk Detection System

In [9], a method for detecting single-talk and double-talk regions from a given speech mixture was proposed. The method was based on multiple hypothesis testing and was speaker-dependent. In this work, we briefly describe the method and generalize the idea to gender-dependent scenario which is more practical in real environments. Consider monaural speech signal with N samples $\mathbf{y} = [y(0), \ldots, y(N-1)]$ composed of J speaker signals as $\mathbf{y} = \mathbf{s}_1 + \mathbf{s}_2 + \cdots + \mathbf{s}_j + \mathbf{e}$, where $j \in [1, J]$ indicates the number of signals in the mixed signal, $\mathbf{s}_j = [s_j(0), \cdots, s_j(N-1)]$ is the jth signal and $\mathbf{e}$ is the noise signal incorporated in the model. In the following, we focus on $J = 2$, that is, a mixture of two speakers.

Assume that we have K candidate models denoted by $\mathbf{M}_k$, for describing $\mathbf{y}$. The double-talk detection addresses the following problem: given the mixed signal, select the model which has the the maximum *a posteriori* (MAP) probability. We consider four models for $\mathbf{y}$ as: $\mathbf{M}_0$: $\mathbf{y} = \mathbf{e}$, $\mathbf{M}_1$:$\mathbf{y} = \hat{\mathbf{s}}_1(\{\boldsymbol{\theta}_1\})$, $\mathbf{M}_2$:$\mathbf{y} = \hat{\mathbf{s}}_2(\{\boldsymbol{\theta}_2\})$, $\mathbf{M}_3$:$\mathbf{y} = \hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\})$. Here $\hat{\mathbf{s}}_i(\{\boldsymbol{\theta}_i\})$ indicates the ith signal modeled by the parameter set $\{\boldsymbol{\theta}_i\}$ and $\hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}) = \sum_{j=1}^{J} \hat{\mathbf{s}}_j(\{\boldsymbol{\theta}_j\})$ is the estimated mixed signal by model $\mathbf{M}_3$. Let $g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}_k)$ be a generic form for class $\mathbf{M}_k$ where $k \in Z_K = \{0, 1, 2, 3\}$. Here, $\boldsymbol{\theta}_k$ is a vector composed of model parameters in a parameter space $\boldsymbol{\theta}_k \in \mathbb{R}^{m_k}$ and m_k is the length of the parameter vector $\boldsymbol{\theta}_k$. Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be the vectors for model parameters for speaker one and two, respectively. Following the model selection approach in [10], we here adopt a MAP criterion for multiple-hypothesis testing to determine double-talk/single-talk regions in segments of a mixed signal. To this end, we need to evaluate the posterior probabilities of $\mathbf{M}_k$ with $k \in Z_K$. The MAP estimate of the most likely hypothesis is,

$$\hat{\mathbf{M}}_k \quad = \quad \arg \max_{\mathbf{M}_k : k \in Z_K} \left\{ \frac{p(\mathbf{y}|\mathbf{M}_k) p(\mathbf{M}_k)}{p(\mathbf{y})} \right\}, \tag{1}$$

where $p(\mathbf{y})$ denotes the marginal density of the observed signal and $p(\mathbf{M}_k)$ is the *a priori* probability of the model $\mathbf{M}_k$. Assuming that the underlying models are equiprobable, $P(\mathbf{M}_k) = \frac{1}{K}$, dropping K and $p(\mathbf{y})$ since they are independent of M_k, the model selection rule becomes

$$\hat{\mathbf{M}}_k = \arg \max_{\mathbf{M}_k : k \in Z_K} \left\{ \int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{M}_k) p(\boldsymbol{\theta}_k|\mathbf{M}_k) d\boldsymbol{\theta}_k \right\} \tag{2}$$

where $\hat{\mathbf{M}}_k$ is the best model which achieves the MAP probability and the argument in (2) is basically $p(\mathbf{y}|\mathbf{M}_k)$. The integral in (2) is a complicated nonlinear minimization problem which can be solved by, for instance, *Laplace*s method for integration. According to [10], instead of numerical integration for the evaluation

of marginal density in (2), we employ *asymptotically* MAP criterion as

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k : k \in Z_K} \left\{ -L(\hat{\boldsymbol{\theta}}_k) + p_c \right\}, \tag{3}$$

where p_c is the penalty of the MAP criterion and $-L(\hat{\boldsymbol{\theta}}_k)$ is log-likelihood term, given $\mathbf{M}_k$. Let $\hat{\boldsymbol{\theta}}_k$ be our feature parameters for the kth model, $\mathbf{M}_k$. As our signal modeling, to find $\hat{\boldsymbol{\theta}}_k$, we use *sinusoidal modeling* described in [7] which is based on selecting one peak per frequency band. Let $\mathbf{e}_i$ be the residual signal due to the sinusoidal modeling error in the ith band indicated by $\mathbf{e}_i = \mathbf{y}_i - \mathbf{s}_{1,i}(\boldsymbol{\theta}_1)$, where σ_i denotes the variance of the error signal in the ith band, $\mathbf{e}_i$, due to the modeling error and $\boldsymbol{\theta}_1$ is the parameter vector of length $3 \times L$ for the first speaker composed of sinusoidal parameters, L being the model order of sinusoids. Given the independence assumption in the frequency bands in subband decomposition, the likelihood function for all Q bands is

$$p(\mathbf{e}|\sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{i=1}^{Q} \sigma_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{Q} \frac{\mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}}{2\sigma_i^2} \right), \tag{4}$$

where $(\cdot)^{\mathrm{H}}$ represents the Hermitian operator. A similar expression goes for the second speaker class, $\mathbf{M}_2$ just by replacing $\mathbf{e}_i = \mathbf{y}_i - \mathbf{s}_{2,i}(\boldsymbol{\theta}_2)$ in (4), where $\boldsymbol{\theta}_2$ is the parametric vector for the second speaker.

We also include the noise model as one of the examined models by setting $g(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}) = \mathbf{e}$ and setting the number of sinusoids equal to zero ($L = 0$). We define $p(\mathbf{e}|\sigma_0^2)$ as the probability density function, with $\mathbf{e}$ considered as zero mean Gaussian noise whose noise variance is estimated by $\hat{\sigma}_0^2 = \frac{1}{N}\mathbf{y}\mathbf{y}^T$ and likelihood function given by (4).

As our last hypothesis, we are required to include the mixture model, $\mathbf{M}_3$ where the residual signal for the ith band is considered as a colored noise not fitted by $\mathbf{M}_3$ denoted by $\mathbf{e}_i = \mathbf{y} - \hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\})$. The negative log-likelihood function for mixture model $\mathbf{M}_3$ is

$$-\ln p(\mathbf{y}|\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}, \hat{\sigma}_i^2, \mathbf{M}_3) = \frac{N}{2} \sum_{i=1}^{Q} \ln(2\pi\hat{\sigma}_i^2) + \frac{1}{2} \sum_{i=1}^{Q} \frac{\mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}}{\hat{\sigma}_i^2}. \tag{5}$$

In order to form the MAP criterion in (3), we employ the MAP criterion [10] for sinusoids composed of amplitude and unknown frequencies and $\hat{\mathbf{M}}_k$ is obtained as

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k \in Z_K} \left\{ \frac{N}{2} \sum_{i=1}^{Q} \ln \hat{\sigma}_i^2 + \frac{5L}{2} \ln N \right\}, \tag{6}$$

where $\hat{\sigma}_i$ is the estimated variance for the modeling error defined for each model. For mixed class, $\hat{\mathbf{M}}_3$, as our mix model denoted by $\hat{\mathbf{s}}^{(J)}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\})$, we use the

minimum mean square error (MMSE) mixture estimate presented recently in [11]. According to (6), the best model, as a result, is the one which yields high log-likelihood and low model order, which is achieved according to (6) [5].

Figure 2 shows the clean signal for two speakers together with their mixture. It is observed that, the double-talk detector effectively finds the boundaries of single-talk regions. Comparing with the ground-truth, it accurately determines for each frame that which speaker(s), if any, is active. It is important to note that, for same gender or same talker scenarios DTD module degenerates into a three-class problem since it only employs one speaker model for these scenarios. Then, the double-talk detector cannot distinguish between $\mathbf{M}_1$ and $\mathbf{M}_2$, since the residual signals of these classes, are the same. The double-talk detector, however, can still identify single/double-talk regions and pass this information to the SID module.

3 Speaker Identification System

The speaker identification module is based on maximum *a posteriori* (MAP) adapted Gaussian mixture models (GMM) [12]. A speaker GMM is a weighted linear combination of M unimodal Gaussian densities where, letting λ denote a model of single speaker, the likelihood function is defined as,

$$\ell(\mathbf{x}) = p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x}), \tag{7}$$

where $p_m(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m)$ and the mixture weights w_m further satisfy the constraints $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$. Speaker-dependent GMMs are adapted from a so called *universal background model* (UBM) [12]. The UBM is a GMM trained on a pool of feature vectors extracted from as many speakers as possible and it serves as *a priori* information for feature distribution. By defining λ_{ig} as the signal-to-signal ratio dependent model for the ith speaker at SSR level g, we use frame-level likelihood and model-level approximate *Kullback-Leibler divergence* (KLD) as the similarity and distance measures respectively. For a feature vector $\mathbf{x}_t$ extracted from a speech segment at time instance t, frame level score for speaker i is defined as $s_{it} = \frac{1}{G}\sum_{g=1}^{G} s_{igt}$, where

$$s_{igt} = \log[p(\mathbf{x}_t|\lambda_{ig})] - \log[p(\mathbf{x}_t|\lambda_{UBM})]. \tag{8}$$

For speaker identification, we average over all SSR levels to make the system less dependent on the SSR level [4]. Meanwhile we normalize all speakers scores at time instance t with the corresponding UBM score. Another approach to measure similarity of a speech segment with a speaker model (λ_i) is to make a

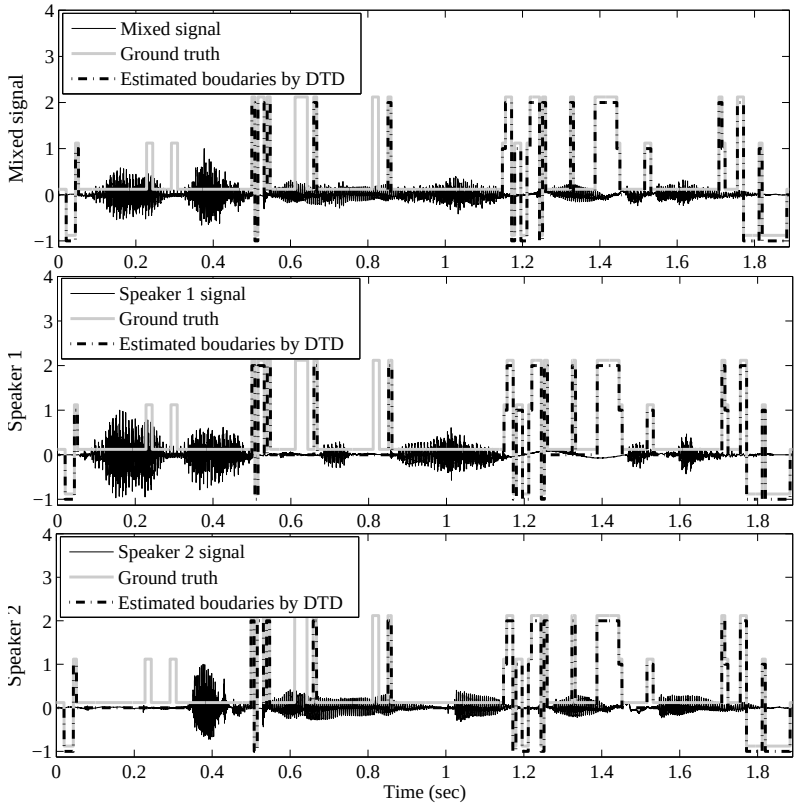

Fig. 2: *Double-talk detection results for a speech mixture of a male and a female speaker mixed at 3 dB SSR. The mixed signal is composed of a male speaker 12 uttering "Lay white with e 8 again" with female speaker 11 uttering "Set green with v 3 soon". Decisions are -1 for no speech, 1 for speaker 1, 2 for speaker 2 and 0 for mixed signal regions.*

model from the test utterance with MAP adaptation (λ_e) and calculate the distance between λ_e and the speaker model. Since KLD distance cannot be directly evaluated for GMMs, we use the upper bound of KLD which has successfully been applied to speaker verification [13]:

$$\begin{aligned} \mathrm{KLD}_{ig} &= \tfrac{1}{2}\textstyle\sum_{m=1}^{M} w_m(\boldsymbol{\mu}_{me}-\boldsymbol{\mu}_{mig})^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1} \\ &\quad (\boldsymbol{\mu}_{me}-\boldsymbol{\mu}_{mig}). \end{aligned} \tag{9}$$

Here g ranges over a discrete set of SSR levels, $\boldsymbol{\mu}_{me}$ is the mth mean vector in λ_e and $\boldsymbol{\mu}_{mig}$ is the mth mean vector in λ_{ig}, whereas w_m and $\boldsymbol{\Sigma}_m$ are the weights and the covariances of the UBM, respectively. To sum up, we consider two different scores for a speaker:

FLL: *Frame level likelihood*, where we are considering number of winning frames that speaker i is the most probable speaker in that frame for speaker identification.

KLD: *Kullback-Leibler divergence* between λ_e and a set of models λ_{ig}, computed using (9). We form an $N \times G$ distance matrix and average over SSR levels to raise the speaker with minimum average distance.

As commonly done in speaker recognition, to enable using benefits from different recognizers, we considered the fusion of the scores with equal weights. Similar to [4], each speaker's decision score is computed as $0.5 \times \text{FLL} + 0.5 \times \text{KLD}$. The frames detected by DTD module to belong to a single speaker only (1 or 2) are collected accordingly and passed to KLD score computation. Since we believe that these frames belong to only one speaker, for the speaker that gets the minimum KLD, we add a bonus score to it's decision score as $score[idx] = score[idx] + \alpha T_1/T$ (or $\alpha T_2/T$) where idx is the identified speaker from single-talk frames. The bonus is made relative to the number of single-talk frames identified to belong to speaker 1 or 2 (T_1 or T_2) respect to total number of frames in a given test signal (T). The stressing factor α is a control parameter. Details of the SID algorithm presented as a pseudocode in Algorithm 3.1.

4 Experimental Results

We evaluate the proposed system on the speech separation database known as GRID corpus [2] composed of 34,000 different utterances. The sentences were originally sampled at 25 kHz with a duration of 2 seconds each. As we usually deal with 8 kHz speech in most of speech applications, we decreased the sampling rate down to 8 kHz. The speaker models used for DTD module are split-VQ codebooks [7] composed of sinusoidal amplitude and frequencies. For training the speaker models, we used 11 bits for amplitude and 3 bits for frequency part. To train gender-dependent models, we selected 10 female and 10 male speakers each producing 35 s of speech signal. Throughout the experiments, a Hamming window of length 32 ms with frame-shift equal to 8 ms is used to segment the speech files both in the training and test phases. As our test data, we used the mixture of target and masker speakers in the test setup of [3] mixed at six SSR levels of $\{-9, -6, -3, 0, 3, 6\}$ dB. The codebook size for split-VQ was M=2048 and the sinusoidal model order was set to 50.

For speaker identification, we extract features from 30 ms frames multiplied by a Hamming window. A 27-channel mel-frequency filterbank is applied on DFT spectrum to extract 12-dimensional mel-frequency cepstral coefficients (MFCCs), followed by appending Δ and Δ^2 coefficients, and using an energy-based voice activity detector (VAD) for extracting the feature vectors. We digitally add the signals with an average frame-level SSR to construct the UBM and

Table G.1: Speaker identification accuracy (% correct) where both speakers are correctly found in the top-3 list. Yes/No indicates whether the proposed DTD method is included. For the ST scenario both of the systems provide 100 % accuracy.

	SG		DG		Average	
DTD	No	Yes	No	Yes	No	Yes
SSR						
-9 dB	92.74	93.30	82.50	86.97	92.00	94.68
-6 dB	96.65	96.65	94.00	95.00	97.00	97.71
-3 dB	99.44	99.44	97.50	98.00	99.00	99.39
0 dB	98.32	98.32	99.00	98.00	99.17	99.39
3 dB	97.21	97.77	93.50	95.00	97.00	98.11
6 dB	93.85	94.41	90.50	89.50	95.00	95.63
Average	96.36	96.65	92.83	93.83	96.53	97.43

the target speakers GMMs. For each of 34 speakers, 50 random files from each speaker were mixed at SSRs levels $\{-9, -6, -3, 0, 3, 6\}$ dB with 50 random files from other speakers which gives us about 180 hour of speech for training UBM. The model order of the GMM is set to 2048.

The speakers' SSR-dependent GMMs, λ_{ig} , trained by mixing 100 random files from each speaker with 100 random files from other speakers yielding about 1.8 hours data for each SSR. Relevance factor was set to 16 for training speaker models, λ_{ig}, where its value was set to 0 in training test model, λ_e, because of availability of only 2 seconds of data for adaptation. For each six test sets of two-talker signal, 600 utterances were provided among which 200 were for same gender (SG), 179 for different gender (DG), and 221 for same talker (ST) where the target and masker signals are from the same speaker. To incorporate the bonus for single-talk detected frames, we used $\alpha = 5$.

Speaker identification results for the combined system presented in table 1. Compared to the previous results without DTD [4], embedding the DTD module enhances performance. The improvement is higher on the different gender (DG) case where the gender-dependent DTD module distinguishes between single-talk areas for two speakers accordingly. Compared to the reported accuracy of 99 % for the *Iroquois* system for detecting target speakers among three most probable cases [2], the proposed system achieves a comparable rate of 97.43%. Given its relatively low complexity, our proposed system could be considered as an alternative or a pre-processing block for *Iroquois* system.

5 Conclusions

We introduced gender-dependent double talk detector for monaural speech and applied it in speaker identification task for. Speaker identification results on

GRID corpus demonstrated the improvement over the system without DTD. Overall speaker identification performance is close to the results of the *Iroquois* system using computationally simple approach.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] M. Cooke, J. Hershey, and S. Rennie, "An audio-visual corpus for speech perception and automatic speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[3] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[4] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. Christensen, S. Jensen, and P. Fränti, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," Aug. 2010.

[5] P. Mowlaee, M. Christensen, Z.-H. Tan, and S. Jensen, "A MAP criterion for double-talk/single-talk detection in model-basded single-channel speech separation," Aug. 2010.

[6] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, 2003, pp. 1009–1012.

[7] P. Mowlaee, R. Saeidi, Z.-H. Tan, M. Christensen, P. Fränti, and S. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4430–4433.

[8] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Proc. Interspeech*, 2006, pp. 97–100.

[9] P. Mowlaee, M. Christensen, Z.-H. Tan, and S. Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," Aug. 2010.

[10] P. M. Djuric, "Asymptotic MAP criteria for model selection," *Signal Processing, IEEE Transactions on*, vol. 46, no. 10, pp. 2726–2735, Oct 1998.

[11] P. Mowlaee, M. Christensen, Z.-H. Tan, and S. Jensen, "The MMSE mixture estimator for single-channel speech separation," Aug. 2010.

[12] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[13] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Sig. Proc. Letters, IEEE*, vol. 13, no. 5, pp. 308–311, May 2006.

Paper H

A Joint Approach for Single-Channel Speaker Identification and Speech Separation

P. Mowlaee, R. Saeidi, M. G. Christensen, Z. -H. Tan, Tomi Kinnunen, P. Fränti, and S. H. Jensen

This paper has been submitted to
IEEE Transactions on Audio, Speech, and Language Processing,
December 2010.

The layout has been revised.

Abstract

In this paper, we present a system for joint speaker identification, signal-to-signal ratio (SSR) estimation and speech separation. For speaker identification and SSR estimation, a novel single-channel speaker identification algorithm is proposed and adopted for SSR estimation. For the separation part, we propose a double-talk/single-talk detector followed by a minimum mean square error estimator of sinusoidal parameters aimed at finding optimal codevectors from each pre-trained speaker codebook. We start from a situation where we have prior information of codebook index, speaker identities and SSR-level, and then, by relaxing these assumptions one by one, we demonstrate the efficiency of the proposed full system. Compared to previous studies that mostly focused on speech recognition accuracy, here, we report the perceived signal quality and intelligibility of the separated signals. To this end, we report perceptual evaluation of speech quality scores, objective speech intelligibility measure and cross-talk measure as objective measures and MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) and intelligibility tests as subjective measurements. The proposed method achieves on average, 41.2 points in MUSHRA and 85.9% in speech intelligibility.

1 Introduction

Human beings have the amazing capability of perceiving multiple speech sources from their mixtures. For machines, however, separating speech mixtures recorded by single microphone is still a rather difficult task. Designing reliable and robust speech processing systems for adverse conditions is a challenging problem since the observed signal is often corrupted by other interfering signals, making the performance significantly lower to that of clean conditions. In extremely noisy environments, a high-quality speech separation algorithm is required as a pre-processing stage before the target application, such as hearing aids, automatic speech recognition, speaker/language recognition and speech coding (see Fig. 1). Being able to separate the desired sources from the interfering ones in the mixture, one would expect a better performance in all these applications.

A single-channel speech separation (SCSS) system aims at recovering the underlying speaker signals from a mixed signal. At first look, SCSS is similar to speech enhancement but the goal in SCSS is to recover *all* the underlying signals rather than enhancing the desired speech signal by filtering out others. In speech separation, the stronger signal can shift its role to a weaker one at some time-frequency regions, and, further, at different signal-to-signal ratios (SSRs) either one of the signals may dominate the other one [1]. It is arguable that one would be interested in separating either of the source signals from their single-channel

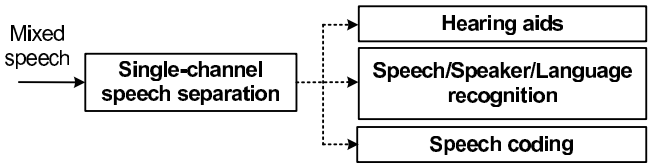

Fig. 1: Block diagram showing how a single-channel speech separation module can be used as a pre-processing stage to enhance the ultimate performance of a target application.

recorded mixture in certain applications, including signal recovery at low signal-to-noise ratios (SNRs), surveillance and tele-conferencing. Another application is speech watermarking where the goal is opposite of separating mixtures, i.e., designing a mixed signal such that the hidden data satisfies high capacity, high perceptual fidelity and robustness against channel attacks [2]. In summary, in all the applications discussed above, it is important to recover all the speaker signals in the mixture.

Conventional speech enhancement methods including spectral subtraction [3] and Wiener filter [4] cannot effectively denoise the mixed signal. More advanced speech enhancement methods e.g. [5–7] are mostly based on second order statistics of the noise signal estimated from preceding noise-only frames. As a consequence, they cannot be used when the interfering signal is another speech signal due to its fast varying statistics [8, ch. 9]. As another classical enhancement approach, subspace-based methods [9–11] aims at estimating the unknown speech signal by decomposing the observed noisy signal into signal and noise subspaces. This decomposition is possible under the assumption of an uncorrelated additive noise interference. However, for the mixture of speech signals, finding such decomposition is difficult since the signal bases of the speakers in the mixture follow similar probability density function, which is mostly super-gaussian [12, 13]. As a variant of subspace method for basis decomposition, non-negative matrix factorization (NMF) has been applied to SCSS. It decomposes short-time Fourier transform (STFT) power of a mixed signal into a product of two low-rank matrices, namely basis vectors and their corresponding weights. According to [14], NMF cannot always separate speech mixtures when the sources largely overlap or when the speakers are of same gender.

Due to these limitations in applying classic speech enhancement methods for separating speech mixtures, dedicated methods to tackle the SCSS problem are needed. The current approaches are divided into two groups, computational auditory scene analysis (CASA) [15–17] and model-driven methods [18–23]. CASA methods use *multi-pitch* estimation methods [24] to extract pitch estimates of the speakers directly from the mixture. The separation performance of CASA-based methods, as a consequence, is predominantly affected by the accuracy of the multi-pitch estimator, especially when the pitch of one of the speakers is masked by the other [25]. To compensate for the inherent problems of

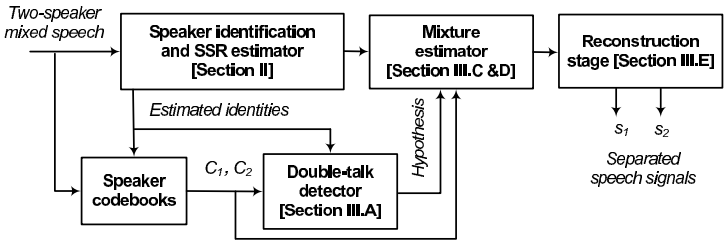

Fig. 2: Block diagram of the proposed joint speaker identification and speech separation system.

multi-pitch estimation methods, an adaptive comb filter was proposed in [26]. It suggested to enhance harmonic signals via eliminating harmonic interference from the noisy observation. More recently, [27] proposed signal-adaptive optimal filters aimed at separating single-channel mixtures of periodic signals.

The second group, model-driven methods, uses pre-trained speaker models as *a priori* information to constrain the solution of the ill-conditioned SCSS problem. In particular, source-specific speaker models are incorporated to capture specific characteristics of individual speakers at each frame. The model-driven methods require mixture estimation stage which aims at finding the most likely codevectors in the speaker models. These found codevectors are then passed to a reconstruction stage which produces the separated signals (see Fig. 2). In terms of how to reconstruct the separated signals, separation methods can be divided into *reconstruction*-based [20–23] and *mask* methods [18, 19, 28]. In the former approach, the codevectors found in the mixture estimation stage are directly used for reconstructing the separated signals. The mask methods, as the name suggests, produce a mask based on the codevectors selected from the speaker models. These masks are then applied to mixture to provide separated speaker signals.

Model-based SCSS approaches are, by definition, speaker-dependent. Speaker-dependency can be avoided by integrating a speaker identification (SID) module to identify the speakers in the mixed signal. In [18], Gaussian mixture models (GMMs) and pitch-dependent method were used for speaker identification and speech re-synthesis, respectively. In [29], interaction of the speaker identification (SID) and speech separation modules was managed in a closed loop to increase the overall performance. The separation system in [21] used max-model and *Algonquin* as their mixture models. Max-model approximates the mixture logarithm power spectrum by the maximum element-wise logarithm of the speaker spectra while *Algonquin* models the combination of log-spectrum models as a sum in the power spectrum. In [21], *Iroquois* system was used for determining the identities of the speakers present in the mixture. The *Iroquois* system uses minimum mean square error (MMSE) estimator reconstruction of the speakers' features, under factorial hidden markov model (HMM), for modeling the mix-

ture frame, using speaker-dependent models, estimated SSR and maximum *a posteriori* (MAP) joint grammar state sequence of the speakers [21, 30].

The contribution of the current study, as illustrated in Fig. 2, is to propose a novel joint speaker identification and speech separation system which is composed of double-talk/single-talk detector followed by MMSE mixture estimator in sinusoidal domain, speaker identification and SSR estimation. First, the mixed signal is input to a joint speaker identification and gain-estimation module (Section II). The estimated identities are then passed to a double-talk/single-talk detector which outputs one of the three possible hypotheses: non-speech, single speaker and mixture (Section III-A). The next block, mixture estimator, aims at finding the magnitude spectra of the underlying two speakers, by utilizing the codebooks of the speakers (Section III-C). The estimated codevectors, provided by the mixture estimator, are then used for generating masks, one for each speaker, to reconstruct the separated signals (Section III-E).

The proposed system is evaluated on the *speech separation challenge* corpus provided in [31]. Our main focus in this work is to assess the speech quality and intelligibility of the separated signals offered by different separation methods. For evaluation purposes, in contrast to previous studies focusing on speech recognition accuracy, in Section IV, we further employ perceptual evaluation of speech quality (PESQ) [32] and the short-time objective intelligibility measure in [33] as objective measures. This alternative evaluation methodology has the benefit that the results are expected to carry on to other applications beyond ASR, as indicated in Fig. 1. In Section V, we employ MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [34] and intelligibility test [35] as subjective measurements. Further, the separation results are compared with two well-known benchmark methods, IBM's *super-human* speech recognition system [21] and speaker-dependent full system [20]. We also assess the speaker identification and SSR estimation accuracy for different mixing scenarios. Section VI features the discussion and Section VII concludes the work.

2 Speaker Identification and Gain Estimation

Speaker identification (SID) is the task of recognizing speaker identity based on observed speech signal [36]. Typical speaker identification systems consist of short-term spectral feature extractor (front-end) and a pattern matching module (back-end). In traditional SID, the basic assumption is that only one target speaker exists in the given signal whereas in *co-channel* SID, the task is to identify two target speakers in a given mixture. Distinct from the so-called *summed channel* speaker recognition task [37], where only one speaker is talking most of the time, in the co-channel SID problem, both speakers talk simultaneously which makes the problem much more challenging. Research on co-channel

speaker identification has been done for more than one decade [38], yet the problem remains largely unsolved.

Most of the current SCSS systems use model-driven *Iroquois* system [21] to identify the speakers in a mixed signal. Recognition accuracy as high as 98% on the speech separation corpus [31] has been reported for *Iroquois* [21], which makes it as a viable choice to be used in SCSS systems [20]. In the *Iroquois* system, a short-list of the most likely speakers are produced based on the frames of the mixed signal that are dominated by one speaker. This short-list is then passed to a *max-based EM algorithm* to find the SSR and the two speakers' identities with an exhaustive search on codebooks created for speech synthesis [21].

As an alternative approach, we propose an SID module for co-channel speech which also produces SSR estimate (assuming constant SSR level when mixing two utterances), as a by-product [39]. Based on the preliminary results reported in [39], we believe that our alternative solution provides faster approach compared to *Iroquois*. We use two complementary methods for speaker recognition: (1) frame-level log-likelihood calculation for a given mixed signal against a speakers' GMM and (2) between-model distance of GMMs.

2.1 Recognition Approach

The frame-level likelihood estimation used here is inspired by the method used in the *Iroquois* system [18, 21] which aims at finding the frames where only a single speaker is present. Maximum likelihood (ML) trained GMMs were used in [21]; however, *maximum a posteriori* (MAP) derived GMMs [40] are much more accurate in speaker verification and we follow this latter approach. Let λ denote a GMM of one speaker. The probability density function is

$$p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x}). \tag{1}$$

The GMM density function is a weighted linear combination of M Gaussian densities $p_m(\mathbf{x})$, where $p_m(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m)$. Here $\mathbf{\Sigma}_m$ is diagonal covariance matrix and the mixture weights w_m further satisfy the constraints $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$. The speaker-dependent GMMs are adapted from a universal background model (UBM) [40]. The UBM is a GMM trained on a pool of feature vectors, extracted from as many speakers as possible, to serve as *a priori* information for the acoustic feature distribution. When adapting the speaker-dependent GMMs, usually only mean vectors are adapted while weights and covariances are shared between all speakers [40]. Considering $\mathbf{x}_t$ as the tth training vector out of T vectors for speaker i and μ_m^{UBM} as the mth mean vector

for UBM, MAP-adapted mean vector for speaker i is calculated as [40]

$$\mu_{mi} = \frac{\beta \mu_m^{\text{UBM}} + \sum_{t=1}^{T} Pr(m|\mathbf{x}_t)\mathbf{x}_t}{\beta + \sum_{t=1}^{T} Pr(m|\mathbf{x}_t)}, \tag{2}$$

where β is a fixed *relevance factor* that controls relative contributions of UBM statistics and training data in GMM adaptation and $Pr(m|\mathbf{x}_t)$ is *probabilistic alignment* defined as,

$$Pr(m|\mathbf{x}_t) = \frac{w_m \mathcal{N}(\mathbf{x}_t; \mu_m^{\text{UBM}}, \mathbf{\Sigma}_m)}{\sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}_t; \mu_m^{\text{UBM}}, \mathbf{\Sigma}_m)}. \tag{3}$$

Another approach to measure the similarity of speaker models, $\{\lambda_i\}$, is to train a model of the test utterance, λ_e, with MAP adaptation and calculate the distance between λ_e and the speaker model. We define λ_{ig} as the SSR-dependent model for the ith speaker at SSR level g and use the *Kullback-Leibler divergence* (KLD) as a distance measure between the two probability distributions [41]. Since this distance cannot be evaluated in closed form for GMMs, we use the upper bound which has successfully been applied to speaker verification [42]:

$$\text{KLD}_{ig} = \frac{1}{2} \sum_{m=1}^{M} w_m (\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig})^T \mathbf{\Sigma}_m^{-1} (\boldsymbol{\mu}_{me} - \boldsymbol{\mu}_{mig}). \tag{4}$$

Here g ranges in a discrete set of pre-defined SSR levels, $\boldsymbol{\mu}_{me}$ is the mth mean vector in λ_e and $\boldsymbol{\mu}_{mig}$ is the mth mean vector in λ_{ig}, whereas w_m and $\mathbf{\Sigma}_m$ are the weights and the covariances of the UBM, respectively.

2.2 Joint Speaker Identification and SSR Estimation

In traditional speaker recognition, the UBM is trained from a pool of data from different speakers. To characterize mixed speech, in this study we propose to train the UBM (λ_{UBM}) from mixed utterance pairs at different SSR levels. For the ith speaker, the gain-dependent models are adapted from the UBM using ith speaker speech files corrupted by other speakers signal at SSR level g. Using SSR-dependent speaker models, the system captures speaker-specific information when it is contaminated by other speakers. Our method is similar to having an SSR-dependent bias in the GMM parameters [21], but we build separate GMMs for each SSR level to utilize the advantages of GMM-UBM system [40].

For a feature vector $\mathbf{x}_t$ extracted from frame t, we define the log-likelihood ratio score for speaker i as $s_{it} = \max_g \{s_{igt}\}$,

$$s_{igt} = \log p(\mathbf{x}_t|\lambda_{ig}) - \log p(\mathbf{x}_t|\lambda_{\text{UBM}}). \tag{5}$$

Given a test utterance, we are interested in the model λ_{ig} which gives the smallest KLD distance (4) and the highest frame likelihood (5). To sum up, we consider two different scores for each speaker:

FLL: Frame level likelihood, where we consider the number of winning frames where speaker i is the most probable speaker. For the frames identified as belonging to speaker i, we look for the most probable SSR level by finding λ_{ig} which maximizes the likelihood score in (5).

KLD: Kullback-Leibler divergence between λ_e and a set of models λ_{ig} calculated using (4). Considering N as the number of speakers, we form an $N \times G$ distance matrix and find the minimum over all SSR levels to detect the speaker with the smallest distance. We find the most probable SSR levels for each speaker by sorting the KLD scores in an ascending order.

To enable using benefits from different recognizers, we combine the two scores with equal weights summation. Although non-equal weights can be estimated from development data [39], we found that using equal weights yields similar accuracy. Note that we normalize the range of scores from two recognizers before fusion. The implementation details of the proposed system is presented in Algorithm.

2.3 Selecting the optimal SID and SSR pair

The joint speaker identification and separation module produces short-lists of speaker identities and the SSR candidates. In our preliminary speaker identification experiments, we found that the dominant speaker was *always* correctly identified and the second speaker also ends up most of the time in the top-3 list. Thus, rather than selecting the top-scoring speaker or the most likely SSR level, we propose the following procedure to refine speaker identification and SSR estimation results.

Let SID_1 denote the estimated identity for the first speaker. Assume that the estimated top-2 identities for the second speaker are $\mathrm{SID}_2 = \{\mathrm{SID}_2^{(1)}, \mathrm{SID}_2^{(2)}\}$. Additionally, we define $\mathrm{SSR} = \{\mathrm{SSR}_1^{(i)}, \mathrm{SSR}_2^{(i)}, \mathrm{SSR}_3^{(i)}\}$ as the short-list for SSR candidates consisting of three most likely SSR levels for combination of speakers SID_1 and $\mathrm{SID}_2^{(i)}$ with $i \in \{1, 2\}$. The search space is shown graphically in Fig. 3. The speaker identity and SSR candidates in the reduced search space are further passed to the separation module which attempts to reconstruct the mixed signal as combinations of both the two top-scoring speakers and the three SSR candidates. The details of the reconstruction error computation are given in Section III-E.

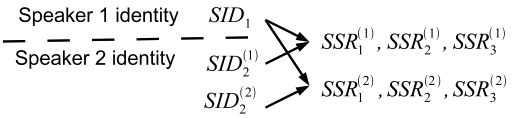

Fig. 3: Demonstration of the reduced search space for speaker-SSR combination. There are $N(N-1)/2 \times G$ possible combination for N speakers and G SSR levels, which is reduced to 2×3 combinations by the proposed joint speaker identification and gain estimation algorithm.

3 Single-channel Speech Separation System

Let $s_z(n)$ denote the nth sample of the observed mixed signal with N' samples composed of B additive source signals as

$$s_z(n) = \sum_{b=1}^{B} g_b s_b(n), \quad n = 0, \ldots, N' - 1. \tag{6}$$

Here, $s_b(n)$ is the bth speaker signal in the mixture, and g_b is its *gain*. Note that the speaker gains are assumed to be fixed over the entire utterance. This assumption, although somewhat unrealistic, is made in most current speech separation systems [31]. For the sake of simplicity and tractability, we consider the case $B = 2$, a mixture of two speakers. The goal of a SCSS system is to estimate the unknown speaker signals based on the observed mixture in (6).

3.1 Double-talk Detection

A mixed speech signal can be classified into single-talk (one speaker), double-talk (speech mixture), and noise-only regions. This information can be used to simplify the computationally expensive separation task since we only need to process the mixed frames with the separation system. To detect double-talk regions with two speakers present, we employ a MAP detector proposed recently in [43]. The proposed method is based on multiple hypothesis test and can be implemented in both speaker-dependent and speaker-independent scenarios. We consider here the speaker-dependent scenario since the information for speaker identities are given by SID module (Section II). We use three candidate models for describing the mixed signal, namely,

$\mathbf{M}_0$: None of the speakers are active (non-speech)
$\mathbf{M}_1$: One of the speakers is active (single-talk)
$\mathbf{M}_2$: Both of the speakers are active (double-talk)

Following the model selection approach in [44], we adopt a MAP criterion for multiple-hypothesis test to determine the double-talk and single-talk regions in

segments of a mixed signal. The double-talk detection approach is summarized in the following steps [43]:

(1) Find the noise variance of each frame.
(2) Assess the log-likelihood of the three models $\mathbf{M}_0$, $\mathbf{M}_1$ and $\mathbf{M}_2$.
(3) Find the MAP estimate among the underlying candidate models.

We use the decision making among $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ to narrow down the separation problem only for the mixed frames. For the single-speaker frames, the observed signal is directly re-synthesized according to the corresponding speaker models. For more details of the method, refer to [43].

3.2 Sinusoidal Signal Modeling

The selected feature for separation needs to meet at least two requirements: (i) high re-synthesized signal quality, and (ii) low number of features for computational and statistical reasons (curse of dimensionality [45]). A vast majority of the previous separation methods are based on short-time Fourier transform (STFT) features which poorly match the logarithmic frequency sensitivity of auditory system [46]. In this paper, we choose sinusoidal parameters which satisfy both of the aforementioned requirements as well as it leads to an improved signal quality compared to the STFT approaches in terms of both objective and subjective measures [23]. Furthermore, in [47], it was shown that applying a sinusoidal coder as speaker model results in a better quantization performance compared to STFT features, in having less outliers.

The proposed separation system transforms the underlying speaker signals into a parametric feature set composed of amplitude, frequency and phase vectors of sinusoidals. After translating the spectral coefficients to the mel-frequency scale, at each frequency band, we select sinusoidal amplitude and frequency corresponding to the peak with the highest amplitude [23]. Taking the highest peak of the amplitude spectrum is, in fact, equivalent to choosing the maximum likelihood estimate for frequency of single sinusoid in white Gaussian noise per band [48, ch. 13].

3.3 MMSE Mixture Estimator

In model-driven speech separation we estimate the codevectors in the speaker models whose combination best matches the mixed signal. This is accomplished by employing a *mixture estimator*. In the following, we present the MMSE mixture estimator for the SCSS problem. Assume that $\sum_{k=1}^{K} S_1^2(\omega_k) = \sum_{k=1}^{K} S_2^2(\omega_k) = 1$ where K is the number of DFT bins. Beginning from the relationship between the mixed signal and the underlying signals in time-domain

given in (6), we have

$$g_z S_z(\omega_k) = \sqrt{g_1^2 S_1^2(\omega_k) + g_2^2 S_2^2(\omega_k) + 2g_1 g_2 S_1(\omega_k) S_2(\omega_k) \cos\theta(\omega_k)}, \quad (7)$$

where $S_1(\omega_k)$, $S_2(\omega_k)$ and $S_z(\omega_k)$ are the kth components of the magnitude spectrum for the first speaker, the second speaker and the mixed signal, respectively and $k \in [1, K]$. We also define $\theta(\omega_k) = \phi_1(\omega_k) - \phi_2(\omega_k)$ as the phase difference between the kth frequency bin of the underlying spectra. Dividing both sides of (7) by $g_1^2 S_1^2(\omega_k) \neq 0$, we arrive at

$$\frac{g_z^2 S_z^2(\omega_k)}{g_1^2 S_1^2(\omega_k)} = 1 + \frac{g_2^2 S_2^2(\omega_k)}{g_1^2 S_1^2(\omega_k)} + \frac{2g_1 g_2 S_2(\omega_k) S_1(\omega_k)}{g_1^2 S_1^2(\omega_k)} \cos\theta(\omega_k). \quad (8)$$

By defining $\tilde{S}_z(\omega_k) \triangleq \ln S_z^2(\omega_k)$ and $\tilde{S}_i(\omega_k) \triangleq \ln S_i^2(\omega_k)$ for $i = \{1, 2\}$, using (8) we get

$$\begin{aligned} \tilde{S}_z(\omega_k) \quad &= \tilde{S}_1(\omega_k) + \ln\left(1 + G_{\text{prior}}^{-1} e^{\tilde{S}_2(\omega_k) - \tilde{S}_1(\omega_k)}\right) \\ &+ \ln G_{\text{prior}} G_{\text{post}}^{-1} + \ln\left(1 + \frac{\cos\theta(\omega_k)}{\cosh\left(\frac{\tilde{S}_2(\omega_k) - \tilde{S}_1(\omega_k)}{2G_{\text{prior}}}\right)}\right), \end{aligned} \quad (9)$$

where we define $G_{\text{prior}} \triangleq g_1^2/g_2^2$ and $G_{\text{post}} \triangleq g_z^2/g_2^2$ as the *a priori* and the *a posteriori* gains similar to *a priori* and *a posteriori* SNRs in speech enhancement [5]. A similar expression can be derived by dividing both sides of (7) to $S_2^2(\omega_k) \neq 0$, which gives

$$\begin{aligned} \tilde{S}_z(\omega_k) \quad &= \tilde{S}_2(\omega_k) + \ln\left(1 + G_{\text{prior}}^{-1} e^{\tilde{S}_2(\omega_k) - \tilde{S}_1(\omega_k)}\right) \\ &- \ln G_{\text{post}} + \ln\left(1 + \frac{\cos\theta(\omega_k)}{\cosh\left(\frac{\tilde{S}_1(\omega_k) - \tilde{S}_2(\omega_k)}{2G_{\text{prior}}}\right)}\right). \end{aligned} \quad (10)$$

The derivation presented here is similar to [49], for representing the relationship among the log-spectra of the noisy signal for speech enhancement, but adopted here for speech mixture of two speakers. In the following, we derive a closed-form solution to the minimum mean square error (MMSE) mixture estimation problem. Let $\hat{S}_z(\omega_k)$ be the MMSE estimate for mixture magnitude spectrum averaging out $\theta(\omega_k)$. The signal spectra of the underlying speakers, $S_1(\omega_k)$ and $S_2(\omega_k)$, are considered to be given but in the experiments, we relax this assumption by choosing the estimates of $S_1(\omega_k)$ and $S_2(\omega_k)$ from the pre-trained codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$ of the two speakers.

Given the speaker's signal spectra, $S_1(\omega_k)$ and $S_2(\omega_k)$, and modeling the mixture phase with uniform distribution [50], the MMSE estimate for the mixed

magnitude spectrum $S_z(\omega_k)$ is

$$\hat{S}_z(\omega_k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g_z S_z(\omega_k) d\theta(\omega_k). \tag{11}$$

Following a similar approach as in [51], (11) simplifies to

$$\hat{S}_z(\omega_k) = \frac{1}{\pi}\left(\sqrt{\frac{G_{\text{prior}}}{G_{\text{post}}}} S_1(\omega_k) + \frac{1}{\sqrt{G_{\text{post}}}} S_2(\omega_k)\right)\mathcal{E}(\gamma(\omega_k)), \tag{12}$$

where $\gamma(\omega_k) = 4(\sqrt{\text{SSR}_{\text{prior}}} + \frac{1}{\sqrt{\text{SSR}_{\text{prior}}}})^{-2}$, $\text{SSR}_{\text{prior}} \triangleq \xi_k = g_1 S_1(\omega_k)/g_2 S_2(\omega_k)$ and $\mathcal{E}(\cdot)$ is the *complete Elliptic integral of the second kind*. This integral can be approximated by selecting some terms of the following series:

$$\mathcal{E}(\alpha) = \pi\left\{1 - \sum_{m=1}^{\infty}\left[\prod_{v=1}^{m}\left(\frac{2v-1}{2v}\right)^2\right]\frac{\alpha^{2m}}{(2m-1)}\right\}. \tag{13}$$

The Elliptic series denoted by $\mathcal{E}(\cdot)$ can also be written as

$$\mathcal{E}(\gamma(\omega_k)) = \frac{\pi}{2}{}_2F_1(-0.5, 0.5; 1; \gamma^2(\omega_k)), \tag{14}$$

where ${}_2F_1(a, b; c; t)$ is *Gauss' hypergeometric function* with α as an argument replaced by $\gamma^2(\omega_k)$. Provided that $|\alpha| \leq 1$, $\mathcal{E}(\gamma(\omega_k))$ will converge absolutely, and since $\gamma(\omega_k) \leq 1$, convergence is indeed guaranteed. Note that the values of ${}_2F_1(\cdot)$ can be found from a look-up table since it depends on a single variable, $\gamma(\omega_k)$. This helps keeping the complexity of the mixture estimator low.

3.4 Relation to Existing mixture Estimators

It is important to note that previous separation systems used either *max-model* [28] or *Algonquin model* [52, 53] as their mixture estimator. A simplified version of the max-model, MAX-vector quantization (MAX-VQ) was used in [18, 20, 28]. In [21], both the Algonquin and the max-model were studied and compared, and Algonquin was found to perform slightly better. Wiener filter has also been employed as MMSE estimator in power-spectrum domain [19, 54]. Comparing to the log-max and Wiener filter, in the MMSE estimator proposed here, we consider the phase term as a random variable and derive the MMSE estimate for the magnitude mixture spectrum. This results in a more general and accurate estimator for mixture magnitude spectrum compared to previous estimators. Furthermore, according to [30], specifying the mixture estimation stage in the log spectral domain is convenient because speech states can be represented efficiently as a mixture of Gaussians in the log-spectrum. For reconstruction

purposes, then, they use anti-logarithmic transformation. However, in this paper, here, we solve the problem directly in spectrum amplitude domain without logarithmic mapping.

Under specific conditions, the MMSE mixture estimator reduces to previous mixture estimators: log-max and Wiener filter, also known as the MMSE estimates for log and power-spectrum domain, respectively. Consider the case when one speaker dominates the other, i.e. $\tilde{S}_1(\omega_k) \gg \tilde{S}_2(\omega_k)$ or $\tilde{S}_2(\omega_k) \gg \tilde{S}_1(\omega_k)$. Then (9) reduces to log-max mixture approximation; this is because the logarithmic terms in (9) will be zero. Another important case is when $\tilde{S}_1(\omega_k)$ and $\tilde{S}_2(\omega_k)$ are orthogonal, i.e., $\cos\theta(\omega_k) = 0$. Then (9) reduces to the Wiener filter mixture estimate; this is because the last logarithmic term will be equal to zero. The derivation presented here was obtained by considering $g_1 \neq 1$ and $g_2 \neq 1$, which are to be estimated by the SSR estimation module presented in Section II.

3.5 MMSE Mixture Estimator in Sinusoidals

Here, we explain how to find $\hat{S}_z(\omega_k)$, the estimated mixture magnitude spectrum, at the kth frequency bin. To implement the mixture estimator in (12), we need the magnitude spectra of the two speakers, $S_1(\omega_k)$ and $S_2(\omega_k)$. The estimates for $S_1(\omega_k)$ and $S_2(\omega_k)$ are obtained from the codebooks of the two speakers, $\mathbb{C}_1 = \{\mathbf{c}_1^{(1)}, \mathbf{c}_2^{(1)}, \ldots, \mathbf{c}_r^{(1)}, \ldots, \mathbf{c}_M^{(1)}\}$ and $\mathbb{C}_2 = \{\mathbf{c}_1^{(2)}, \mathbf{c}_2^{(2)}, \ldots, \mathbf{c}_q^{(2)}, \ldots, \mathbf{c}_M^{(2)}\}$, respectively, where $\mathbf{c}_r^{(1)}$ and $\mathbf{c}_q^{(2)}$ refer to the rth and qth codevector in the codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$, respectively, and M is the model order of sinusoidal speaker models [47]. The mixture estimation is carried out by searching for the optimal codevectors of the codebooks by minimizing the *mixture estimation error*,

$$J_{r,q} = \sum_{l=1}^{L} |S_z(\omega_l) - \hat{S}_z^{r,q}(\omega_l)|^2, \tag{15}$$

where ω_l is the peak selected from the speech mixture at the lth band with $l \in [1, L]$ and $\omega_l \in \{\omega_k\}_{k=1}^{K}$. Here, $\hat{S}_z^{r,q}(\omega_l)$ denotes the estimated mixture produced by the codevectors r and q. In (15), r and q are the codebook indices selected from $\mathbb{C}_1$ and $\mathbb{C}_2$, respectively. To minimize (15), we are required to do search on pairs of codevectors to determine the optimal pair for signal reconstruction that is,

$$\{r^*, q^*\} = \arg\min_{\{r,q\}} J_{r,q}. \tag{16}$$

Note that, even after knowing the estimated SSR level and identities of the speakers, exhaustive search of (16) requires $\mathcal{O}(M^2)$ evaluations of the cost function (15) for *all frames*, which is impractical. Considerable time saving, still retaining high separation quality, can be obtained by using an EM-like search as

follows: We start with random r, and keep it fixed while optimizing with respect to q, then switching the roles. This requires a total number of $\mathcal{O}(M \times I)$ evaluations of (15), where we typically use $I = 3$ iterations. This leads to practical speed-up factors of 700:1 for a codebook size $M = 2048$.

3.6 Signal Reconstruction using Sinusoidal Wiener Mask

Wiener filter is a classical speech enhancement method that relies on the MMSE estimation to restore the underlying clean signal [55]. Previous studies utilizing the Wiener filter [8] operate in the STFT domain. Here we propose to use a similar constrained optimization problem already studied in speech enhancement [8, ch. 6]. Instead of noise distortion, we balance a trade-off between minimizing the speech distortion of the target signal and keeping *cross-talk* of the other speaker below a given tolerance threshold. This is well-justified, as in speech separation we are required to have no trace of cross-talk from the other speaker. Indeed, according to our preliminary experiments in [56], the masks defined in the sinusoidal domain, including both the binary and Wiener masks, improve the separation quality as compared to their STFT counterparts. The *sinusoidal Wiener mask* in [56] is only manipulated in sinusoidal frequencies estimated from the speaker signal. The sinusoidal Wiener mask is similar to the so-called *parametric Wiener filter* [8, p. 158], which is a general case of other Wiener filters like square-root Wiener filter. Differently from [8], where a Gaussian assumption was used in modeling the noisy signal, here, we have a mixture of two speakers rather than a mixture of speech and noise. For synthesizing the separated signals, we use the sinusoidal Wiener mask to recover the unknown signals in the given mixture. Like other separation methods reported in [31], we employ the mixture phase for re-synthesizing the separated outputs.

It can be shown that the sinusoidal Wiener mask is similar to the well-known *Ephraim and Malah* noise suppression rule used in speech enhancement [57]. To this end, from sinusoidal Wiener mask for the first speaker and using $\cos\theta(\omega_k) = 0$, we have

$$G_1(\omega_k) = \sqrt{\frac{\xi_k}{\xi_k + 1} \frac{S_1^2(\omega_k) + S_2^2(\omega_k)}{S_z^2(\omega_k)}} = \sqrt{\frac{\xi_k}{\xi_k + 1} \left(\frac{1 + \nu_k}{\zeta_k} \right)} \tag{17}$$

Similarly to [57, 58] let $\nu_k = \mathrm{SSR}_{\mathrm{post}} \frac{\xi_k}{\xi_k + 1}$ be the *instantaneous SSR*. We further define $\mathrm{SSR}_{\mathrm{post}} \triangleq \zeta_k = g_z S_z(\omega_k) / g_2 S_2(\omega_k)$. From (20) and using the fact that at

every ω_k we have $G_1(\omega_k) + G_2(\omega_k) = 1$, we obtain:

$$G_1(\omega_k) = \frac{1}{2}\sqrt{\frac{\xi_k}{\zeta_k}}\,{}_2F_1(-0.5, 0.5; 1; \gamma^2(\omega_k)), \quad (18)$$

$$G_2(\omega_k) = \frac{1}{2\sqrt{\zeta_k}}\,{}_2F_1(-0.5, 0.5; 1; \gamma^2(\omega_k)). \quad (19)$$

These expressions are similar to the classical Ephraim and Malah rule given by [57],

$$G_1(\omega_k) = \Gamma(1.5)\frac{\sqrt{\nu'_k}}{\gamma'_k}M(-0.5; 1; -\nu'_k), \quad (20)$$

where ν'_k and γ'_k are the instantaneous and *a posteriori* SNRs as defined in [57], $\Gamma(\cdot)$ is the Gamma function with $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$ and $M(a; c; x) = {}_1F_1(a; c; x)$ is the confluent hyper-geometric function which is a limiting case of the more general hypergeometric function as [59, p. 158]: $M(a, c, h) = \lim_{d\to\infty} {}_2F_1(a, b; c; h/d)$. Replacing h/d with $\gamma^2(\omega_k)$, the condition $\gamma^2(\omega_k) \to 0$ indicates the situation when we have only one speaker (i.e. $S_1(\omega_k) = 0$ or $S_2(\omega_k) = 0$).

4 Objective Evaluations

4.1 Dataset and System Setup

The proposed speech separation system is evaluated on the speech separation corpus provided in [31]. This corpus consists of 34,000 distinct utterances from 34 speakers (18 males and 16 females). The sentences follow a command-like structure with a unique grammatical structure as six word commands such as "*bin white at p nine soon*" and "*set blue at z five please*". Each sentence in the database is composed of verb, color, preposition, letter, digit and coda. The keywords emphasized for speech intelligibility or recognition task in challenge are the items in position 2, 4, and 5 referring to color, letter and digit, respectively. The possible choices for color are green, blue, red, and white. The possible letters are 25 English alphabet letters and finally the digits are selected from 0 to 9.

For each speaker, 500 clean utterances are provided for training purposes. The test data is a mixture of target and masker speakers mixed at six SSR levels ranging from -9 dB to 6 dB. For each of the six test sets, 600 utterances are provided of which 200 are for same gender, 179 for different gender, and 221 for same talker. The sentences were originally sampled at 25 kHz. We decrease the sampling rate to 16 kHz (Some additional experiments are also carried on at 8 kHz). The results presented here are averaged over all the utterances in the dataset.

For speaker identification, we extract features from 30 ms Hamming-windowed frames using a frame shift of 15 ms. A 27-channel mel-frequency filterbank is applied on DFT spectrum to extract 12-dimensional mel-frequency cepstral coefficients (MFCCs), followed by appending Δ and Δ^2 coefficients, and using an energy-based voice activity detector for extracting the feature vectors. We add the signals with an average frame-level SSR to construct the universal background model (UBM) and the target speaker GMMs. For each of the 34 target speakers, 50 randomly chosen files from each speaker are mixed at SSR levels $g \in \{-9, -6, -3, 0, 3, 6\}$ dB with 50 random files from all other speakers, which gives us about 180 hours of speech for UBM training. The number of Gaussians is set to M=2048.

Each SSR-dependent GMM, λ_{ig}, is trained by mixing 100 random files from the ith speaker with 100 random files from all other speakers which gives about 1.8 hours data for training. The relevance factors in MAP adaptation were set to β=16 for the training of speaker models and β=0 for the training of the test utterance models, respectively. The choice of β=0 for the test utterance was done due to short length of data for adaptation. For SID and SSR estimation, the fusion of the FLL and KLD was done by employing equal weights. Table H.1 shows the accuracy of the proposed speaker identification module for listing two target speakers in the 3-best list where one of the speakers in the mixed signal is always identified. An average accuracy of 97% is achieved using the proposed SID module.

For separation, we extract features by employing a Hann window of length 32 ms and shift of 8 ms. We use split-VQ based on sinusoidal parameters. The source models are divided into magnitude spectrum and frequency parts where each entry is composed of a sinusoidal amplitude vector and several sinusoidal frequency vectors as its candidates. According to previous experiments, we set the sinusoidal model order to L=100 for 16 kHz and L=50 for 8 kHz [23]. For speaker modeling, we use 11 bits for amplitude and 3 bits for frequency part in the sinusoidal coder. The pre-trained speaker codebooks are then used in the test phase to guide the speech separation. The codebooks are used for both mixture estimator and double-talk detection blocks (Fig. 2). For the mixture estimator given in (12), we used the first 5 terms of the Elliptic series in (13).

4.2 Mixture Estimation Results

To evaluate the effectiveness of the MMSE estimator, we compare it to the log-max [28], Wiener filter [54] and sinusoidal mixture estimators [23]. We use the mixture estimation MSE in (15) to evaluate the separation performance of each mixture estimator. For each experiment, the mixture estimation error is calculated at each frame and averaged over 250 utterances selected from the test corpus. The resulting MSE scores obtained by each method and its 95%

Table H.1: Speaker identification accuracy (% correct) where both speakers are correctly found in the top-3 SID list.

SSR (dB)	-9	-6	-3	0	3	6	Average
Same Talker	100	100	100	100	100	100	100
Same Gender	93	97	100	100	97	94	96
Different Gender	83	94	98	99	94	91	93
Average	92	97	99	100	97	95	97

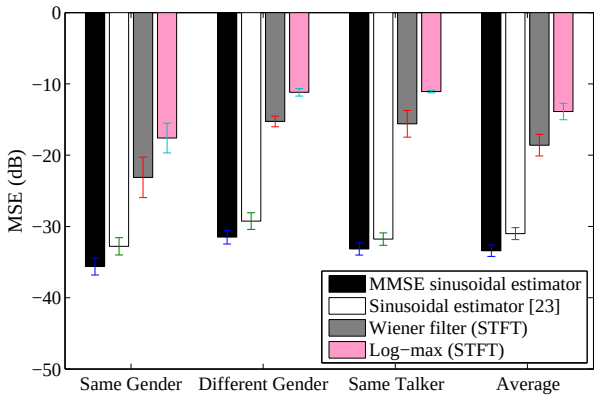

Fig. 4: The MSE results for mixture magnitude spectrum estimation measured in dB averaged over different SSRs for different mixing scenarios (same gender, different gender, same talker and their average.)

confidence interval are shown in Fig. 4. The sinusoidal MMSE mixture estimator achieves the lowest MSE values. It outperforms log-max and Wiener filter mixture estimators in the STFT domain by a wide margin. We remark that the confidence intervals for the MSE results of sinusoidal and MMSE sinusoidal estimators are close but they do not overlap. The MMSE sinusoidal estimator leads to a slightly better statistical result compared to the sinusoidal mixture estimator in [23].

4.3 Separation System Set-up and Benchmark Methods

To study the performance of the proposed speech separation system, we consider six different setups, covering the situations from all parameters known to all parameters estimated. These six setups are shown in the legend of Fig. 5 as scenarios 1, 2 and 3 with their corresponding upper bounds (which we call *known codebook index*). Parameters that we consider include codebook index, speaker

identities and SSR level. The scenarios are defined as:

- Scenario 1: known SID and SSR,
- Scenario 2: estimated SID and known SSR,
- Scenario 3: estimated SID and SSR.

Estimated codebook index refers to the situation where r and q are jointly estimated from the mixed signal using (16). In turn, if we estimate r and q from the original spectra, $S_1(\omega)$ and $S_2(\omega)$, using

$$r^* = \arg \min_{\mathbf{c}_r^{(1)} \in \mathbb{C}_1} \|S_1(\omega_l) - \hat{S}_1^r(\omega_l)\|_2^2, \quad (21)$$

$$q^* = \arg \min_{\mathbf{c}_q^{(2)} \in \mathbb{C}_2} \|S_2(\omega_l) - \hat{S}_2^q(\omega_l)\|_2^2, \quad (22)$$

we call the set-up *known codebook index*. In (21), $\{\hat{S}_1^r(\omega_l)\}_{l=1}^L = \mathbf{c}_r^{(1)}$ and $\{\hat{S}_2^q(\omega_l)\}_{l=1}^L = \mathbf{c}_q^{(2)}$ are the estimated magnitude spectra of speaker one and two selected from $\mathbb{C}_1$ and $\mathbb{C}_2$, respectively. *Known codebook index* is the best possible performance obtainable by the model-driven speech separation approach [60]. Similar to *known/estimated codebook indices*, we also consider degradations caused by erroneous speaker identities and SSR estimation.

We include both objective and subjective measures to assess separation quality. For objective measurement, we use PESQ [32], and *speech intelligibility* [33] since they correlate well with subjective listening scores [32, 61]. For subjective measurements, we conduct MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test as described in [34] to assess the perceived quality obtained by different separation methods. Furthermore, we conduct speech intelligibility test according to the routine suggested in [35].

As our benchmark methods, we use two systems in [20, 21]. We report the separation results obtained by the *super-human* multi-talker speech recognition system [21] as one of the top-performing separation systems in the single-channel speech separation challenge (SSC), and outperforming even human listeners in some of the speech recognition tasks [31]. We also use "speaker-dependent full system" proposed in [20] (see Table II in [20]) as another benchmark method where *Iroquois* [21] system was used for estimating the speaker identity and the SSR level.

4.4 Separation Results on PESQ and Intelligibility Scores

Figure 5 shows the separation results in terms of PESQ and intelligibility score obtained for different scenarios. The results obtained from mixture and PESQ

scores calculated for the separated wave files of [20] are also shown for comparative purposes. Figure 5 suggests that the proposed method improves the quality of the separated signals compared to the mixture. According to the masking theorem [62], at low SSR levels, *energetic masking* occurs and the separation system successfully performs in compensating this effect by separating the underlying speakers for each frame.

At high SSR levels, *informational masking* is more dominant and the mixed signal itself is more intelligible than the separated signals obtained by separation module. The mixed signal itself achieves higher intelligibility score compared to the separated target signal since the target speaker becomes more dominant. At high SSR levels, the proposed method asymptotically reaches the best possible performance with *known codebook index*.

The proposed method outperforms the method in [20] in terms of PESQ at all SSR levels. It also improves the intelligibility of the target speaker significantly at low SSR levels (lower than -3 dB). However, the speaker-dependent full system in [20] achieves slightly higher intelligibility scores which is not easily audible by listening to the resulting separated signals. By comparing the results of the known and the estimated speaker identities, the results are generally close to each other. The same conclusion holds also for the *known and estimated SSR levels*. This confirms that the SID and SSR estimates were relatively accurate as suggested by Table H.1.

Studying different scenarios, the proposed system performs better for *different gender* compared to the *same gender*. A similar observation was reported in [20]. This can be explained by the different time-frequency masking patterns and physiological differences in the vocal characteristics of male and female speakers. Thus, the underlying sources are less overlapped compared to other scenarios.

4.5 Comparing the Results with Benchmark Methods

An ideal separation system would filter out any trace of the interfering speaker signal in the mixture. As a proof of concept, we use the amount of *cross-talk* remaining in the separated output signal for comparing different separation methods. For measuring the cross-talk, we employ the method first proposed in [63] and afterwards used in [64]. The method produces ideal binary masks constructed from the separated output signals, since ideal binary mask theoretically excludes all the cross-talk from the interfering speaker. We compare the proposed method to those proposed in [20, 21] in Fig. 6. The selected clips from test dataset are composed of same gender, different gender and two same talker scenarios at -3, 0, -6 and 0 SSR levels, respectively. The proposed SCSS method often introduces less cross-talk compared to [20]. Compared to the super-human speech recognition system [21], the proposed method leads to relatively less or

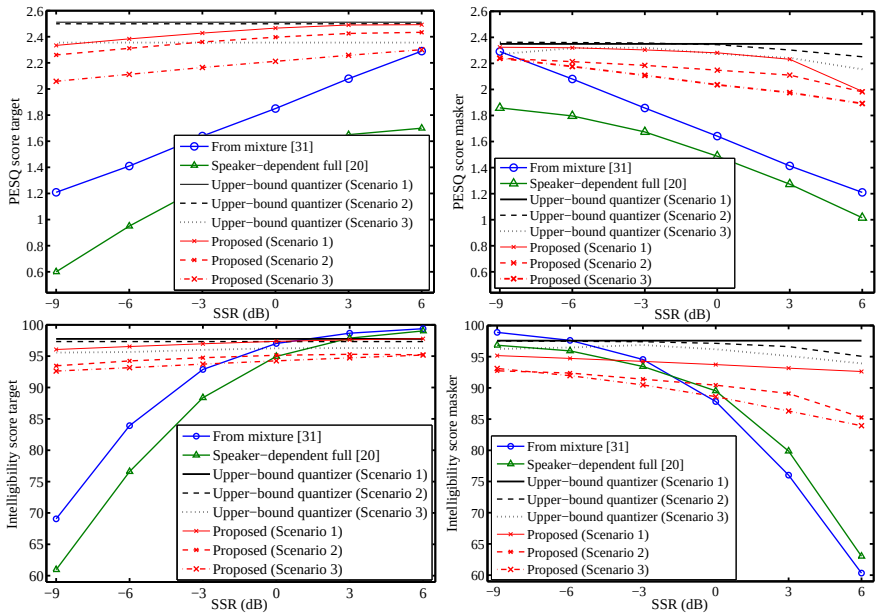

Fig. 5: (Top) PESQ, and (Bottom) intelligibility scores for target and masker. According to [32], for normal subjective test material the PESQ values lie between 1.0 (bad) and 4.5 (no distortion). According to [33], the intelligibility score lies between 0 (bad) and 100 (no distortion). All the results are reported on the speech separation challenge test data provided in [31].

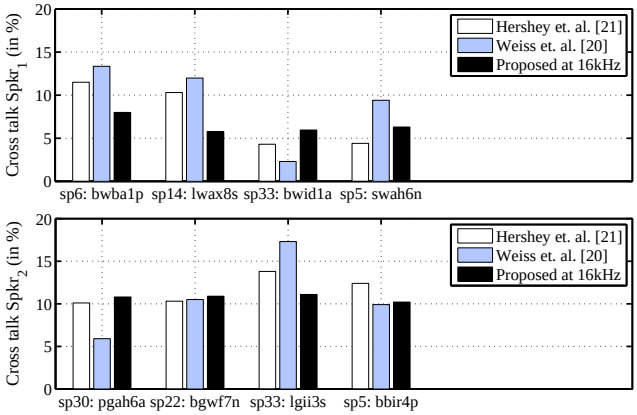

Fig. 6: Comparing the cross-talk results of the proposed method to Hershey et. al. [21] and those of Weiss et. al. [20] for test dataset.

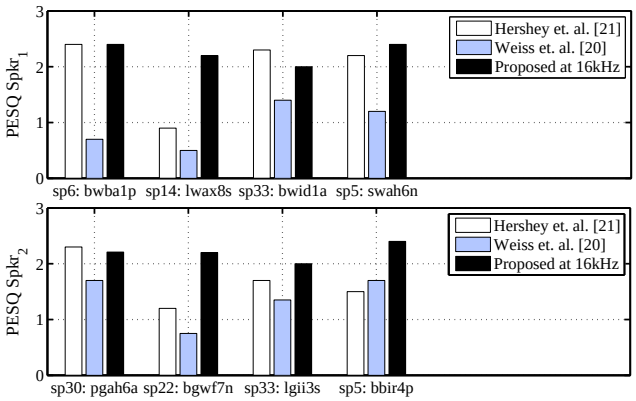

Fig. 7: Comparison of the PESQ values of the proposed method with [20, 21].

comparable amount of cross-talk in most of the cases.

We also report the PESQ values for different methods in Fig. 7. The proposed system yields comparable and improved results over the methods in [21] and [20], respectively.

5 Subjective Evaluation

To assess the perceived speech quality of the separated output signals, we conduct subjective test using the so-called MUSHRA test [34]. The MUSHRA test is a double blind test for the subjective assessment of intermediate quality level benefits obtained by different methods (via displaying all stimuli at the same time). The MUSHRA test enables simultaneous comparison of different separation methods directly.

We conducted the listening experiments in a silent room using high quality audio device with firewire interface for digital-to-analog conversion and AKG K240 MKII headphones. To ease the test procedure, we prepared a graphical user interface (GUI) in MATLAB$^{\text{TM}}$. Seven untrained listeners participated in the test (none of the authors were included). The excerpts consisted of the hidden reference (HR) showing the known quality on the scale; it is used to check the consistency of the responses of a subject. A high score is expected for HR. We also include the mixed signal (without any separation) as an anchor point to enable comparison of separated signal and mixture qualities. This reflects how hard it was to perceive the reference signal when listening to the mixture. The remaining four excerpts are the separated signals obtained by *super-human* speech recognition system [21], speaker-dependent full system [20], and our pro-

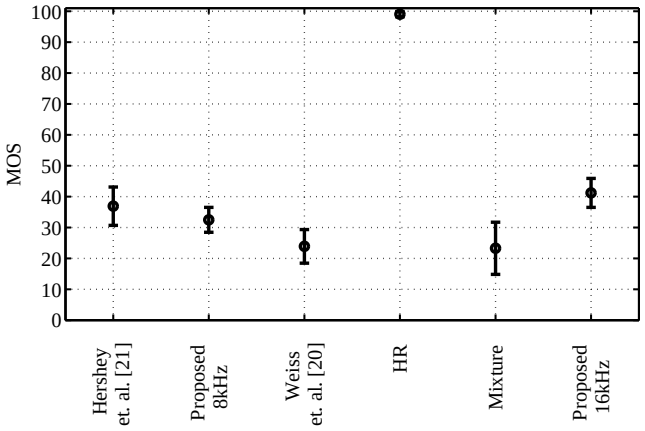

Fig. 8: Results of the MUSHRA listening test for different separation methods averaged over all excerpts and listeners. Error bars indicate 95% confidence intervals.

posed methods configured for both 8 kHz and 16 kHz sampling frequencies. The excerpts were randomly chosen and played for each subject[1]. The listeners were asked to rank eight separated signals relative to a known reference on a scale of 0 to 100.

Figure 8 shows the mean opinion score (MOS) for comparing the separation results obtained by different separation methods discussed in this paper. We observe that the maximum and minimum scores were obtained at hidden reference and speech mixture, respectively, as expected. Further, the proposed method at 16 kHz achieves the best performance compared to the other two separation methods. The proposed method at 8 kHz also achieves comparable result with [21] and is better than the one reported in [20].

Following the principle and standard described in [35], here, we conduct a speech intelligibility test to assess speech intelligibility of the separated signals obtained by different methods. We chose seven listeners and eight segments to be played for each listener. We asked the listeners to identify color, alphabet letter, and digit number spoken during each of the played segments. The listeners were required to enter their results using a GUI in MATLAB$^{\text{TM}}$, which enabled listeners to enter their results both accurately and comfortably. On average, it took 15 minutes per listener to complete the test.

Figure 9 shows the results of the intelligibility test averaged over all excerpts and listeners. We observe that the proposed method at 16 kHz achieves higher speech intelligibility compared to the methods in [21] and [20]. The mixed signal

[1]The excerpts used in our subjective tests are downloadable from our webpage: `http://kom.aau.dk/~pmb/JointIdentificationAndSeparationpaper.htm`

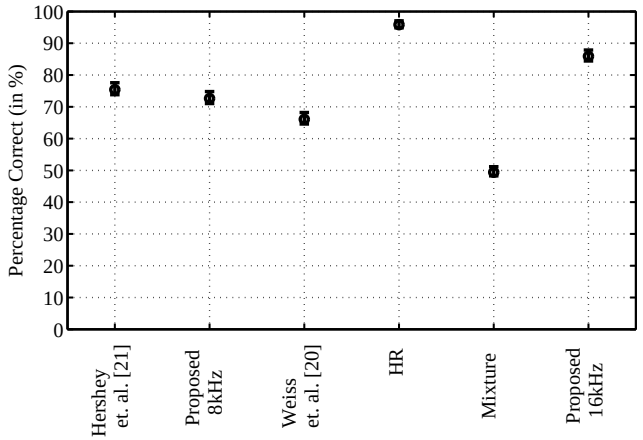

Fig. 9: Speech intelligibility test results. The calculated percentage of correct keywords is averaged over all excerpts and all listeners. Error bars indicate 95% confidence intervals.

also has the lowest score while the hidden reference signal achieves the highest intelligibility score.

6 Discussion

Both the objective and subjective results show that fairly good separation quality was achieved, in comparison to other methods in the field. In particular, the subjective measurements indicated that the proposed method improves both quality and intelligibility of the signal and achieves a performance comparable to the super-human system of [21] and better than [20]. Our proposed separation system separates the mixed signal frame-by-frame and is appropriate for low-delay applications, such as speech coding. Considering N speakers, M Gaussians and G SSR-levels, the number of Gaussian pdf evaluations for speaker recognition system are $\mathcal{O}(NM^N)$ for the *Iroquois* system in [21] and $\mathcal{O}(NGM)$ for the proposed approach. Therefore, the proposed SID module is much faster in operation in exchange of slightly reduced accuracy.

The proposed system, like other current separation systems, still has some limitations. The training samples used to train the speaker models are noise-free and relatively large and the evaluation corpus consists of only digitally added mixtures. Additionally, the gains of the underlying speakers in the mixture are assumed to be constant and that we have only a mixture of two speakers. We also neglected the environmental or background noise effects, as well as the reverberation problem. In practice, each one of these issues and their effect on the

overall separation performance should be carefully studied. Future work should systematically address how these simplifying yet restrictive and impractical pre-assumptions could be relaxed. As an example, recently in [65], a new corpus was provided for noise-robust speech processing research where the goal was to prepare realistic and natural reverberant environments using many simultaneous sound sources.

The improvement gain using the MMSE sinusoidal mixture estimator over our previous sinusoidal mixture estimator [23] can be elaborated as follows. The sinusoidal estimator ignores the cross term components between the underlying speakers' spectra at each frame, as well as the phase differences which, in some situations, plays a critical role and can change the position of spectral peaks completely [23]. The MMSE mixture estimator considers the cross terms and integrates out the phase difference based on uniformity assumption of the speech phase. This explains why the proposed MMSE sinusoidal mixture estimator presented here achieves improved MSE compared to the sinusoidal mixture estimator of [23].

The presented system showed high separated perceived signal quality and intelligibility. The results obtained in the speech intelligibility test can be interpreted as the speech recognition results offered by the separated signals. In our preliminary experiment, we configured an automatic speech recognition system using mean subtraction, variance normalization, and ARMA filtering (MVA) [66], which gave an overall recognition accuracy of 52.3%. Comparing the result with those reported by other participants in the separation challenge [31, Table. 1], we observed that our system is on the range of median over all methods. There are two possible reasons why the ASR results are in disagreement with our signal quality scores. Firstly, the metric used in ASR does not correlate with those used for assessing the signal quality. Secondly, evaluating the separation performance using ASR systems depends on the speech recognizer configuration, features, training of acoustic and language models. It is not trivial to configure an ASR-system optimized for STFT-like features, to work well on sinusoidally coded speech. Therefore, improvement of the automatic speech recognition performance of the proposed system is left as a future work.

7 Conclusion

We presented a joint speaker identification and speech separation system as a novel approach to solve single-channel speech separation problem. For the separation part, we proposed a double-talk/single-talk detector followed by a minimum mean square error mixture estimator for mixture magnitude spectrum operating in the sinusoidal domain. The proposed method does not require pitch estimates and is based on sinusoidal parameters. We relaxed the *a priori* knowl-

edge of speaker identities and the underlying signal-to-signal ratio (SSR) levels in the mixture by proposing a novel speaker identification and SSR estimation method. The proposed system was evaluated on the test dataset provided in *speech separation challenge*. Compared to previous studies that report speech recognition accuracy, we focused on reporting the signal quality performance obtained by different separation methods. To this end, we used PESQ and intelligibility scores as objective measures and MUSHRA and intelligibility tests as subjective measures. From the experimental results, we conclude that the proposed method improves the signal quality of the underlying speakers compared to the mixed signals. It also provides somewhat better signal quality compared to two well-known benchmark methods. Finally, in many cases, the method offered separated signals with less cross-talk and higher intelligibility compared to the benchmark methods.

References

[1] Y. Gu and W. van Bokhoven, "Co-channel speech separation using frequency bin non-linear adaptive filtering," vol. 2, 1991, pp. 949 –952.

[2] K. Hofbauer, G. Kubin, and W. B. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 8, pp. 1624–1637, 2009.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 355–358, Apr 1993.

[6] 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, Sep 1998.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[8] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[9] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov 1995.

[10] P. C. Hansen and S. H. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, Oct. 2005.

[11] ——, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Processing*, vol. 1, p. 24, March. 2007.

[12] R. C. Hendriks and R. Martin, "MAP estimators for speech enhancement under normal and rayleigh inverse Gaussian distributions," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 918–927, March 2007.

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

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

[15] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[16] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77 – 93, 2010.

[17] J. Barker, M. Ning, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 94 – 111, 2010.

[18] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[19] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67 – 76, 2010.

[20] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[21] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Superhuman multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[22] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single-channel separation performance in transform domain," *Journal of Zhejiang University-SCIENCE C, Computers and Electronics*, vol. 11, no. 3, pp. 160–174, Jan. 2010.

[23] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 8, 2010.

[24] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. Synthesis Lectures on Speech and Audio Processing: Morgan and Claypool publishers, 2009.

[25] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the em algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Apr. 1993, pp. 728–731.

[26] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 34, no. 5, pp. 1124–1138, 1986.

[27] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *Signal Processing, IEEE Transactions on*, vol. 58, no. 12, pp. 5969–5983, 2010.

[28] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, 2003, pp. 1009–1012.

[29] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4430–4433.

[30] S. Rennie, J. Hershey, and P. Olsen, "Single-channel multitalker speech recognition," *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 66 –80, nov. 2010.

[31] M. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[32] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Elsevier speech communication*, vol. 2, pp. 749–752, Aug. 2001.

[33] C. Taal, R. 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 Processing*, March 2010, pp. 4214–4218.

[34] "Method for the subjective assessment of intermediate quality level of coding systems." iTU-R BS.1534-1, 2003.

[35] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Commun.*, vol. 49, no. 5, pp. 402–417, 2007.

[36] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[37] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Elsevier Computer Speech and Language*, vol. 20, no. 3, pp. 128–158, 2006.

[38] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Co-channel speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1995, pp. 828–831.

[39] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. Christensen, P. Fränti, and S. H. Jensen, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2010, pp. 4545–4548.

[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Elsevier Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[41] J. Hershey and P. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2007, pp. 317–320.

[42] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[43] P. Mowlaee, M. Christensen, Z.-H. Tan, and S. Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2010.

[44] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct 1998.

[45] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Math Challenges of the 21st Century*, 2000, pp. 1–33.

[46] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, May 2006, pp. 957–960.

[47] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[48] S. M. Kay, *Modern Spectral Estimation: Theory and Application.* Prentice Hall, 1999.

[49] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[50] H. Pobloth and W. B. Kleijn, "Squared error as a measure of perceived phase distortion," *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 1081–1094, 2003.

[51] P. Mowlaee, A. Sayadiyan, and M. Sheikhan, "Optimum mixture estimator for single-channel speech separation," *Proc. IEEE Int. Symposium on Telecommunications*, pp. 543–547, Aug. 2008.

[52] J. F. Brendan, T. T. Kristjansson, L. Deng, and A. Acero, "Learning dynamic noise models from noisy speech for robust speech recognition," in *NIPS14*, 2001.

[53] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2004, pp. 817–820.

[54] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[55] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[56] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Sinusoidal masks for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4262–4266.

[57] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[58] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP J. on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003.

[59] J. Spanier and K. B. Oldham, *An atlas of functions*. Bristol, PA, USA: Taylor Francis/Hemisphere, 1987.

[60] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[61] H. Y. and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 15-20 2007, pp. 561–564.

[62] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. San Diego: Academic Press, 1997.

[63] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sept. 2004.

[64] M. Radfar, R. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 1, p. 15, March 2007.

[65] H. Christensen, J. Barker, N. Ma, and P. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, 2010.

[66] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 257–270, 2007.