Pitch Estimation for Non-Stationary Speech

Christensen, Mads Græsbøll; Jensen, Jesper Rindom

Published in:
Asilomar Conference on Signals, Systems and Computers. Conference Record

DOI (link to publication from Publisher):
10.1109/ACSSC.2014.7094691

Publication date:
2014

Document Version
Early version, also known as pre-print

PITCH ESTIMATION FOR NON-STATIONARY SPEECH

Mads Græsbøll Christensen and Jesper Rindom Jensen

Audio Analysis Lab, AD:MT
Aalborg University, Denmark
email: {mgc,jrj}@create.aau.dk

ABSTRACT

Recently, parametric methods have proven capable of overcoming the problems of correlation-based methods for pitch estimation. However, the argument against such methods is that the underlying model is wrong, particularly for non-stationary signals, like speech. To investigate whether this is true, we propose a new, non-stationary harmonic chirp model for pitch estimation, and we derive an estimator for determining its parameters. Experimental results show that the new model and the estimator lead to both improved pitch estimates and reconstruction quality, but also that the improvements in pitch are usually quite small, typically in the order of a few Hertz.

Index Terms— Pitch estimation, chirp model, speech analysis, non-stationary speech

1. INTRODUCTION

Pitch estimation is a classical problem in speech processing and remains an active research area today. It is an important problem because many signal processing and machine learning tasks rely on pitch information, some examples being speech coding, diagnosis of certain illnesses, speech enhancement and speech separation. The de facto standard remains the nonparametric correlation-based methods, the likely reasons being that these are conceptually simple, fast, and quite mature, and that implementations are freely available on the Internet. However, they suffer from a number of problems, including that the underlying assumptions are not clear, and, as a result, they are not easy to improve, and they are not particularly robust towards noise [1]. In recent years, parametric methods, e.g., [1–4] have been demonstrated to be capable of overcoming many of the shortcomings of the aforementioned nonparametric methods (see, e.g., [1, 5]). However, an argument against the parametric methods, which are most often based on the harmonic model, is that there are important aspects of speech signals that they often do not take into account, such as modulations in pitch and amplitude, especially for long segments.

In this paper, we attempt to address this criticism by proposing a new harmonic chirp model (HCM) for pitch estimation that explicitly takes into account that the pitch is time-varying. Moreover, we derive an estimator for this new model. In its exact form, this proves to be a difficult, multidimensional, nonlinear problem, and we propose a simple, iterative approach for this. We then use the model and its estimator to investigate the importance of taking the non-stationarity of the pitch into account in speech analysis.

Much work has, of course, previously been devoted in the past to analysis of the time-varying aspects of speech, i.e., modulations, including AM-FM models [6–8]. However, these models are too general for the purpose of estimating the pitch, as they do not explicitly model the modulation as being generated by the same process. Polynomial models of modulation have also been considered in, for example, [9–12], however these all model the amplitude modulation, which is only an approximation of the harmonic chirp model proposed herein, and some do not consider its impact on pitch estimation. It should also be mentioned that the notion of using chirp-like basis functions also has been explored in the context of time-frequency analysis, even specifically aimed at harmonic signal and speech [13, 14]. However, these can generally be classified as being nonparametric and are hence fundamentally different from the proposed approach. Chirp-like models have also been studied for single sinusoids, e.g., [15], in the context of sonar, radar, communications, etc. For speech and audio signals, chirp models (or polynomial phase models) have also been considered in [16, 17], however these differ from the proposed model in that different modulations are allowed for the different harmonics. Compared to these, the proposed model is likely to be more robust towards noise and, hence, lead to more accurate estimates, as fewer parameters have to be estimated.

The rest of the present paper is organized as follows: First, the new model, the harmonic chirp model, is introduced and discussed in Section 2. Then, in Section 3, an estimator for

M. G. Christensen was supported, in part, by the Villum Foundation and the Danish Council for Strategic Research of the Danish Agency for Science, Technology, and Innovation under the CoSound project, case number 11-115328.

J. R. Jensen was supported by the Danish Council for Independent Research, grant ID: DFF 1337-00084.

This publication only reflects the authors' views.

finding the parameters of the new model is introduced based on the nonlinear least squares estimator. In Section 4, the properties of the model and the estimator are explored in detail before Section 5 concludes on the work.

2. HARMONIC CHIRP MODEL

For a segment of a speech signal with $n = n_0, \ldots, n_0+N-1$ (with n_0 being the start of the segment) the new harmonic chirp model is given by

$$x(n) = \sum_{l=1}^{L} A_l e^{j\theta_l(n)} + e(n) \tag{1}$$

where L is the number of harmonics (which is here assumed known or found using some other method, e.g., [2, 18]), A_l the lth amplitude and $\theta_l(n)$ is the instantaneous phase of the lth harmonic while $e(n)$ represents all stochastic parts of the observed signal, i.e., background noise, unvoiced speech, etc. Note that $\theta_l(\cdot)$ is a continuous function. To stress this, we write it now as a function of t. It is given by

$$\theta_l(t) = \int_0^t l\omega_0(\tau)d\tau + \phi_l, \tag{2}$$

where $\omega_0(t)$ is the time-varying pitch and ϕ_l is the initial phase of the lth harmonic. We confirm that the instantaneous frequency of the lth harmonic is

$$\omega_l(t) = \frac{d\theta_l(t)}{dt} = l\omega_0(t). \tag{3}$$

In pitch estimation, it is most often assumed (explicitly or implicitly) that the pitch is constant, i.e., $\omega_l(t) = l\omega_0$. We refer to this case as the harmonic model (HM). This is also often the case for the nonparametric methods, such as those based on correlations, since it would not be possible to estimate the correlation sequence from time-averaging otherwise. If we accept that the pitch is slowly and smoothly varying as a function of time, then an appropriate model would be

$$\omega_0(t) = \alpha_0 t + \omega_0, \tag{4}$$

which yields a second-order polynomial instantaneous phase model for the lth harmonic, i.e.,

$$\theta_l(t) = \frac{1}{2}\alpha_0 l t^2 + \omega_0 l t + \phi_l, \tag{5}$$

where $\alpha_0 l$ is then the chirp rate of the lth harmonic. We term α_0 the fundamental chirp rate. We refer to this model as the harmonic chirp model (HCM). The problem considered in this paper is then the joint estimation of α_0 and ω_0 from N samples of a noisy signal $x(n)$ for the purpose of obtaining more accurate estimation of the pitch function $\omega_0(n)$. The model in (4) can be seen as a first-order Taylor approximation to a more complicated and possibly nonlinear pitch function, and the shorter the segments, the more appropriate this model can be expected to be. Similarly, assuming $\alpha_0 = 0$ corresponds to assuming that the pitch is constant over n, something that would be more accurate for even shorter segments yet. When (5) is inserted into (1), we obtain the proposed harmonic chirp model.

3. PROPOSED ESTIMATOR

It must be stressed that since the fundamental chirp rate and the fundamental frequency drive the instantaneous phase of all harmonics, all harmonics should be exploited when estimating these parameters. This means that the application of methods derived for a single sinusoid, e.g., [15], are not optimal for the problem at hand. Consequently, we proceed to derive an optimal estimator for the harmonic chirp model that exploits all harmonics. First, we introduce some useful vectors and matrices. Define a vector containing the observed signal as

$$\mathbf{x} = \begin{bmatrix} x(n_0) & x(n_0+1) & \ldots & x(n_0+N-1) \end{bmatrix} \tag{6}$$

and a vector containing the complex amplitudes, comprised of the amplitudes A_l and the initial phases ϕ_l as

$$\mathbf{a} = \begin{bmatrix} A_1 e^{j\phi_1} & A_2 e^{j\phi_2} & \ldots & A_L e^{j\phi_L} \end{bmatrix}. \tag{7}$$

Then, we define a matrix containing the individual harmonics in the columns as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\omega_0, \alpha_0) & \mathbf{z}(2\omega_0, 2\alpha_0) & \ldots & \mathbf{z}(L\omega_0, L\alpha_0) \end{bmatrix}, \tag{8}$$

where we have omitted the dependencies on ω_0 and α_0 for notational simplicity. The columns of $\mathbf{Z}$ are given by

$$\mathbf{z}(l\omega_0, l\alpha_0) = \begin{bmatrix} e^{j(\frac{1}{2}\alpha_0 l n_0^2 + \omega_0 l n_0)} \\ e^{j(\frac{1}{2}\alpha_0 l (n_0+1)^2 + \omega_0 l (n_0+1))} \\ \vdots \\ e^{j(\frac{1}{2}\alpha_0 l (n_0+N-1)^2 + \omega_0 l (n_0+N-1))} \end{bmatrix}. \tag{9}$$

Assuming that the stochastic parts of the observed signal in (1) are white and Gaussian, the maximum likelihood estimator of the model parameters is the nonlinear least squares (NLS) estimator:

$$\{\hat{\mathbf{a}}, \hat{\alpha}_0, \hat{\omega}_0\} = \arg\min_{\mathbf{a}, \alpha_0, \omega_0} \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|^2. \tag{10}$$

Since the amplitudes in $\mathbf{a}$ are not of interest, we substitute them by their least squares estimate, which yields the concentrated estimator

$$\{\hat{\alpha}_0, \hat{\omega}_0\} = \arg\min_{\alpha_0, \omega_0} \|\mathbf{x} - \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}\|^2. \tag{11}$$

This involves two-dimensional optimization over the nonlinear parameters α_0 and ω_0. For convenience, we introduce the orthogonal projection matrix as

$$\mathbf{\Pi}(\omega_0, \alpha_0) = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H, \tag{12}$$

and its orthogonal complement as $\mathbf{\Pi}^\perp(\omega_0, \alpha_0) = \mathbf{I} - \mathbf{\Pi}(\omega_0, \alpha_0)$ which are written as functions of ω_0 and α_0 to stress their dependencies on these parameters. To solve the above difficult optimization problem in a computationally efficient manner, we propose to proceed as follows.

Let $\hat{\omega}_0^{(i)}$ denote the estimate of ω_0 in iteration i. Then, first obtain an estimate of α_0, denoted $\hat{\alpha}_0^{(i)}$, from a previous estimate of the fundamental frequency $\hat{\omega}_0^{(i-1)}$ for $i = 1, 2, \ldots$ as

$$\hat{\alpha}_0^{(i)} = \arg\min_{\alpha_0} \mathbf{x}^H\mathbf{\Pi}^\perp(\hat{\omega}_0^{(i-1)}, \alpha_0)\mathbf{x}. \tag{13}$$

and then the fundamental frequency, ω_0, given this estimate as

$$\hat{\omega}_0^{(i)} = \arg\min_{\omega_0} \mathbf{x}^H\mathbf{\Pi}^\perp(\omega_0, \hat{\alpha}_0^{(i)})\mathbf{x}. \tag{14}$$

These iterations are then repeated until convergence, which can be defined in terms of either the cost function or the estimates. Regarding the initialization of this procedure, we note that the fundamental chirp rate is generally expected to be low, while the fundamental frequency can be any number in the interval $\omega_0 \in (0, 2\pi/L)$. Therefore, it is natural to initialize the fundamental chirp rate as $\alpha_0^{(0)} = 0$ and then $\hat{\omega}_0^{(0)}$ is simply the fundamental frequency estimate obtained using the HM model, which can be found using any of the methods in [2]. It is possible to implement (14) and (13) efficiently and in a robust manner, because 1) they involve only one-dimensional searches, albeit nonlinear ones, 2) once the initial fundamental frequency has been found using the harmonic model, only small changes in each iteration are expected. In practice, (14) and (13) are implemented via a grid search to locate the minimum in a region near the previous estimate followed by a dichotomous search [19] in the convex region around that minimum. In our experience, this is much less error-prone than gradient-based methods in nonlinear problems. Moreover, we also note that usually only a few iterations are needed for convergence, since the fundamental chirp rates are usually quite small. Regarding implementation issues, it was shown in [15] that to obtain the minimum estimation error for chirp models (and others), then we should choose $n_0 = -(N-1)/2$ (assuming an odd N), and this is also what we do here.

4. EXPERIMENTAL RESULTS

We will start the experimental part of the paper by exploring the differences between the HM, HCM, and a commonly used approximation of HCM [11, 12], where the assumption $x \approx 0$ is used to obtain $e^x \approx (1+x)$. For the HCM, this would mean

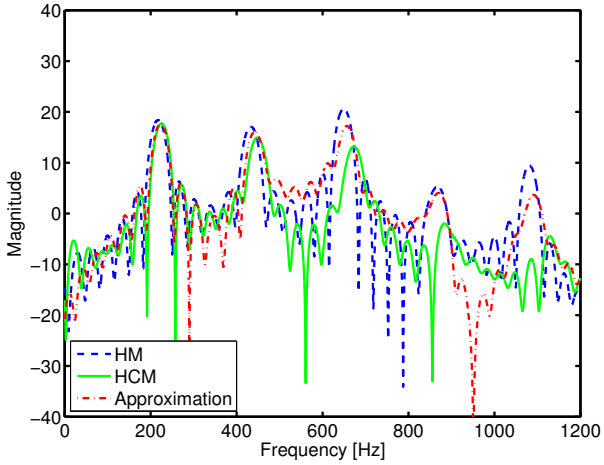

Fig. 1. Spectra for the harmonic model (HM), the harmonic chirp model (HCM), and its approximation.

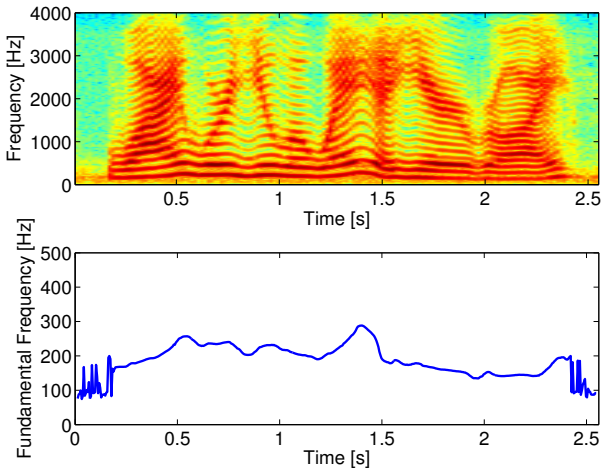

Fig. 2. Spectrogram of the speech signal and fundamental frequency estimates obtained using the proposed method.

that $e^{j(\frac{1}{2}\alpha_0 l n^2 + \omega_0 l n)} \approx (1 + j\frac{1}{2}\alpha_0 l n^2)e^{j\omega_0 l n}$. However, while α_0 may be small, the chirp rate for the higher harmonics are given by $\alpha_0 l$, which means that such an approximation will get progressively worse for higher harmonics. In Figure 1, the spectra of HM, HCM and its approximation are shown for $\omega_0 = 0.2232$ and $\alpha_0 = 1\times10^{-4}$ with $L = 5$. Even though the number of harmonics is quite low, it can be still be seen that the spectra of the higher harmonics of the approximate model do not look much like that of the HCM. Indeed, it appears that this approximation is quite inaccurate, which is why we here use the exact model.

We now present some experimental results with the new model and its estimator. First, an example is shown for the all-voiced female utterance "why were you away a year, roy?" sampled at 8 kHz. The signal is converted to the complex analytic signal using the Hilbert transform. In Figure 2, the

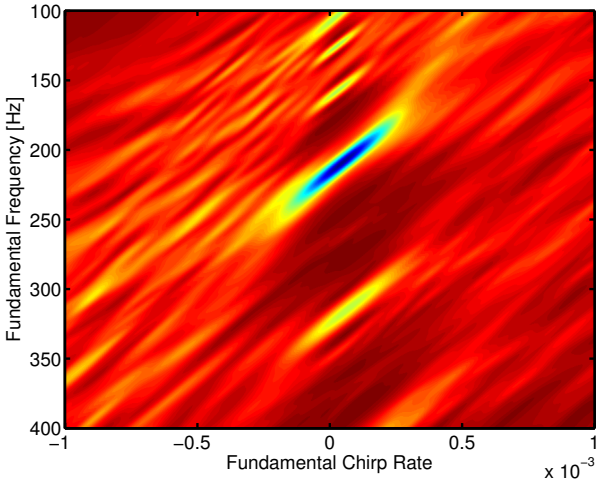

Fig. 3. Example of cost function for speech signal as a function of fundamental frequency and chirp rate.

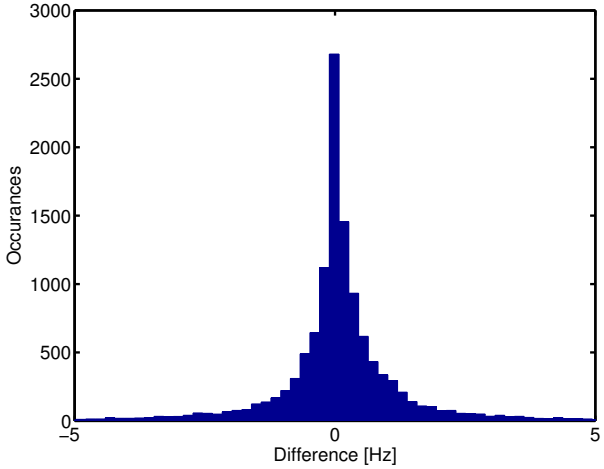

Fig. 4. Histogram of differences (in Hz) between fundamental frequency estimates obtained using the harmonic model and the harmonic chirp model.

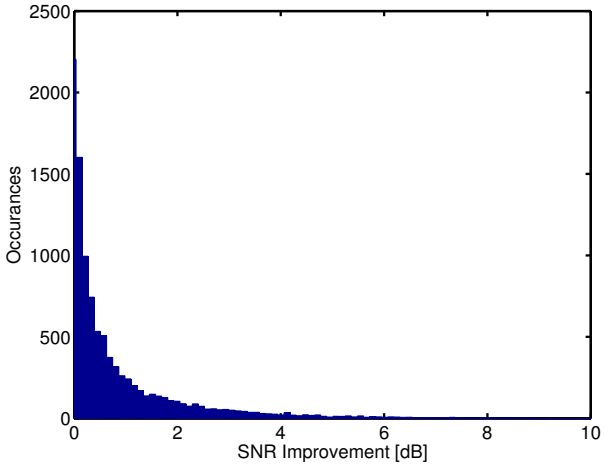

Fig. 5. Histogram of reconstruction SNR improvements (in dB) of using the HCM over the HM.

spectrogram of the speech signal is depicted along with the fundamental frequency estimates obtained using the proposed method. The method is initialized using the MATLAB function `joint_anls()` from the toolbox of [2], which estimates both the pitch and the number of harmonics L, and the estimates are obtained for segments of 30 ms shifted 5 ms. As can be seen, the pitch is varying continuously throughout the signal. In Figure 3, an example of the cost function in (11) is shown for part of the speech signal in Figure 2, namely 30 ms at about 1.3 s where the characteristics of the signal changes rapidly. From the figure, the nonlinear nature of the problem can be seen. Moreover, it can clearly be seen how a fundamental frequency estimate would be highly dependent on the fundamental chirp rate for this part of the signal. As can be seen, assuming a higher chirp rates yields a lower fundamental frequency estimate while a lower one yields a higher one. Hence, if the HM model is used and the the pitch is rising, the estimated pitch will be too high, i.e., it will be biased. It can also be seen that the convex region around the optimal values contain $\alpha_0 = 0$, which means that the presented optimization procedure initialized with the HM is likely to converge even though the problem is not convex.

Next, we will study the impact of the new model on the resulting pitch estimates. We do this based on all 30 clean speech sentences from the NOIZEUS database [20], which are processed in 30 ms segments with in steps of 5 ms. As before, the signals are mapped to complex ones using the Hilbert transform and an initial pitch estimate is found using the harmonic model along with the model order L using `joint_anls()`. This estimate is then refined using a dichotomous search with an exact NLS cost function and subsequently used for initialization of the proposed method for finding the parameters of the harmonic chirp model. In Figure 4, a histogram of the differences (in Hz) between the initial estimates obtained for the harmonic model and the final estimates obtained using the harmonic chirp model. As can be seen, the differences are quite small most of the time, usually in the order of a few Hz, but depending on the application, such differences might be important. And, if these estimates are used for synthesis, differences of this size might be audible in some cases. It should be stressed that the longer the segments, the bigger the difference can be expected between the estimates obtained using the HM and HCM models, as the HCM may better model longer segments. To study the ability of the new model to capture more complicated behavior of speech signals, we also compare the reconstruction SNR of the the two models. The improvements gained with the proposed model are shown in Figure 5 in the form of a histogram of the SNR improvements for individual segments. For this

experiment, we use the maximum number of possible harmonics. Note that for both Figure 4 and Figure 5, the results only include segments that are detected to be voiced. This is done with the the generalized likelihood ratio test (GLRT) for deterministic signals with a linear model with unknown parameters and unknown noise parameters [21] with a false alarm probability of 1×10^{-5} to ensure that only segments that are very likely to be voiced are included. This resulted in about 12,000 voiced segments for the 30 signals.

5. CONCLUSION

In this paper, a new chirp model, called the harmonic chirp model, for pitch estimation has been proposed along with an estimator. The new model captures the non-stationary nature of speech signals using a linear model, parametrized by a fundamental frequency and chirp rate, of the change of the pitch over a segment of speech. The estimator is a nonlinear least squares estimator, which is equivalent to the maximum likelihood estimator for white Gaussian noise. To find the parameters, we propose to first find the pitch and the model order using the usual harmonic model and then use this to initialize the new estimator, which then finds refined estimates of the fundamental chirp rate and the fundamental frequency in an iterative fashion. The resulting method is simple to implement, fast, and provides very accurate pitch estimates. In simulations on speech signals, it has been demonstrated that the proposed model and estimator result in both improved pitch estimates, but also that the improvements are usually quite small, typically in the order of a few Hertz. Moreover, it has been shown that the reconstruction signal-to-noise ratio is improved with the new model compared to the harmonic model. For applications where very accurate pitch estimates are desired, the new model and the estimator may be of interest.

6. REFERENCES

[1] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 12(1), pp. 76–87, 2004.

[2] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, ser. Synthesis Lectures on Speech & Audio Processing. Morgan & Claypool Publishers, 2009, vol. 5.

[3] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14(2), pp. 502–510, 2006.

[4] J. K. Nielsen, M. G. Christensen, ans S. H. Jensen, "Default Bayesian estimation of the fundamental frequency," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21(3), pp. 598–610, 2013.

[5] M. G. Christensen, "Accurate estimation of low fundamental frequencies," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21(10), pp. 2042–2056, 2013.

[6] B. Santhanam and P. Maragos, "Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation," *IEEE Trans. Commun.*, vol. 48(3), pp. 473–490, 2000.

[7] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech and Audio Process.*, vol. 8(3), pp. 240–254.

[8] M. G. Christensen, S. V. Andersen and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. Springer-Verlag, 2003, vol. 2773, pp. 1334–1342, invited.

[9] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 2, 2002, pp. 1769–1772.

[10] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 8(3), pp. 353–357, 2000.

[11] M. Zivanovic and J. Schoukens, "Single and piecewise polynomials for modeling of pitched sounds," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20(4), pp. 1270–1281, 2012.

[12] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, 2009, pp. 3985–3988.

[13] M. Képesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *Speech Communication*, vol. 48(5), pp. 474–492, 2006.

[14] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87(6), pp. 1504–1522, 2007.

[15] P. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38(12), pp. 2118–2126, 1990.

[16] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, 1987, pp. 1641–1644.

[17] F. Myburg, A. C. den Brinker, and S. van Eijndhoven, "Multicomponent chirp analysis in parametric audio coding," in *Fourth IEEE Benelux Signal Processing Symposium*, 2004.

[18] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE Trans. Signal Process.*, vol. 62(1), pp. 225–238, 2014.

[19] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer Verlag, 2007.

[20] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[21] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice-Hall, 1998.