Hindawi Publishing Corporation
Research Letters in Signal Processing
Volume 2008, Article ID 364674, 5 pages
doi:10.1155/2008/364674

Research Letter

Time Domain Method for Precise Estimation of Sinusoidal Model Parameters of Co-Channel Speech

Y. A. Mahgoub and R. M. Dansereau

Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Correspondence should be addressed to Y. A. Mahgoub, ymahgoub@sce.carleton.ca

Received 14 November 2007; Accepted 18 February 2008

Recommended by Tyseer Aboulnasr

A time domain method to precisely estimate the sinusoidal model parameters of cochannel speech is presented. The method does not require the calculation of the Fourier transform nor the multiplication by a window function. It incorporates a least-squares estimator and an iterative technique to model and separate the cochannel speech into its individual speakers. The application of this method on speech data demonstrates the effectiveness of this method in separating cochannel speech signals in different target-to-interference ratios. This method is capable of producing accurate and robust parameter estimation in low signal-to-noise ratio situations compared to other existing algorithms.

1. INTRODUCTION

Separation of mixed speech signals is still one of the major challenges in speech processing. This problem is commonly referred to as co-channel speech separation. The main goal of co-channel speech separation is to automatically process the mixed signal in order to recover each talker's original speech. Minimizing artifacts in the processed speech is a key concern, especially if the final goal is to use the recovered speech in machine-based applications such as automatic speech recognition and speaker identification systems.

Several previous studies have developed signal processing algorithms for modeling and separating co-channel speech. The primary approaches have taken the harmonic structure of voiced speech as the basis for separation and have used either frequency-domain spectral analysis and reconstruction [1–3] or time-domain filtering [4]. One promising approach to address co-channel speech separation is to exploit a speech analysis/synthesis system based on sinusoidal modeling of speech. For example, in [1, 2] a voiced segment of co-channel speech is modeled as the sum of harmonically related sine waves with constant amplitudes, frequencies, and phases. In the sinusoidal modeling approach, the speech parameters of individual talkers are estimated by applying a high-resolution short-time Fourier transform (STFT) to the windowed speech waveform. The frequencies of underlying sine waves are assumed to be known a priori from the individual speech waveforms, or they are determined by using a simple frequency domain peak-picking algorithm. The amplitudes and phases of the component waves are then estimated at these frequencies by performing a least-squares (LS) algorithm. This technique has the following drawbacks:

(1) the accuracy of the estimate is limited by the frequency resolution of the STFT;

(2) error is introduced due to edge effects of the window function used for the STFT.

This paper presents a time domain method to precisely estimate the sinusoidal model parameters of co-channel speech. The method does not require the calculation of the STFT nor the multiplication by a window function. It incorporates a time-domain least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers. The performance of the proposed method is evaluated using a database consisting of a wide variety of mixed male and female speech signals at different target-to-interference ratios (TIRs).

This paper is organized as follows. In Section 2, the sinusoidal model of co-channel speech consisting of K speakers is presented. The proposed time-domain method for

estimating the sinusoidal model parameters is discussed in Section 3. In Section 4, experimental results and comparisons with other techniques are reported and discussed. Finally, the results are summarized and conclusions given in Section 5.

2. SINUSOIDAL MODELING OF CO-CHANNEL SPEECH

According to the speech analysis/synthesis approach based on the sinusoidal model [1], a short segment of co-channel speech (about 20 to 30 milliseconds) can be represented as the sum of harmonically related sinusoidal waves with constant amplitudes, frequencies, and phases as follows:

$$x(n) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} c_\ell^{(k)} \cos\left(\ell\omega_k n - \phi_\ell^{(k)}\right) \tag{1}$$

$$= \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} \left[a_\ell^{(k)} \cos(\ell\omega_k n) + b_\ell^{(k)} \sin(\ell\omega_k n)\right], \tag{2}$$

where $n = 0, \ldots, N-1$ is the discrete time index, ω_k is the fundamental frequency for that segment of the kth talker, and $c_\ell^{(k)}$, $\ell\omega_k$, and $\phi_\ell^{(k)}$ denote the amplitude, frequency, and phase, respectively, of the ℓth harmonic of the kth talker. The total number of harmonics in each talker's model is given as L_k for $k = 1, \ldots, K$. The quadrature amplitudes $a_\ell^{(k)}$ and $b_\ell^{(k)}$ in (2) are related to $c_\ell^{(k)}$ and $\phi_\ell^{(k)}$ in (1) as follows [1]:

$$c_\ell^{(k)} = \sqrt{\left(a_\ell^{(k)}\right)^2 + \left(b_\ell^{(k)}\right)^2}, \qquad \phi_\ell^{(k)} = \tan^{-1}\left(\frac{b_\ell^{(k)}}{a_\ell^{(k)}}\right). \tag{3}$$

A precise estimate of the sinusoidal-model parameters is essential for separating the co-channel speech into its individual components. The basic problem addressed in this paper can be stated as follows. Given the real observed N samples of the co-channel speech sequence $x(n)$, find the parameters $\hat{L}_k$, $\hat{\omega}_k$, $\{\hat{a}_\ell^{(k)}\}_{\ell=1}^{\hat{L}_k}$, and $\{\hat{b}_\ell^{(k)}\}_{\ell=1}^{\hat{L}_k}$ that form the sequence,

$$\hat{x}(n) = \sum_{k=1}^{K} \sum_{\ell=1}^{\hat{L}_k} \left[\hat{a}_\ell^{(k)} \cos(\ell\hat{\omega}_k n) + \hat{b}_\ell^{(k)} \sin(\ell\hat{\omega}_k n)\right], \tag{4}$$

that best fits $x(n)$ by minimizing the mean-squared error (MSE)

$$E = \frac{1}{N} \sum_{n=0}^{N-1} \left[x(n) - \hat{x}(n)\right]^2. \tag{5}$$

In the following sections, we will consider the case of two talkers ($K = 2$) to represent the co-channel speech without loss of generality.

3. TIME-DOMAIN ESTIMATION OF MODEL PARAMETERS

3.1. Estimation setup

In a matrix notation, we may write (4) as

$$\hat{\mathbf{x}} = \mathbf{Q}\mathbf{h}, \tag{6}$$

where $\hat{\mathbf{x}}$ is the vector

$$\hat{\mathbf{x}} = \left[\hat{x}(0), \hat{x}(1), \ldots, \hat{x}(N-1)\right]^T, \tag{7}$$

and $\mathbf{h}$ is given as

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{bmatrix}, \tag{8}$$

with

$$\mathbf{h}^{(k)} = \left[\hat{a}_1^{(k)}, \hat{a}_2^{(k)}, \ldots, \hat{a}_{\hat{L}_k}^{(k)}, \hat{b}_1^{(k)}, \hat{b}_2^{(k)}, \ldots, \hat{b}_{\hat{L}_k}^{(k)}\right]^T. \tag{9}$$

$\mathbf{Q}$ is a matrix of the form

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{(1)} & \mathbf{Q}^{(2)} \end{bmatrix}, \tag{10}$$

where the matrix elements are given as

$$Q_{ij}^{(k)} = \begin{cases} \cos(ij\hat{\omega}_k) & \text{for } j = 1, 2, \ldots, \hat{L}_k, \\ \sin\left(i(j - \hat{L}_k)\hat{\omega}_k\right) & \text{for } j = \hat{L}_k + 1, \ldots, 2\hat{L}_k, \end{cases} \tag{11}$$

with $i = 0, 1, \ldots, N-1$ and $k = 1, 2$. The MSE in (5) can now be written as

$$E = \left|\left|\mathbf{x} - \hat{\mathbf{x}}\right|\right|_2^2 = \mathbf{x}^T\mathbf{x} + \hat{\mathbf{x}}^T\hat{\mathbf{x}} - 2\hat{\mathbf{x}}^T\mathbf{x}, \tag{12}$$

where

$$\mathbf{x} = \left[x(0), x(1), \ldots, x(N-1)\right]^T. \tag{13}$$

Substituting (6) into (12) gives

$$E = \mathbf{x}^T\mathbf{x} + \mathbf{h}^T\mathbf{Q}^T\mathbf{Q}\mathbf{h} - 2\mathbf{h}^T\mathbf{Q}^T\mathbf{x}. \tag{14}$$

The estimation criteria are to seek the minimization of (14) over the parameters $\hat{L}_k$, $\hat{\omega}_k$, $\{\hat{a}_\ell^{(k)}\}_{\ell=1}^{\hat{L}_k}$, and $\{\hat{b}_\ell^{(k)}\}_{\ell=1}^{\hat{L}_k}$.

The most important and difficult part in the estimation process is to estimate the fundamental frequencies $\{\omega_k\}_{k=1,2}$. Unfortunately, without a priori knowledge of the frequency parameters, direct minimization of (14) is a highly nonlinear problem that is very difficult to solve. If these frequencies were known a priori or can be estimated precisely, one can easily find the optimum values of the other parameters accordingly.

3.2. Estimating the number of harmonics

If the fundamental frequencies $\hat{\omega}_k$ are assumed to be known, the total number of harmonics in each signal can be estimated simply as

$$\hat{L}_k = \left\lfloor \frac{\pi}{\hat{\omega}_k} \right\rfloor. \tag{15}$$

Practically, $\hat{L}_k$ is chosen much smaller than the value calculated by (15) since most of the energy of voiced speech is concentrated below 2 kHz. Using this assumption can reduce dramatically the computational complexity of the system.

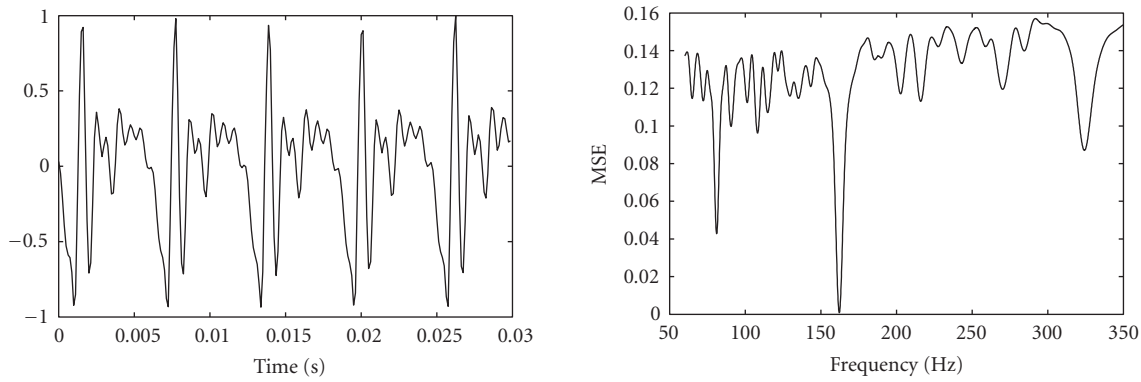

Figure 1: An example of MSE surface for a single-talker: (a) 30-millisecond single voiced speech segment in the time domain, (b) MSE performance versus fundamental frequency based on (18).

3.3. Estimating the amplitude parameters

The optimum values of the quadrature parameters $\{a_\ell^{(k)}\}_{\ell=1}^{L_k}$ and $\{b_\ell^{(k)}\}_{\ell=1}^{L_k}$ can be estimated directly (assuming the availability of the fundamental frequencies) by finding the standard linear LS solution to (14) as follows [5]:

$$\mathbf{h}_{\text{opt}} = \left(\mathbf{Q}^T\mathbf{Q}\right)^{-1}\mathbf{Q}^T\mathbf{x} = \mathbf{R}^{-1}\mathbf{P}, \tag{16}$$

where

$$\mathbf{R} = \mathbf{Q}^T\mathbf{Q}, \qquad \mathbf{P} = \mathbf{Q}^T\mathbf{x}. \tag{17}$$

The minimum MSE corresponding to $\mathbf{h}_{\text{opt}}$ is given by substituting (16) into (14) to give

$$E_{\min} = \mathbf{x}^T\mathbf{x} - \mathbf{P}^T\mathbf{R}^{-1}\mathbf{P}. \tag{18}$$

3.4. Estimating the fundamental frequencies

Since in practical applications, the fundamental frequencies of the individual speech waveforms are not known a priori, they must be estimated from the mixed data. A direct approach to solve this problem is to search the K-dimensional MSE surface for its minimum with respect to the fundamental frequencies. The initial estimate can be determined either from the previous frame or by applying a simple rough multipitch estimation method such as the one proposed in [6], which is a time-domain method that depends on the average magnitude difference function. After finding an initial guess for the $\widehat{\omega}_k$, the optimum fundamental frequencies can be estimated by searching the MSE surface of (18) by the method of steepest descent [7]. Using a weight vector $\mathbf{w} = [\widehat{\omega}_1, \widehat{\omega}_2]^T$, we describe the steepest descent algorithm by

$$\mathbf{w}(i+1) = \mathbf{w}(i) - \frac{1}{2}\mu\nabla\mathbf{E}(i), \tag{19}$$

where

$$-\nabla\mathbf{E}(i) = \begin{bmatrix} -\nabla\mathbf{E}^{(1)}(i) \\ -\nabla\mathbf{E}^{(2)}(i) \end{bmatrix}, \tag{20}$$

and μ is a positive scalar that controls both the stability and the speed of convergence. The gradient of the MSE is calculated by differentiating (18) with respect to each fundamental frequency as follows:

$$\begin{aligned} -\nabla\mathbf{E}^{(k)} &= -\frac{\partial E_{\min}}{\partial\widehat{\omega}_k} \\ &= \mathbf{x}^T\dot{\mathbf{Q}}\mathbf{h}_{\text{opt}} - \mathbf{h}_{\text{opt}}^T\dot{\mathbf{R}}\mathbf{h}_{\text{opt}} + \mathbf{h}_{\text{opt}}^T\dot{\mathbf{Q}}^T\mathbf{x}, \end{aligned} \tag{21}$$

where

$$\dot{\mathbf{Q}} = \frac{\partial\mathbf{Q}}{\partial\widehat{\omega}_k}, \qquad \dot{\mathbf{R}} = \frac{\partial\mathbf{R}}{\partial\widehat{\omega}_k}. \tag{22}$$

Differentiating (10) and substituting into (21) give

$$-\nabla\mathbf{E}^{(k)} = 2\left(\mathbf{x} - \mathbf{Q}\mathbf{h}_{\text{opt}}\right)^T\dot{\mathbf{Q}}^{(k)}\mathbf{h}_{\text{opt}}^{(k)}. \tag{23}$$

The fundamental frequencies are updated iteratively using (19). After each iteration, the optimum amplitude parameters corresponding to the estimated frequencies are calculated using (16). Note that even by using (19), final estimates of fundamental frequencies may still have small inaccuracies because frequencies may vary slightly within the speech frame. The use of exact gradient to update the fundamental frequencies in (19) gives an advantage compared to [1], where an approximation of the gradient is used. Gradient calculation is an integrated process in the time-domain method since the components on the right-hand side of (23) are already part of the previous steps in the algorithm.

An example of the MSE surface obtained for the single talker ($K = 1$) case is shown in Figure 1. Figure 1(a) shows a 30-millisecond speech frame for a single talker, while Figure 1(b) shows the corresponding MSE surface using (18) as the cost function. From Figure 1(b), the optimal fundamental frequency is approximately 165 Hz. For the two-talker case ($K = 2$), the MSE surface would instead be two-dimensional.

3.5. The ill-conditioned estimation problem

In some instances, the harmonics of the two speakers can be very close to each other. When the harmonics overlap, the

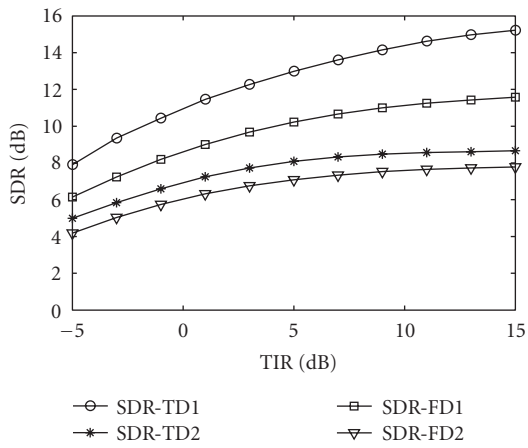

Figure 2: SDR results; SDR-TD1 and SDR-TD2 for the proposed time-domain method, and SDR-FD1 and SDR-FD2 for the frequency-domain method [1], with precise and initial frequency estimates of $\{\omega_k\}_{k=1,2}$, respectively.

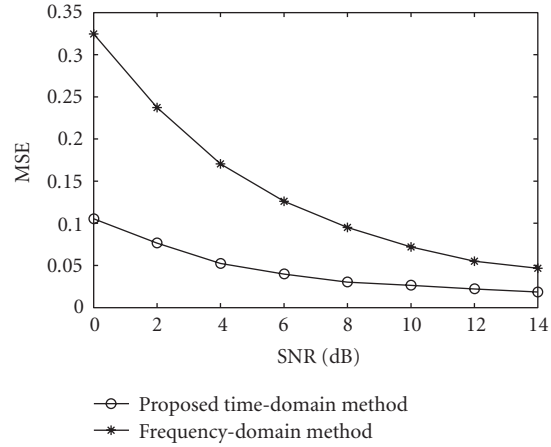

Figure 3: MSE results for AWGN for both the proposed time-domain technique compared with the standard frequency-domain method [1].

matrix $\mathbf{R}$ in (17) will be singular, and the parameter estimation process in (16) becomes ill-conditioned. To handle this problem, the spacing between adjacent harmonics is continuously calculated. If two adjacent harmonics are found to be closely spaced, that is, less than 25 Hz apart, only one sinusoid is used to represent these two harmonics. The amplitude parameters of this single component are then estimated and shared equally between the two speakers [1].

4. SIMULATION RESULTS

The performance of the proposed method is evaluated using a speech database consisting of 200 frames of mixed speech. All-voiced speech segments of length 30 milliseconds were randomly chosen from the TIMIT dataset [8] for male and female speakers and mixed at different TIRs. The speech data were sampled at a rate of 16 kHz.

Two sets of simulations were conducted to compare the performance of the proposed method with the *frequency sampling* approach presented in [1]. As suggested by the authors, a Hann window and a high-resolution STFT of length $M = 4096$ were used in the frequency-domain technique. To avoid errors due to the multipitch detection algorithm, the initial guess of the fundamental frequency of each talker was calculated directly from the original speech frames before mixing, using a simple autocorrelation method.

In the first set of simulations, the comparison was carried out in terms of the signal-to-distortion ratio (SDR) versus TIR as shown in Figure 2 for TIRs ranging from −5 to +15 dB. The SDR measure is defined as [9]

$$\mathrm{SDR}_{\mathrm{dB}} = 10\log_{10}\frac{\sum_n s(n)^2}{\sum_n \left[s(n)-\hat{s}(n)\right]^2}, \tag{24}$$

where $s(n)$ is the original target signal before mixing, and $\hat{s}(n)$ is the reconstructed signal after separation from the mixture $x(n)$. Each point in the plot of Figure 2 presents the ensemble average of the SDRs over all 200 test frames. Two cases are considered for each algorithm. In case 1, precise estimation of the fundamental frequencies is done using (19), and in case 2 only the initial guess of the fundamental frequencies is used. Plots SDR-TD1 and SDR-TD2 are the results for the proposed time-domain algorithm in case 1 and case 2, respectively, while the plots SDR-FD1 and SDR-FD2 depict the results for the frequency-domain method. As can be seen from Figure 2, the SDR increases monotonically for both algorithms with an increase of the TIR in all cases.

More importantly, we see from Figure 2 that the proposed technique outperforms the frequency-domain techniques in both case 1 and case 2. At TIR = −5 dB, SDR-TD1 and SDR-TD2 are greater than SDR-FD1 and SDR-FD2 by about 2 and 1 dB, respectively. This difference is greater for larger TIR. As suggested in Section 1, analysis of the resulting estimates using voiced speech segments has revealed that the discrepancies are due to the limited frequency resolution of the STFT (even with $M = 4096$) and due to the choice of window function and resulting edge effects. Other window functions such as rectangular and Hamming windows had similar discrepancies when tested.

The robustness against background noise was examined in a second set of simulations using MSE versus signal-to-noise ratio (SNR). Speech segments were corrupted by additive white Gaussian noise (AWGN) with SNR varied from 0 to 15 dB. The results are presented in Figure 3. As shown in the figure, the proposed algorithm has a superior performance in low SNR compared to the frequency-domain technique. The AWGN causes additional frequency resolution problems after even a high-resolution STFT. If the proposed time-domain estimation approach is used instead, then the effect of the AWGN is not as severe.

5. CONCLUSIONS

A time-domain method to precisely estimate the sinusoidal model parameters of co-channel speech is presented. The method does not require calculation of the STFT nor

multiplication by a window for the primary model parameters. The proposed method incorporates a least-squares estimator and an adaptive technique to model and separate the co-channel speech into its individual speakers, all in the time domain.

The application of this time-domain method on real data demonstrates the effectiveness of this method in separating co-channel speech signals at different TIRs. Overall, an improvement of 1–3 dB in SDR is obtained over the frequency-domain method, dependent on the accuracy of the fundamental frequency estimates of the talkers in the tested two-talker scenario. Note that these time-domain results are compared with the frequency-domain approach with an $M = 4096$-point STFT. Changes in M would affect the precision in the estimates of the frequency-domain technique.

We also note that the time-domain method is not as sensitive to additive white Gaussian noise as is the frequency-domain method for sinusoidal modeling. This result is particularly true for lower-SNR situations.

ACKNOWLEDGMENT

The authors wish to thank the National Capital Institute of Telecommunications (NCIT) for partially funding this research.

REFERENCES

[1] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 38, no. 1, pp. 56–69, 1990.

[2] F. M. Silva and L. B. Almeida, "Speech separation by means of stationary least-squares harmonic estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, vol. 2, pp. 809–812, Albuquerque, NM, USA, April 1990.

[3] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Co-Channel speaker separation by harmonic enhancement and suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 407–424, 1997.

[4] A. Bánhalmi, K. Kovács, A. Kocsor, and L. Tóth, "Fundamental frequency estimation by least-squares harmonic model fitting," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech'05)*, pp. 305–308, Lisbon, Portugal, September 2005.

[5] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results and an application," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 338–352, 2000.

[6] A. de Cheveigné, "A mixed speech F_0 estimation algorithm," in *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech '91)*, pp. 445–448, Genova, Italy, September 1991.

[7] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[8] J. Garofolo, L. Lamel, W. Fisher, et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.

[9] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.