University of Wollongong
Research Online

Faculty of Informatics - Papers (Archive)

Faculty of Engineering and Information Sciences

1-11-2005

Blind speech separation using a joint model of speech production

Daniel Smith
University of Wollongong

Jason Lukasiak
University of Wollongong, jl01@ouw.edu.au

Ian Burnett
University of Wollongong, ianb@uow.edu.au

Follow this and additional works at: https://ro.uow.edu.au/infopapers

 Part of the Physical Sciences and Mathematics Commons

Recommended Citation

Smith, Daniel; Lukasiak, Jason; and Burnett, Ian: Blind speech separation using a joint model of speech production 2005.
https://ro.uow.edu.au/infopapers/14

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Blind speech separation using a joint model of speech production

Abstract

We propose a new blind signal separation (BSS)technique, developed specifically for speech, that exploits a priori knowledge of speech production mechanisms. In our approach, the autoregressive (AR) structure and fundamental frequency (0) production mechanisms of speech are jointly modeled. We compare the separation performance of our joint AR-F0 algorithm to existing BSS algorithms that model either speech's AR structure [1] or 0 [2] individually. Experimental results indicate that the joint algorithm demonstrates superior separation performance to both the individual AR algorithm (up to 77% improvement) and F0 (up to 50% improvement) algorithms. This suggests that speech separation performance is improved by employing a BSS model with a more realistic description of the speech production process.

Keywords

autoregressive (AR) process and fundamental frequency, blind signal separation (BSS), speech, temporal modeling

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was originally published as: Smith, D, Lukasiak, J & Burnett, I, Blind speech separation using a joint model of speech production, IEEE Signal Processing Letters, November 2005, 12(11), 784-787.

Copyright IEEE 2005.

This journal article is available at Research Online: https://ro.uow.edu.au/infopapers/14

Blind Speech Separation Using a Joint Model of Speech Production

Daniel Smith, Jason Lukasiak, and Ian Burnett

***Abstract*—We propose a new blind signal separation (BSS) technique, developed specifically for speech, that exploits *a priori* knowledge of speech production mechanisms. In our approach, the autoregressive (AR) structure and fundamental frequency $(F0)$ production mechanisms of speech are jointly modeled. We compare the separation performance of our joint AR-F0 algorithm to existing BSS algorithms that model either speech's AR structure [1] or $F0$ [2] individually. Experimental results indicate that the joint algorithm demonstrates superior separation performance to both the individual AR algorithm (up to 77% improvement) and F0 (up to 50% improvement) algorithms. This suggests that speech separation performance is improved by employing a BSS model with a more realistic description of the speech production process.**

***Index Terms*—autoregressive (AR) process and fundamental frequency $(F0)$, blind signal separation (BSS), speech, temporal modeling.**

I. Introduction

BLIND signal separation (BSS) has been a major area of interest in audio research, with the application of BSS to speech signals being of particular importance. The interest in BSS for audio is motivated by its use in developing adaptive, intelligent solutions to the "cocktail party problem," a problem in which any speaker in an acoustic environment can be independently retrieved (or made the focus of listening attention) amidst other concurrent speakers and noise [3].

Conventional BSS techniques attempt to solve the "cocktail party problem" using independent component analysis (ICA); this operates without any prior knowledge of the signals (or mixing process) other than the assumption that the signals are non-Gaussian and statistically independent [3]. Although BSS algorithms that use ICA have broad application, when employed specifically for speech separation, their performance may be limited by failure to utilize contextual or *a priori* information about the speech signal. Although there have been a number of BSS approaches that exploit the temporal structure of signals [1]–[3], [4]–[6], these are only capable of modeling the autoregressive (AR) structure [1], [3], [4], [5][1] or fundamental frequency $(F0)$ [2], [6] of speech individually. None of these approaches employs a model that describes both the short-term and long-term speech production process.

Consequently, the objective of this letter is to develop a BSS algorithm that describes speech with a more complete production model. This is achieved by employing a joint model that exploits both AR structure (short-term temporal correlation) and $F0$ delay (long-term temporal correlation). The joint model is combined with gradient descent adaptation, or gradient descent merged with optimal solutions, to enable speech signals to be blindly separated. We compare the performance of this joint model approach to two BSS algorithms that exploit either the AR structure [1] or long-term correlations [2] exclusively.

Manuscript received February 7, 2005; revised April 25, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yue (Joseph) Wang.

The authors are with the Whisper Laboratories, School of Electrical, Computer and Telecommunication Engineering, University of Wollongong, New South Wales 2522, Australia (e-mail: dsmith@titr.uow.edu.au).

Digital Object Identifier 10.1109/LSP.2005.856869

[1]Although [1] used a relatively long AR filter of 50 taps to model the temporal structure of speech, it will only guarantee that the short-term correlation is captured. An AR filter that is 150 taps long is required to ensure that the long-term correlation (a period) of voiced speech (sampled at 8 kHz) is captured [7].

II. Problem Formulation

The BSS problem can be formulated as follows: The vector of sensor signals $(X(t))$ contains observations of the vector of signals $(S(t))$ linearly mixed according to the system A

$$X(t) = A \cdot S(t) \tag{1}$$

where $X(t) = [X_1, , \ldots, , X_M]^T$ is a $M \times 1$ vector of mixed observations, $S(t) = [S_1, \ldots, S_N]^T$ is an unknown $N \times 1$ vector of signals, and A is an unknown $M \times N$ nonsingular matrix. In this approach, it is assumed that A contains scalar elements (instantaneous mixing) and the system is square, i.e., the number of signals is equal to the number of sensors.

Given only mixed observations $X(t)$, an $N \times M$ separation matrix W(estimating A^{-1}) must be computed and then multiplied by $X(t)$ in order to obtain a scaled permutation of the original signals $c \cdot S(t)$. In contrast to simultaneous estimation of the entire separation matrix, the method presented in this letter is a sequential approach in which each column of the separation matrix (W_j) and the separated signal $(S_{je}(t) = W_j^T \cdot X(t))$ is estimated individually.

III. Separation of Speech Signals

The BSS approaches of [1] and [3] have demonstrated that speech signals can be extracted from a mixture by exploiting the following assumption.

a) *A single speaker has more temporal correlation than any linear combination of mixed speakers.*

It is the temporal correlation generated by the production mechanisms of speech that make assumption a) hold true [7]. The BSS approach developed in this letter utilizes assumption a) by modeling these production mechanisms. First, the short-term temporal correlation (i.e., correlation between adjacent samples) of speech is modeled by an AR process [shown in (2)],

1070-9908/$20.00 © 2005 IEEE

such that speech is predicted as a linear combination of its previous P samples

$$S_j(t) = \sum_{i=1}^{P} b_{ji} \cdot S_j(t-i) \qquad j = 1, \ldots, N \tag{2}$$

where $b_j = [b_{j1} \cdots b_{jP}]$ is a $1 \times P$ vector of short-term prediction coefficients. In addition, the long-term temporal correlation of voiced speech, generated by a quasi-periodic excitation source [7], is represented by the $F0$ delay $(1/F0)$. A normalized auto-correlation method [7] is used to estimate $F0$.

In the proposed model, the AR structure of (2) and periodicity $(1/F0)$ are jointly represented in the cost function $C(W_j, b_j, B_j)$ as

$$\begin{aligned} C(W_j, b_j, B_j) &= 1/2 * E[\xi(t)^2] \\ \xi(t) &= W_j^T \cdot X_a(t) - B_j \cdot W_j^T \cdot X_l(t) \end{aligned} \tag{3}$$

where $\hat{X}(t) = [X(t-1), \ldots, X(t-P)]$ is an $M \times P$ matrix, $X_a(t) = X(t) - \hat{X}(t) \cdot b_j^T$ is the short-term temporal prediction error of the mixtures, and $X_l(t) = X(t-(1/F0)) - \hat{X}(t-(1/F0)) \cdot b_j^T$ is the short-term period-delayed prediction error of the mixtures. $E[\cdot]$ is the expected value of the function. $\xi(t)$ is the error function jointly describing the short-term and long-term temporal prediction error of the estimated speech. The first term in $\xi(t)$ [containing $X_a(t)$] represents the short-term prediction model, and the second term [containing $X_l(t)$] represents the long-term prediction model.

A. Derivation of the Learning Algorithm

As the sole objective of a separation approach is to learn W_j, we present two different approaches to adapt W_j to the minima of the cost function of (3). The first approach (GradDes) uses a stochastic gradient descent to derive adaptation rules for the parameter set W_j, b_j and B_j. The second approach (ComGradOpt) employs the stochastic gradient descent to develop the adaptation rule for W_j and an optimum solution to derive the rules of the other parameters b_j and B_j.

In order to minimize the cost function in (3), the initial step in deriving the adaptation rules for GradDes and ComGradOpt involves computing the partial derivatives of $C(W_j, b_j, B_j)$ with respect to each of the parameters W_j, b_j and B_j. The partial derivatives are calculated as

$$\begin{aligned} \frac{\delta C(W_j, b_j, B_j)}{\delta W_j} &= E[\xi(t) \cdot (X_a(t) - B_j \cdot X_l(t))] \\ \frac{\delta C(W_j, b_j, B_j)}{\delta b_j} &= -E\left[\xi(t) \cdot \left(W_j^T \cdot \hat{X}(t) - B_j \cdot W_j^T \cdot \hat{X}\left(t - \frac{1}{F0}\right)\right)\right] \\ \frac{\delta C(W_j, b_j, B_j)}{\delta B_j} &= -E\left[\xi(t) \cdot W_j^T \cdot X_l(t)\right]. \end{aligned} \tag{4}$$

The learning rules of GradDes, shown in (5), are then derived by substituting the derivatives from (4) into the stochastic gradient descent approach

$$\begin{aligned} W_{j+1} &= W_j - \Delta W \cdot E[\xi(t) \cdot (X_a(t) - B_j \cdot X_l(t))] \\ b_{j+1} &= b_j + \Delta b \cdot E\left[\xi(t) \cdot \left(W_j^T \cdot \hat{X}(t) - B_j \cdot W_j^T \cdot \hat{X}\left(t - \frac{1}{F0}\right)\right)\right] \\ B_{j+1} &= B_j + \Delta B \cdot E\left[\xi(t) \cdot W_j^T \cdot X_l(t)\right] \end{aligned} \tag{5}$$

where $\Delta W, \Delta b$, and ΔB are the step sizes, and W_{j+1}, b_{j+1}, and B_{j+1} are the parameters for the next iteration of the gradient descent.

In ComGradOpt, we utilize the learning rule for W derived in (5), while B_j and b_j are updated as the optimal solutions of (3), by solving the expressions $(\delta C(W_j, b_j, B_j))/(\delta b_j) = 0$ and $(\delta C(W_j, b_j, B_j))/(\delta B_j) = 0$ in terms of b_j and B_j, respectively

$$\begin{aligned} B_j &= \left(W_j^T \cdot R_{X_l X_l} \cdot W_j\right)^{-1} \cdot \left(W_j^T \cdot R_{X_l X_a} \cdot W_j\right) \\ b_j &= \left(W_j^T \cdot R_{\hat{x}\hat{x}} \cdot W_j\right)^{-1} \cdot \left(W_j^T \cdot R_{\hat{x}x} \cdot W_j\right) \end{aligned} \tag{6}$$

where $R_{X_l X_l}, R_{X_l X_a}, R_{\hat{x}\hat{x}}$, and $R_{\hat{x}x}$ are correlation matrix estimates $x = X(t) - B_j \cdot X(t-(1/F0))$ and $\hat{x} = \hat{X}(t) - B_j \cdot \hat{X}(t-(1/F0))$.

B. Outline of the AR-F0 Algorithm

The proposed AR-F0 algorithm involves the following steps.

- Step 1) The mixed observations $X(t)$ are broken into frames, with each frame being applied to steps 2)–6) sequentially. For the first frame, W_j is randomly initialized. For all preceding frames, W_j is set to the separation column from the previous frame.
- Step 2) The analysis frame is whitened, so that the separation matrix is constrained to the space of orthonormal matrices. This is particularly beneficial in ill-conditioned problems [3]. Steps 3)–5) are then repeated until the minima of the cost function $C(W_j, b_j, B_j)_{\min}$ is reached.
- Step 3) The $F0$ of the current clean speech estimate $S_{je}(t) = W_j^T \cdot X(t)$ is obtained using the normalized autocorrelation pitch detection method [7]. $F0$ is calculated during every iteration of the gradient descent to ensure that the algorithm is relatively insensitive to $F0$ estimation errors. As the gradient descent steps toward a clean speech solution, $F0$ errors that may occur during the initial iterations of the gradient descent are replaced by $F0$ estimates of greater accuracy.
- Step 4) The parameters W_j, b_j and B_j are updated with the gradient descent of (5), or alternatively, W_j is updated with the gradient descent, and b_j and B_j are updated with the optimal solutions of (6).
- Step 5) W_j is then normalized, i.e., $(W_j / \| W_j \|^2)$, such that the estimated signal is constrained to $E[S_{je}^2] = 1$. This ensures that the trivial solution $S_{je} = 0$ is avoided when finding $C(W_j, b_j, B_j)_{\min}$.
- Step 6) The separated speech signal is estimated by W_j at the point at which the cost function converges to $(C(W_j, b_j, B_j)_{\min}$. Under the assumption (a), $C(W_j, b_j, B_j)_{\min}$ will estimate a scaled version of one of the original signals $S_{je} = c \cdot S_j$.

IV. Results

We compared the performance of our joint AR-F0 algorithms to two other algorithms. The first was a short-term correlation approach (AR algorithm) given in [1], which applies a gradient descent optimization to the cost function $W_j \cdot X_a$ [the first term of $\xi(t)$ in (3)]. The second approach ($F0$ algorithm) was similar to that reported in [2], exploiting the long-term correlation

between $S_{je}(t)$ and $B \cdot S_{je}(t-(1/F0)))$. The algorithm in [2], however, exploits the long-term correlation of signals using an optimal solution. In our analysis, using a gradient descent approach in [2] provided a better comparison to the other models, as the ComGradOpt, GradDes, and AR algorithms all employed gradient descent adaptation of W_j. Therefore, in this experiment, gradient descent adaptation of the cost function from [2] was used, replacing the optimal solution.

We applied all four algorithms to a data set consisting of eight different pairs of sustained vowels (pure voiced speech) 1.5 s in duration and ten different pairs of natural speech segments 2.5 s in length. All vowels and speech signals were sampled at 8000 Hz. The simulation was conducted over a range of frame sizes extending from 10 to 200 ms. Furthermore, the simulation was repeated three times, with a different stationary mixing system A being applied to the data set on each occasion. An AR filter b_j of order 10 was used in both the AR-F0 and AR algorithms, and step sizes $\Delta W = \Delta b = \Delta B = 0.05$ were employed in all algorithms. In this analysis, only a single speaker was extracted from the mixture. Although a deflationary technique as in [3] can be used to enable the removal of additional speakers from the mixture, in the context of this analysis, it was unnecessary, as it provided no further information regarding the model's separation performance.

The separation performance measure used in this analysis was an interference measure (IM), which is defined as $\text{IM} = \frac{(p \cdot p^T - \max(p)^2)^{\frac{1}{2}}}{\max(p)}$, where $p = W_j^T \cdot A$. IM is the inverse of the measure used in [8]. An IM $= 0$ corresponded to ideal signal separation, that is, without any interference from other signals in the mixture. Informal listening tests, however, indicated that for the speech mixtures in this experiment, an IM < 0.03 related to a level of separation where interference was inaudible. In addition, the minimum mean-squared error (MMSE) corresponds to $C(W_j, b_j, B_j)_{\min}$, the criteria used to model (3). It is presented in the results to demonstrate the estimated signal's adherence to the joint model of (3).

Fig. 1 compares the MMSE and separation performance of the joint AR-F0 models (GradDes, ComGradOpt), AR algorithm and F0 algorithm, averaged over eight pairs of voiced speech and three different mixing systems A. As voiced speech can be modeled by an AR process and periodic excitation simultaneously, it is the mode of speech that should be best modeled by our joint AR-F0 algorithms. The results in Fig. 1(a) support this statement, as both the joint AR-F0 algorithms (solid line, solid line with circles) have a lower average MMSE than both the AR (dashed line) and FO (dotted line) algorithms across all frame sizes. The MMSE of the joint model is 48%–65% less than the AR algorithm and 88%–92% less than the FO algorithm.

The MMSE advantage of the joint AR-F0 algorithms correlates with their significant separation performance advantage over the AR algorithm, as displayed in Fig. 1(b). This shows that the average IM of the joint AR-F0 algorithms is 55%–77% less than the average IM of the AR algorithm across all frame sizes. We can hypothesize that it is the inclusion of long-term correlation (pitch period) into the joint model that provides this separation improvement, as when the IM of the F0 algorithm saturates at a frame size of around 60 ms, the IM of the AR-F0 joint algorithms monotonically increase at a similar rate to the AR algorithm.

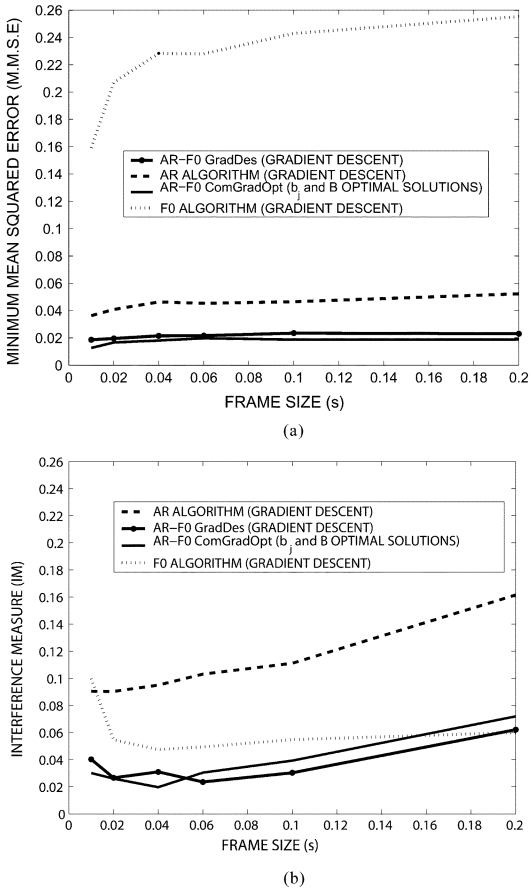

Fig. 1. MMSE and separation performance IM [(a) and (b), respectively] of the joint AR-F0, AR, and F0 models, averaged over eight pairs of sustained vowels and three mixing simulations. In each simulation, the sustained vowels were mixed by a different mixing system A. (a) Average MMSE. (b) Average IM.

The joint AR-F0 algorithm's IM advantage over the F0 algorithm is present for frame sizes less than 0.15 s; however, this advantage declines with an increase in frame size. The IM of the F0 algorithm is reasonably constant for longer frames of sustained vowels, as they possess a relatively stable pitch. This ensures that F0 can be estimated with a consistent level of accuracy across the longer frames. The monotonically decreasing separation performance of the AR-F0 joint models for frame sizes greater than 60 ms can be attributed to the underlying sustained vowels becoming less stationary [7] as the frame size increases. This characteristic results in a weakening of the underlying vowel's conformance to the imposed AR structure, and hence, assumption a) becomes increasingly invalid. The same decrease in performance, however, is not evident in the AR-F0 joint model's MMSE for frame sizes greater than 60 ms. This is a consequence of the MMSE criteria employed in the AR modeling [7]. Under the constraints of this criterion, the AR model parameters will be selected to minimize the overall MMSE, whether or not the formants modeled by these parameters conform to a single speech signal. Thus, as the speech signals become less stationary, the AR model may simply combine formants from each of the underlying signals into the error minimization process.

Fig. 2 compares the MMSE and separation performance of the algorithms averaged over ten pairs of natural speakers and three different mixing systems A. Natural speech is less

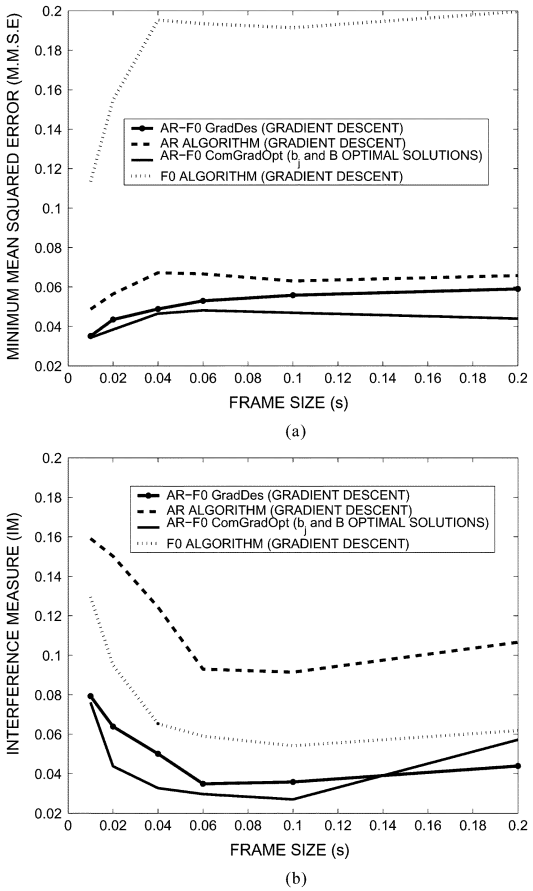

Fig. 2. MMSE and separation performance IM [(a) and (b), respectively] of the joint AR-F0, AR, and F0 models, averaged over ten pairs of speech and three mixing simulations. In each simulation, the speech was mixed by a different mixing system A. (a) Average MMSE. (b) Average IM.

stationary than sustained vowels, consisting of some nonperiodic portions (unvoiced and transient) that are inapplicable to the long-term component (F0) of the joint model. In an average sense, however, the joint AR-F0 models still provide a significantly better representation of speech than the AR and F0 models. Fig. 2(a) shows that the joint AR-F0 algorithms offer between 10%–33% MMSE improvement upon the AR algorithm and a 70%–77% MMSE improvement over the F0 algorithm across all frame sizes.

Fig. 2(b) indicates that the average separation performance (IM) of the joint AR-F0 model is superior to both the AR and F0 separation models for natural speech. The average IM of the AR-F0 algorithm is 50%–70% less than the AR algorithm and 7%–50% less than the F0 algorithm across all frame sizes. Fig. 2(b) also shows that ComGradOpt exhibits an IM advantage (of up to 33%) over GradDes for frame sizes less than 0.14 s. ComGradOpt's separation performance increasingly degrades for frame sizes longer than 0.14 s, such that GradDes approach outperforms ComGradOpt by 23% at a frame size of 0.2 s. We conclude from these results that ComGradOpt has a performance advantage over GradDes approach when speech is reasonably stationary. This is because stationary speech conforms to assumption a), and an optimum approach models the AR structure of the underlying speech signal better than a gradient technique. For longer, less stationary frames of speech (>0.15 s), however, the separation performance of GradDes is superior to ComGradOpt. This is because the nonstationary speech frames do not conform to assumption a), and an optimal solution is more likely to incorrectly model the underlying AR structure of a speech signal than a gradient descent approach. When assumption a) is not completely valid, the gradient descent approach of stepping toward the MMSE after each iteration provides it with a greater ability to track the underlying AR structure of a speech signal.

V. Conclusion

In this letter, we have developed a BSS approach that jointly models the AR and periodic $(1/F0)$ production mechanisms of speech. Experimental results with both voiced and natural speech verified that the joint algorithm achieves significant separation improvement over algorithms that model either the AR structure (up to 77% improvement) or $F0$ (up to 50% improvement) individually. The superior separation performance of the joint approach suggests that a more inclusive model of *a priori* knowledge of speech, in the form of its production mechanisms, is beneficial in BSS.

In addition, two different optimization approaches to the joint algorithm were compared: GradDes and ComGradOpt. Results showed that ComGradOpt provided better separation performance when the assumptions of our model were closely met; otherwise, GradDes outperformed ComGradOpt, as ComGradOpt was more susceptible to introducing errors into the modeling of the AR structure of a speech signal.

References

[1] R. Thawonmas and A. Cichoki, "Blind signal extraction of arbitrary distributed but temporally correlated signals-Neural network approach," *IEICE Trans. Fundam.*, vol. E82-A, no. 9, pp. 1834–1844, Sep. 1999.

[2] A. Barros and A. Cichoki, "Extraction of specific signals with temporal structure," *Neural Comput.*, vol. 13, no. 9, pp. 1995–2003, Sep. 2001.

[3] A. Cichoki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New York: Wiley, 2002.

[4] B. Pearlmutter and L. Parra, "Maximum likelihood blind source separation: A context-sensitive generalization of ICA," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, Dec. 1996, vol. 9, pp. 613–619.

[5] A. Acero, S. Altschuler, and L. Wu, "Speech/noise separation using two microphones and a VQ model of speech signals," in *Proc. Int Conf. Spoken Language Processing*, vol. 4, Beijing, China, Oct. 2000, pp. 532–535.

[6] F. Tordini and F. Piazza, "A semi-blind approach to the separation of real world speech mixtures," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Honolulu, HI, May 2002, pp. 1293–1298.

[7] A. Kondoz, *Digital Speech Coding for Low Bit Rate Communications Systems*. New York: Wiley, 1994.

[8] K. Hild, D. Erdogmus, and J. Principe, "Blind source separation using Renyi's mutual information," *IEEE Signal Process. Lett.*, vol. 8, no. 6, pp. 174–176, Jun. 2001.