An Extended Model for Speech Segregation

Guoning Hu
The Ohio State University
Biophysics Program
2015 Neil Ave.
Columbus, Ohio, 43210
hu.117@osu.edu

DeLiang Wang
The Ohio State University
Department of Computer & Information Science
& Center of Cognitive Science
2015 Neil Ave.
Columbus, Ohio, 43210
dwang@cis.ohio-state.edu

Abstract

Speech segregation is an important task of auditory scene analysis (ASA), in which the speech of a certain speaker is separated from other interfering signals. Wang and Brown proposed a multistage neural model for speech segregation, the core of which is a two-layer oscillator network. In this paper, we extend their model by adding further processes based on psychoacoustic evidence to improve the performance. These processes include estimation of the pitch of target speech and refined generation of a target speech stream with the estimated pitch. Our model is systematically evaluated and compared with the Wang-Brown model, and it yields significantly better performance.

1 Introduction

In an environment with all kinds of audible signals, the auditory system is able to segregate signals from different sources or events and represent them separately. This auditory process is described as *Auditory Scene Analysis* (ASA) [1]. In practice, it is a challenging task to develop a computational system to separate signals from acoustic mixtures. Blind sources separation [2, 7] provides a general way for signal separation, which assumes that signals from different sources are statistically independent. However, this method only works when information of more than one acoustic mixture is available [9]. Another approach is to develop a system utilizing psychoacoustic cues to mimic ASA [3-5, 8], which generally works in the monaural condition.

An important task of ASA is to segregate speech from other interfering signals. An efficient speech segregation process is required for robust *automatic speech recognition* (ASR) in a noisy environment. Wang and Brown proposed a multistage neural network model for speech segregation [8]. The schematic diagram of their model is shown in Fig. 1. The core of their model is a two-layer neural oscillator network that performs speech segregation in two stages: segmentation and grouping. In the segmentation stage, the acoustic mixture is decomposed into segments. The corresponding signals of those oscillators in the same segment are likely to come from the same source. In the grouping stage, segments that are likely to contain signals mainly from the same source or event are grouped together.

The main psychoacoustic cues used in their model are global pitch and temporal continuity. For an acoustic mixture of target speech and intrusion, the global pitch serves as a good grouping cue only when it is close to either the pitch of target speech or the pitch of intrusion (if pitched). When target speech mixed with wideband intrusions, the global pitch is meaningless sometimes and cannot provide useful information for grouping (Fig. 2). Using the temporal continuity condition helps in generating large segments across time. However, when target speech and intrusion have a lot of components with close frequencies, i.e., target speech and intrusion overlap significantly in their spectra, some segments may contain strong signals from both sources. Due to the above reasons, the Wang-Brown model performs poorly when intrusion is wideband. For example, when the intrusion is the “cocktail party” noise (N3, see Table 1), a lot of speech signals are missed from the segregated target speech. In another example when the intrusion is a female voice (N9), a lot of intrusion signals remains in the segregated target speech.

In this paper, we extend their model by introducing two further processes to the oscillator network part. The first process is to estimate the pitch of target speech, which is a much better grouping cue than the global pitch of the mixture. The pitch of target speech is difficult to obtain from the mixture when the intrusion signal is strong compared with the target speech signal. However, it is much easier to obtain a good approximation of it when target speech is dominant. Since the target speech segregated by the Wang-Brown

0-7803-7044-9/01/$10.00 ©2001 IEEE

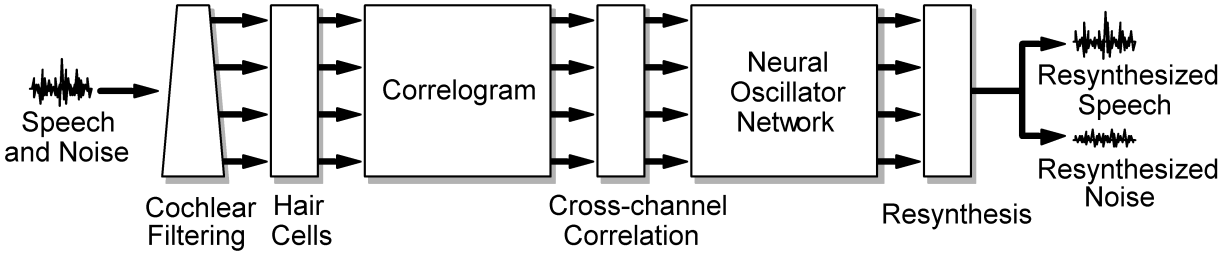

Figure 1: The schematic diagram of the Wang-Brown model. First, simulated auditory nerve activity is obtained by passing the input through a model of the auditory periphery (cochlear filtering and hair cells). Mid-level auditory representations are then formed (correlogram and cross-channel correlation map). Subsequently, a two-layer oscillator network performs grouping of acoustic components. Finally, the resynthesis path allows the separation performance to be evaluated.

model contains dominant target speech signals, the pitch of target speech can be estimated from it. The second process refines the generation of a new target speech stream. In this process, with the estimated pitch, those segments that are likely to contain strong signals from both sources are divided into smaller segments so that each segment is more likely to arise from one source. Then these smaller segments are grouped into a target speech stream.

Detailed explanations of these two processes are given in Section 2 and Section 3. In Section 4, the performance of our model is systematically evaluated and compared with that of the Wang-Brown model. Discussions are given in the last section.

2 Target Pitch Estimation

For any input signal, the Wang-Brown model is first applied to generate a target speech stream, referred to as S_{WB}, and a background stream. Let τ_j represent the estimated target pitch period at time frame j. Note that signals are divided into time frames and every time frame is 20 *ms* long with 10 *ms* overlap between consecutive time frames. The target pitch is obtained from S_{WB} as follows:

First, τ_j is obtained by searching the peaks in the pooled correlogram of S_{WB} in the range $[2\ ms, 12.5\ ms]$. The pooled correlogram of a stream is the summation of the autocorrelation functions (obtained in the correlogram part in the Wang-Brown model) of the oscillators in S_{WB}, which is similar to the local summary autocorrelation computed by Brown and Cooke [4]. These obtained pitch periods are checked with oscillators in S_{WB}. In the Wang-Brown model, signals are analyzed by an auditory filterbank. Every channel corresponds to an auditory filter with a certain passband. For an oscillator of channel *i* at time frame *j*, let $A(i, j, \tau)$ represent the corresponding autocorrelation function. The oscillator agrees with τ_j if

$$A(i, j, \tau_j) / A(i, j, \tau_m) > \theta_d \qquad (1)$$

where $\theta_d = 0.95$, τ_m is the lag where $A(i, j, \tau)$ is maximum for $\tau \in [2\ ms, 12.5\ ms]$. If more than half of the oscillators in S_{WB} at time frame *j* agree with τ_j, this pitch period is marked as reliable.

Because in most cases the pitch of speech changes smoothly, we stipulate that the difference between the pitch periods of nearby time frames is no greater than 20% of the pitch periods themselves. Using this criterion, these pitch periods are checked again. Let

$$d_{\tau_1}(j) = \begin{cases} |\tau_j - \tau_{j-1}| & if\ \tau_{j-1}\ reliable \\ 0 & otherwise \end{cases} \qquad (2)$$

$$d_{\tau_2}(j) = \begin{cases} |\tau_j - \tau_{j+1}| & if\ \tau_{j+1}\ reliable \\ 0 & otherwise \end{cases} \qquad (3)$$

If $d_{\tau_1}(j) > 0.2\tau_j$ or $d_{\tau_2}(j) > 0.2\tau_{j,}$ τ_j will be treated as unreliable.

Among all the intervals where all τ_j's are reliable, the longest streak is selected. That is, let j_s - j_e represent a streak: for all $j_s \le j \le j_e$, τ_j is reliable and both τ_{j_s-1} and τ_{j_e+1} are unreliable. Among all these streaks, let j_{ms} - j_{me} be the selected streak. For $j < j_{ms}$, each τ_j is checked and determined one by one. Starting with $j = j_{ms} - 1$, let

$$d_{\tau_3}(j) = \begin{cases} |\tau_j - \tau_{j+1}| & if\ \tau_{j+1}\ is\ reliable \\ \tau_{j+1} & otherwise \end{cases} \qquad (4)$$

If $d_{\tau_3}(j) > 0.2\tau_{j+1}$, τ_j will be changed as follows: let *f* be $1/\tau_{j+1}$, those oscillators in S_{WB} are selected if they correspond to channels with center frequencies close to *f* or 2*f*. If more than 3 oscillators are selected, the autocorrelation functions of these oscillators are added. The value of τ_j is determined by searching the peak in this summary autocorrelation in the range $[0.8\tau_{j+1}, 1.2\tau_{j+1}]$, and τ_j is marked as reliable. Otherwise, let $\tau_j = \tau_{j+1}$, and τ_j is treated as unreliable. Subsequently, τ_j is determined for $j = j_{ms} - 2,\ j_{ms} - 3,\ \cdots, 1$. For all $j > j_{me}$, τ_j is determined similarly. Finally, every unreliable τ_j is determined by a linear interpolation from reliable τ_j's at earlier and later time frames.

As an example, in Fig. 2(a), the global pitch periods obtained from the mixture are quite different from the pitch periods obtained from clean target speech. In Fig. 2(b), the estimated pitch periods obtained from the same acoustic mixture match that obtained from clean target speech well except at the several time frames in the beginning and end of target speech.

3 Stream Generation

Based on the estimated pitch, the target speech stream is generated as follows. First, with the estimated pitch, Eq. (1) is used to determine whether an oscillator agrees with the estimated pitch with the following modifications. Due to wide bandwidths for high-frequency channels (>1 *k*Hz), the responses of these channels usually contain several harmonic components when the input is speech. Therefore, these responses and the corresponding autocorrelation functions are likely to be amplitude modulated. As an example, Fig. 3(a) shows amplitude-modulated response of a high-frequency channel. In 3(b), the corresponding autocorrelation function is also amplitude modulated. The pitch period corresponds to a local maximum, but not the global maximum for $\tau \in [2\,ms, 12.5\,ms]$. Therefore, the range for searching maximum to determine $A(i, j, \tau_m)$ is changed accordingly. Let $A_{\min}$ represent the minimum of $A(i, j, \tau)$ for $\tau \in [0, 12.5\,ms]$. If $A_{\min}$ is greater than 50, $A(i, j, \tau_m)$ will be the maximum for $\tau \in [\tau_j / 2, 12.5\,ms]$. Otherwise, $A(i, j, \tau_m)$ is still the maximum for $\tau \in [2\,ms, 12.5\,ms]$. Furthermore, the threshold θ_d is changed to 0.85.

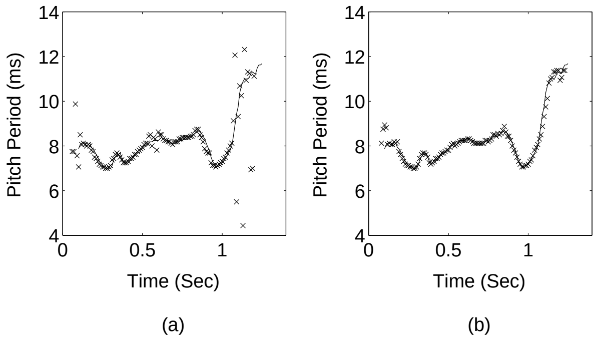

Figure 2: The global pitch and the estimated pitch of the speech obtained from the mixture of a male voice and the "cocktail party" noise. In both (a) and (b), the line contour represents the pitch obtained from clean speech. In (a), symbol 'x' represents the global pitch periods. In (b), symbol 'x' represents the estimated pitch periods.

We say that a segment agrees with the estimated pitch period at a particular time frame if more than half of the oscillators in this segment at this time frame agree with the estimated pitch period [8]. Furthermore, the segment agrees with the estimated pitch contour if it agrees with the estimated pitch periods at more than half of its total time frames. Oscillators are marked as follows:

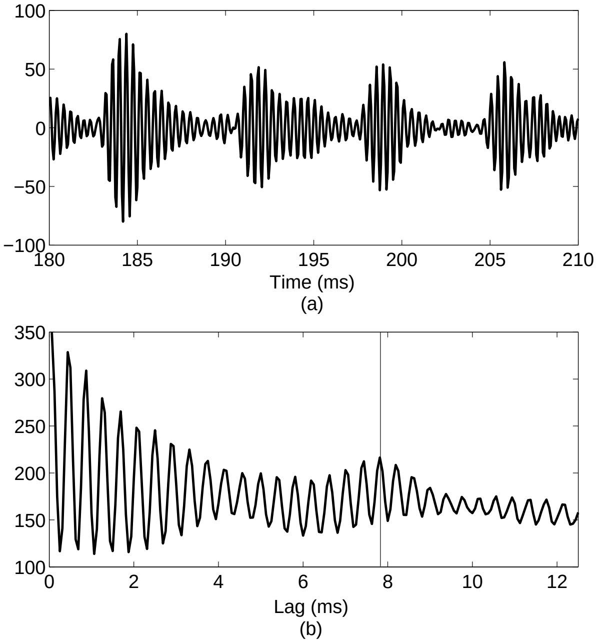

Figure 3: (a) The response from the auditory filter with center frequency 2.6 *k*Hz. The input contains only the male voice used in Fig. 2. (b) The corresponding autocorrelation function. The vertical line marks the position of the corresponding lag of the pitch period.

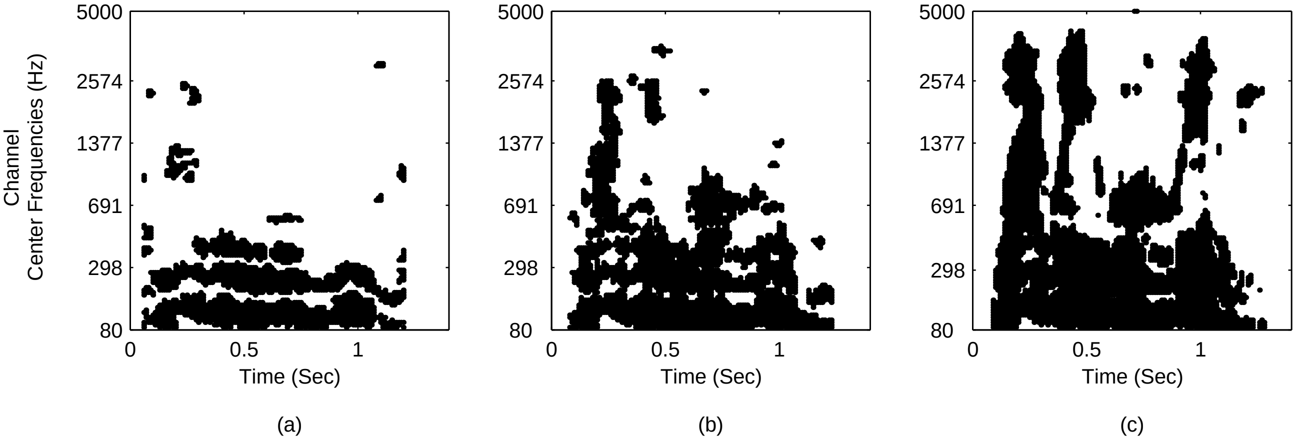

Figure 4: (a) Black part - the speech stream obtained from the Wang-Brown model. (b) The speech stream obtained from our model. (c) The stream corresponds to the ideal mask. These are generated from the same mixture as used in Figure 2.

a. All the oscillators in the segments that do not agree with the estimated pitch contour are marked -1.
b. For those in the segments that agree with the estimated pitch contour, if the oscillators themselves agree with the estimated pitch periods, they are marked 1; otherwise, they are marked 2.
c. For those that do not belong to any segment, they are marked 3 if they agree with the estimated pitch.
d. Other oscillators are marked 0.

New segments are formed by putting nearby oscillators together if they are marked the same. Based on the lengths of these segments across time, we adjust their marks as follows. If a new segment with oscillators marked 2 is longer than 50 *ms*, all the oscillators in this segment will be marked -1. For a new segment with oscillators marked 1 is shorter than 50 *ms*, all the oscillators in this segment will be marked 2. Furthermore, if a new segment with oscillators marked 3 is longer than 50 *ms*, all the oscillators in this segment will be marked 1.

All the new segments containing oscillators marked 1 are grouped together into a new target speech stream, referred as S_{NEW}. All the new segments containing oscillators marked –1 are grouped into a background stream. The background stream expands in the following way: for every oscillator in the background stream, any nearby oscillator marked 2 will be added into it. The background stream keeps on expanding until no additional oscillator can be put in. Then all the oscillators marked 2 are added into S_{NEW}. S_{NEW} expands in the following way: for all the oscillators in S_{NEW}, any nearby oscillator marked 3 will be added, and it keeps on expanding until no additional oscillators can be put in. S_{NEW} represents the target speech stream generated by our model. As an example, Fig. 4(a) shows the target speech stream generated by the Wang-Brown model. Fig. 4(b) shows the target speech stream generated by our model from the same mixture, which is much closer to the speech stream corresponding to the ideal mask, which will be explained later.

4 Results

Our model is evaluated with the same corpus of mixtures-10 voiced utterances mixed with 10 intrusions-as used to evaluate the Wang-Brown model [8]. The speech signal resynthesized [4] from the target speech stream is used for evaluation. In resynthesis, the target speech stream provides a binary mask, which guides the formation of the segregated speech.

Because target speech and intrusion are available, before mixing it in the corpus, we generate an "ideal mask" for every mixture by comparing the energies of the target speech signal and the intrusion signal corresponding to each oscillator. The ideal mask corresponds to a stream consisting of all the oscillators with stronger target speech signals. Here, we use the speech resynthesized from the ideal mask as ground truth of target speech. This evaluation methodology is supported by the following observations. First, it is well known that in a critical band, a weak signal is masked by a stronger one [6]. Second, the ideal mask is very similar to the prior mask used in a recent study that employs a missing data technique for ASR [10], and the study yields excellent recognition performance.

Let $S(t)$ represent the speech resynthesized from S_{NEW} and $I(t)$ the corresponding speech resynthesized using the ideal mask. Let

$$e(t) = I(t) - S(t) \tag{5}$$

which includes two parts. The first part consist of the signal present in $I(t)$, but not in $S(t)$. This part is the lost speech and let $e_1(t)$ represent this part. The second part consists of the signal present in $S(t)$, but not in $I(t)$. This part is the noise residue in $S(t)$, and let $e_2(t)$ represent this part.

Now we define the energy loss ratio R_{EL} and noise residue ratio R_{NR} as follows:

$$R_{EL} = \sum_t e_1^2(t) \Big/ \sum_t I^2(t) \tag{6}$$

$$R_{NR} = \sum_t e_2^2(t) \Big/ \sum_t S^2(t) \tag{7}$$

Table 1 shows the R_{EL} and R_{NR} values for the 10 kinds of noise intrusions. Each value is the average of 10 voiced utterances mixed with a certain intrusion. Table 1 also shows the R_{EL} and R_{NR} values of the resynthesized speech obtained by the Wang-Brown model.

Table 1: R_{EL} and R_{NR} of resynthesized speech from both the Wang-Brown model and the proposed model. Here, N0 = 1 *k*Hz tone, N1 = random noise, N2 = noise bursts, N3 = “cocktail party” noise, N4 = rock music, N5 = siren, N6 = trill telephone, N7 = female speech, N8 = male speech, N9 = female speech.

Intrusions	**Wang-Brown model**		**Proposed model**	
	R_{EL}	R_{NR}	R_{EL}	R_{NR}
N0	6.99%	0%	3.93%	0.0019%
N1	28.96%	1.61%	8.16%	0.75%
N2	5.77%	0.71%	3.13%	0.75%
N3	21.92%	1.92%	6.88%	1.42%
N4	10.22%	1.41%	6.19%	0.97%
N5	7.47%	0%	4.58%	0.0055%
N6	5.99%	0.48%	3.46%	0.22%
N7	8.61%	4.23%	5.88%	2.30%
N8	7.27%	0.48%	3.91%	0.83%
N9	15.81%	33.03%	11.93%	26.20%
Average	11.9%	4.39%	5.81%	2.73%

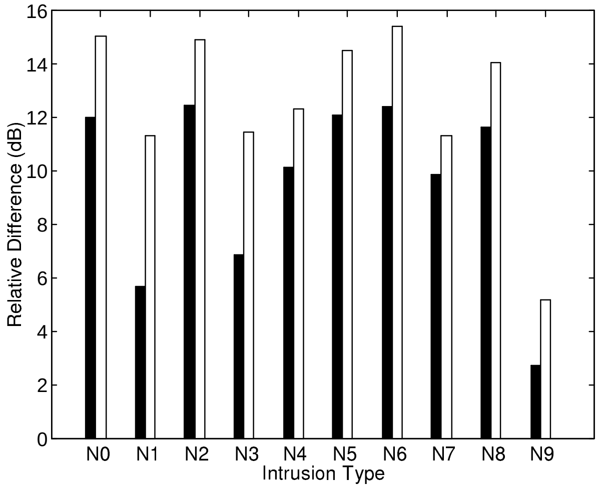

Figure 5: Black bar-The relative difference between the target speech resynthesized from the Wang-Brown model and that resynthesized from the ideal binary mask. White bar-The relative difference between the resynthesized speech from the proposed model and the speech resynthesized from the ideal mask. The different intrusion types are shown in Table 1.

R_{EL} obtained from our model is significantly smaller than from the Wang-Brown model, especially for random noise (N1) and the cocktail party noise (N3). For most wideband intrusions (N1, N3, N4, N7, N9), R_{NR} is decreased in our model, especially for N9. On the other hand, R_{NR} is also increased for some other intrusions (N0, N2, N5, N8), but the increase is rather small. Overall, the pattern of results from our model is substantially better.

To measure the waveform directly, we also calculate the relative difference in decibel between $I(t)$ and $S(t)$ as follows:

$$D = 10\log_{10}[\sum_t I^2(t) \Big/ \sum_t e^2(t)] \tag{8}$$

D is an evaluation that combines both R_{EL} and R_{NR}. The average D for each intrusion is shown in Fig. 5. Each value is again the average of 10 voiced utterances mixed with a certain kind of intrusion. The results of the Wang-Brown model are also shown in Fig. 5. For all the intrusions, we observe an improvement, and the average increase is around 3 dB.

5 Discussion

In this paper, we have extended the Wang-Brown model to improve the performance of speech segregation. Our model includes two novel processes, both of which can be implemented by oscillator neural networks similar to the two-layer oscillator network employed by Wang and Brown. Neural implementation will be addressed in future research.

Our estimation for the pitch of target speech is based on the outcome of the Wang-Brown model. For all the mixtures in the evaluation corpus, most estimated pitch contours are close to the ones obtained from clean target speech. With the estimated pitch, most oscillators of low-frequency channels (<1 *k*Hz) are grouped correctly for most intrusions. One exception is the intrusion N9, which is a female voice with fundamental frequency (F0) close to the doubles of the F0s of target speech. Therefore, the spectra of N9 and target speech overlap considerably. Although the performance of our model on N9 is still relatively poor, the amount of the residue noise is significantly reduced.

For two oscillators of nearby high-frequency channels (>1 *k*Hz), the corresponding responses may not be highly correlated even when they mainly come from the same source. These oscillators are put to the background in the Wang-Brown model, though many of them containing target speech signals. In our model, segments are generated with less constraint by cross-correlation between adjacent oscillators. These segments will be grouped into the new target speech stream if they agree with the estimated pitch contour and are sufficiently long. As a result, our model is able to recover more target speech signals in the high-frequency domain.

In generally, the oscillators corresponding to high-frequency channels (>1 *k*Hz) are difficult to group by pitch alone. There are three reasons:

1. Amplitude modulation (AM). Our model handles AM with our new grouping criterion.
2. The fundamental frequency of target changes with time. As a result, the peaks of the autocorrelation functions in the high-frequency domain do not always align with those in the low-frequency domain.
3. The responses in the high-frequency domain change rapidly, and the peaks in the autocorrelation functions are steep. As a result, a small difference between the estimated pitch period and the ideal pitch period may cause incorrect grouping.

The grouping method in our model does not deal with the last two issues, which will be addressed in future research.

In summary, our model mainly includes the following innovations:

- Estimate the pitch contour of target speech directly and use it for grouping.
- Further divide segments into smaller ones. The target speech stream is generated by grouping these segments. This helps in dealing with situations where target speech and intrusion overlap significantly in their spectra.
- Further group the oscillators whose corresponding signal is not highly correlated with that of nearby oscillators.
- A new grouping criterion is proposed to deal with AM.

Acknowledgements. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027).

References

[1] A. S. Bregman, *Auditory Scene Analysis*, Cambridge, MA: MIT press, 1990.

[2] C. Jutten and J. Herault, “Blind Separation of Sources, Parts I - III,” *Signal Processing*, Vol. 24, 1991, pp. 1-29.

[3] M. P. Cooke, *Modeling Auditory Processing and Organization*, U.K.: Cambridge University, 1993.

[4] G. J. Brown and M. P. Cooke, “Computational Auditory Scene Analysis,” *Computer Speech and Language*, Vol. 8, 1994, pp. 297-336.

[5] D. P. W. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D Dissertation, MIT Department of Electrical Engineering and Computer Science, 1996.

[6] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th Ed. San Diego, CA: Academic, 1997.

[7] T. W. Lee, M. S. Lewicki, and T. J. Sejnowski, “Blind Source Separation of More Sources Than Mixtures Using Overcomplete Representations,” *IEEE Signal Processing Letters*, Vol. 6, 1999, pp. 87-90.

[8] D. L. Wang and G. J. Brown, “Separation of Speech from Interfering Sounds Based on Oscillatory Correlation,” *IEEE Trans. Neural Network*, Vol. 10, 1999, pp. 684-697.

[9] A. J. W. van der Kouwe, D. L. Wang, and G. J. Brown, “A Comparison of Auditory and Blind Separation Techniques for Speech Segregation,” *IEEE Trans. Speech and Audio Processing*, Vol. 9, 2001, pp. 189-195.

[10] M. P. Cooke, P. D. Green, L. Josifovski, and A. Vizinho, “Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data,” *Speech Communication*, Vol. 34, 2001, pp. 267-285.