SPEECH SEGREGATION BASED ON PITCH TRACKING AND AMPLITUDE MODULATION

Guoning Hu

The Ohio State University
Biophysics Program2015 Neil Ave. Columbus, Ohio, 43210
hu.117@osu.edu

DeLiang Wang

The Ohio State University
Department of Computer & Information Science, and 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 pitch tracking and grouping based on amplitude modulation (AM). Our model is systematically evaluated and compared with the Wang-Brown model, and it yields significantly better performance.

1. INTRODUCTION

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]. It is a challenging task to develop a computational system to separate signals from acoustic mixtures. Blind sources separation [2] [3] provides a general way for signal separation, which only works when information of more than one acoustic mixture is available [4]. Another approach is to develop a system utilizing psychoacoustic cues to mimic ASA [5] [6] [7] [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 certain acoustic mixtures, the global pitch is meaningless (Fig. 2) and cannot provide useful information for grouping. The temporal continuity condition helps to generate large segments across time. However, when target speech and intrusion overlap significantly in their spectra, some segments may contain strong signals from both sources. As a result, the Wang-Brown model performs poorly when intrusion is wideband.

In this paper, we extend their model by introducing two further processes. The first process is to estimate the pitch contour of target speech, which is a much better grouping cue than the global pitch contour. Generally, it is difficult to obtain a good approximation of the pitch contour of target speech except that target speech is dominant. The target speech segregated by the Wang-Brown model provides a good basis for its estimation. The second process refines the generation of a new target speech stream. In this process, with the estimated pitch contour, 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. A new grouping criterion is proposed to deal with AM of the responses in high-frequency domain.

Detailed explanations of these two processes are given in Section 2 and Section 3. Evaluations and discussions are given in Section 4 and Section 5.

2. TARGET PITCH CONTOUR ESTIMATION

For any input signal, the Wang-Brown model is first applied to generate a target speech stream, referred to as S_{WB}. Let τ_j represent the estimated pitch period of target speech 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. 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 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 [6].

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

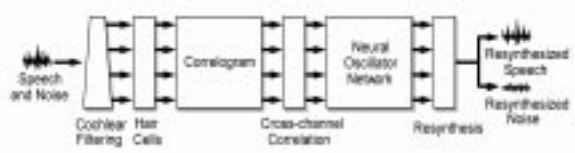

Figure 1. *The schematic diagram of the Wang-Brown model*

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

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. Furthermore, 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. Otherwise, they will be treated as unreliable.

Let j_s - j_e represent a streak that 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 longest one. For $j < j_{ms}$, τ_j is determined as follows. Starting with $j = j_{ms} - 1$, let

$$d_{\tau_3}(j) = \begin{cases} \sum \tau_j - \tau_{j+1} | & \text{if } \tau_{j+1} \text{ is reliable} \\ \tau_{j+1} & \text{otherwise} \end{cases} \quad (2)$$

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, τ_j is determined by searching the peak in the summary autocorrelation of these oscillators in the range $[0.8\tau_{j+1}, 1.2\tau_{j+1}]$ and 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 contour, the target speech stream is generated as follows. First, Eq. (1) is used to determine whether an oscillator agrees with the estimated pitch contour with the following modifications. Because the responses of high-frequency channels ($>1\,k$Hz) usually contain several harmonic components, the corresponding autocorrelation functions are likely to be amplitude modulated. As an example, Fig. 3(a) shows the amplitude-modulated response of a high-frequency

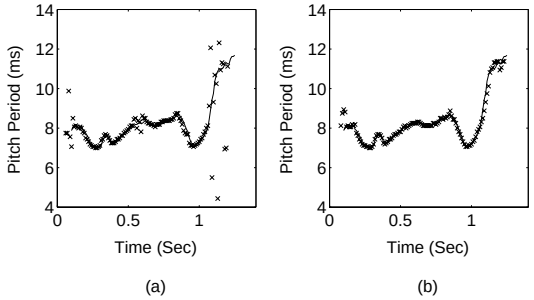

Figure 2. *The global and the estimated pitch contour of the speech obtained from the mixture of a male voice and the "cocktail party" noise. In both (a) and (b), the line contour represents the pitch contour obtained from clean speech. In (a), symbol 'x' represents the global pitch periods. In (b),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

channel. In 3(b), the corresponding autocorrelation function is also amplitude modulated, and the pitch period corresponds to a local maximum, but not the global maximum for $\tau \in [2\ ms, 12.5\ ms]$. Therefore, $A(i, j, \tau_m)$ is determined by searching the maximum only for $\tau \in [\tau_j / 2, 12.5\ ms]$. Furthermore, the threshold θ_d is changed to 0.85.

A segment agrees with the estimated pitch period at a particular time frame if more than half of its oscillators at this time frame agree with the estimated pitch period. 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 [8]. Then all the oscillators in the segments disagree with the estimated pitch contour are marked -1. For those in the segments agree with the estimated pitch contour, if themselves agree with the estimated pitch periods, they are marked 1; otherwise, they are marked 2. For those that do not belong to any segment, they are marked 3 if they agree with the estimated pitch periods. Other oscillators are marked 0.

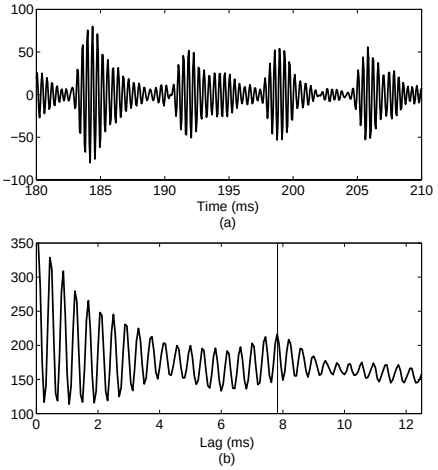

Figure 3. *(a) The response from the auditory filter with center frequency* 2.6 *kHz. The input is the male voice used in Fig.* 2. *(b) The corresponding autocorrelation function. The vertical line marks the corresponding position of the pitch period.*

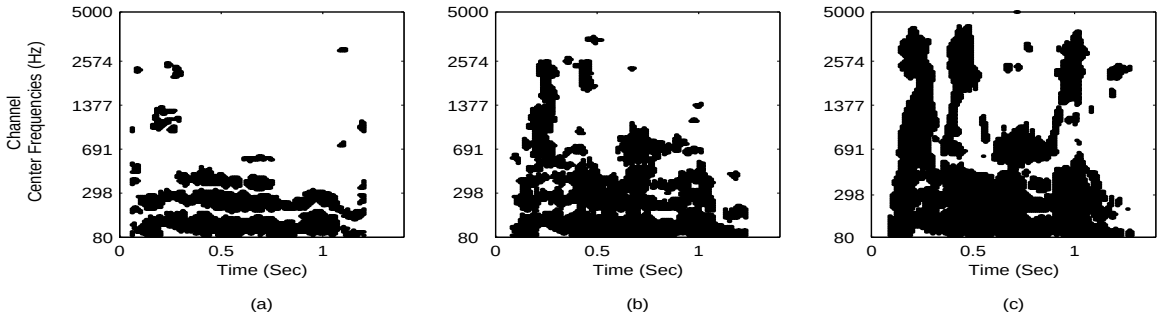

Figure 4. *(a) Black part - the speech stream obtained from the Wang-Brown model. (b) The speech stream obtained from our model. (c) The stream corresponds to the ideal mask. These are generated from the same mixture as used in Figure* 2.

New segments are formed by putting nearby oscillators together if they are marked the same. 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, and it 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 (Fig. 4(c)), which will be explained later.

4. Results

Our model is evaluated on the same corpus of mixtures-10 voiced utterances mixed with 10 intrusions-as used to evaluate the Wang-Brown model [8]. The speech signal are resynthesized [6] 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 [9]. 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 from the ideal mask. Let $e(t)$ be the difference between $I(t)$ and $S(t)$, 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. We define the energy loss ratio R_{EL} and noise residue ratio R_{NR} as follows:

$$R_{EL} = \sum_t e_1^2(t) \Big/ \sum_t I^2(t) \tag{3}$$

$$R_{NR} = \sum_t e_2^2(t) \Big/ \sum_t S^2(t) \tag{4}$$

Table 1. R_{EL} and R_{NR} of resynthesized speech from both the Wang-Brown model and the proposed model. Here, N0 = 1 kHz 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, and 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%

Table 1 shows the R_{EL} and R_{NR} for the 10 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} for the resynthesize speech from the Wang-Brown model. 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 between $I(t)$ and $S(t)$ in decibels as follows:

$$D = 10\log_{10}[\sum_t I^2(t) \Big/ \sum_t e^2(t)] \quad (5)$$

D is an evaluation that combines both R_{EL} and R_{NR}. The average D for each intrusion is shown in Fig. 5. 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

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 contour, 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 arise 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 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.

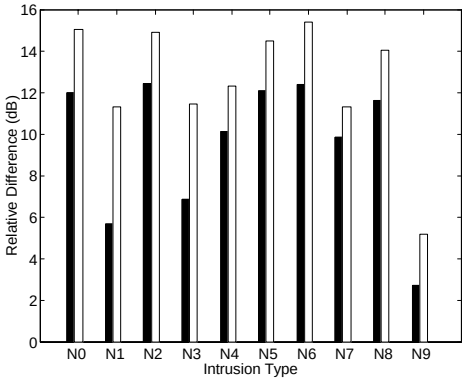

Figure 5. *Black bar-The relative difference between the target speech resynthesized from the original model and that resynthesized from the ideal binary mask. White bar-The relative difference between the resynthesized speech from the proposed model and the speech resynthesized from the ideal mask.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

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

6. 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, pp. 1-29, 1991.

[3] 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, pp. 87-90, 1999.

[4] 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, pp. 189-195, 2001.

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

[6] G. J. Brown and M. P. Cooke, "Computational Auditory Scene Analysis," *Computer Speech and Language*, vol. 8, pp. 297-336, 1994.

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

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

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

[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, pp. 267-285, 2001