MONAURAL SPEECH SEGREGATION BASED ON PITCH TRACKING AND AMPLITUDE MODULATION

Guoning Hu

Biophysics Program,
The Ohio State University
hu.117@osu.edu

DeLiang Wang

Department of Computer and Information Science
& Center of Cognitive Science
The Ohio State University
dwang@cis.ohio-state.edu

ABSTRACT

Monaural speech segregation remains a computational challenge for auditory scene analysis (ASA). A major problem for existing computational auditory scene analysis (CASA) systems is their inability to deal with signals in the high-frequency range. Psychoacoustic evidence suggests that different perceptual mechanisms are involved to handle resolved and unresolved harmonics. We propose a system for speech segregation that deals with low-frequency and high-frequency signals differently. For low-frequency signals, our model generates segments based on temporal continuity and cross-channel correlation, and groups them according to periodicity. For high-frequency signals, the model generates segments based on common amplitude modulation (AM) in addition to temporal continuity, and groups them according to AM repetition rates. Underlying the grouping process is a pitch contour that is first estimated from segregated speech based on global pitch and then verified by psychoacoustic constraints. Our system is systematically evaluated, and it yields substantially better performance than previous CASA systems, especially in the high-frequency range.

1. INTRODUCTION

In the real-world environment, target speech usually occurs simultaneously with acoustic interference. An effective speech segregation system will greatly facilitate many applications, including automatic speech recognition (ASR) and speaker identification. Many systems have been proposed to deal with speech segregation, primarily using blind source separation (BSS) [1] or speech enhancement techniques [2]. BSS performs well when there are enough sensors and the mixing signals satisfy some statistical independence. However, BSS techniques require at least two sensors, while many applications such as telecommunication and audio retrieval need a monaural (one sensor) solution. Speech enhancement techniques perform well in certain environments where some prior knowledge about target/interference is available. However, no system can efficiently separate speech from a variety of acoustic intrusions with one sensor.

While monaural segregation remains a difficult challenge for computational systems, the auditory system shows an impressive capacity for monaural segregation. ASA is the perceptual process in which an acoustic mixture is analyzed and separated into streams, corresponding to the acoustic sources [3]. Considerable research has been carried out to build monaural CASA systems [4-7]. Almost all existing systems rely on periodicity as a main grouping cue. However, the performance of these systems is limited and progress has stagnated in recent years. A main problem with the current systems is that they lack the ability to deal with high-frequency signals.

We study monaural speech segregation with particular emphasis on the high-frequency problem. For voiced signal, we note that the auditory system can resolve the first few harmonics in the low-frequency range but higher harmonics are unresolved unless they are much more intense than adjacent ones [8]. Psychophysical evidence suggests different mechanisms to deal with resolved and unresolved harmonics [9]. Consequently, our model employs different methods to segregate target speech in the low-frequency range and in the high-frequency range. According to Bregman [3], ASA takes place in two stages: segmentation (or analysis) and grouping. In segmentation, the acoustic input is decomposed into sensory segments, each of which would belong to one source. In grouping, those segments that are likely to respond to the same source are grouped together. Inspired by this suggestion, our model performs segregation in the two corresponding stages across all frequency channels. More specifically, for low-frequency channels, our system generates segments based on temporal continuity and cross-channel correlation between responses from nearby channels. These segments are grouped by comparing the periodicity of these responses with the estimated pitch of the target speech. On the other hand, high-frequency channels due to the wide bandwidths tend to respond to multiple harmonics of voiced speech, which are usually unresolved. These high-frequency responses are amplitude modulated, and their envelopes fluctuate at the frequency corresponding to the fundamental frequency (F0) [10]. Our model generates segments in the high-frequency range based on common AM and temporal continuity. These segments are grouped by comparing AM repetition rates with estimated F0 of the target speech. In addition, we propose a new method for pitch tracking. First, a foreground stream is segregated from the mixture according to

0-7803-7402-9/02/$17.00 ©2002 IEEE

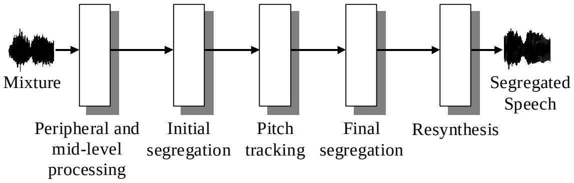

Fig. 1. The schematic diagram of the proposed system.

the global pitch of the mixture [7]. Then the pitch contour of the target speech is estimated from the foreground stream. Finally, the estimated pitch is checked according to two psychoacoustically-inspired constraints: 1) An accurate pitch period should be consistent with the periodicity of responses in the channels where the target speech dominates; 2) Pitch periods should vary smoothly in time.

Section 2 describes the overall system. In section 3, systematic results and a comparison with an existing CASA system are given. Section 4 concludes the paper.

2. MODEL DESCRIPTION

Our model is a multistage system, as shown in Fig. 1. Description for each stage is given below.

2.1 Peripheral and mid-level processing, and initial segregation

First, an acoustic input is analyzed by a peripheral model comprising cochlear filtering with a bank of 128 gammatone filters and subsequent hair cell transduction. This peripheral processing is done in time frames of 20 *ms* long and 10 *ms* overlap between consecutive ones. As a result, the input signal is decomposed into a group of cells. Each time-frequency cell contains the response of a certain channel in a certain frame. The envelope of the response is obtained by a lowpass filter with passband [0, 1 *k*Hz] and a Kaiser window of 18.25 *ms*. Mid-level processing is performed by computing a correlogram (autocorrelation function) of the individual responses and their envelopes. The global pitch contour is obtained from the summary correlogram.

Initial segregation takes place in two steps. First, segments are formed by grouping neighboring time-frequency cells based on temporal continuity and cross-channel correlation. In general, segments correspond to resolved components of the input signal, and most of them lie in the low-frequency range. Then according to global pitch, segments are grouped into a foreground stream, which corresponds to the target speech, and a background stream, which corresponds to the intrusion. A similar process is described in the oscillatory correlation model of Wang and Brown [7].

2.2 Target pitch tracking

First, the pitch periods of target speech are estimated from the foreground stream. In each frame, the autocorrelation functions of cells in the foreground stream are summated. The pitch period is the lag corresponding to the maximum of the summation in the range [2 *ms*, 12.5 *ms*].

Since the foreground stream still contains intrusions, some of the estimated pitch periods are not accurate. Our system reestimates target pitch with two constraints. First, an accurate pitch period in a frame should be consistent with the periodicity of those cells in this frame in the foreground stream. Let $\tau(j)$ represent the estimated pitch period in frame j, and $A(i, j, \tau)$ the autocorrelation function of the cell of channel *i*. This cell agrees with $\tau(j)$ if

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

Here $\theta_d = 0.95$, τ_m is the lag corresponding to the maximum of $A(i, j, \tau)$ for $\tau \in$ [2 *ms*, 12.5 *ms*]. If more than half of the cells in the foreground stream in frame j agree with $\tau(j)$, $\tau(j)$ is marked as reliable. Second, pitch periods should vary smoothly in time. We stipulate the difference between reliable pitch periods in consecutive frames be smaller than 20%, which is justified from empirical data. Otherwise, they are marked as unreliable.

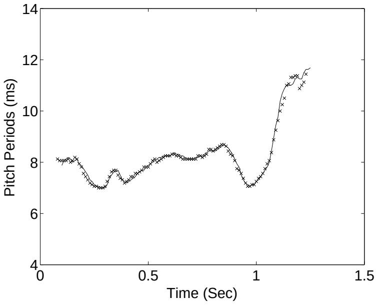

Fig. 2. "x" marks pitch periods of target speech estimated from a mixture of a voiced utterance and "cocktail party" noise. The solid line is the pitch contour obtained from clean speech.

Unreliable pitch periods are replaced by new values obtained through temporal continuity. Suppose in two consecutive frames j and $j+1$ that $\tau(j)$ is reliable while $\tau(j+1)$ is unreliable. All the channels corresponding to the cells agreeing with $\tau(j)$ in frame j are selected. $\tau(j+1)$ is obtained from the summation of the autocorrelations of the cells in frame $j+1$ that correspond to those selected channels. Then it is verified with the second constraint. Finally, every unreliable pitch period is replaced by a linear interpolation of reliable pitch periods from nearby frames. As an example, Fig. 2 shows the estimated pitch periods from a mixture of a voiced utterance and the cocktail party noise, which match the pitch periods obtained from clean speech well.

2.3 Pitch-based labeling

With the estimated pitch periods, (1) provides a criterion to label whether target speech dominates in a cell or not. This criterion compares an estimated pitch period with the periodicity of the response in a cell. It works well in the low-frequency range where harmonics are resolved. However, it is not suitable for high-frequency channels because their responses are likely to contain multiple harmonics and therefore are amplitude modulated. As shown in Fig. 3, for a response with strong AM, the pitch period corresponds to a local maximum in a correlogram instead of the global maximum. In addition, the peaks of the correlogram are steep, which makes (1) too sensitive to interference.

For high-frequency responses where speech dominants, response envelopes fluctuate at the rate of F0 [10]. Based on this

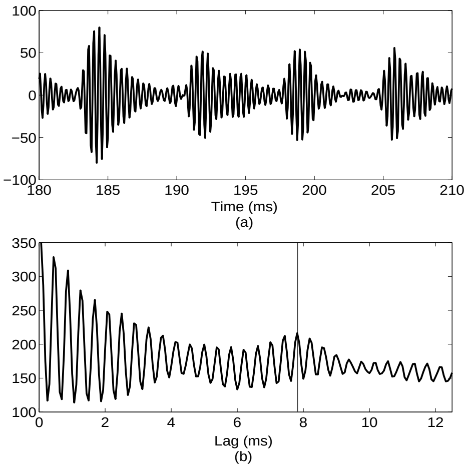

Fig. 3. (a) The response of a channel with center frequency 2.6 *k*Hz. The input is the clean speech used in Figure 2. (b) The corresponding autocorrelation function. The vertical line marks the position of the pitch period.

phenomenon, we propose a new criterion by comparing AM repetition rate with estimated instantaneous F0, which is obtained by interpolating the estimated pitch periods of target speech. To obtain the AM repetition rate, first, the response of each channel is half-wave rectified and then bandpass filtered to remove DC component and harmonics except for the one corresponding to F0. A filter with passband $[0.9\bar{f}, 1.2\bar{f}\ ]$ and a Kaiser window of $50\,ms \sim 100\,ms$ is used for response in every $100\,ms$ period. $\bar{f}$ is the average of estimated instantaneous F0 in the $100\,ms$ and it determines the window size. The instantaneous frequency (IF) of the rectified and filtered signal, $f_I(i,t)$, obtained through a linear prediction algorithm in the spectral domain [11], indicates the AM repetition rate of the response. To measure the relative difference between the estimated instantaneous F0 and AM repetition rate, let

$$D(i,j)=\sqrt{\frac{1}{M}\sum_{k=0}^{M-1}[\log f_0(jT-k)-\log f_I(i,jT-k)]^2}\ . \tag{2}$$

Where $f_0(t)$ is the estimated instantaneous F0, M spans $20\,ms$, and $T=10\,ms$. The smaller $D(i,j)$ is, the more likely it is for target speech to dominate the cell. The following criterion is used to label whether target speech dominates in a cell or not:

$$D(i,j)<\theta_f\ . \tag{3}$$

Listeners can discriminate two simultaneous sounds with unresolved harmonics if the difference in F0 is more than 10% [9]. $D(i,j)$ is about 0.1 if there is a constant 10% difference between $f_I(i,t)$ and $f_0(t)$. We set θ_f to 0.15 with the consideration that the difference between $f_I(i,t)$ and $f_0(t)$ is more likely to be Gaussian distributed.

2.4 Final segregation

First, segments in the high-frequency range are generated based on temporal continuity and common AM repetition rate for cells that satisfy (3). In this process, only the cells that are neither in the foreground stream nor in the background stream are considered for the following reasons. There should be no conflict between this segmentation process and the one in initial segregation. Furthermore, the segments generated in initial segregation tend to reflect resolved components, and therefore shall be retained. The similarity of AM repetition rates between the responses of nearby cells is measured by the cross-channel correlation of response envelopes. Segments are formed by grouping neighboring cells satisfying the above criteria. Most of them are in the high-frequency range.

Then these segments are grouped into the foreground stream. Besides them, other segments in the foreground stream are separated so that all the cells in one segment either satisfy or violate (1). Some segments are removed from the foreground stream as a result, and they are put into the background stream if they contain cells violating (1) only.

Other cells that do not belong to either stream are grouped according to temporal and spectral continuity. More specifically, first, the background stream expands iteratively by grouping neighboring cells violating (1) or (3) until no more cell can be added. Then the foreground stream expands by grouping neighboring cells satisfying (1) or (3) iteratively.

Finally, segregated target speech can be resynthesized from the foreground stream. In resynthesis, the signals of the cells in the foreground stream are retained, while other signals are removed [5].

3. RESULTS AND COMPARISON

Our system is evaluated with a corpus of 100 mixtures composed of 10 voiced utterances mixed with 10 intrusions collected by Cooke [4]. The speech waveform resynthesized from the segregated speech stream is used for evaluation. For every mixture, the speech waveform resynthesized from an ideal binary stream, which is composed of all the cells where target speech dominates, is used as the ground truth of target speech. Theoretically speaking, an ideal binary mask gives the ceiling of performance for all binary masks. This evaluation methodology is supported by the following observations. In a critical band, a weak signal is masked by a stronger one [8]. In addition, the ideal stream is similar to the prior mask used in a recent study for ASR [12], which yields excellent recognition performance.

Fig. 4 illustrates the speech stream segregated from a mixture of a voiced utterance and the cocktail party noise. It matches the ideal binary stream shown in Fig. 4(b) well.

Let $S(t)$ be the resynthesized waveform by our system, $I(t)$ the waveform from the ideal stream, $e_1(t)$ the signal present in

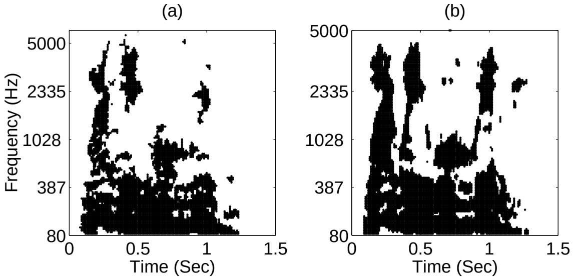

Fig. 4. (a) The segregated speech stream. (b) The ideal binary stream. The input signal is a voiced utterance mixed with the cocktail party noise.

$I(t)$ but missing from $S(t)$, and $e_2(t)$ the signal present in $S(t)$ but missing from $I(t)$. We measure the ratio of energy loss, R_{EL}, and the ratio of noise residue R_{NR} as follows:

$$R_{EL} = \sum_t e_1^2(t) \Big/ \sum_t I^2(t) \qquad (4)$$

$$R_{NR} = \sum_t e_2^2(t) \Big/ \sum_t S^2(t)\,. \qquad (5)$$

The results are shown in Table 1. The table also shows for comparison the results from the Wang-Brown system [7], which has the best performance on the same corpus. Each value is the average of a certain intrusion type. Compared with the Wang-Brown system, our system generates significantly smaller ratios of energy loss, especially for N1 and N3. Similar ratios of noise residue are obtained from both systems except for N9 where our result is much better. We note that our overall improvement comes mainly from high-frequency channels.

To compare waveforms directly, we also measure a form of signal to noise ratio (SNR) in decibels using $I(t)$ as ground truth:

$$SNR = 10\log_{10}[\sum_t I^2(t) \Big/ \sum_t (I(t) - S(t))^2]\,. \qquad (6)$$

The average SNR for each intrusion is shown in Fig. 5. Compared with the Wang-Brown model, our model increases SNR for all the intrusions. The average gain is about 4.5 dB.

Table 1: R_{EL}'s and R_{NR}'s of segregated speech from the proposed system and the Wang-Brown system. Each value is the average of each intrusion over 10 voiced utterances. (Intrusion types are: N0-pure tone, N1-white noise, N2-noise bursts, N3-the cocktail party noise, N4-rock music, N5-siren, N6-trill telephone, N7-female speech, N8-male speech, and N9-female speech.)

Intrusion	**Proposed system**		**Wang-Brown system**	
	R_{EL} (%)	R_{NR} (%)	R_{EL} (%)	R_{NR} (%)
N0	2.52	0.01	6.99	0
N1	7.76	1.08	28.96	1.61
N2	1.96	0.11	5.77	0.71
N3	5.27	1.64	21.92	1.92
N4	5.45	0.91	10.22	1.41
N5	3.38	0.02	7.47	0
N6	2.20	0.09	5.99	0.48
N7	4.14	1.71	8.61	4.23
N8	2.56	1.34	7.27	0.48
N9	10.00	19.08	15.81	33.03
Average	4.52	2.60	11.91	4.39

4. CONCLUSION

Our model estimates target pitch from initial segregation based on global pitch. Estimated pitch periods are corrected by psychoacoustically-motivated constraints. As a result, most estimated pitch contours are close to those obtained from clean speech. With reliable pitch, our system deals with low-frequency and high-frequency signals differently. AM repetition rate is used for segregation in the high-frequency range. Our monaural model has been systematically evaluated on a mixture corpus, and it yields very good results. The performance of our system is significantly better than a pervious CASA system evaluated on the same corpus. Our study demonstrates that computational investigation that incorporates ASA principles is a promising direction for monaural segregation, given the auditory system's remarkable ability for the task.

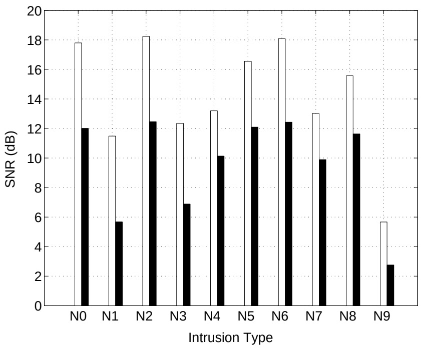

Fig. 5. SNR of segregated target speech. White bar: results from our system, and black bar: results from the Brown-Wang system. Different intrusion types are shown in Table 1.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] V. Zarzoso and A. K. Nandi, "Blind Source Separation," *Blind Estimation Using Higher-order Statistics*, Boston: Kluwer Academic Publishers, 1999, pp. 167-252.

[2] D. O'Shaughnessy, *Speech Communications, Human and Machine*, 2nd Ed. New York: IEEE Press, 2000, pp. 323-336.

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

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

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

[6] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*, Mahwah, NJ: Lawrence Erlbaum, 1998.

[7] 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.

[8] C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th Ed. Academic Press, 1997.

[9] C. J. Darwin and R. P. Carlyon, "Auditory Grouping," *Hearing*, 2nd Ed. Academic Press, 1995, pp. 387-424.

[10] H. Helmholtz, *On the Sensations of Tone*, Braunschweig: Vieweg & Son, 1863. (A.J. Ellis, English Trans., Dover, 1954).

[11] R. Kumaresan and A. Rao, "Model-based Approach to Envelope and Positive Instantaneous Frequency Estimation of Signals with Speech Applications," *J. Acoust. Soc. Am.*, Vol. 105, 1999, pp. 1912-1924.

[12] 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.