Technical Report: OSU-CISRC-3/02-TR6
Department of Computer and Information Science
The Ohio State University
Columbus, OH 43210-1277

Ftp site: **ftp.cis.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2002**
File in ps format: **TR6.ps**
Web site: **http://www.cis.ohio-state.edu/research/tech-report/html**

Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation

Guoning Hu and DeLiang Wang

Abstract—**Speech segregation in the monaural condition has proven to be very challenging. Monaural speech segregation has been studied in previous systems that incorporate auditory scene analysis principles. A major problem for these systems is their inability to deal with speech in the high-frequency range. Psychoacoustic evidence suggests that different perceptual mechanisms are involved in handling resolved and unresolved harmonics. We propose a system that deals with resolved and unresolved harmonics differently. For resolved harmonics, the system generates segments based on temporal continuity and cross-channel correlation, and groups them according to their periodicities. For unresolved harmonics, it generates segments based on common amplitude modulation (AM) in addition to temporal continuity and groups them according to AM repetition rates derived from sinusoidal modeling. Underlying the segregation process is a pitch contour that is first estimated from speech segregated according to global pitch and then adjusted according to psychoacoustic constraints. Our system is systematically evaluated, and it yields substantially better performance than previous systems, especially in the high-frequency range.**

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

G. Hu and D. L. Wang are with the Department of Computer and Information Science and Center for Cognitive Science, the Ohio State University, Columbus, OH43210-1277, USA (email: {hu, dwang}@cis.ohio-state.edu).

Index Terms—**Monaural speech segregation, computational auditory scene analysis, pitch tracking, amplitude modulation, segmentation, grouping.**

I. INTRODUCTION

In a natural environment, speech usually occurs simultaneously with acoustic interference. An effective system for attenuating acoustic interference would greatly facilitate many applications, including automatic speech recognition (ASR) and speaker identification. General methods for signal separation or enhancement, such as blind source separation using independent component analysis [13] or sensor arrays for spatial filtering [16], require multiple sensors. However, many applications such as telecommunication and audio retrieval need a monaural solution. In the monaural condition (one microphone), intrinsic properties of speech or interference must be considered. Various algorithms have been proposed for monaural speech enhancement, and they are generally based on some analysis of speech or interference and subsequent speech amplification or noise reduction. For example, methods have been proposed to estimate the short-time spectra of interference [18] [27] or speech modeling [10], [14] in the context of speech enhancement. Another way to deal with interference is to transform the acoustic mixture into vectors in a vector space and then apply subspace analysis to remove part of the interference [8] [15]. Hidden Markov models have also

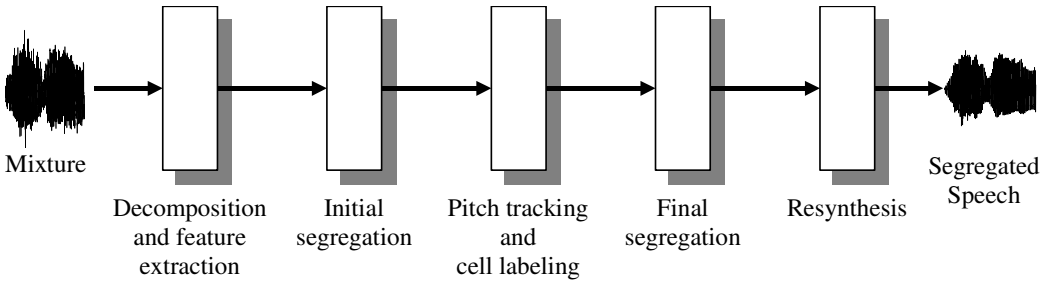

Figure 1. Schematic diagram of the proposed multistage system.

been used to model the spectral characteristics of speech and interference and then separate them [9] [26]. These methods usually assume certain properties of interference and have difficulty in dealing with general acoustic interference, because the variety of both interference and natural speech make them very difficult to model.

While monaural speech enhancement remains a challenge, the auditory system shows a remarkable capacity for monaural speech segregation. According to Bregman [1], the auditory system segregates the acoustic signal into streams, corresponding to different sources, according to *auditory scene analysis* (ASA) principles. Research in ASA has inspired considerable work to build *computational auditory scene analysis* (CASA) systems for sound segregation [2] [4] [7] [25] [28]. Such systems generally approach speech segregation without making strong assumptions on the acoustic properties of interference, and follow two main stages: segmentation (analysis) and grouping (synthesis) [1]. In segmentation, the acoustic input is decomposed into sensory segments, each of which is likely to originate from a single source. In grouping, those segments that likely come from the same source are grouped together, based mostly on periodicity.

A recent CASA model proposed by Wang and Brown consists of three stages: auditory periphery, mid-level representation, and a two-layer oscillator network for segmentation and grouping [28]. Segments are formed on the basis of similarity between adjacent filter responses (cross-channel correlation) and temporal continuity, while grouping among segments is performed according to the global pitch extracted within each time frame. Their system has been tested on a common corpus of acoustic mixtures of voiced utterances and different types of interference, including white noise, "cocktail party" noise, and competing speech [4]. They have reported good results in comparison with previous CASA systems. In most situations, the model is able to remove intrusions and recover low-frequency (below $1\,k$Hz) energy of target speech. However, this model cannot handle high-frequency (above $1\,k$Hz ~ $5\,k$Hz) signals well, and it loses much of target speech in the high-frequency range. In fact, the inability to deal with speech in the high-frequency range is a common problem for CASA systems.

We propose a model for monaural speech segregation with particular emphasis on the high-frequency problem in CASA. The auditory system can resolve the first 5 to 7 harmonics of complex tones with F0 ranging from 44 Hz to 2000 Hz [23]. Higher harmonics are unresolved unless they are much more intense than adjacent ones. It is likely that different perceptual mechanisms are used to deal with resolved and unresolved harmonics [3]. Consequently, our model employs different methods to segregate resolved and unresolved harmonics of target speech. More specifically, we generate segments for resolved harmonics based on temporal continuity and cross-channel correlation, and these segments are grouped according to common periodicity. For unresolved harmonics, it is well known that filter responses are strongly amplitude-modulated and the response envelopes fluctuate at the fundamental frequency (F0) of target speech [11]. Therefore, our model generates segments for unresolved harmonics based on common AM in addition to temporal continuity. These segments are further grouped based on AM repetition rates. By observing pitch characteristics in natural speaking, we propose a method to obtain an accurate pitch track for target speech. Guided by a target pitch track we compute AM repetition rates using sinusoidal modeling and gradient descent. Also, we suggest the notion of an ideal binary mask as a psychoacoutically motivated criterion to measure the performance of CASA systems that produce binary time-frequency masks.

Section II gives an overview of our model, and Sections III-VI explain model components in detail. In Section VII, our system is systematically evaluated and results are compared with those of the Wang-Brown model. Further discussions are given in Section VIII.

II. MODEL OVERVIEW

Our model is a multistage system, as shown in Fig. 1. In the first stage, the input mixture is analyzed by an

auditory filterbank in consecutive time frames. This processing results in a decomposition of the input into a collection of time-frequency responses, which are called T-F units. Then certain auditory features are extracted. In the initial segregation stage, T-F units are merged into segments and then these segments are grouped into an initial foreground stream and a background stream, corresponding to target speech and intrusion respectively. In the third stage, target pitch is estimated from the initial foreground stream, and it is used to perform unit labeling. In the final segregation stage, some T-F units are merged into segments corresponding to unresolved harmonics of target speech, and these segments are added to the foreground stream. Then both streams are modified according to unit labels. Finally, a speech waveform is resynthesized from the final foreground stream. Here, the foreground stream works as a binary mask. It is used to retain the acoustic energy from the mixture that corresponds to 1's and reject the mixture energy corresponding to 0's; for more details of this stage, see [2]. The first four stages are explained in the following sections.

III. DECOMPOSITION AND FEATURE EXTRACTION

In this stage, our system decomposes the input signal and extracts following auditory features: the autocorrelation of filter responses, the autocorrelation of response envelopes, cross-channel correlation, and global pitch.

A. Signal Decomposition

First, the input signal passes through an auditory filterbank. Here we use a 128-channel "gammatone" filterbank [22] whose center frequencies are logarithmically spaced from 80 Hz to 5000 Hz. The gammatone filterbank is a standard model of cochlear filtering. The impulse response of a gammatone filter is:

$$g(t) = \begin{cases} t^{n-1}\exp(-2\pi bt)\cos(2\pi ft) & t \ge 0 \\ 0 & else \end{cases}, \qquad (1)$$

where $n = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, and f is the center frequency of the filter. The input signal is divided into $20\,ms$ time frames with $10\,ms$ overlapping between consecutive frames. This frame size is sufficiently long to estimate the pitch contour of target speech. Therefore, the input is decomposed into a collection of T-F units. Unit u_{ij} contains the response of the filter channel i at time frame j. Two units are neighbors when they are in the same channel but at consecutive frames, or they are in adjacent channels but at the same frame.

The response of each channel is further transduced by the Meddis model of inner hair cells [19]. This model stimulates well-known properties of hair cells, such as rectification, saturation, and phase locking. Its output represents the firing rate of an auditory nerve fiber. The envelope of the hair cell output is obtained through lowpass filtering, which contains AM information to be utilized later to segregate unresolved harmonics. The cutoff frequency of the lowpass filter needs to be below the frequency range where harmonics are unresolved but above the plausible F0 range of target speech. Here we use a FIR filter with passband [0, $1\,k$Hz] and a Kaiser window of $18.25\,ms$, but results do not change significantly for a range of cutoff frequencies.

B. Auditory Feature Extraction

A well-established mechanism for pitch extraction employs a correlogram - a running autocorrelation of each filter response across an auditory filterbank [20]. Previous models have shown that correlograms provide an effective mid-level auditory representation [2] [28]. Within u_{ij}, the autocorrelation function of the hair cell response is given by

$$A_H(i, j, \tau) = \frac{1}{K}\sum_{k=0}^{K-1} h(i, jT-k)h(i, jT-k-\tau)\,. \qquad (2)$$

Here, time lag $\tau \in [0, 12.5\,ms]$. The maximum delay corresponds to 80 Hz, and it is an appropriate choice since the F0 of target speech in our test corpus does not fall below this frequency. $h(i, t)$ indicates the hair cell output of channel i at time step t. T is the progression period from one frame to the next. Let f_i be the center frequency of channel i. The window size of the autocorrelation, K, is chosen to be the longer of $4/f_i$ and $20\,ms$, the latter being the width of a frame.

When a periodic sound is presented, the autocorrelations of the activated filters in a correlogram all exhibit a peak at the delay corresponding to the period. Using this property, the correlogram method for pitch extraction simply pools autocorrelations across all the channels and then identifies a global peak in the pooled correlogram. Let $s_G(j, \tau)$ be the summary correlogram at frame j,

$$s_G(j, \tau) = \sum_i A_H(i, j, \tau)\,. \qquad (3)$$

The global pitch at frame j, $\tau_G(j)$, is the lag corresponding to the maximum of $s_G(j, \tau)$ in the range [2 ms, 12.5 ms] - the plausible pitch range of target speech. For those channels where target speech dominates, their autocorrelations have peaks consistent with the target pitch and their summary generally shows a dominant peak corresponding to the pitch period. With the acoustic interference, a global pitch is usually not an accurate description of the target pitch, but it is reasonably close.

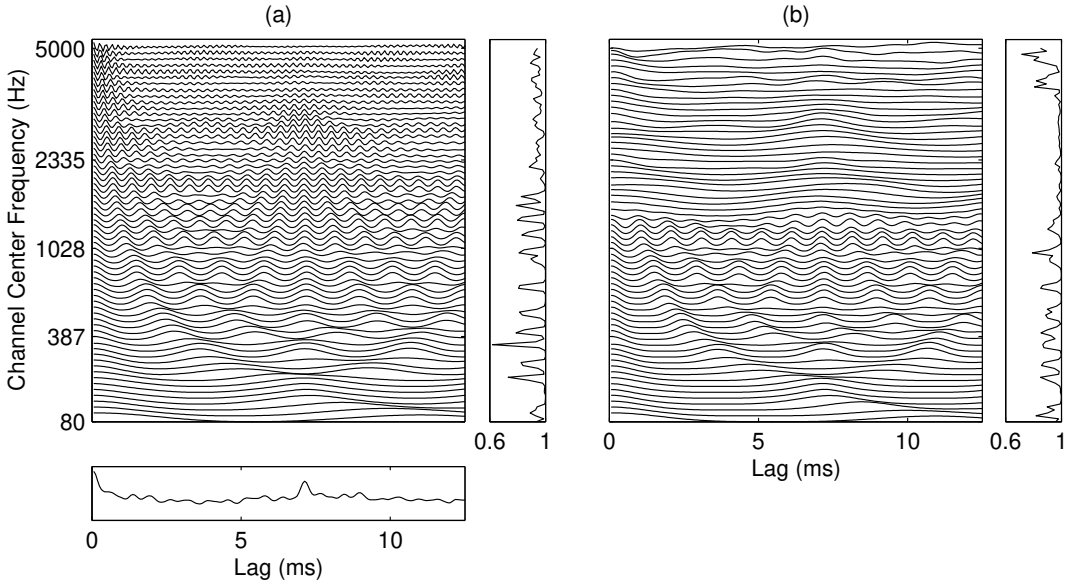

Figure 2. Auditory features in response to a mixture of speech and cocktail-party noise. (a) A correlogram at frame 25 (i.e. 0.25 second after the onset). For clarity, every other channel is shown. The corresponding cross-channel correlations are given in the right panel, and the summary correlogram in the bottom panel. (b) An envelope correlogram. The corresponding cross-channel correlations are shown in the right panel.

We introduce an envelope correlogram by computing the autocorrelation of the response envelope,

$$A_E(i, j, \tau) = \frac{1}{K}\sum_{k=1}^{K} h_E^2(i, jT - k) h_E^2(i, jT - k - \tau) . \quad (4)$$

Here, $h_E(i, t)$ is the envelope of the hair cell output in channel i at time step t. Because these envelopes are relatively flat, they are squared to yield sharper peaks. These autocorrelation functions reveal response periodicity as well as AM repetition rates.

As demonstrated by Wang and Brown [28], cross-correlation between adjacent filter channels indicates whether the filters respond to the same harmonic or formant, thus providing a useful feature in subsequent segmentation processes. For the same T-F unit, the cross-channel correlation is calculated as:

$$C_H(i, j) = \sum_{\tau=0}^{L-1} \hat{A}_H(i, j, \tau) \hat{A}_H(i+1, j, \tau) , \quad (5)$$

where $\hat{A}_H(i, j, \tau)$ denotes normalized $A_H(i, j, \tau)$ (zero mean and unity variance), and L spans $12.5\,ms$ - the maximum delay for A_H. Similarly, the cross-channel correlation of envelopes is calculated as follows:

$$C_E(i, j) = \sum_{\tau=0}^{L-1} \hat{A}_E(i, j, \tau) \hat{A}_E(i+1, j, \tau) , \quad (6)$$

where $\hat{A}_E(i, j, \tau)$ denotes normalized $A_E(i, j, \tau)$. C_H measures the response similarity between adjacent channels, and C_E the corresponding similarity between AM patterns.

Fig. 2(a) and 2(b) illustrate the correlogram and the envelope correlogram at a particular frame, in response to a mixture of a voiced utterance and the "cocktail party" noise (see Sect. VII). Their respective cross-channel correlations are given in the right panels. The figure indicates that hair cell responses are not highly correlated for high-frequency channels whose center frequencies are above $1\,k$Hz. On the other hand, the response envelopes are highly correlated when they have similar fluctuation patterns. The bottom panel in Fig. 2(a) shows the corresponding summary correlogram, where the dominant peak at $7\,ms$ corresponds to the global pitch period at this frame.

IV. INITIAL SEGREGATION

In this stage, units are merged into segments based on temporal continuity and cross-channel correlation. Using global pitch, these segments are grouped into an initial foreground stream and an initial background stream, corresponding to target speech and intrusion respectively. This stage is similar to the segmentation and grouping stage of the Wang-Brown system [28].

A. Initial Segmentation

Because a harmonic extends for a period of time and its frequency changes smoothly, target speech likely activates contiguous T-F units. This is an instance of the

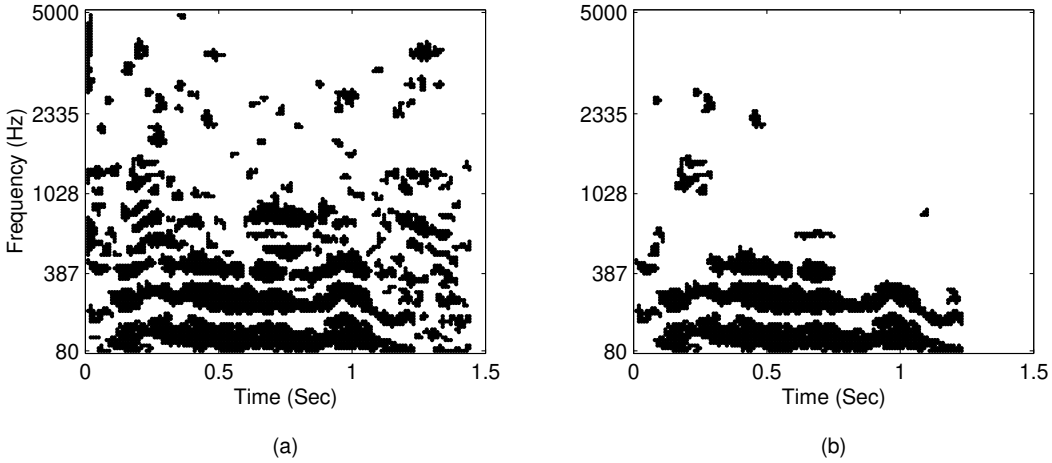

Figure 3. Results of initial segregation for the speech and cocktail-party mixture. (a) Segments formed. Each segment corresponds to a contiguous black region. (b) Foreground stream.

temporal continuity principle. In addition, since the passbands of adjacent channels overlap, a resolved harmonic usually activates adjacent channels, which leads to high cross-channel correlations. Therefore, we form segments by merging T-F units based on temporal continuity and cross-channel correlation.

First, only units with some response energy and sufficiently high cross-channel correlation are considered. More specifically, u_{ij} is selected if $A_H(i, j, 0) > \theta_e^2$ and $C_H(i, j) > \theta_c$. Note that the autocorrelation at zero-lag measures the energy. $\theta_e = 50$, which is close to the spontaneous firing rate of the auditory nerve, and $\theta_c = 0.985$, chosen to be the same as in [28]. Selected neighboring units are iteratively merged into segments. Finally, segments shorter than $30\,ms$ are removed since they unlikely arise from target speech.

For units that respond to unresolved components, their cross-channel correlations are relatively small, and these units are hence not selected. Therefore, segments generally reflect resolved components of either target speech or intrusion.

B. Initial Grouping

Initial grouping is done by comparing the periodicities of unit responses with global pitch. The response period of a T-F unit is obtained by finding the global maximum of its autocorrelation within the plausible pitch range. If this period and the corresponding global pitch period are close, the unit is said to agree with the global pitch. That is, u_{ij} agrees with the global pitch if

$$A_H(i, j, \tau_G(j)) / A_H(i, j, \tau_m) > \theta_p \,. \tag{7}$$

Here, $\theta_p = 0.985$ and τ_m is the lag corresponding to the maximum of $A_H(i, j, \tau)$ within [2 ms, 12.5 ms].

For any segment, if more than half of its units at a certain frame agree with the global pitch, this segment is said to agree with the global pitch at this frame. For the segments of target speech, if the global pitch at a certain frame is really close to the true target pitch, all of these segments tend to agree with the global pitch at this frame, and vice verse. Hence, segments are grouped into two streams as follows. First, the longest segment is selected as a seed stream. Since target speech in this study is a voiced utterance, the longest segment extends through most of the frames of the entire utterance duration (See Fig. 3(a)). For any other segment, if its agreement with the global pitch is the same as that of the longest segment for more than half of their overlapping frames, the segment is grouped into the same stream. Otherwise, the segment is grouped into the competing stream. The longest segment is also used to determine which stream corresponds to target speech. If it agrees with the global pitch for more than half of its frames, it is likely to contain dominant target speech. In this case, we refer to the stream containing the longest segment as the foreground stream, S_F^0, and the competing stream as the background stream, S_B^0. Otherwise, the names of the two streams are swapped.

Fig. 3(a) and Fig. 3(b) illustrate the segments and the foreground stream. The input is again the mixture of speech and cocktail party noise. Because the intrusion is not strongly structured, most segments correspond to target speech. In addition, most segments are in the low-frequency range where harmonics are resolved, and the high-frequency range contains only small segments. The initial foreground stream successfully groups most of the major segments in the low-frequency range.

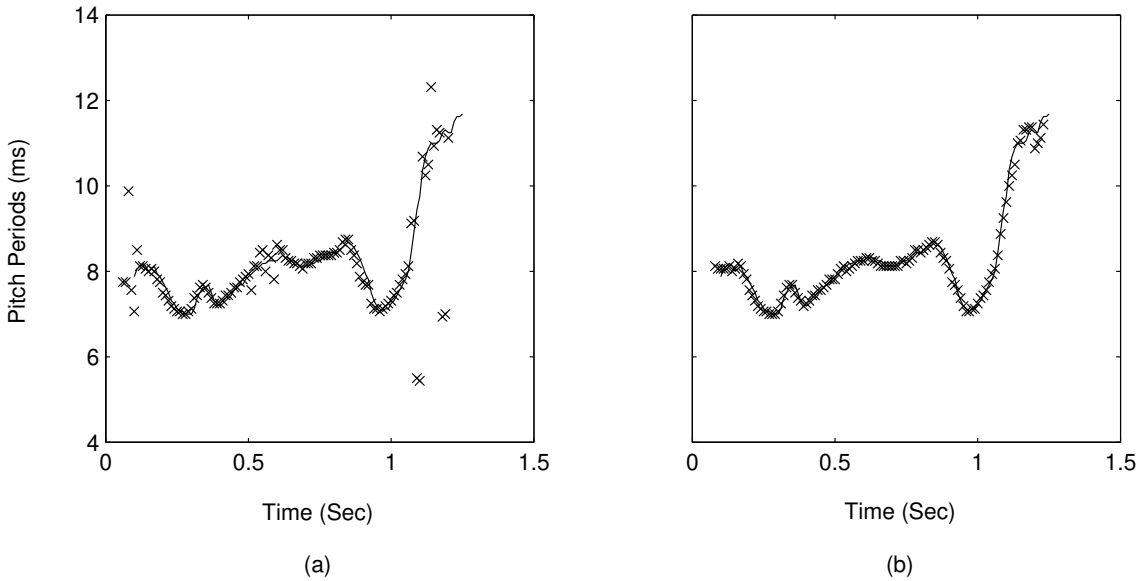

Figure 4. Results of pitch tracking for the speech and cocktail-party mixture. (a) Global pitch periods, marked by "x". (b) Estimated target pitch contour, marked by "x". The solid line indicates the pitch contour obtained from clean speech before mixing.

V. PITCH TRACKING AND UNIT LABELING

Although the global pitch estimated in the last section is generally close the true target pitch, they do not match well in many frames. As illustrated in Fig. 4(a) for the speech and cocktail-party mixture, the global pitch periods in a number of frames significantly deviate from the target pitch contour obtained from clean speech. To obtain more accurate pitch, we verify the target pitch obtained from S_F^0 with two psychoacoustically-motivated constraints and re-estimate subsequently the target pitch contour. With the new target pitch contour, units are labeled according to whether target speech dominates. Different criteria are used to deal with resolved and unresolved harmonics.

A. Pitch Tracking

First, our pitch tracking starts from the target pitch contour obtained from the summation of the autocorrelation functions of the T-F units in S_F^0. More specifically, let

$$s_S(j,\tau) = \sum_i A_H(i,j,\tau) \quad for \;\; u_{ij} \in S_F^0. \tag{8}$$

Then the pitch period of target speech at frame j, $\tau_S(j)$, is the lag corresponding to the maximum of $s_S(j,\tau)$ in the plausible range [2 ms, 12.5 ms].

Usually S_F^0 contains some intrusion, and $\tau_S(j)$ may not be accurate. Here we employ the following two constraints to check their reliabilities:

Constraint 1: An accurate target pitch is consistent with the periodicities of individual constituent units. Similar to (7), we check whether unit u_{ij} agrees with $\tau_S(j)$ by the following inequality

$$A_H(i,j,\tau_S(j)) / A_H(i,j,\tau_m) > \theta_p. \tag{9}$$

If an estimated pitch period is reliable, at least half of the units in the foreground stream at the corresponding frame must agree with it. Since the units in the foreground stream mainly correspond to target speech, $\tau_S(j)$ reflects target pitch better than global pitch. Note that the threshold in (9) is the same as that for initial grouping (see (7)). A smaller threshold could be used; however, in order to increase the reliability of the estimated target pitch, the same threshold is used.

Constraint 2: Pitch contour of speech changes slowly [17]. We stipulate that the difference between the reliable pitch periods at frame j and j+1 be less than 20% of $\tau_S(j)$. Constraint 2 applies to only consecutive pitch periods satisfying Constraint 1.

With these two constraints, our model re-estimates the target pitch contour as follows. First, $\tau_S(j)$'s are checked according to Constraint 1. Then the two streams are adjusted using only the reliable pitch periods satisfying the constraint: any segment that agrees with the reliable pitch periods for more than half of its length is grouped into a new foreground stream, S_F^1, and the others are grouped into a new background stream, S_B^1. The target pitch periods, $\tau_S(j)$, are estimated from S_F^1, and are further checked with both

constraints. Among all the intervals where all of the estimated pitch periods satisfy the constraints, the longest streak is selected. We found that the pitch periods within this streak provide the most reliable estimate of the target pitch. The pitch periods at the frames before this streak are re-estimated by the following algorithm, which operates only on the first 40 channels whose center frequencies range from 80 Hz to about 500 Hz, the plausible F0 range.

1. Let $j = j_0$, the first frame of the longest streak. Mark those channels in which T-F units of S_F^1 at frame j agree with $\tau_S(j)$.
2. Iterate until $j = 1$:
   2.1. $j = j - 1$.
   2.2. Check $\tau_S(j)$ with the two constraints.
   2.3. If $\tau_S(j)$ is reliable go to step 2.4. Otherwise, summate the autocorrelations of the units in frame j and the channels most recently marked. Replace $\tau_S(j)$ with the lag corresponding to the maximum of the summation in the range [2 *ms*, 12.5 *ms*]. Check new $\tau_S(j)$ with the two constraints. If $\tau_S(j)$ is reliable go to step 2.4; otherwise let $\tau_S(j) = \tau_S(j+1)$ and go to step 2.
   2.4. Mark those channels in which T-F units of S_F^1 at frame j agree with $\tau_S(j)$.

Step 2.3 in the above algorithm propagates, through a marking process, a reliable pitch estimate using the temporal continuity principle. The pitch periods after the longest streak are re-estimated in a similar way, starting from the frame immediately following the streak.

This algorithm results in interleaving streaks of reliable and unreliable pitch estimates. Finally, for any unreliable streak between the two reliable ones, the pitch periods of its frames are obtained by a simple linear interpolation from the last frame of the preceding streak and the first frame of the succeeding one. The remaining unreliable pitch periods are set to 0, indicating no pitch at these frames. Because the target speech is a voiced utterance, this procedure always produces a continuous pitch contour.

Fig. 4(b) shows the re-estimated pitch target contour from the speech and cocktail-party mixture, together with the pitch contour obtained from clean target. Except for a few frames, the resulting pitch contour matches that of the clear speech very well, significantly better than the global pitch contour shown in Fig. 4(a).

With re-estimated pitch, the two streams are subsequently adjusted in the same way; that is, any segment that agrees with new target pitch for more than half of its length is grouped into the foreground stream, now denoted as S_F^2, and the remaining ones are put to the background stream, S_B^2.

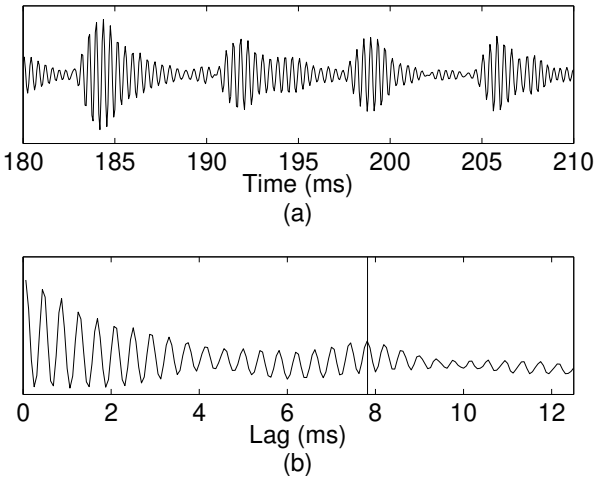

Figure 5. AM effects for the speech and cocktail-party mixture. (a) Response of a gammatone filter in the high-frequency range, with center frequency 2.6 *k*Hz. (b) Corresponding autocorrelation of the hair cell response. The vertical line marks the position corresponding to the pitch period of target speech.

B. Unit Labeling

The pitch contour computed above is used to label T-F units according to whether target speech dominates the unit responses or not. We label a unit by comparing its response periodicity with the estimated target pitch. Similar to (9), u_{ij} is labeled as target speech if

$$A_H(i, j, \tau_S(j)) / A_H(i, j, \tau_m) > \theta_t . \quad (10)$$

With more accurate pitch, the threshold θ_t can be smaller than θ_p; here θ_t is set to 0.85. The above criterion, referred as the periodicity criterion, works well for resolved harmonics, and is used to label the units belonging to the segments generated in the initial segmentation (See Sect. IV.A).

However, for units responding to multiple harmonics, their responses are amplitude-modulated. As a result, target pitch does not necessarily correspond to the global maximum of the autocorrelation of such a unit in the plausible pitch range. The periodicity criterion is thus not suitable for these units. This problem is illustrated in Figure 5. Fig. 5(a) shows over two consecutive frames a response of a filter with the center frequency of 2.6 *k*Hz, and the corresponding autocorrelation function for the second frame is given in Fig. 5(b). The filter response is strongly amplitude-modulated, and in this case, the target pitch corresponds to a local maximum, indicated by a vertical line in Fig.

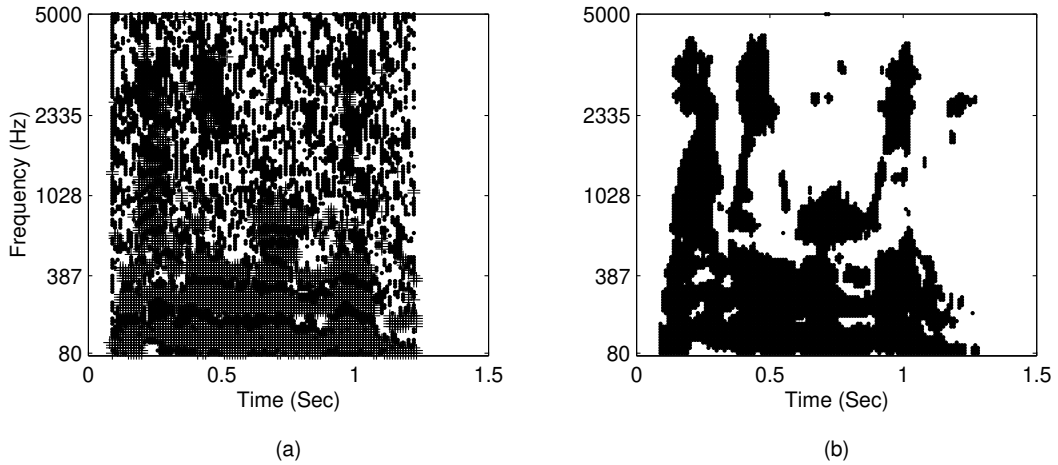

Figure 6. Results of unit labeling for the speech and cocktail-party mixture. (a) Units labeled as target speech. Symbol "+" marks units labeled by the periodicity criterion, and "•" marks units labeled by the AM criterion. (b) Units where target speech is stronger than intrusion.

5(b), in the autocorrelation instead of the global maximum.

To deal with this problem we propose a new criterion for labeling units corresponding to unresolved harmonics based on the following fact: for a filter responding to multiple harmonics of a single harmonic sound source the response envelope fluctuates at the rate of F0 of the source [11]. The criterion is referred to as the AM criterion. For each such unit, the AM criterion compares AM repetition rate with estimated pitch.

To obtain an AM repetition rate, the entire response of a gammatone filter is half-wave rectified and then band-passed to remove the DC component and other possible harmonics except for the F0 component. For every 10 frames, we use a filter with passband [0.9 $\bar{f}$, 1.2 $\bar{f}$] and stopbands [0, 0.5 $\bar{f}$] and [1.6 $\bar{f}$, +∞). Here, $\bar{f}$ is the average of estimated F0 over the 10 frames, which is the inverse of the average estimated pitch period in these frames. This filter uses a Kaiser window and the ripples of the passband and the stopbands are all 0.01. The rectified and filtered signal is then normalized by its envelope to remove the intensity fluctuations of the original mixture. That is,

$$\hat{r}(i,t) = r(i,t) / r_E(i,t), \tag{11}$$

where $r(i,t)$ is the rectified and filtered signal of channel i, and $r_E(i,t)$ is the envelope of $r(i,t)$, obtained via a Hilbert transform. Observe that generally the pitch of natural speech does not change noticeably within a single frame. We thus model the corresponding normalized signal within a T-F unit by a single sinusoid, in order to obtain the AM repetition rate. Specifically,

$$f_{ij}, \phi_{ij} = \arg\min_{f,\phi} \sum_{k=1}^{M} [\hat{r}(i, jT-k) - \sin(2\pi k f / f_S + \phi)]^2, \quad \text{for } f \in [80 \text{ Hz}, 500 \text{ Hz}], \tag{12}$$

where a square error measure is used. f_S is the sampling frequency, M spans a frame and $T = 10ms$ (see (2)). In the above equation, f_{ij} gives the AM repetition rate.

Sinusoidal modeling is a common technique for speech coding and synthesis, where the harmonic structure of voiced speech is generally modeled by a set of sinusoidals [21]. We use sinusoidal modeling for a different purpose - to derive an AM repetition rate - and our modeling is based on a single sinusoid. Note that in the discrete case, there are only limited signal samples within each frame, and a single sinusoid with a sufficiently high frequency can always match these samples perfectly. In our case, however, we are interested in modeling with one sinusoid with a frequency within the plausible pitch range. Hence, the solution does not reduce to a degenerate case.

We found that, with appropriately chosen initial values, the optimization problem in (12) can be solved effectively using an iterative gradient descent technique [24]:

$$\begin{cases} \Delta f = \eta \dfrac{M}{f_S} \displaystyle\sum_{k=1}^{M} k\{\hat{r}(i, jT-k) - \sin(2\pi k f / f_S + \phi)\} \\ \qquad\qquad \cdot \cos(2\pi k f / f_S + \phi) \\ \Delta \phi = \eta \displaystyle\sum_{k=1}^{M} \{\hat{r}(i, jT-k) - \sin(2\pi k f / f_S + \phi)\} \\ \qquad\qquad \cdot \cos(2\pi k f / f_S + \phi) \end{cases}, \tag{13}$$

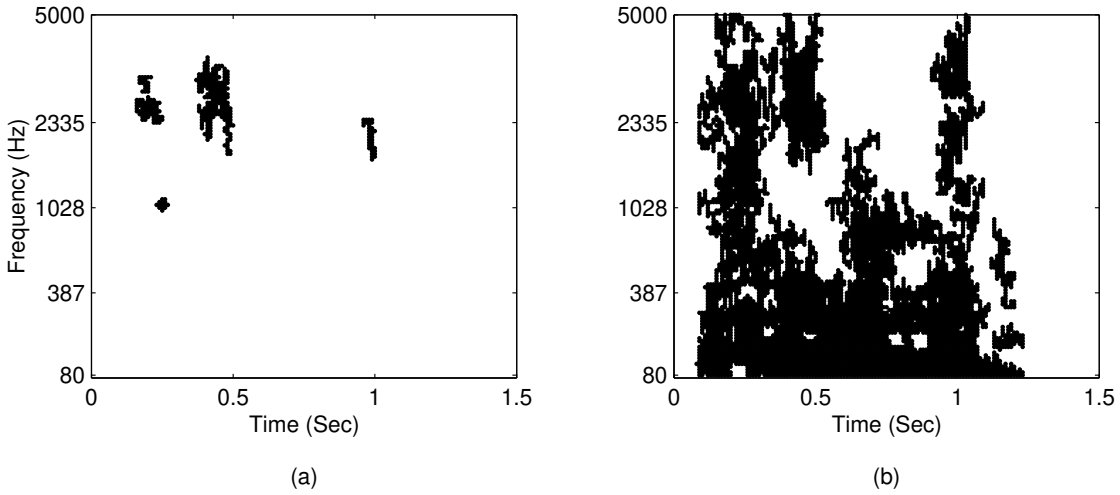

Figure 7. Results of final segregation for the speech and cocktail-party mixture. (a) New segments formed in the final segregation. (b) Final foreground stream.

The initial conditions are: $f = 1/\tau_S(j)$ and $\phi = 0$. The step size, η, is set to 0.005. The iterative procedure is stopped when the change in the square error is less than 0.0001.

The AM criterion is used to label T-F units that do not belong to any segments generated in the initial segregation; such segments, as discussed earlier, tend to miss unresolved harmonics. Specifically, unit u_{ij} is labeled as target speech if the sinusoid found by gradient descent fits the corresponding signal well. That is, the final square error is less than half of the total energy of the corresponding signal and the AM repetition rate is close to the estimated target pitch:

$$|f_{ij}\tau_S(j) - 1| < \theta_f . \quad (14)$$

Psychoacoustic evidence suggests that separating sounds with overlapping spectra requires 6-12% difference in F0 [6]. Accordingly we choose $\theta_f = 0.12$.

For the speech and cocktail-party mixture, Fig. 6(a) gives the units labeled as target speech according to either criterion. Fig. 6(b) shows all the units where target speech is stronger than intrusion. With the AM criterion, most units where target speech signal is stronger are correctly labeled. However, some units with stronger intrusion are also labeled as target speech, especially in the high-frequency range. Therefore, using unit labels alone tends to over-group the foreground stream. In the next stage, those units labeled as target speech according to the AM criterion will be first merged into segments.

VI. FINAL SEGREGATION

In this stage, segments corresponding to unresolved harmonics are generated based on temporal continuity and cross-channel envelope correlation. These segments are further grouped into the foreground stream. Then, both the foreground stream and the background stream are adjusted according to unit labels.

A. Further Segmentation

For adjacent channels responding to unresolved harmonics, although their responses may be quite different, they exhibit similar AM patterns and their response envelopes are highly correlated (see Fig. 2(b)). Therefore, further segmentation is based on cross-channel envelope correlation in addition to temporal continuity. First, T-F units are selected if they are labeled as target speech but do not belong to any segment generated in initial segmentation and their C_E's are greater than 0.985, the same threshold used in initial segmentation. Afterwards, selected neighboring units both in time and frequency are iteratively merged into segments. Finally, to reduce the influence of noise intrusion, segments shorter than 50 *ms* are removed. All the generated segments are added to S_F^2.

Fig. 7(a) shows the new segments generated in this process for the speech and cocktail-party mixture. Although many units in the high-frequency range are incorrectly labeled as target speech (see Fig. 6(a)), these segments contain only units where target speech is stronger.

B. Final Grouping

The spectra of target speech and intrusion often overlap and, as a result, some segments generated in initial segmentation (Sect. IV.A) contain both units where target dominates and units where intrusion dominates. Given unit labels generated in Sect. V.B, we further divide a segment in S_F^2 into smaller ones so that

all the units in a segment have the same label. Then the segments in S_F^2 are adjusted as follows:

1. Segments with the target label are retained in S_F^2 if they are no shorter than 50 *ms*.
2. Segments with the intrusion label are added to S_B^2 if they are no shorter than 50 *ms*.
3. The remaining segments are removed from S_F^2, becoming undecided.

Then S_B^2 expands iteratively to include undecided segments in its neighborhood. All the remaining undecided segments are added back to S_F^2.

The above adjustment applies to only the foreground stream. Segments in the background stream could be processed in a similar way, which may help to recover some target units. However, this procedure can also lead to over-grouping of the foreground since unit labeling is not error-free as discussed in Sect. V.B. Therefore, the segments in the background stream are not adjusted.

Finally, for individual units that do not belong to either stream, they are grouped into the foreground stream iteratively if they are labeled as target speech and in the neighborhood of the foreground stream. The result of this is the final segregated stream of target speech, denoted as S_F^3. The remaining units are added to the background stream, yielding S_B^3.

Fig. 7(b) illustrates S_F^3 segregated from the speech and cocktail-party mixture. Comparing with Fig. 6(b), this stream contains most of the units where target speech is stronger. In addition, only a small number of units where intrusion is stronger are incorrectly grouped into S_F^3.

VII. EVALUATION AND COMPARISON

Our model is evaluated with a corpus of 100 mixtures composed of 10 voiced utterances mixed with 10 intrusions collected by Cooke [4], which is commonly used to test CASA systems and hence facilitates our comparison. As mentioned in Sect. I, the intrusions have a considerable variety. Specifically, 10 intrusions are: N0 - 1 *k*Hz pure tone, N1 - white 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.

An unsolved issue in CASA research is how to quantitatively assess the performance of a CASA system [25]. Given our decomposition of a signal into T-F units, we here suggest the use of an ideal binary mask as the ground truth of target speech. The ideal binary mask is constructed as follows: a unit in the mask is assigned 1 if the target energy in the corresponding T-F unit is greater than the intrusion energy and 0 otherwise. With the availability of target and intrusion before mixing, as is the case for our evaluation corpus, ideal binary masks can be readily constructed. Theoretically speaking, an ideal binary mask gives a performance ceiling for all binary masks. Figure 6(b) illustrates the ideal mask for the speech and cocktail-party mixture. Ideal masks also suit well the situations where more than one target need to be segregated or the target changes dynamically. The use of ideal masks is supported by the auditory masking phenomenon: within a critical band a weaker signal is masked by a stronger one [20]. In addition, an ideal mask is similar to a priori mask used in a recent ASR study [5], which yields excellent recognition performance.

The speech waveform resynthesized from S_F^3, which is treated as a binary mask, is used for evaluation, and it is denoted by $S(t)$ The speech waveform resynthesized from the ideal binary mask is denoted by $I(t)$. Furthermore, let $e_1(t)$ denote the signal present in $I(t)$ but missing from S(t), and $e_2(t)$ the signal present in $S(t)$ but missing from $I(t)$. Then, the relative energy loss, R_{EL}, and the relative noise residue R_{NR} are calculated as follows:

$$R_{EL} = \sum_t e_1^2(t) \Big/ \sum_t I^2(t), \tag{15a}$$

$$R_{NR} = \sum_t e_2^2(t) \Big/ \sum_t S^2(t). \tag{15b}$$

The results from our model are shown in Table 1. Each value in the table represents the average result of one intrusion with 10 voiced utterances. A further average across all intrusions is also shown in the table. On average, our system retains 96.60% of target speech energy, and the relative residual noise is kept at 3.32%. The relative noise residue for the original mixtures is 36.05%, as shown in the table; energy loss is obviously zero for the original mixtures. As indicated by the table, our model achieves very good performance across the corpus. In particular, the errors measured R_{EL} and R_{NR} are balanced in our systems.

As discussed earlier, the performance of the Wang-Brown model on monaural speech segregation is representative of current CASA systems. As a comparison, Table 1 also shows the results from the Wang-Brown system [28]. Our model produces an improvement in both R_{EL} and R_{NR} measures. In particular, our system has cut R_{EL} from 11.91% in their system to 3.40% in our system; this is especially noticeable for N1 and N3. Both models achieve a low level of R_{NR}, except for N9 where our result is much better. This reflects the conservative strategies employed by the models in dealing with intrusions: both tend to exclude uncertain units from the foreground stream.

TABLE 1: R_{EL} AND R_{NR}

Intrusion	Proposed model		Wang-Brown model		Mixture
	R_{EL} (%)	R_{NR} (%)	R_{EL} (%)	R_{NR} (%)	R_{NR} (%)
N0	2.12	0.02	6.99	0	67.76
N1	4.66	3.55	28.96	1.61	57.16
N2	1.38	1.30	5.77	0.71	5.04
N3	3.83	2.72	21.92	1.92	18.15
N4	4.00	2.27	10.22	1.41	27.17
N5	2.83	0.10	7.47	0	78.84
N6	1.61	0.30	5.99	0.48	39.24
N7	3.21	2.18	8.61	4.23	16.68
N8	1.82	1.48	7.27	0.48	7.37
N9	8.57	19.33	15.81	33.03	43.09
Average	3.40	3.32	11.91	4.39	36.05

Since our model mainly focuses on unresolved harmonics in the high-frequency range, it is instructive to present the performance of our model in this frequency range separately. For this purpose, we calculate R_{EL} and R_{NR} for only the filter channels with center frequencies greater than 1 *k*Hz. Note that since the F0 range of target speech is from 80 Hz to 160 Hz, target harmonics in the frequency region above 1 *k*Hz are generally unresolved. The corresponding results are shown in Table 2. It is clear from the average R_{NR} of the original mixtures that intrusions are much stronger in the high-frequency range. The particularly bad results from the Wang-Brown system quantitatively illustrate our point that CASA systems handle high-frequency signals poorly (see Sect. I). Our model, as a result of handling resolved and unresolved harmonics with different ASA cues, achieves large improvements in both R_{EL} and R_{NR}.

TABLE 2: R_{EL} AND R_{NR} IN THE HIGH-FREQUENCY RANGE

Intrusion	Proposed model		Wang-Brown model		Mixture
	R_{EL} (%)	R_{NR} (%)	R_{EL} (%)	R_{NR} (%)	R_{NR} (%)
N0	9.07	0.36	64.54	0	96.82
N1	41.60	44.05	83.66	66.08	96.00
N2	6.90	20.03	50.80	32.75	44.02
N3	21.04	9.46	74.99	15.00	42.57
N4	31.28	39.35	78.22	60.45	81.31
N5	15.59	2.69	63.26	0	97.90
N6	8.09	6.29	48.21	14.50	91.26
N7	15.88	12.40	64.25	24.78	43.49
N8	10.00	7.15	49.52	2.38	31.07
N9	38.58	7.21	69.17	25.88	27.72
Average	19.80	14.90	64.65	25.17	65.24

Finally, to compare waveforms directly we measure a form of signal-to-noise ratio (SNR) in decibels using the resynthesized signal from the ideal binary mask as ground truth:

$$SNR = 10\log_{10}[\sum_t I^2(t) \Big/ \sum_t (I(t) - S(t))^2]. \quad (16)$$

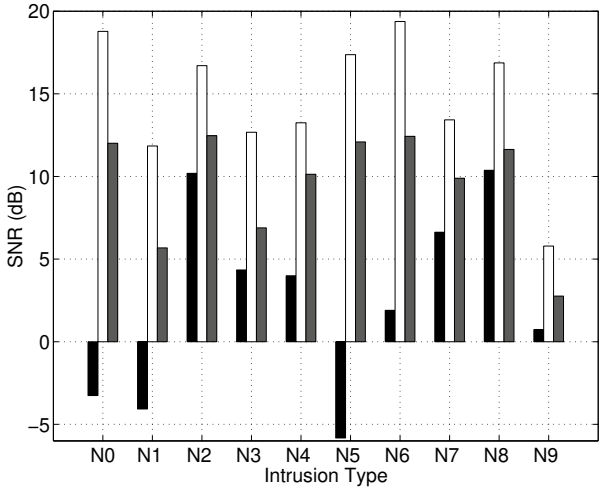

Figure 8. SNR results for segregated speech. White bars show the results from the proposed model, gray bars those from the Wang-Brown system, and black bars those of original mixtures.

The SNR for each intrusion averaged across 10 target utterances is shown in Fig. 8, together with the results from the Wang-Brown system and the SNR of the original mixtures without processing. Both of the systems show improvements over mixtures. Compared with the Wang-Brown model, our system yields at least a 3 dB SNR improvement for every intrusion type. The average improvement for the entire corpus is about 5 dB.

VIII. DISCUSSION

Our model computes AM repetition rates by sinusoidal modeling and subsequent gradient descent. In a previous study, we obtained AM repetition rates by extracting the instantaneous frequencies of filter responses [12]. The results there were similar to Table 1 and Figure 8. One advantage of the current method is that it is much more efficient computationally.

Currently, our model decomposes input signal with a 128-channel gammatone filterbank. We have also tried a 256-channel filterbank, but seen no significant improvement in performance. In addition, our model uses a 4-unit neighborhood for segment formation and grouping. We have also tried an 8-unit neighborhood, but results were similar.

Consistent with [28], our study suggests that segments formed on the basis of cross-channel correlation and cross-channel envelope correlation

provide an intermediate structure for speech segregation. As shown in Figure 6(a), individual unit labels are not well organized, and if taken directly they would lead to a target stream that contains too much intrusion. Performing segmentation before final grouping takes advantage of local correlation between filter responses and the temporal continuity principle. As a result, the system is more robust to interference.

An important feature of our system lies in using different mechanisms to deal with resolved and unresolved harmonics. Unresolved harmonics are grouped using the AM criterion. The AM of filter responses has been explored by Cooke's model [4]. Our model differs from his both in the way AM is computed and the use of AM for speech segregation. Specifically, Cooke's model first generates local elements based on filter response frequencies and temporal continuity. Segments are merged into groups based on common harmonicity and AM. Then a pitch contour is obtained from each group and groups with similar pitch contours are put into the same stream. In contrast, our model obtains a target pitch contour from resolved harmonics first, which is consistent with the psychoacoustic evidence that pitch is determined primarily by resolved harmonics [20]. The target pitch contour is then used to identify the unresolved harmonics of target speech according to AM repetition rates, which is derived by sinusoidal modeling within a limited range guided by the pitch contour. Unresolved harmonics then form new segments and these segments are grouped into the foreground stream. As a result, our model is able to recover much more target speech and reduce noise interference in the high-frequency range where harmonics of target speech are generally unresolved.

The performance of the proposed model depends on the accuracy of estimated target pitch. For most mixtures in the evaluation corpus, estimated pitch contours match those obtained from clean target speech well. However, our pitch tracking method is designed for voiced targets and its utility as a general pitch determination algorithm is uncertain. Recently, Wu et al. [29] proposed a robust algorithm for multipitch tracking for speech utterances in the presence of other interfering sources. Such general pitch determination algorithms may be incorporated into our system to broaden its scope for monaural speech segregation.

The proposed system considers the pitch contour of the target source only. However, it is possible to track the pitch contour of the intrusion if it has a harmonic structure. With two pitch contours, one could label T-F units more accurately by comparing whether their periodicities are more consistent with one or the other. Such a method is expected to lead to better performance for the two-speaker situation, e.g. N7 through N9. As indicated in Fig. 8, the performance gain of our system for such intrusions is relatively limited.

Our model performs grouping based only on pitch. As a result, it is limited to segregation of only voiced speech. In our view, unvoiced speech poses the biggest challenge for monaural speech segregation. Other grouping cues, such as onset, offset, and timbre, have been demonstrated to be effective for human ASA [1], and may play a role in grouping unvoiced speech. Also, it appears that one must consider the acoustic and phonetic characteristics of individual unvoiced consonants. We plan to investigate these issues in future study.

Acknowledgement

We wish to thank G. J. Brown and M. Wu for helpful comments.

Reference

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

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

[3] R. P. Carlyon and T. M. Shackleton, "Comparing the Fundamental Frequencies of Resolved and Unresolved Harmonics: Evidence for Two Pitch Mechanisms?" *J. Acoust. Soc. Am.*, Vol. 95, pp. 3541-3554, 1994.

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

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

[6] C. J. Darwin and R. P. Carlyon, "Auditory Grouping," in *Hearing,* B. C. J. Moore Ed. San Diego, CA: Academic Press, 1995.

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

[8] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 3, pp. 251-266, 1995.

[9] Y. Ephraim, D. Malah, and B. H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 37, pp. 1846-1856, 1989.

[10] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to

Speech Recognition," *IEEE. Trans. Signal Processing,* Vol. 39, pp.795-805, 1991.

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

[12] G. Hu and D. L. Wang, "Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation", *ICASSP* 2002, in press.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: Wiley, 2001

[14] J. Jensen and J. H. L. Hansen, "Speech Enhancement Using a Constrained Iterative Sinusoidal Model," *IEEE Trans. Speech and Audio Processing,* Vol. 9, pp. 731-740, 2001.

[15] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of Broad-Band Noise in Speech by Truncated QSVD," *IEEE Trans. Speech and Audio Processing*, Vol. 3, pp. 439-448, 1995.

[16] H. Krim and M. Viberg, "Two Decades of Array Signal Processing Research: the Parametric Approach," *IEEE Signal Processing Magazine,* Vol. 13, pp. 67-94, 1996.

[17] W. J. M. Levelt, *Speaking: From Intention to Articulation*, Cambridge, MA: MIT Press, 1989.

[18] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, Vol. 9, pp. 504-512, 2001.

[19] R. Meddis, "Simulation of Auditory-Neural Transduction: Further Studies", *J. Acoust. Soc. Am.* Vol. 83, pp. 1056-1063, 1988.

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

[21] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd Ed. NY: IEEE Press, 2000.

[22] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, *APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function*, Cambridge: Applied Psychology Unit. 1988.

[23] R. Plomp and A. M. Mimpen, "The Ear as a Frequency Analyzer II", *J. Acoust. Soc. Am*., Vol. 43, pp. 764-767, 1968.

[24] J. C. Principe, N. R. Euliano, and W. C. Lefebvre, *Neural and Adaptive Systems: Fundamentals through Simulations*, NY: John Wiley, 2000.

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

[26] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise," *IEEE Trans. Speech and Audio Processing*, Vol. 6, pp. 445-455, 1998.

[27] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Trans. Speech and Audio Processing*, Vol. 7, pp. 126-137, 1999.

[28] 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.

[29] M. Wu, D. L. Wang and G. J. Brown, "A Multi-pitch Tracking Algorithm for Noisy Speech," *ICASSP* 2002, in press.