Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation

Guoning Hu and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Segregating speech from one monaural recording has proven to be very challenging. Monaural segregation of voiced speech has been studied in previous systems that incorporate auditory scene analysis principles. A major problem for these systems is their inability to deal with the high-frequency part of speech. Psychoacoustic evidence suggests that different perceptual mechanisms are involved in handling resolved and unresolved harmonics. We propose a novel system for voiced speech segregation that segregates 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 rates. Underlying the segregation process is a pitch contour that is first estimated from speech segregated according to dominant pitch and then adjusted according to psychoacoustic constraints. Our system is systematically evaluated and compared with pervious systems, and it yields substantially better performance, especially for the high-frequency part of speech.**

***Index Terms*—Amplitude modulation (AM), computational auditory scene analysis, grouping, monaural speech segregation, pitch tracking, segmentation.**

I. Introduction

In a natural environment, speech often 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 [1] or sensor arrays for spatial filtering [21], require multiple sensors. However, many applications such as telecommunication and audio retrieval need a monaural solution. For a monaural (one microphone) signal, 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 [23] or to extract speech based on speech modeling [20]. Another way to deal with interference is to perform eigen-decomposition on an acoustic mixture and then apply subspace analysis to remove part of the interference [13]. Hidden Markov models have also been used to model the spectral characteristics of speech and interference and then separate them [30]. 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 human auditory system shows a remarkable capacity for monaural speech segregation. According to Bregman [4], 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 [5], [8], [12], [29], [33], [34]. Such systems generally approach speech segregation without making strong assumptions about the acoustic properties of interference, and follow two main stages: segmentation (analysis) and grouping (synthesis) [4]. In segmentation, the acoustic input is decomposed into sensory segments, each of which should originate from a single source. In grouping, those segments that likely come from the same source are grouped together.

Recently, Wang and Brown proposed a CASA model to segregate voiced speech based on oscillatory correlation, where a stream is represented by an assembly of synchronized oscillators and different streams are represented by desynchronized oscillator assemblies [33]. The Wang–Brown model uses harmonicity and temporal continuity as major grouping cues and consists of three stages: auditory periphery, mid-level representation, and segmentation and grouping. The auditory periphery performs spectral analysis through an auditory filterbank. 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 dominant 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 [8]. They have reported good results in comparison with previous CASA systems. In most situations, the model is able to remove intrusions and recover the low-frequency (below 1 kHz) energy of target speech, i.e., voiced utterances. However, this model cannot handle the high-frequency (above 1 kHz) part of target speech well, and it loses much of it. In fact, the inability to deal with the high-frequency part of speech is a common problem for CASA systems.

Manuscript received August 2003; revised December 2003. This research was supported in part by the National Science Foundation (NSF) under Grant IIS-0081058 and in part by the Air Force Office of Scientific Research (AFOSR) under Grant F49620-01-1-0027.

G. Hu is with the Biophysics Program, The Ohio State University, Columbus, OH 43210 USA (e-mail: hu.117@osu.edu).

D. Wang is with the Department of Computer and Information Science and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cis.ohio-state.edu).

Digital Object Identifier 10.1109/TNN.2004.832812

1045-9227/04$20.00 © 2004 IEEE

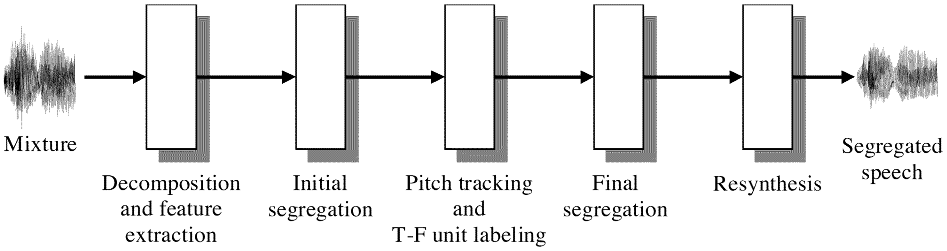

Mixture
Decomposition and feature extraction
Initial segregation
Pitch tracking and T-F unit labeling
Final segregation
Resynthesis
Segregated speech

Fig. 1. Schematic diagram of the proposed multistage system.

The design of an effective segregation algorithm depends on an initial analysis of input signal. Most CASA systems, including the Wang–Brown model, perform initial signal decomposition in the frequency domain with a bank of auditory filters, whose bandwidths increase quasi-logarithmically with center frequencies. These filters are usually derived from psychophysical observations of the auditory periphery. The main motivation for using an auditory filterbank is to mimic cochlear filtering. There are also studies reporting that auditory-based front-ends are more robust than traditional Fourier-based analysis in the presence of background interference [14], [19]. For harmonic signal, in the low-frequency range, an auditory filter has a narrow passband, which generally contains only one harmonic. In the high-frequency range, an auditory filter has a wide passband, which usually contain multiple harmonics. This structure of cochlea filtering limits the abilities of human listeners to resolve harmonics; they can resolve only the first few harmonics, whereas higher harmonics are unresolved unless these harmonics are much more intense than adjacent ones [7], [28]. Consequently, with a model of auditory filterbank, a harmonic series is divided into resolved and unresolved harmonics. A harmonic is called resolved if there exists an auditory filter channel that responds primarily to it; otherwise, it is called unresolved. For voiced speech, harmonics less than 1 kHz are generally resolved while others may be unresolved. An important fact is that the filter responses to unresolved harmonics are strongly amplitude-modulated and the response envelopes fluctuate at the fundamental frequency (F0) of target speech [15]. Hence, it may be computationally more appropriate to segregate resolved and unresolved harmonics using different methods. In addition, psychophysical evidence suggests that the human auditory system uses different mechanisms to deal with resolved and unresolved harmonics [7], and resolved and unresolved harmonics contribute differently to speech segregation for human listeners [2]. However, previous CASA systems use the same method to segregate resolved and unresolved harmonics [5], [8], [33], [34], and we think this is a main reason why these systems cannot segregate voiced speech well in the high-frequency range.

Based on the above observations, we propose a new model for monaural segregation of voiced speech. Similar to previous CASA systems, our model performs initial spectral analysis using an auditory filterbank, and forms segments on the basis of similarity between adjacent filter responses (cross-channel correlation) and temporal continuity. Different from previous systems, 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. Since the amplitude modulation (AM) of filter responses reveals the periodicities of unresolved harmonics, we generate segments for unresolved harmonics based on common AM in addition to temporal continuity. These segments are further grouped based on AM rates, which are obtained from the temporal fluctuations of the corresponding response envelopes. In addition, by building on an initial segregation stage and observing pitch characteristics in natural speech, we propose a method to obtain an accurate pitch contour for target speech.

With signal decomposition using a filterbank and over successive time frames, the computational goal of our model is to retain time–frequency regions where target speech is more intense than interference and cancel those regions where interference is more intense. In other words, the goal is to identify a binary mask where 1 indicates that target is stronger than interference and 0 otherwise. In a sense, this objective amounts to maximizing the signal-to-noise ratio (SNR) of the output where binary decisions are made for local time–frequency regions. To maximize the SNR, we want to minimize the denominator—the sum of missing target energy that is discarded by the mask and interference energy that gets through the mask along with the target—by treating the numerator, the clean target, as a constant. This would imply to include local time–frequency regions where target is stronger than interference and exclude the other regions. Further justifications of our computational goal are given in Section VII.

Section II gives an overview of our model and further motivation, and Sections III–VI explain model components in detail. In Section VII, our system is systematically evaluated and compared with other systems for speech segregation or enhancement. Further discussion is given in Section VIII. The Appendix lists all the symbols used in this paper and their definitions.

II. Model Overview

The main idea of our model is to employ different segregation methods for resolved and unresolved harmonics. The overall model is a multistage system, as shown in Fig. 1.

In the first stage, an input mixture is analyzed by an auditory filterbank in consecutive time frames. The sampling frequency is 16 kHz. This processing results in a decomposition of the input into a two-dimensional time–frequency map. Each unit of the map is called a T-F unit, corresponding to a certain filter at a certain time frame. Then the following features are extracted: autocorrelation of a filter response, autocorrelation of the envelope of a filter response, cross-channel correlation, and dominant pitch within each time frame. These features are used in the following stages.

In the initial segregation stage, T-F units are merged into segments. Segmentation has been performed in previous CASA systems [5], [8], [33]. A segment is a larger component of an auditory scene than a T-F unit, and it is meant to capture a perceptually-relevant acoustic component of a single source. An auditory segment is composed of a spatially contiguous region of T-F units. This segment structure encodes the basic proximity principle in human ASA that applies to both frequency and time dimensions [4]. Conceptually, segments provide an intermediate level representation between T-F units and streams (or sound sources). Computationally, performing segmentation before segregation exploits local correlation between adjacent filter responses and, hence, it is reasonable to expect that segments are less sensitive to interference than T-F units. This robustness is confirmed by the evaluation in Section VII. Segments are then grouped into an initial foreground stream and a background stream based on dominant pitch extracted in the previous stage; the two streams roughly correspond to target speech and intrusion, respectively. Due to the intrusion, the dominant pitch may not be an accurate description of the target pitch. For example, if the intrusion has a strong pitch at a certain frame, the obtained dominant pitch at this frame may be close to the intrusion pitch instead of the target pitch. As a result, the foreground stream will usually miss some target speech and include some intrusion.

In the third stage, the pitch of target speech is estimated from the initial foreground stream, and it is used to label units as speech dominant or interference dominant. In the final segregation stage, according to unit labels, segments formed in the initial segregation stage are regrouped into foreground and background stream. This stage corrects some errors of initial grouping due to the inaccuracy of the dominant pitch. In addition, some T-F units are merged into segments that correspond to unresolved harmonics of target speech, and these segments are added to the foreground stream. Then the foreground stream expands to include neighboring T-F units labeled as speech dominant. Finally, a speech waveform is resynthesized from the resulting foreground stream using a method described by Weintraub [34]. Here, the foreground stream works as a binary mask, where 1 indicates T-F units within which target speech dominates and 0, otherwise. The mask is used to retain the acoustic energy from the mixture that corresponds to 1's in the mask and reject the mixture energy corresponding to 0; for more details of this stage, see [5], [33], [34]. The first four stages in Fig. 1 are explained in detail in the following sections.

For ease of comparison, let us reiterate the terms that have been introduced so far. A T-F unit is a very local time–frequency region corresponding to a certain filter at a certain time frame. We use u_{cm} to refer to the T-F unit corresponding to filter channel c at time frame m. A segment is a contiguous time–frequency region that corresponds to a component of a single sound source, and it is a set of connected T-F units. A stream is a group of segments that corresponds to an entire sound source. The target speech, or the target stream, is an utterance we aim to segregate from an acoustic mixture. What constitutes target speech is obviously task-dependent. In this study, target speech refers to an entirely voiced utterance in a sound mixture.

III. Decomposition and Feature Extraction

In this stage, our system decomposes the input signal and extracts the following features: the autocorrelation of filter responses, the autocorrelation of response envelopes, cross-channel correlation, and dominant pitch within each time frame.

A. Signal Decomposition

First, the input signal passes through an auditory filterbank. Here we use a 128-channel "gammatone" filterbank [27] whose center frequencies are quasi-logarithmically spaced from 80 to 5000 Hz. The gammatone filterbank is a standard model of cochlear filtering [3]. The impulse response of a gammatone filter is

$$g(t) = \begin{cases} t^{l-1}\exp(-2\pi bt)\cos(2\pi ft), & t \geq 0 \\ 0, & \text{else} \end{cases} \tag{1}$$

where $l = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, and f is the center frequency of the filter.

The response of each gammatone filter is further transduced by the Meddis model of inner hair cells [24]. This model simulates 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 low-pass filtering, which contains AM information to be utilized later to segregate unresolved harmonics. The cutoff frequency of the low-pass filter needs to be below the frequency range where harmonics are unresolved but above the plausible F0 range of target speech. Here we use an FIR filter with passband [0, 1 kHz] and a Kaiser window of 18.25 ms, but results do not change significantly for cutoff frequencies ranging from 800 Hz to 1.2 kHz.

In each filter channel, the output is divided into 20-ms time frames with 10-ms overlapping between consecutive frames. This frame size is commonly used for speech analysis and is sufficiently long for pitch estimation within each frame since pitch periods of speech are shorter than 20 ms. As a result of bandpass filtering and short-time windowing, the input is decomposed into a two-dimensional time–frequency representation, or a collection of T-F units. Two units are neighbors when they are in the same channel but at consecutive frames, or when they are in adjacent channels but at the same frame.

B. Auditory Feature Extraction

1) Correlogram: A well established mechanism for pitch extraction employs a correlogram—a running autocorrelation of each filter response across an auditory filterbank [18], [31]. Previous models have shown that correlograms provide an effective mid-level auditory representation between auditory periphery and segregation [5], [33]. For T-F unit u_{cm}, its autocorrelation function of the hair cell response is given by

$$A_H(c,m,\tau) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h(c,mT-n)h(c,mT-n-\tau). \tag{2}$$

Here, delay $\tau \in [0,\ 12.5\,\text{ms}]$. 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. $h(c,n)$ indicates the hair cell output of channel c at time step n. $T = 160$ corresponds to 10 ms, the time shift from one frame to the next. Let f be the center frequency of channel c. The window size of autocorrelation is chosen to be the longer of $4/f$ and 20 ms, the latter being the width of a frame, and N_c is the corresponding number of samples.

2) Dominant Pitch: 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 summary correlogram. Let $s(m,\tau)$ be the summary correlogram at frame m

$$s(m,\tau) = \sum_c A_H(c,m,\tau). \tag{3}$$

We define the dominant pitch period at frame m, $\tau_D(m)$, to be the lag corresponding to the maximum of $s(m,\tau)$ in the plausible pitch range of target speech [2 ms, 12.5 ms], or from 80 Hz to 500 Hz. For those channels where target speech dominates, their autocorrelations have peaks consistent with the pitch of target speech and the summation of these autocorrelations generally shows a dominant peak corresponding to the pitch period. With acoustic interference, a dominant pitch is a good description of a target pitch at those frames where target speech is stronger, but not for other frames.

3) Envelope Correlogram: We introduce an envelope correlogram by computing the autocorrelation of a response envelope

$$A_E(c,m,\tau) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h_E(c,mT-n)h_E(c,mT-n-\tau). \tag{4}$$

Here, $h_E(c,n)$ is the envelope of the hair cell output in channel c at time step n. The autocorrelation functions reveal response periodicities as well as AM rates.

4) Cross-Channel Correlation: As demonstrated by Wang and Brown [33], cross correlation between adjacent filter channels indicates whether the filters mainly respond to the same source or not, thus, providing a useful feature for subsequent segmentation. For the same T-F unit, the cross-channel correlation is calculated as

$$C_H(c,m) = \sum_{\tau=0}^{L-1} \hat{A}_H(c,m,\tau)\hat{A}_H(c+1,m,\tau) \tag{5}$$

where $\hat{A}_H(c,m,\tau)$ denotes $A_H(c,m,\tau)$ normalized to zero mean and unity variance, and $L = 201$ corresponds to 12.5 ms, the maximum delay for A_H. Similarly, the cross-channel correlation of envelopes is calculated as follows:

$$C_E(c,m) = \sum_{\tau=0}^{L-1} \hat{A}_E(c,m,\tau)\hat{A}_E(c+1,m,\tau) \tag{6}$$

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

In the low-frequency range, C_H gives a good indication of whether two adjacent frequency channels respond primarily to the same source. In the high-frequency range, a channel responds to multiple unresolved harmonics, leading to AM in the channel response (see related discussion in Section I). Now consider two adjacent channels in the high-frequency range. Because individual harmonics contribute differently to each of the two responses, the corresponding C_H is in general lower than those in the low-frequency range. Since C_E measures the similarity of the corresponding AM patterns, we find that it yields a range of values that is comparable with that of C_H in the low-frequency range. In other words, C_E is a better indicator of whether two channels respond to the same periodic source in the high-frequency range.

Fig. 2(a) and (b) illustrates the correlogram and the envelope correlogram at a specific frame, in response to a voiced utterance,"Why were you all weary," mixed with a "cocktail party" noise, which is a recording in a bar. The mixture has an SNR of about 6.5 dB. Their respective cross-channel correlations are given in the right panels. The bottom panel in Fig. 2(a) shows the corresponding summary correlogram, where the dominant peak at 7.2 ms gives the dominant pitch period at this frame. Fig. 2(c) and (d) shows the corresponding correlogram, cross-channel correlation, and summary correlogram for the voiced utterance alone. As shown in the figure, the autocorrelation of a hair cell response generally reflects the periodicity of a single harmonic for a channel in a low-frequency range, where harmonics are resolved. The autocorrelation is strongly amplitude-modulated in some high-frequency channels, where harmonics are unresolved. These autocorrelations are not highly correlated for frequency channels whose center frequencies are above 1 kHz. On the other hand, the autocorrelations of the response envelopes are highly correlated when they have similar fluctuation patterns.

IV. Initial Segregation

In this stage, units are merged into segments based on temporal continuity and cross-channel correlation. Using dominant pitch, these segments are grouped into an initial foreground stream and an initial background stream, roughly corresponding to target speech and intrusion, respectively. This stage is similar to the segmentation and grouping stage of the Wang–Brown system [33].

A. Initial Segmentation

A voiced section usually lasts for more than 50 ms. In addition, because the passbands of adjacent channels have significant 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 correlations are considered. More specifically, u_{cm} is selected for consideration if $A_H(c,m,0) > \theta_H^2$ and $C_H(c,m) > \theta_C$. Note that the autocorrelation at zero-lag measures the response energy. $\theta_H = 50$, which is close to the spontaneous firing rate of the auditory nerve from the Meddis

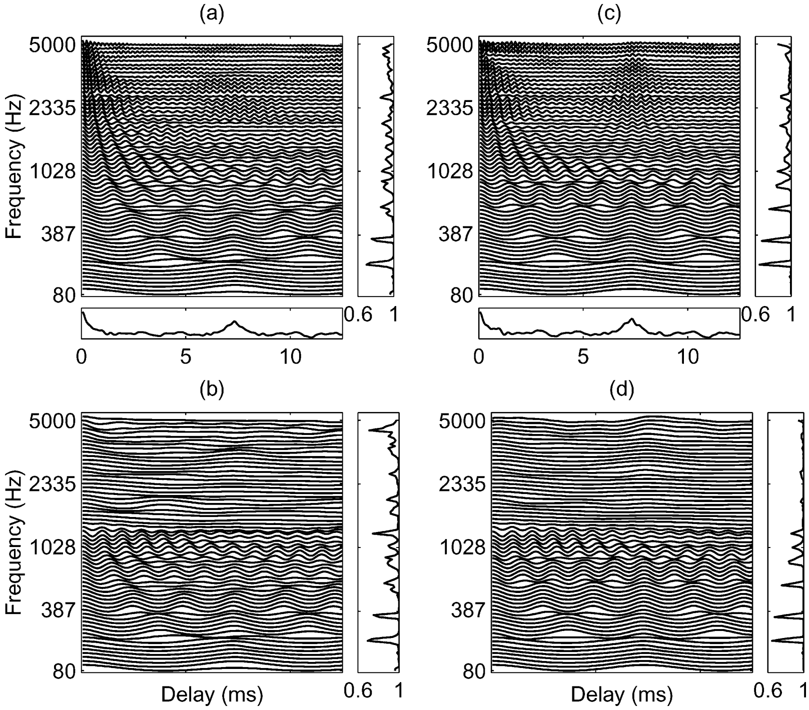


Fig. 2. Auditory features. (a) Correlogram at frame 40 (i.e., 0.4 s after the onset) for a mixture of speech and cocktail-party noise. For clarity, every other channel is shown. The corresponding cross-channel correlation is given in the right panel, and the summary correlogram in the bottom panel. (b) Corresponding envelope correlogram for the mixture. The corresponding cross-channel envelope correlation is shown in the right panel. (c) Correlogram at frame 40 for the clean speech. The corresponding cross-channel correlation is given in the right panel, and the summary correlogram in the bottom panel. (d) Corresponding envelope correlogram for the clean speech. The corresponding cross-channel envelope correlation is shown in the right panel.

model [24], and $\theta_C = 0.985$, chosen to be the same as in [33]. 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 formed this way generally reflect resolved components of either target speech or intrusion.

B. *Initial Grouping*

Initial grouping is done through comparing the periodicities of unit responses with dominant pitch. The response period of a T-F unit is obtained by finding the maximum of the corresponding autocorrelation within the plausible pitch range. If this period is compatible with the corresponding dominant pitch period, the unit is said to agree with the dominant pitch. That is, u_{cm} agrees with $\tau_D(m)$ if $A_H(c, m, \tau_D(m))$ is close to the maximum of $A_H(c, m, \tau)$ within the plausible pitch range

$$\frac{A_H\left(c, m, \tau_D(m)\right)}{A_H\left(c, m, \tau_P(c, m)\right)} > \theta_P. \tag{7}$$

Here, $\theta_P = 0.95$, chosen to be the same as in [33], and $\tau_P(c, m)$ is the delay corresponding to the maximum of $A_H(c, m, \tau)$ within the plausible pitch range [2 ms, 12.5 ms].

For any segment, if more than half of its units at a certain frame agree with the dominant pitch, this segment is said to agree with the dominant pitch at this frame. For the segments of target speech, if the dominant pitch at a certain frame is very close to the true pitch of target speech, all of these segments tend to agree with the dominant pitch at this frame. 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 all voiced, the longest segment extends through most of the frames of the entire utterance duration [see Fig. 3(b)]. At a certain frame, a segment is said to agree with the longest segment if both segments agree or both disagree with the dominant pitch. If a segment agrees with the longest segment for more than half of their overlapping frames, its T-F units within the duration of the longest segment is grouped into the seed stream. Otherwise, this 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 dominant 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.

For the speech and cocktail-party mixture, Fig. 3(a) shows the hair cell response energy (zero-lag autocorrelation) of each T-F unit. Fig. 3(b) and (c) shows the segments and the foreground stream. Because the intrusion is not strongly harmonic, most segments correspond to target speech. In addition, most segments are in the low-frequency range where harmonics are resolved, and the high-frequency part contains only small segments. The initial foreground stream successfully groups most of the major segments in the low-frequency range.

V. Pitch Tracking and Unit Labeling

Although the dominant pitch is generally close to the true pitch of target speech, they do not match well in many frames.

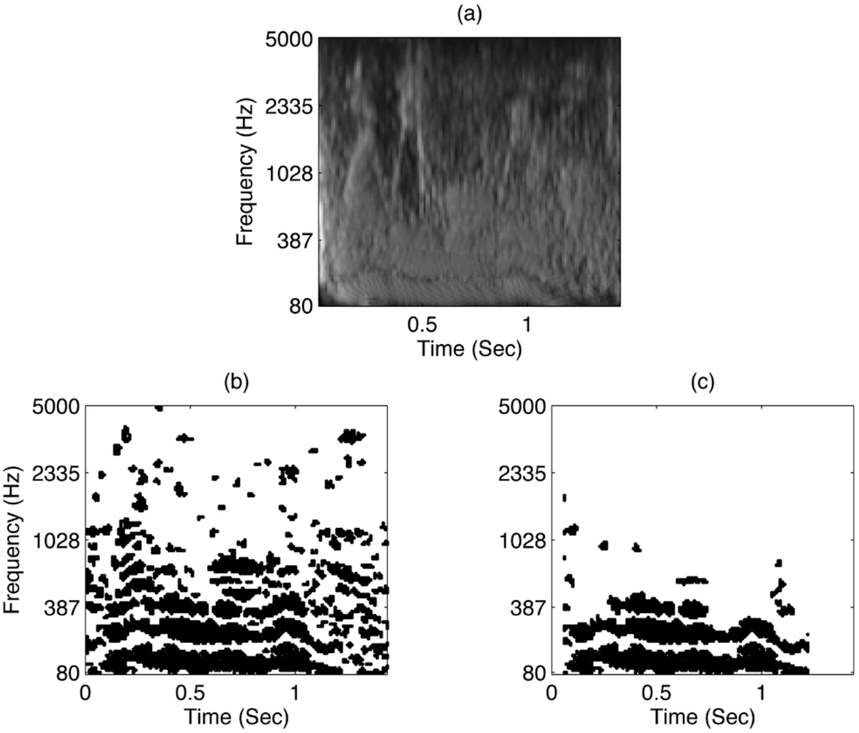


Fig. 3. Results of initial segregation for the speech and cocktail-party mixture. (a) Hair cell response energy of each T-F unit. The figure shows a two-dimensional plot with 128 frequency channels and 144 time frames. (b) Segments formed. Each segment corresponds to a contiguous black region. (c) Foreground stream.

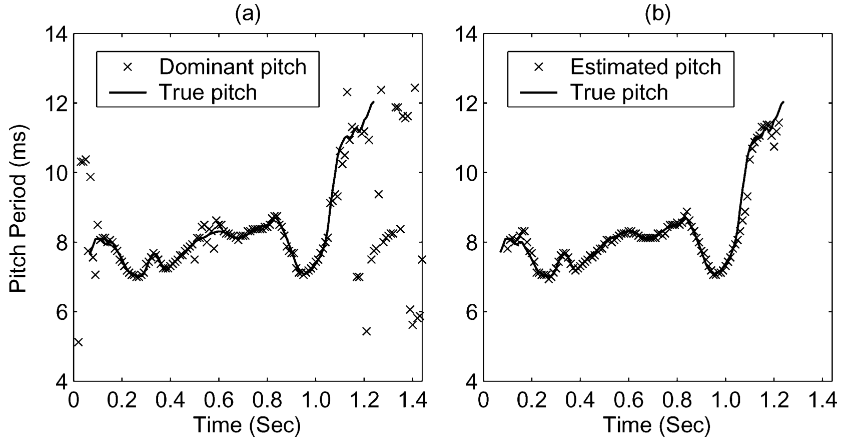


Fig. 4. Results of pitch tracking for the speech and cocktail-party mixture. (a) Dominant pitch periods, marked by "x." (b) Estimated pitch contour of target speech, marked by "x." The solid line indicates the pitch contour obtained from clean speech before mixing.

As illustrated in Fig. 4(a), for the speech and cocktail-party mixture, the dominant pitch periods in a number of frames significantly deviate from the pitch contour of target speech obtained from clean speech. To obtain a more accurate pitch contour, we verify the pitch contour obtained from S_F^0 with two psychoacoustically-motivated constraints and subsequently re-estimate the obtained pitch contour. With the new pitch contour of target speech, units are labeled according to whether target speech dominates. Here, different criteria are used to deal with resolved and unresolved harmonics.

A. Pitch Tracking

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

$$s_F(m,\tau) = \sum_c A_H(c,m,\tau), \quad \text{for } u_{cm} \in S_F^0. \tag{8}$$

Then the pitch period of the target speech at frame m, $\tau_S(m)$, is the lag corresponding to the maximum of $s_F(m,\tau)$ in the plausible pitch range [2 ms, 12.5 ms]. Differing from the dominant pitch period, $\tau_D(m)$, $\tau_S(m)$ is obtained only from signals likely from target speech. Therefore, it is much less affected by the intrusion and describes the target pitch more accurately. However, because S_F^0 contains some intrusion, occasionally $\tau_S(m)$ may not be accurate. Here we employ the following two constraints to check its reliability:

Constraint 1: An accurate pitch period is consistent with the periodicities of individual constituent units. Similar to

(7), we check whether unit u_{cm} agrees with $\tau_S(m)$ by the following inequality:

$$\frac{A_H\left(c, m, \tau_S(m)\right)}{A_H\left(c, m, \tau_P(c, m)\right)} > \theta_P. \tag{9}$$

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, $\tau_S(m)$ reflects the pitch of target speech better than dominant pitch. Note that the threshold in (9) is the same as that for the initial grouping [see (7)].

Constraint 2: The pitch contour of speech changes slowly [22]. We stipulate that the difference between the reliable pitch periods at frame m and $m+1$ be less than 20% of themselves. Constraint 2 applies to only consecutive pitch periods satisfying Constraint 1.

With these two constraints, our model re-estimates the pitch contour of target speech as follows. First, $\tau_S(m)$s are checked according to Constraint 1. Then the two streams are adjusted using only the pitch periods satisfying the constraint: any segment that agrees with these 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 pitch periods, $\tau_S(m)$, are estimated from S_F^1, and are further checked with both constraints. If pitch periods in consecutive frames satisfy both constraints, they are combined into a pitch streak. Therefore, for every interval where all the estimated pitch periods satisfy the constraints, we obtain a pitch streak. Among them, the longest streak is selected. We found that the pitch periods within this streak provide the most reliable estimate of the pitch periods of target speech. 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, where harmonics of speech are generally resolved.

```
1) let $m = m_0$, be the first frame of the longest streak. Mark those channels in which T-F units of $S_F^1$ at frame $m$ agree with $\tau_S(m)$;
2) iterate until $m = 1$
    a. $m = m - 1$;
    b. check $\tau_S(m)$ with the two constraints;
    c. if $\tau_S(m)$ satisfies both constraints, mark it as reliable and go to step 2-d Otherwise, pool the autocorrelations of the units at frame $m$ and in the channels most recently marked. Replace $\tau_S(m)$ with the lag corresponding to the maximum of the summation in the plausible pitch range [2 ms, 12.5 ms]. Check new $\tau_S(m)$ with the two constraints. If $\tau_S(m)$ is reliable, go to step 2-d; otherwise, let $\tau_S(m) = \tau_S(m+1)$ and go to step 2;
    d. Mark those channels in which T-F units of $S_F^1$ at frame $m$ agree with $\tau_S(m)$.
```

Step 2-c 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 sometimes results in interleaving intervals of reliable and unreliable pitch estimates. Finally, for any interval with unreliable pitch estimates between two reliable estimates, the pitch periods within this interval are obtained by a simple linear interpolation from the last frame of the preceding reliable interval and the first frame of the succeeding one. The remaining unreliable pitch periods are set to 0, indicating no pitch at these frames. Because target speech in this research is a voiced utterance, this procedure always produces a continuous pitch contour.

Fig. 4(b) shows the re-estimated target pitch contour from the speech and cocktail-party mixture, together with the pitch contour obtained from the clean target. Except for a few frames, the resulting pitch contour matches that of the clean speech very well, significantly better than the dominant pitch contour, shown in Fig. 4(a). This improvement demonstrates the benefit of applying iterative steps for pitch and target stream estimation.

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 pitch period. Similar to (9), u_{cm} is labeled as target speech if $A_H(c, m, \tau_S(m))$ is close to the maximum of $A_H(c, m, \tau)$ within the plausible pitch range

$$\frac{A_H\left(c, m, \tau_S(m)\right)}{A_H\left(c, m, \tau_P(c, m)\right)} > \theta_T. \tag{10}$$

We find that $\theta_T = 0.85$ is an appropriate choice. The above criterion, referred to 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 Section IV-A). Foreground and background streams are subsequently adjusted. A segment is grouped into the foreground stream, now denoted as S_F^2, if it agrees with the new pitch contour of target speech, according to (10), for more than half of its length; otherwise, it is put into the background stream, S_B^2.

For units responding to multiple harmonics, their responses are amplitude-modulated. As a result, the pitch of target speech 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 Fig. 5. Fig. 5(a) shows over two consecutive frames the response of a gammatone filter with the center frequency of 2.6 kHz, 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, due to the intensity changes of individual harmonics and the interaction between multiple harmonics, the pitch of target speech corresponds to a local maximum in the autocorrelation, indicated by a vertical line in Fig. 5(b), instead of the global maximum within the plausible pitch range.

To deal with this problem we propose a new criterion for labeling units corresponding to unresolved harmonics. Our new

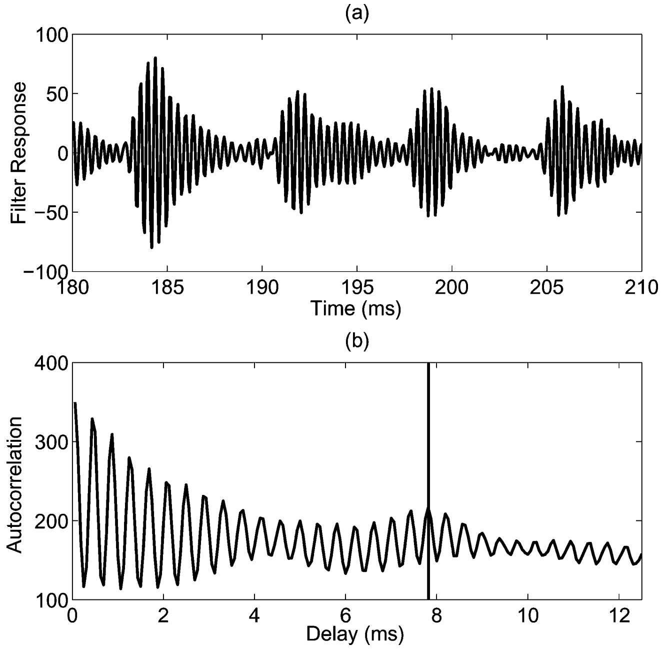


Fig. 5. AM effects for the speech and cocktail-party mixture. (a) Response of a gammatone filter in the high-frequency range, with center frequency 2.6 kHz. (b) Corresponding autocorrelation of the hair cell response. The vertical line marks the position corresponding to the pitch period of target speech.

criterion is 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 [15]. The criterion is referred to as the AM criterion. For each unit, the AM criterion compares AM rate with estimated pitch as follows.

First, the response of a gammatone filter over the entire utterance duration is half-wave rectified and then band-passed to remove the DC component and all possible harmonics except for the F0 component. For every five 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 $\bar{f}$, $+\infty$). Here, $\bar{f}$ is the average of the estimated F0 over the five 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

$$\hat{r}(c,n) = \frac{r(c,n)}{r_E(c,n)} \tag{11}$$

where $r(c,n)$ is the rectified and filtered output in channel c at time step n, and $r_E(c,n)$ is the envelope of $r(c,n)$, obtained via the 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 u_{cm} by a single sinusoid with the specified period of $\tau_S(m)$, in order to compare the AM rate with the estimated pitch period. Specifically

$$\phi_{cm} = \arg\min_{\phi} \sum_{n=0}^{2T-1} \left| \hat{r}(c, mT-n) - \exp\left[j\left(\frac{2\pi n}{\tau_S(m) f_S} + \phi\right)\right]\right|^2 \tag{12}$$

where a square error measure is used. j is the imaginary unit, and $f_S = 16$ kHz is the sampling frequency. Setting to 0 the derivative of the square error with respect to ϕ, we have

$$\tan\phi_{cm} = -\frac{\sum_{n=0}^{2T-1} \hat{r}(c, mT-n) \sin\left(\frac{2\pi n}{\tau_S(m) f_S}\right)}{\sum_{n=0}^{2T-1} \hat{r}(c, mT-n) \cos\left(\frac{2\pi n}{\tau_S(m) f_S}\right)}. \tag{13}$$

Note that within $[0, 2\pi)$, there are two solutions for (13). ϕ_{cm} is the one minimizing the square error, while the other maximizes the square error. Unit u_{cm} is labeled as target speech if the corresponding signal can be well described by the obtained sinusoidal function, i.e., if the following square error is below a certain percentage of the total energy of the corresponding signal

$$\frac{\sum_{n=0}^{2T-1} \left[\hat{r}(c, mT-n) - \cos\left(\frac{2\pi n}{\tau_S(m) f_S} + \phi_{cm}\right)\right]^2}{\sum_{n=0}^{2T-1} \hat{r}^2(c, mT-n)} < \theta_{\mathrm{AM}}. \tag{14}$$

θ_{AM} is chosen to be 0.2, but values of θ_{AM} between 0.1 and 0.4 give very similar results. 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, generally correspond to resolved components.

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 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 group some intrusion-dominant units into the foreground stream. An evaluation using unit labels alone is given in Section VII. To address this issue, those units labeled as target speech according to the AM criterion will be first merged into segments. Then these segments are grouped into streams. This is the task of the next stage.

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

As explained before, the responses of adjacent channels to unresolved harmonics exhibit very similar AM patterns and their envelopes are highly correlated [see Fig. 2(b) and (d)]. 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 (Section V-B) 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

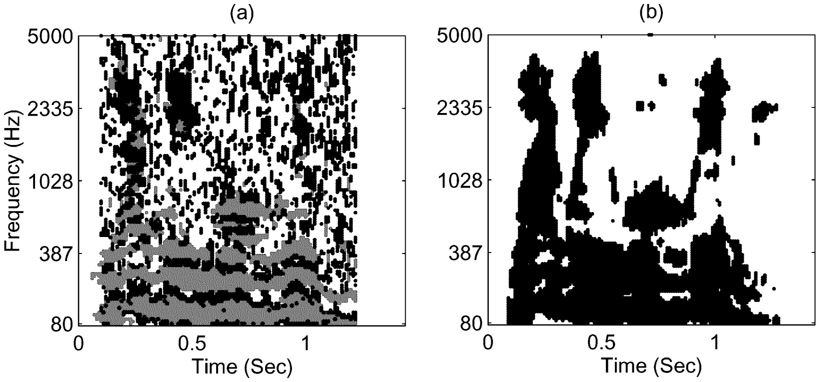


Fig. 6. Results of unit labeling for the speech and cocktail-party mixture. (a) Units labeled as target speech. Gray regions: units labeled by the periodicity criterion; black regions: units labeled by the AM criterion. (b) Units where target speech is stronger than intrusion. Note that the target pitch contour (i.e., target speech) does not extend to the beginning or ending time frames, hence, no processing in those periods.

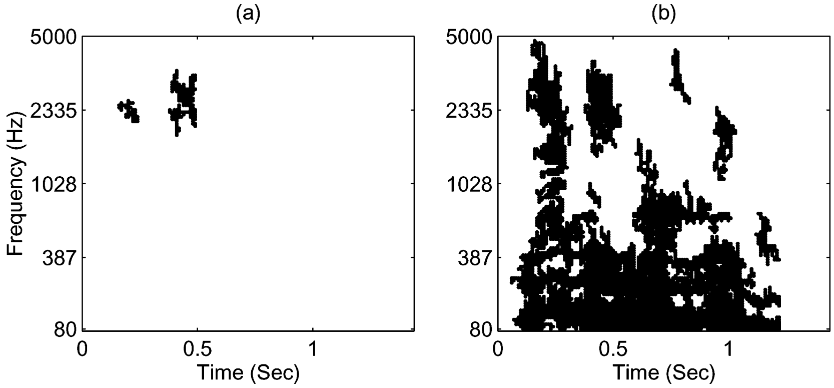


Fig. 7. Results of final segregation for the speech and cocktail-party mixture. (a) New segments formed in the final segregation. (b) Final foreground stream.

units 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)], the segments in Fig. 7(a) contain mainly 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 (Section IV-A) contain units where target dominates as well as those where intrusion dominates. Given unit labels generated in Section 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) remaining segments are removed from S_F^2, and they become 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 Section V-B. Therefore, the segments in the background stream are not adjusted.

Finally, individual units that do not belong to either stream 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. Fig. 8 illustrates the segregation result in waveform format for the speech and cocktail party mixture. The clean speech is shown in Fig. 8(a), the mixture in Fig. 8(b), and the segregated speech in Fig. 8(c). To facilitate comparison between these waveforms, an all-one mask is used to synthesize the waveforms in Fig. 8(a) and (b). One can easily see that the segregated speech waveform is much more similar to the clean speech than the mixture waveform.

VII. Evaluation and Comparison

Our model is evaluated with a corpus of 100 mixtures composed of 10 voiced utterances mixed with 10 intrusions collected

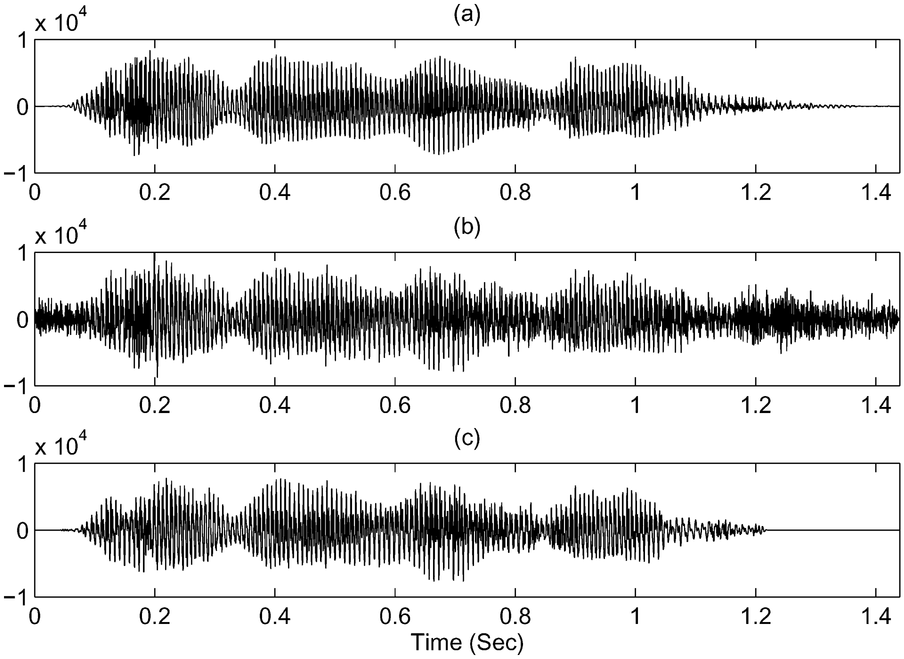


Fig. 8. Waveform results. (a) Clean speech. (b) Mixture of the speech and cocktail-party noise. (c) Speech segregated from the mixture.

TABLE I
SNR Results

Intrusion	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Average
Mixture	-3.26	-4.07	10.19	4.34	3.99	-5.82	1.90	6.62	10.37	0.73	2.50
Proposed model	16.34	7.83	16.71	8.32	10.88	14.41	16.89	11.97	14.44	5.27	12.30
Pitch-labeled mask	4.94	5.14	16.68	8.27	9.60	7.50	7.87	10.56	14.67	3.15	8.83
A_E-based mask	16.06	0.38	16.90	8.07	9.06	14.65	16.10	11.83	14.19	5.34	11.26
True pitch	16.33	8.35	17.71	8.79	11.56	15.06	17.76	12.31	15.32	6.04	12.92
Narrow band	9.88	6.74	11.44	6.94	8.95	8.33	11.31	9.15	10.60	3.98	8.73
Comb filter	3.12	3.01	13.28	8.72	8.32	2.25	6.56	10.57	13.19	5.39	7.44
Wang-Brown system	11.31	4.93	11.19	5.65	8.72	10.44	11.15	9.22	10.84	2.66	8.61
Spectral subtraction	18.35	3.05	16.00	6.14	8.32	-5.51	4.85	8.23	10.90	2.46	7.28
Ideal binary mask	20.76	9.04	22.90	9.72	13.19	18.40	21.53	15.78	18.10	10.05	15.95

by Cooke [8], which has been used to test CASA systems [5], [8], [10], [12], [33] and, hence, facilitates our comparison. The sampling frequency is 16 kHz. The intrusions have a considerable variety. Specifically, the 10 intrusions are: N0, 1-kHz 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 [29]. An objective and straightforward criterion is to measure SNR before and after segregation, using target speech before mixing as signal. As discussed later, this conventional SNR criterion has undesirable properties for measuring CASA performance. Since it is commonly used, we first report our results using this criterion. To compensate for amplification and distortion effects introduced in the resynthesis process, we use resynthesized target speech with an all-one mask as signal to compute SNR for evaluation cases that involve masks. Table I gives a variety of SNR results, including those of our model and original mixtures. Each value in the table represents the average SNR for one intrusion mixed with 10 target utterances. A further average across all intrusions is shown in the last column of the table. As can be seen in the table, our system improves the SNR for every intrusion, producing a gain of 9.8 dB over the original mixtures. Large SNR improvements are obtained for intrusions whose spectra do not significantly overlap with those of target utterances (e.g., N0 and N5), whereas improvements are modest for intrusions with significant overlap (e.g., N3 and N8).

We choose two broadband intrusions: N1 and N3, for closer examination of model performance with respect to varying SNR levels. The target speech, 10 voiced utterances, remains the same, and we vary the level of intrusion to obtain mixture SNR levels ranging from -15 dB to 20 dB. The SNR values of input and output are shown in Fig. 9. When the input SNR is lower than -10 dB, intrusion is stronger than target speech almost everywhere, and our system basically rejects almost the entire input signal and the output SNR is around 0 dB. As the input SNR increases, our system groups more target speech,

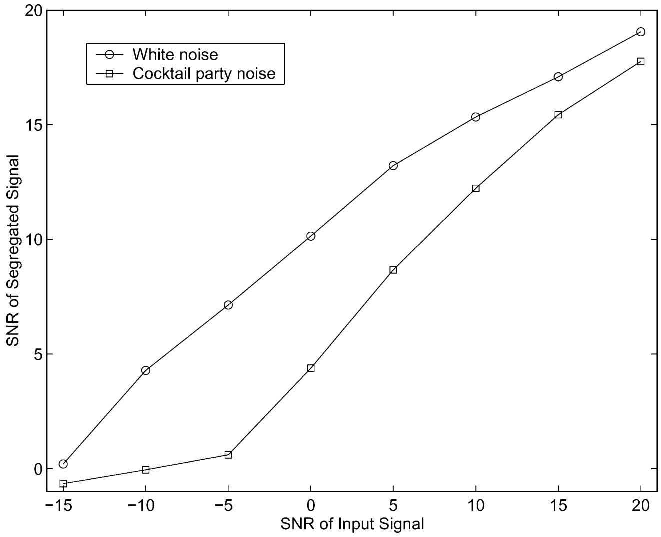


Fig. 9. SNR results of segregated speech with respect to different SNR levels of input mixtures. Two noise intrusions are N1 (white noise) and N3 (cocktail party noise).

resulting in significant SNR improvements. When the input SNR is very high (greater than 15 dB), target speech is stronger than intrusion in a majority of T-F units. Although our system retains much of target speech, it still loses some target energy and therefore cannot improve the SNR. Note that our system performs uniformly better for N1 than for N3 since the latter overlaps more extensively with target speech.

To illustrate the effect of segmentation and final segregation (see Fig. 1), we measure the SNR performance for pitch-labeled binary masks, i.e., the masks resulting from pitch-based unit labeling described in Section V-B. The SNR results for pitch-labeled masks are given in the third row of Table I. It is clear that the performance of a system using pitch information only is not as good, especially for the intrusions of N0, N5, and N6, which have strong components with frequencies close to harmonics of target speech. On average, there is an SNR drop of more than 3 dB compared to the overall model.

Our system labels T-F units in the high-frequency range through sinusoidal modeling. A simpler alternative is to use the autocorrelations of response envelopes, i.e., A_E, much like A_H is used in the low-frequency range. The fourth row of Table I reports the results of this alternative labeling with the rest of the system kept the same. The performance of A_E-based masks is not much worse except for N1, white noise, where sinusoidal modeling gives significantly better labeling.

When a CASA system like ours makes a segregation error, it may be caused by errors in pitch estimation or pitch-based grouping. To examine more closely the type of error, we employ the use of true pitch information for speech segregation. True pitch is obtained from premixing target speech and further verified manually to ensure high quality. The fifth row of Table I gives the SNR results for our system using true pitch instead of estimated pitch. With true pitch, the system performs only slightly better. This suggests that estimated pitch of our system is quite accurate.

A main novel aspect of our model is a careful treatment of unresolved harmonics in the high-frequency range, caused by broader bandwidths of higher frequency channels in the gammatone filterbank. A legitimate question is, are we solving a problem created by our choice of a filterbank in the first place? To examine this question, we employ an alternative filterbank with a fixed narrow bandwidth. The alternative filterbank contains 128 gammatone filters equally spaced from 80 Hz to 5 kHz, the same number of filters covering the same frequency range as the auditory filterbank used in our model. These filters have the same bandwidth and adjacent filters are half-overlapped. This filterbank can resolve all the harmonics of target speech so there is no need to address the issue of unresolved harmonics. The segregation is then performed in the same way as that for the low-frequency range in our model using the same estimated pitch. The SNR results are given in Table I in the sixth row indicated by "Narrow band." The results are not as good as those using the original gammatone filterbank. We think that the drop in performance can be attributed to two factors. First, in the high-frequency range, an error in estimated F0 gets multiplied in higher harmonics and, hence, treating multiple harmonics together in the same channel may enhance the robustness to pitch estimation errors. Second, although some intrusion components have frequencies near those of target harmonics, these components are less likely to produce an AM pattern similar to that of target speech.

We also compare our system with the comb filtering method, which is another narrow-band algorithm to extract a harmonic series using pitch information [11]. Given a target pitch, the filter retains target speech and attenuates interference whose frequency components are incompatible with the target harmonic series. To check how well harmonic extraction works we supply the true pitch contour of target speech to a comb filter with 3 coefficients at every 20-ms frame. The resulting SNR is shown in the seventh row of Table I. The comb filter results are worse than those of the narrow-band filterbank. Note that when the input is target speech alone, the resulting SNR from comb filtering is consistently higher than 20 dB. This suggests that poor performance of comb filtering stems from retaining too much interference. This should not come as a surprise since the comb filter passes though all frequency components close to the multiples of target F0.

The above comparisons with narrow-band gammatone filterbank and comb filter suggest that the use of an auditory filterbank produces real performance gains. As far as we know, this is the first time an auditory filterbank is compared quantitatively with alternative filtering methods in the context of speech segregation. Our results provide further credence to previous reports that an auditory-based front-end is less sensitive to background intrusions [14], [19].

The CASA model by Wang and Brown is representative of recent CASA systems [33]. The processing of the Wang–Brown model is similar to the first two stages of our model (see Fig. 1). The SNR results of the model are shown in Table I. Our system performs consistently better than the Wang–Brown system.

Table I also shows a comparison with the spectral subtraction method [17], which is a standard method for speech enhancement. The method is applied as follows. For each intrusion, we find its duration and obtain its average power spectrum within

TABLE II
P_{EL} AND P_{NR} RESULTS

Intrusion	**Proposed model**		**Wang-Brown system**		**Mixture**
	P_{EL} (%)	P_{NR} (%)	P_{EL} (%)	P_{NR} (%)	P_{NR} (%)
N0	2.00	0.02	6.99	0	67.76
N1	5.87	2.43	28.96	1.61	57.16
N2	1.34	1.05	5.77	0.71	5.04
N3	4.38	2.20	21.92	1.92	18.15
N4	3.84	1.75	10.22	1.41	27.17
N5	2.65	0.04	7.47	0	78.84
N6	1.52	0.37	5.99	0.48	39.24
N7	3.02	1.83	8.61	4.23	16.68
N8	1.71	1.34	7.27	0.48	7.37
N9	10.11	17.27	15.81	33.03	43.09
Average	3.64	2.83	11.91	4.39	36.05

the duration. This average is used as the estimate of the intrusion. Then for the corresponding mixtures, spectral subtraction is applied within the corresponding duration. The intrusions in our test corpus are generally continuous except for N2, which contains a sequence of short noise bursts. For this intrusion, subtraction is applied within each burst. Note that it is a difficult task to detect the onset and offset and obtain the power spectrum of a nonstationary intrusion. In this comparison, such information is already made available to the spectral subtraction method. The SNR results are shown in Table I. The spectral subtraction method performs significantly worse than our system. This is because of its well known deficiency in dealing with nonstationary interference.

Despite its common use, the conventional SNR criterion does not take into consideration related perceptual effects such as auditory masking and the ear's insensitivity to phase spectrum [15]. Furthermore, when segregated target is different from original target, the criterion does not provide much information about how they are different. Given our objective of identifying T-F regions with stronger target (see Section I), we here suggest to use an ideal binary mask as the ground truth of target stream. An ideal binary mask is constructed as follows: a T-F unit in the mask is assigned 1 if the target energy in the 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. We call such a mask "ideal" because it represents our computational objective and it is an *a priori* mask constructed using premixing target and intrusion. Fig. 6(b) displays the ideal mask for the speech and cocktail-party mixture, where black regions correspond to 1. In addition to the SNR justification given in Section I, 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 [26]. An ideal mask is also similar to an a priori mask used in a recent ASR study [9], which yields excellent recognition performance. The SNR results from ideal binary masks are shown in Table I, and they are uniformly better than those of our model—on average 3.65 dB higher. This gives an indication on how much our model could be further improved in terms of conventional SNR.

Let $O(n)$ donate the resulting speech from our system. The speech waveform resynthesized from the ideal binary mask is denoted by $I(n)$. Let $e_1(n)$ denote the signal present in $I(n)$ but missing from $O(n)$, and $e_2(n)$ the signal present in $O(n)$ but missing from $I(n)$. Then, the percentage of energy loss, P_{EL}, and the percentage of noise residue, P_{NR}, are calculated as follows:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \tag{15a}$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}. \tag{15b}$$

P_{EL} indicates the percentage of target speech excluded from segregated speech, and P_{NR} the percentage of intrusion included. They provide complementary error measures of a segregation system and a successful system needs to achieve low errors in both measures. To obtain $e_1(n)$, a mask is constructed as follows. A T-F unit is assigned 1 if and only if it is 1 in the ideal binary mask but 0 in the segregated target stream. $e_1(n)$ is then obtained by resynthesizing the input mixture from the obtained mask. $e_2(n)$ is obtained in a similar way.

The results from our model are shown in Table II. As in Table I, each value in the table represents the average result of one intrusion with 10 voiced utterances, and a further average across all intrusions is also shown. On average, our system retains 96.36% of target speech energy, and the percentage of noise residue is kept at 2.83%. The percentage of 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 by P_{EL} and P_{NR} are balanced in our system.

As a comparison, Table II also shows the results from the Wang–Brown model. Our model produces an improvement in both P_{EL} and P_{NR} measures. In particular, our system has cut P_{EL} from 11.91% with their system to 3.64% with our system; the improvement is especially noticeable for N1 and N3. Both models achieve a low level of P_{NR}, except for N9, where our result is much better. The good performance in P_{NR} reflects the conservative strategies employed by the models in dealing with intrusions: both models tend to exclude uncertain units from the foreground stream.

TABLE III
P_{EL} AND P_{NR} RESULTS IN THE HIGH-FREQUENCY RANGE

Intrusion	Proposed model		Wang-Brown system		Mixture
	P_{EL} (%)	P_{NR} (%)	P_{EL} (%)	P_{NR} (%)	P_{NR} (%)
N0	21.11	0.35	64.54	0	96.82
N1	59.16	46.20	83.66	66.08	96.00
N2	13.50	16.79	50.80	32.75	44.02
N3	26.35	5.30	74.99	15.00	42.57
N4	34.84	33.88	78.22	60.45	81.31
N5	20.94	0.83	63.26	0	97.90
N6	17.07	6.92	48.21	14.50	91.26
N7	22.60	9.80	64.25	24.78	43.49
N8	13.21	5.56	49.52	2.38	31.07
N9	28.16	4.92	69.17	25.88	27.72
Average	25.70	13.05	64.65	24.17	65.22

Since our model applies different mechanisms to segregate resolved and unresolved harmonics, it is instructive to present the performance in the high-frequency range separately. For this purpose, we calculate P_{EL} and P_{NR} for only the filter channels with center frequencies greater than 1 kHz. Note that since the F0 range of target speech is from 80 Hz to 160 Hz, target harmonics in the frequency range above 1 kHz are generally unresolved. The corresponding results are shown in Table III. It is clear from the average P_{NR} of the original mixtures that intrusions are much stronger in the high-frequency range. Our model achieves large improvements in both P_{EL} and P_{NR}.

To compare waveforms directly we also measure SNR in decibels using the resynthesized speech from the ideal binary mask as ground truth

$$\text{SNR} = 10\log_{10}\left[\frac{\sum\limits_{n} I^2(n)}{\sum\limits_{n}(I(n)-O(n))^2}\right]. \tag{16}$$

The SNR for each intrusion averaged across 10 target utterances is shown in Fig. 10, together with the SNR of the original mixtures and the results from the Wang–Brown system and spectral subtraction. Note that for an original mixture and an output from spectral subtraction, an all-one mask is used against the corresponding ideal binary mask. The corresponding SNRs are shown in Fig. 10, which are a little higher than that in Table I. All three systems show improvements compared to original 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 5.2 dB. The Wang–Brown model in turn performs 1.2 dB better on average than spectral subtraction. As expected spectral subtraction produces uneven results for the intrusions; for example, its performance is the best for N0 (pure tone) among all the methods.

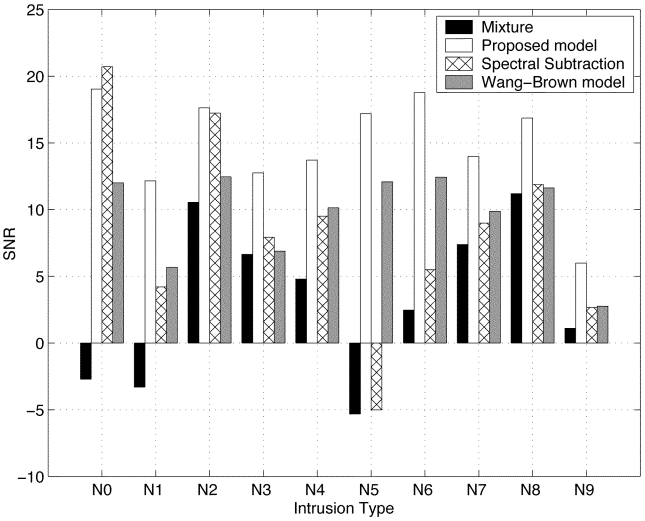


Fig. 10. SNR results against ideal binary masks for segregated speech and original mixtures. White bars show the results from the proposed model, gray bars those from the Wang–Brown system, cross bars those from the spectral subtraction method, and black bars those of original mixtures.

VIII. Discussion

Our system segregates voiced speech based on the analysis of temporal information in the input, the temporal fine structure of a resolved harmonic and the temporal envelope of an unresolved harmonic. There is evidence suggesting that the auditory system uses the temporal patterns of neural spikes to code the input sound [6]. Models based on temporal coding of the input, such as correlogram, have been employed to model auditory perception, especially pitch perception, and have successfully explained many observed perceptual phenomena [6], [25], [31].

Like previous CASA systems, our system exploits the grouping cues of harmonicity and temporal continuity to segregate voiced speech [5], [8], [33], [34]. However, our system is substantially different from previous studies in the way such ASA cues are utilized. First, our system applies different mechanisms to deal with resolved and unresolved harmonics, and uses AM of filter responses to segregate unresolved harmonics. Second, target pitch is estimated from a segregated speech stream, which in turn is based on dominant pitch estimation; estimated target pitch is used for final segregation. This, in fact, can be viewed as an instance of iteratively performing pitch estimation and harmonic segregation. This iterative estimation-segregation process enables us to obtain accurate pitch information. Extensive evaluation results in the previous section demonstrate that our model performs substantially better than other systems aiming at similar objectives. The evaluation also compares a range of alternative methods for components of the model.

TABLE IV

Symbol	Definition
t	Time
f	Center frequency of a filter
n	Index of time sample
f_S	Sampling frequency
c	Index of filter channel
m	Index of time frame
u_{cm}	T-F unit in channel c at frame m
g	Impulse response of a gammatone filter
l	Order of a gammatone filter
b	Bandwidth of a gammatone filter
A_H, $\hat{A}_H$	Autocorrelation and normalized autocorrelation of filter response
A_E, $\hat{A}_E$	Autocorrelation and normalized autocorrelation of response envelope
C_H	Cross-channel correlation of $\hat{A}_H$
C_E	Cross-channel correlation of $\hat{A}_E$
h	Hair cell output
h_E	Envelope of hair cell output
N_c	Number of samples in the time window for computing A_H and A_E in channel c
T	Number of samples in a frame shift (10 ms)
L	Number of samples in the maximum delay for A_H (12.5 ms)
τ	Time delay
τ_D	Dominant pitch delay
τ_S	Estimated target pitch delay
τ_P	Delay corresponding to the maximum of A_H within the plausible pitch range
s	Summary correlogram at a certain frame
s_F	Summary correlogram of the foreground stream at a certain frame
θ_H	Spontaneous firing rate of the auditory nerve (set to 50)
θ_C	Threshold for determining segment formation using cross-channel correlation (set to 0.985)
θ_P	Threshold for determining the agreement between a T-F unit and estimated pitch (set to 0.95)
θ_T	Threshold for labeling units using the time criterion (set to 0.85)
θ_{AM}	Threshold for labeling units using the AM criterion (set to 0.2)
S_F^i, S_B^i	Obtained foreground and background stream at stage i (i = 0, 1, 2, 3)
$\bar{f}$	Average of estimated F0 over 5 frames
$r, r_E, \hat{r}$	Rectified and filtered output of gammatone filter, its envelope, and normalized version
ϕ_{cm}	Estimated phase for the matching sinusoid of $\hat{r}$ in channel c at frame m
I	Clean target speech
O	Output of segregated speech
e_1	Signal present in I but missing from O
e_2	Signal present in O but missing from I
P_{EL}	Percentage of energy loss
P_{NR}	Percentage of noise residue

Amplitude modulation has been explored by Weintraub [34] and Cooke [8]. Specifically, Weintraub used a coincidence function, a version of autocorrelation, to capture periodicity as well as AM. Then pitch contours of multiple utterances are tracked from coincidence functions. Sound separation is achieved through an iterative spectral estimation according to pitch and temporal continuity. 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. Both of these studies use AM primarily for grouping, whereas we use it to deal with unresolved harmonics in both segmentation and grouping. As a result, our model performs significantly better.

As mentioned before, the first two stages of our system (see Fig. 1) are similar to the Wang–Brown system [33]. Their model, however, does not perform subsequent processing, and its performance as a result is substantially worse. Nonetheless, the oscillatory correlation mechanism [32] employed in their model can be similarly used to implement the stages introduced in our model. These new processes include further segmentation for unresolved harmonics and further grouping that adjusts the foreground and background stream according to extracted target pitch.

The AM criterion plays an essential role in labeling T-F units corresponding to unresolved harmonics. Our model compares an AM rate with estimated pitch by sinusoidal modeling and subsequent gradient descent. In a previous study, we obtained AM rates by extracting the instantaneous frequencies of filter responses [16]. The results there were similar to those in Table II and Fig. 10. One advantage of the current method is that it is much more efficient computationally. Our model decomposes the 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 tried an 8-unit neighborhood, but results are similar.

The performance of the proposed model depends on the accuracy of an estimated target pitch contour. 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.* proposed a robust algorithm for multipitch tracking for speech utterances in the presence of other interfering sources [35]. 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 a target source only. However, it is possible to track the pitch contour of intrusion if it has a harmonic structure. With two pitch contours, one could label a T-F unit more accurately by comparing whether its periodicity is 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, N8, and N9. As indicated in Table I, Table II, and Fig. 10, the performance of our system for this kind of 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 [4], and may play a role in grouping unvoiced speech. Also, it appears that one must consider acoustic and phonetic characteristics of individual unvoiced consonants. We plan to investigate these issues in future study.

Appendix
Symbols and Their Definitions

Please see Table IV, shown on the previous page.

Acknowledgment

The authors wish to thank G. Brown, M. Wu, and two anonymous reviewers for helpful comments. Preliminary versions of this work were presented in 2001 IEEE WASSPA, 2002 IEEE ICASSP, and 2002 NIPS conferences.

References

[1] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Trans. Neural Networks*, vol. 13, pp. 888–893, July 2002.

[2] J. Bird and C. J. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, Eds. London, U.K.: Whurr, 1997.

[3] E. de Boer and H. R. de Jongh, "On cochlear encoding: potentialities and limitations of the reverse-correlation techniques," *J. Acoust. Soc. Amer.*, vol. 63, pp. 115–135, 1978.

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

[5] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Comput. Speech Language*, vol. 8, pp. 297–336, 1994.

[6] P. Cariani, "Temporal coding of periodicity pitch in the auditory system: an overview," *Neural Plasticity*, vol. 6, pp. 147–172, 1999.

[7] R. P. Carlyon and T. M. Shackleton, "Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms?," *J. Acoust. Soc. Amer.*, vol. 95, pp. 3541–3554, 1994.

[8] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[9] M. P. Cooke, P. D. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[10] L. A. Drake, "Sound source separation via computational auditory scene analysis (CASA)-enhanced beamforming," Ph.D. dissertation, Dept. Elect. Comput. Eng., Northwestern Univ., Evanston, IL, 2001.

[11] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[12] D. P. W. Ellis, "Prediction-Driven Computational Auditory Scene Analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, 1996.

[13] Y. Ephraim and H. L. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.

[14] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 115–132, Jan. 1994.

[15] H. Helmholtz, *On the Sensations of Tone*. Braunschweig, Germany: Vieweg & Son, 1863.

[16] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," in *Int. Conf. Acoustics, Speech and Signal Processing*, 2002, pp. 553–556.

[17] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[18] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–134, 1951.

[19] C. R. Jankowski, H. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286–293, July 1995.

[20] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 731–740, Oct. 2001.

[21] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 67–94, July 1996.

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

[23] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.

[24] R. Meddis, "Simulation of auditory-neural transduction: further studies," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.

[25] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Amer.*, vol. 102, pp. 1811–1820, 1997.

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

[27] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Appl. Psychol. Unit, Cambridge Univ., Cambridge, U.K., APU Rep. 2341, 1988.

[28] R. Plomp and A. M. Mimpen, "The ear as a frequency analyzer II," *J. Acoust. Soc. Amer.*, vol. 43, pp. 764–767, 1968.

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

[30] 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 Audio Processing*, vol. 6, pp. 445–455, Sept. 1998.

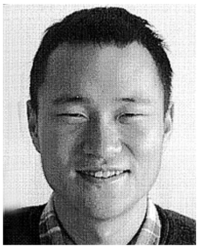
[31] M. Slaney and R. F. Lyon, "On the importance of time—a temporal representation of sound," in *Visual Representations of Speech Signals*, M. P. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993, pp. 95–116.

[32] D. L. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cogn. Sci.*, vol. 20, pp. 409–456, 1996.

[33] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, May 1999.

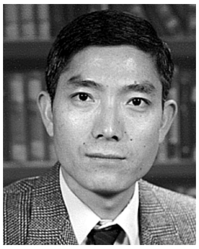
[34] M. Weintraub, "A Theory and Computational Model of Auditory Monaural Sound Separation," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1985.

[35] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 229–241, May 2003.



Guoning Hu received the B.S. and M.S. degrees in physics from Nanjing University, Nanjing, China, in 1996 and 1999, respectively. He is currently working toward the Ph.D. degree in biophysics at The Ohio State University, Columbus, OH.

His research interests include speech segregation, computational auditory scene analysis, and acoustic and phonetic properties of speech.



DeLiang Wang (M'90–F'04) received the B.S. and M.S. degrees in computer science from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, in 1991.

From 1986 to 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing, China. Since 1991, he has been with the Department of Computer and Information Science and the Center for Cognitive Science, The Ohio State University, Columbus, where he is currently a Professor. From 1998 to 1999, he was a Visiting Scholar in the Department of Psychology, Harvard University, Cambridge, MA. His research interests include machine perception and neurodynamics.

Dr. Wang currently chairs the IEEE Neural Networks Society Neural Networks Technical Committee, is a member of the Governing Board of the International Neural Network Society and the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee. He is a recipient of the U.S. Office of Naval Research Young Investigator Award.