A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation

Guoning Hu, *Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—A lot of effort has been made in computational auditory scene analysis (CASA) to segregate speech from monaural mixtures. The performance of current CASA systems on voiced speech segregation is limited by lacking a robust algorithm for pitch estimation. We propose a tandem algorithm that performs pitch estimation of a target utterance and segregation of voiced portions of target speech jointly and iteratively. This algorithm first obtains a rough estimate of target pitch, and then uses this estimate to segregate target speech using harmonicity and temporal continuity. It then improves both pitch estimation and voiced speech segregation iteratively. Novel methods are proposed for performing segregation with a given pitch estimate and pitch determination with given segregation. Systematic evaluation shows that the tandem algorithm extracts a majority of target speech without including much interference, and it performs substantially better than previous systems for either pitch extraction or voiced speech segregation.**

***Index Terms*—Computational auditory scene analysis (CASA), iterative procedure, pitch estimation, speech segregation, tandem algorithm.**

I. Introduction

SPEECH segregation, or the cocktail party problem, is a well-known challenge with important applications. For example, automatic speech recognition (ASR) systems perform substantially worse in the presence of interfering sounds [27], [36] and could greatly benefit from an effective speech segregation system. Background noise also presents a major difficulty to hearing aid wearers, and noise reduction is considered a great challenge for hearing aid design [12]. Many methods have been proposed in monaural speech enhancement [28]. These methods usually assume certain statistical properties of interference and tend to lack the capacity to deal with a variety of interference. While voice separation has proven to be difficult, the human auditory system is remarkably adept in this task. The perceptual process is considered as *auditory scene analysis* (ASA) [6]. Psychoacoustic research in ASA has inspired considerable work in developing computational auditory scene analysis (CASA) systems for speech segregation (see [39] for a comprehensive review).

Natural speech contains both voiced and unvoiced portions, and voiced portions account for about 75%–80% of spoken English [19]. Voiced speech is characterized by periodicity (or harmonicity), which has been used as a primary cue in many CASA systems for segregating voiced speech (e.g., [8] and [16]). Despite considerable advances in voiced speech separation, the performance of current CASA systems is still limited by pitch (F0) estimation errors and residual noise. Various methods for robust pitch estimation have been proposed [11], [24], [33], [40]; however, robust pitch estimation under low signal-to-noise ratio (SNR) situations still poses a significant challenge. Since the difficulty of robust pitch estimation stems from noise interference, it is desirable to remove or attenuate interference before pitch estimation. On the other hand, noise removal depends on accurate pitch estimation. As a result, pitch estimation and voice separation become a "chicken and egg" problem [11].

We believe that a key to resolve the above dilemma is the observation that one does not need the entire target signal to estimate pitch (a few harmonics can be adequate), and without perfect pitch one can still segregate some target signal. Thus, we suggest a strategy that estimates target pitch and segregates the target in tandem. The idea is that we first obtain a rough estimate of target pitch, and then use this estimate to segregate the target speech. With the segregated target, we should generate a better pitch estimate and can use it for better segregation, and so on. In other words, we propose a new algorithm that achieves pitch estimation and speech segregation jointly and iteratively. We call this method a *tandem algorithm* because it alternates between pitch estimation and speech segregation. This idea was present in a rudimentary form in our previous system for voiced speech segregation [16] which contains two iterations. Besides this idea, novel methods are proposed for segregation and pitch estimation; in particular, a classification based approach is proposed for pitch-based grouping.

The separation part of our tandem system aims to identify the *ideal binary mask* (IBM). With a time–frequency (T-F) representation, the IBM is a binary matrix along time and frequency where 1 indicates that the target is stronger than interference in the corresponding T-F unit and 0 otherwise (see Fig. 5 later for an illustration). To simplify notations, we refer to T-F units labeled 1 and those labeled 0 as *active* and *inactive* units, respectively. We have suggested that the IBM is a reasonable goal for CASA [16], [37], and it has since been used as a measure of ceiling performance for speech separation [26], [31], [32]. Recent psychoacoustic studies provide strong evidence that the

Manuscript received October 24, 2008; revised October 18, 2009. Date of publication January 26, 2010; date of current version September 08, 2010. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-01-0155 and in part by the National Science Foundation (NSF) under Grant IIS-0534707. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Susanto Rahardja.

G. Hu is with the Biophysics Program, The Ohio State University, Columbus, OH 43210 USA, and also with AOL Truveo Video Search, San Francisco, CA 94104 USA (e-mail: ghu@truveo.com).

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

Digital Object Identifier 10.1109/TASL.2010.2041110

IBM leads to large improvements of human speech intelligibility in noise [9], [25].

This paper is organized as follows. Section II describes T-F decomposition of the input and feature extraction. The tandem algorithm has two key steps: estimating the IBM given an estimate of target pitch and estimating the target pitch given an estimated IBM. We describe these two steps in Sections III and IV. The tandem algorithm is then presented in Section V. Systematic evaluation on pitch estimation and speech segregation is given in Section VI, followed by discussion in Section VII and conclusion in Section VIII.

II. T-F Decomposition and Feature Extraction

We first decompose an input signal in the frequency domain with a bank of 128 gammatone filters [30], with their center frequencies equally distributed on the equivalent rectangular bandwidth rate scale from 50 to 8000 Hz (see [39] for details). In each filter channel, the output is divided into 20-ms time frames with 10-ms overlap between consecutive frames. The resulting T-F representation is known as a cochleagram [39]. At each frame of each channel, we compute a correlogram, a running autocorrelation function (ACF) of the signal, within a certain period of time delay. Each ACF represents the periodicity of the filter response in the corresponding T-F unit. Let u_{cm} denote a T-F unit for channel c and frame m and $x(c,t)$ the filter response for channel c at time t. The corresponding ACF of the filter response is given by (1) shown at the bottom of the page. Here, τ is the delay and n denotes discrete time. $T_m = 10$ ms is the frame shift and T_n is the sampling time (e.g., $T_n = 0.0625$ ms for the signal sampling frequency of 16 kHz used in this study). The above summation is over 20 ms, the length of a time frame. The periodicity of the filter response is indicated by the peaks in the ACF, and the corresponding delays indicate the periods. We calculate the ACF within the following range: $\tau T_n \in [0, 15\,\text{ms}]$, which includes the plausible pitch frequency range from 70 to 400 Hz [29].

It has been shown that, cross-channel correlation, which measures the similarity between the responses of two adjacent filters, indicates whether the filters are responding to the same sound component [8], [38]. Hence, we calculate the cross-channel correlation of u_{cm} with $u_{c+1,m}$ by (2) shown at the bottom of the page, where $\overline{A}$ denotes the average of A.

When the input contains a periodic signal, high-frequency filters respond to multiple harmonics of the signal and these harmonics are called *unresolved*. Unresolved harmonics trigger filter responses that are amplitude-modulated, and the response envelope fluctuates at the F0 of the signal [14]. Here, we extract envelope fluctuations corresponding to target pitch by half-wave rectification and bandpass filtering, and the passband corresponds to the plausible F0 range of target speech. Then, we compute the envelope ACF, $A_E(c,m,\tau)$, and the cross-channel correlation of response envelopes, $C_E(c,m)$, similar to (1) and (2).

III. IBM Estimation Given Target Pitch

A. Unit Labeling With Information Within Individual T-F Units

We first consider a simple approach: a T-F unit is labeled 1 if and only if the corresponding response or response envelope has a periodicity similar to that of the target. As discussed in Section II, the periodicity of a filter response is indicated by the peaks in the corresponding ACF. Let $\tau_S(m)$ be the estimated pitch period at frame m. When a response has a period close to $\tau_S(m)$, the corresponding ACF will have a peak close to $\tau_S(m)$. Previous work [16] has shown that $A(c,m,\tau_S(m))$ is a good measure of the similarity between the response period in u_{cm} and estimated pitch.

Alternatively, one may compare the instantaneous frequency of the filter response with the estimated pitch directly. However, in practice, it is extremely difficult to accurately estimate the instantaneous frequency of a signal [3], [4], and we found that labeling T-F units based on estimated instantaneous frequency does not perform better than using the ACF-based measures.

We propose a different approach to pitch-based labeling. Specifically, we construct a classifier that combines these two kinds of measure to label T-F units. Let $\bar{f}(c,m)$ denote the estimated average instantaneous frequency of the filter response within unit u_{cm}. If the filter response has a period close to $\tau_S(m)$, then $\bar{f}(c,m)\cdot\tau_S(m)$ is close to an integer greater than or equal to 1. Similarly, let $\bar{f}_E(c,m)$ be the estimated average instantaneous frequency of the response envelope within u_{cm}. If the response envelope fluctuates at the period of $\tau_S(m)$, then $\bar{f}_E(c,m)\cdot\tau_s(m)$ is close to 1. Let (3), shown at the bottom of

$$A(c,m,\tau)=\frac{\sum_n x(c,mT_m-nT_n)x(c,mT_m-nT_n-\tau T_n)}{\sqrt{\sum_n x^2(c,mT_m-nT_n)\sum_n x^2(c,mT_m-nT_n-\tau T_n)}}. \tag{1}$$

$$C(c,m)=\frac{\sum_\tau [A(c,m,\tau)-\overline{A(c,m)}][A(c+1,m,\tau)-\overline{A(c+1,m)}]}{\sqrt{\sum_\tau [A(c,m,\tau)-\overline{A(c,m)}]^2\sum_\tau [A(c+1,m,\tau)-\overline{A(c+1,m)}]^2}} \tag{2}$$

the page, be a set of six features, the first three of which correspond to the filter response and the last three to the response envelope. In (3), the function $\mathrm{int}(x)$ returns the nearest integer. This 6-dimensional vector incorporates both autocorrelation and instantaneous-frequency features calculated from both filter responses and response envelopes. Note that the feature vector is a function of a given pitch period.

Let H_0 be the hypothesis that a T-F unit is target dominant and H_1 otherwise. u_{cm} is labeled as target if and only if

$$P(H_0|r_{cm}(\tau_S(m))) > P(H_1|r_{cm}(\tau_S(m))). \quad (4)$$

Since

$$P(H_0|r_{cm}(\tau_S(m))) = 1 - P(H_1|r_{cm}(\tau_S(m))). \quad (5)$$

Equation (4) becomes

$$P(H_0|r_{cm}(\tau_S(m))) > 0.5. \quad (6)$$

In this paper, we estimate the instantaneous frequency of the response within a T-F unit simply as half the inverse of the interval between zero-crossings of the response [4], assuming that the response is approximately sinusoidal. Note that a sinusoidal function crosses zero twice within a period.

For classification, we use a multilayer perceptron (MLP) with one hidden layer [34] to compute $P(H_0|r_{cm}(\tau))$ for each filter channel. The desired output of the MLP is 1 if the corresponding T-F unit is target dominant and 0 otherwise (i.e., the IBM). When there are sufficient training samples, the trained MLP yields a good estimate of $P(H_0|r_{cm}(\tau))$ [7]. In this paper, the MLP for each channel is trained with a corpus that includes all the utterances from the training part of the TIMIT database [13] and 100 intrusions. These intrusions include crowd noise and environmental sounds, such as wind, bird chirp, and ambulance alarm.[1]. Utterances and intrusions are mixed at 0-dB SNR to generate training samples; the target is a speech utterance and interference is either a nonspeech intrusion or another utterance. We use *Praat* [5] which is a standard pitch estimation algorithm, to estimate the target pitch from a premixed target utterance. The number of units in the hidden layer is determined using cross-validation. Specifically, we divide the training samples equally into two sets, one for training and the other for validation. The number of units in the hidden layer is chosen to be the minimum such that adding more units in the hidden layer will not yield any significant performance improvement on the validation set. Since most obtained MLPs have five units in their hidden layers, we let every MLP have five hidden units for uniformity.

[1]The intrusions are posted at http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

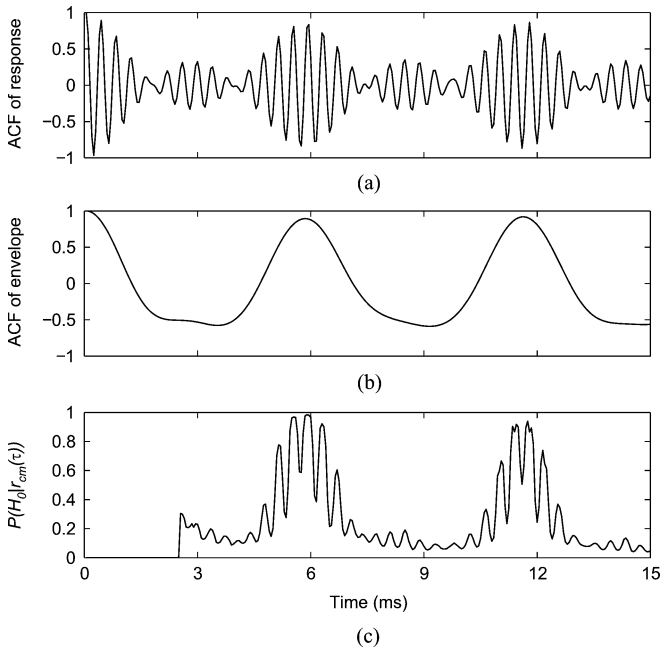


Fig. 1. Autocorrelation function and target probability given pitch. (a) ACF of the filter response within a T-F unit in a channel centered at 2.5 kHz. (b) Corresponding ACF of the response envelope. (c) Probability of the unit being target dominant given target pitch period τ.

Fig. 1(a) and (b) shows the sample ACFs of a filter response and the response envelope in a T-F unit. The input is a female utterance, "That noise problem grows more annoying each day," from the TIMIT database. This unit corresponds to the channel with the center frequency of 2.5 kHz and the time frame from 790 to 810 ms. Fig. 1(c) shows the corresponding $P(H_0|r_{cm}(\tau))$ for different τ values. The maximum of $P(H_0|r_{cm}(\tau))$ is located at 5.87 ms, the pitch period of the utterance at this frame.

The obtained MLPs are used to label individual T-F units according to (6). Fig. 2(a) shows the resulting error rate by channel for all the mixtures in a test corpus (see Section V-B). The error rate is the average of false acceptance and false rejection. As shown in the figure, with features derived from individual T-F units, we can label about 70%–90% of the units correctly across the whole frequency range. In general, T-F units in the low-frequency range are labeled more accurately than those in the high-frequency range. Fig. 2 also shows the error rate by using only subsets of the features from the feature set, $r_{cm}(\tau)$. As shown in this figure, the ACF values at the pitch point and instantaneous frequencies provide complementary information. The response envelope is more indicative than the response itself in the high-frequency range. Best results are obtained when all the six features are used.

It is worth noting that this study represents the first classification-based approach (specifically MLP) to pitch-based

$$r_{cm}(\tau) = \big(A(c,m,\tau), \quad \bar{f}(c,m)\tau - \mathrm{int}(\bar{f}(c,m)\tau), \quad \mathrm{int}(\bar{f}(c,m)\tau), \\ A_E(c,m,\tau), \quad \bar{f}_E(c,m)\tau - \mathrm{int}(\bar{f}_E(c,m)\tau), \quad \mathrm{int}(\bar{f}_E(c,m)\tau)\big) \quad (3)$$

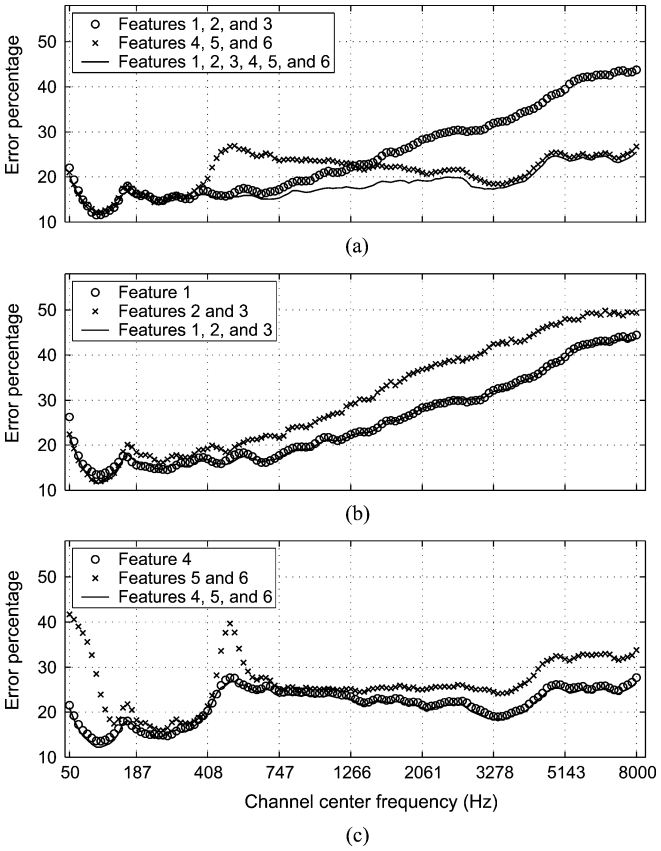


Fig. 2. Error percentage in T-F unit labeling using different subsets of six features (see text for definitions) given target pitch. (a) Comparison between all six features. (b) Comparison between the first three features. (c) Comparison between the last three features.

grouping. Casting grouping as classification affords us the flexibility of using a 6-D feature vector that combines autocorrelation and instantaneous-frequency measures. Besides using MLPs, we have considered modeling the distribution of $r_{cm}(\tau)$ using a Gaussian mixture model as well as a support vector machine based classifier [15]. However, the results are not better than using the MLPs.

B. Multiple Harmonic Sources

When interference contains one or several harmonic signals, there are time frames where both target and interference are pitched. In such a situation, it is more reliable to label a T-F unit by comparing the period of the signal within the unit with both the target pitch period and the interference pitch period. In particular, u_{cm} should be labeled as target if the target period not only matches the period of the signal but also matches better than the interference period, i.e.,

$$\begin{cases} P(H_0|r_{cm}(\tau_S(m))) > P(H_1|r_{cm}(\tau'_S(m))) \\ P(H_0|r_{cm}(\tau_S(m))) > 0.5 \end{cases} \quad (7)$$

where $\tau'_S(m)$ is the pitch period of the interfering sound at frame m. We use (7) to label T-F units for all the mixtures of two utterances in the test corpus. Both target pitch and interference pitch are obtained by applying *Praat* to clean utterances. Fig. 3 shows the corresponding error rate by channel, compared with using only the target pitch to label T-F units. As shown in the figure, better performance is obtained by using the pitch values of both speakers.

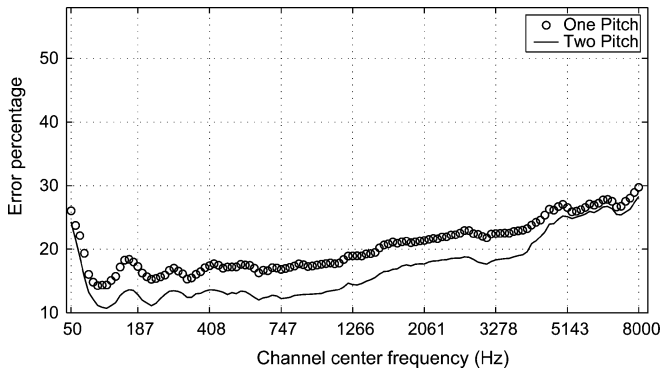


Fig. 3. Percentage of error in T-F unit labeling for two-voice mixtures using target pitch only or both target and interference pitch.

C. Unit Labeling With Information From a Neighborhood of T-F Units

Labeling a T-F unit using only the local information within the unit still produces a significant amount of error. Since speech signal is wideband and exhibits temporal continuity, neighboring T-F units potentially provide useful information for unit labeling. For example, a T-F unit surrounded by target-dominant units is also likely target dominant. Therefore, we consider information from a local context. Specifically, we label u_{cm} as target if

$$P(H_0|\{P(H_0|r_{c'm'}(\tau_S(m')))\}) > 0.5, \quad |c'-c| \le N_c, \quad |m'-m| \le N_m \quad (8)$$

where N_c and N_m define the size of the neighborhood along frequency and time, respectively, and $\{P(H_0|r_{c'm'}(\tau_S(m')))\}$ is the vector that contains the $P(H_0|r_{cm}(\tau_S(m)))$ values of the T-F units within the neighborhood. Again, for each frequency channel, we train an MLP with one hidden layer to calculate the probability $P(H_0|\{P(H_0|r_{c'm'}(\tau_S(m')))\})$ using the $P(H_0|r_{cm}(\tau_S(m)))$ values within the neighborhood as features.

The key here is to determine the appropriate size of a neighborhood. Again, we divide the training samples equally into two sets and use cross-validation to determine N_c and N_m. This cross-validation procedure suggests that $N_c = 8$ and $N_m = 2$ define an appropriate size of the neighborhood. By utilizing information from neighboring channels and frames, we reduce the average percentage of false rejection across all channels from 20.8% to 16.7% and the average percentage of false acceptance from 13.3% to 8.7% for the test corpus. The hidden layer of such a trained MLP has two units, also determined by cross-validation. Note that when both target and interference are pitched, we label a T-F unit according to (7) with probability $P(H_0|\{P(H_0|r_{c'm'}(\tau_S(m')))\})$ and $P(H_1|\{P(H_1|r_{c'm'}(\tau'_S(m')))\})$.

Since $P(H_0|r_{cm}(\tau_S(m)))$ is derived from $r_{cm}(\tau_S(m))$, we have also considered using $r_{cm}(\tau_S(m))$ directly as features. The resulting MLPs are much more complicated, but yield no performance gain.

IV. Pitch Determination Given Target Mask

A. Integration Across Channels

Given an estimated mask of the voiced target, the task here is to estimate target pitch. Let $L(m) = \{L(c,m), \forall c\}$ be the set of binary mask labels at frame m, where $L(c,m)$ is 1 if u_{cm} is active and 0 otherwise. A frequently used method for pitch determination is to pool autocorrelations across all the channels and then identify a dominant peak in the summary correlogram—the summation of ACFs across all the channels [11]. The estimated pitch period at frame m, $\tau_S(m)$, is the lag corresponding to the maximum of the summary ACF in the plausible pitch range. This simple method of pitch estimation is not very robust when interference is strong because the autocorrelations in many channels exhibit spurious peaks not corresponding to the target period. One may solve this problem by removing interference-dominant T-F units, i.e., calculating the summary correlogram only with active T-F units:

$$A(m,\tau) = \sum_c A(c,m,\tau)L(c,m). \tag{9}$$

Similar to the ACF of the filter response, the profile of the probability that unit u_{cm} is target dominant given pitch period τ, $P(H_0|r_{cm}(\tau))$, also tends to have a significant peak at the target period when u_{cm} is truly target dominant [see Fig. 1(c)]. One can use the corresponding summation of $P(H_0|r_{cm}(\tau))$

$$SP_m(\tau) = \sum_c P(H_0|r_{cm}(\tau))L(c,m) \tag{10}$$

to identify the pitch period at frame m as the maximum of the summation in the plausible pitch range.

We apply the above two methods for pitch estimation to two utterances from the test corpus, one from a female speaker and the other from a male speaker. These two utterances are mixed with 20 intrusions at 0-dB SNR. In this estimation, we use the IBM at the voiced frames of the target utterance to estimate a pitch period at each frame. The percentages of estimation error for the two methods are shown in the first two columns of the first row of Table I. We use the pitch obtained by applying *Praat* to the clean target as the ground truth of the target pitch. An error occurs when the estimated pitch period and the pitch period obtained from *Praat* differ by more than 5%. As shown in the table, using the summation of $P(H_0|r_{cm}(\tau))$ performs much better than using the summary ACF for the female utterance. Both methods, especially the one using the summary ACF, perform better on the male utterance. This is because the ACF and $P(H_0|r_{cm}(\tau))$ in target-dominant T-F units all exhibit peaks not only at the target pitch period, but also at time lags multiple the pitch period. As a result, their summations have significant peaks not only at the target pitch period, but also at its integer multiples, especially for a female voice, making pitch estimation difficult.

B. Differentiating True Pitch Period From its Integer Multiples

To differentiate a target pitch period from its integer multiples for pitch estimation, we need to take the relative locations of possible pitch candidates into consideration. Let τ_1 and τ_2 be two pitch candidates. We train an MLP-based classifier that selects the better one from these two candidates using their relative locations and $SP_m(\tau)$ as features, i.e., $(\tau_1/\tau_2,\ SP_m(\tau_1),\ SP_m(\tau_2))$. The training set is the same as described in Section III-A In constructing the training data, we obtain $SP_m(\tau)$ at each time frame from all the target-dominant T-F units. In each training sample, the two pitch candidates are the true target pitch period and the lag of another peak of $SP_m(\tau)$ within the plausible pitch range. Without loss of generality, we let $\tau_1 < \tau_2$. The desired output is 1 if τ_1 is the true pitch period and 0 otherwise. The obtained MLP has three units in the hidden layer. We use the obtained MLP to select the better one from the two candidates as follows: if the output of the MLP is higher than 0.5, we consider τ_1 as the better candidate; otherwise, we consider τ_2 as the better candidate.

TABLE I
Error Rate of Different Pitch Estimation Given Ideal Binary Mask

Method	Summary ACF		Summary $P(H_0\|r_{cm}(\tau))$		Classifier	
	F	M	F	M	F	M
Without temporal continuity	39.6	17.1	18.1	17.2	15.6	17.6
With temporal continuity	31.8	16.3	14.8	15.8	12.7	16.8

The target pitch is estimated with the classifier as follows.

- Find all the local maxima in $SP_m(\tau)$ within the plausible pitch range as pitch candidates. Sort these candidates according to their time lags from small to large and let the first candidate be the current estimated pitch period $\tau_S(m)$.
- Compare the current estimated pitch period with the next candidate using the obtained MLP and update the pitch estimate if necessary.

The percentage of pitch estimation errors with the classifier is shown in the last column of the first row in Table I. The classifier reduces the error rate on the female utterance but slightly increases the error rate on the male utterance.

C. Pitch Estimation Using Temporal Continuity

Speech signals exhibit temporal continuity, i.e., their structure, such as frequency partials, tends to last for a certain period of time corresponding to a syllable or phoneme, and the signals change smoothly within this period. Consequently, the pitch and the ideal binary mask of a target utterance tend to have good temporal continuity as well. We found that less than 0.5% of consecutive frames have more than 20% relative pitch changes for utterances in our training set [15]. Thus, we utilize pitch continuity to further improve pitch estimation as follows.

First, we check the reliability of the estimated pitch based on temporal continuity. Specifically, for every three consecutive frames, $m-1$, m, and $m+1$, if the pitch changes are all less than 20%, i.e.,

$$\begin{cases} |\tau_S(m) - \tau_S(m-1)| < 0.2\min(\tau_S(m), \tau_S(m-1)) \\ |\tau_S(m) - \tau_S(m+1)| < 0.2\min(\tau_S(m), \tau_S(m+1)) \end{cases} \tag{11}$$

the estimated pitch periods in these three frames are all considered reliable.

Second, we reestimate unreliable pitch points by limiting the plausible pitch range using neighboring reliable pitch points. Specifically, for two consecutive time frames, $m-1$ and m, if $\tau_S(m)$ is reliable and $\tau_S(m-1)$ is unreliable, we reestimate $\tau_S(m-1)$ by limiting the plausible pitch range for $\tau_S(m-1)$ to be $[0.8\tau_S(m), 1.2\tau_S(m)]$, and vice versa. Another possible situation is that $\tau_S(m)$ is unreliable while both $\tau_S(m-1)$ and $\tau_S(m+1)$ are reliable. In this case, we use $\tau_S(m-1)$ to limit the plausible pitch range of $\tau_S(m)$ if the mask at frame m is more similar to the mask at frame $m-1$ than the mask at frame $m+1$, i.e.,

$$\sum_c L(c,m)L(c,m-1) > \sum_c L(c,m)L(c,m+1) \quad (12)$$

otherwise, $\tau_S(m+1)$ is used to reestimate $\tau_S(m)$. Then the reestimated pitch points are considered as reliable and used to estimate unreliable pitch points in their neighboring frames. This reestimation process stops when all the unreliable pitch points have been reestimated.

The second row in Table I shows the effect of incorporating temporal continuity in pitch estimation as described above. Using temporal continuity yields consistent performance improvement, especially for the female utterance.

V. Iterative Procedure

Our tandem algorithm first generates an initial estimate of pitch contours and binary masks for up to two sources; a pitch contour refers to a consecutive set of pitches that is considered to be produced by the same sound source. The algorithm then improves the estimation of pitch contours and masks in an iterative manner.

A. Initial Estimation

In this step, we first generate up to two estimated pitch periods in each time frame. Since T-F units dominated by a periodic signal tend to have high cross-channel correlations of filter responses or response envelopes, we only consider T-F units with high cross-channel correlations in this estimation. Let $\tau_{S,1}(m)$ and $\tau_{S,2}(m)$ represent two estimated pitch periods at frame m, and $L_1(m)$ and $L_2(m)$ the corresponding labels of the estimated masks. We first treat all the T-F units with high cross-channel correlations as dominated by a single source. That is,

$$L_1(c,m) = \begin{cases} 1, & C(c,m) > 0.985 \text{ or } C_E(c,m) > 0.985 \\ 0, & \text{else.} \end{cases} \quad (13)$$

We then assign the time delay supported by most active T-F units as the first estimated pitch period. A unit u_{cm} is considered supporting a pitch candidate τ if the corresponding $P(H_0|r_{cm}(\tau))$ is higher than a threshold. Accordingly we have

$$\tau_{S,1}(m) = \arg\max_\tau \sum_c L_1(c,m) \cdot \mathrm{sgn}(P(H_0|r_{cm}(\tau)) - \theta_P) \quad (14)$$

where

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

and θ_P is a threshold. Intuitively, we can set θ_P to 0.5. However, such a threshold may not position the estimated pitch period close to the true pitch period because $P(H_0|r_{cm}(\tau))$ tends to be higher than 0.5 in a relatively wide range centered at the true pitch period [see Fig. 1(c)]. In general, θ_P needs to be much higher than 0.5 so that we can position $\tau_{S,1}(m)$ accurately. On the other hand, θ_P cannot be too high, otherwise most active T-F units cannot contribute to this estimation. We found that 0.75 is a good compromise that allows us to accurately position $\tau_{S,1}(m)$ without ignoring many active T-F units.

The above process yields an estimated pitch at many time frames where the target is not pitched. The estimated pitch point at such a frame is usually supported by only a few T-F units unless the interference contains a strong harmonic signal at this frame. On the other hand, estimated pitch points corresponding to target pitch are usually supported by many T-F units. In order to remove spurious pitch points, we discard a detected pitch point if the total number of channels supporting this pitch point is less than a threshold. We found that an appropriate threshold is 7 from analyzing the training data set (see Section III-A). Most spurious pitch points are thus removed. At the same time, some true pitch points are also removed, but most of them will be recovered in the following iterative process.

With the estimated pitch period $\tau_{S,1}(m)$, we reestimate the mask $L_1(m)$ as

$$L_1(c,m) = \begin{cases} 1, & P(H_0|r_{cm}(\tau_{S,1}(m))) > 0.5 \\ 0, & \text{else.} \end{cases} \quad (15)$$

Then we use the T-F units that do not support the first pitch period $\tau_{S,1}(m)$ to estimate the second pitch period, $\tau_{S,2}(m)$. Specifically, we use (16) shown at the bottom of the next page. We let

$$\tau_{S,2}(m) = \arg\max_\tau \sum_c L_2(c,m) \cdot \mathrm{sgn}(P(H_0|r_{cm}(\tau)) - \theta_P). \quad (17)$$

$$L_2(c,m) = \begin{cases} 1, & P(H_0|r_{cm}(\tau_{S,1}(m))) \le \theta_P \textit{ and } (C(c,m) > 0.985 \textit{ or } C_E(c,m) > 0.985) \\ 0, & \text{else.} \end{cases} \quad (16)$$

Again, if fewer than seven T-F units support $\tau_{S,2}(m)$, we set it to 0. Otherwise, we reestimate $L_2(m)$ as

$$L_2(c,m) = \begin{cases} 1, & P(H_0|r_{cm}(\tau_{S,2}(m))) > 0.5 \\ 0, & \text{else.} \end{cases} \tag{18}$$

Here, we estimate up to two pitch points at one frame; one can easily extend the above algorithm to estimate pitch points of more sources if needed.

After the above estimation, our algorithm combines the estimated pitch periods into pitch contours based on temporal continuity. Specifically, for estimated pitch periods in three consecutive frames, $\tau_{S,k_1}(m-1)$, $\tau_{S,k_2}(m)$, and $\tau_{S,k_3}(m+1)$, where k_1, k_2, and k_3 are either 1 or 2, they are combined into one pitch contour if they have good temporal continuity and their associated masks also have good temporal continuity. That is, see (19) shown at the bottom of the page. The remaining isolated estimated pitch points are considered unreliable and set to 0. Note that requiring only the temporal continuity of pitch periods cannot prevent connecting pitch points from different sources, since the target and interference may have similar pitch periods at the same time. However, it is very unlikely that the target and interference have similar pitch periods *and* occupy the same frequency region at the same time. In most situations, pitch points that are connected according to (19) do correspond to a single source. As a result of this step, we obtain multiple pitch contours and each pitch contour has an associated T-F mask.

B. Iterative Estimation

In this step, we first reestimate each pitch contour from its associated binary mask. A key step in this estimation is to expand estimated pitch contours based on temporal continuity, i.e., using reliable pitch points to estimate potential pitch points at neighboring frames. Specifically, let τ_k be a pitch contour and $L_k(m)$ the associated mask. Let m_1 and m_2 be the first and the last frame of this pitch contour. To expand τ_k, we first let $L_k(m_1-1) = L_k(m_1)$ and $L_k(m_2+1) = L_k(m_2)$. Then we reestimate τ_k from this new mask using the algorithm described in Section IV-B. Reestimated pitch periods are further verified according to temporal continuity described in Section IV-C except that we use (19) instead of (11) for continuity verification. If the corresponding source of contour τ_k is pitched at frame m_1-1, our algorithm likely yields an accurate pitch estimate at this frame. Otherwise, the reestimated pitch period at this frame usually cannot pass the continuity check, and as a result it is discarded and τ_k still starts from frame m_1. The same applies to the estimated pitch period at frame m_2+1. After expansion and reestimation, two pitch contours may have the same pitch period at the same frame and therefore they are combined into one pitch contour.

Then, we reestimate the mask for each pitch contour as follows. First, we compute the probability of each T-F unit dominated by the corresponding source of a pitch contour k, as described in Section III-C. Then, we estimate the mask for contour k according to the obtained probabilities, as shown in (20) at the bottom of the page.

Usually, the estimation of both pitch and mask converges after a small number of iterations, typically smaller than 20. Sometimes this iterative procedure runs into a cycle where there are slight cyclic changes for both estimated pitch and estimated mask after each iteration. In our implementation, we stop the procedure after it converges or 20 iterations.

C. Incorporating Segmentation

So far, unit labeling does not take into account of T-F segmentation, which refers to a stage of processing that breaks the auditory scene into contiguous T-F regions each of which contains acoustic energy mainly from a single sound source [39]. By producing an intermediate level of representation between individual T-F units and sources, segmentation has been demonstrated to improve segregation performance [16]. Here, we apply a segmentation step after the iterative procedure stops. Specifically, we employ a multiscale onset/offset based segmentation algorithm [18] that produces segments enclosed by detected onsets and offsets. After segments are produced, we form T-segments each of which is a subset of a T-F segment within an individual frequency channel or the longest section of consecutive T-F units within the same frequency channel of a T-F segment. T-segments strike a reasonable balance between accepting target and rejecting interference [15], [19]. With obtained T-segments, we label the T-F units within a T-segment wholly as target if 1) more than half of T-segment energy is included in the voiced frames of the target, and 2) more than half of the T-segment energy in the voiced frames is included in the

$$\begin{cases} |\tau_{S,k_2}(m) - \tau_{S,k_1}(m-1)| < 0.2\min(\tau_{S,k_2}(m), \tau_{S,k_1}(m-1)) \\ |\tau_{S,k_2}(m) - \tau_{S,k_3}(m+1)| < 0.2\min(\tau_{S,k_2}(m), \tau_{S,k_3}(m+1)) \\ \sum_c L_{k_2}(c,m)L_{k_1}(c,m-1) > 0.5\max(\sum_c L_{k_2}(c,m), \sum_c L_{k_1}(c,m-1)) \\ \sum_c L_{k_2}(c,m)L_{k_3}(c,m+1) > 0.5\max(\sum_c L_{k_2}(c,m), \sum_c L_{k_3}(c,m+1)) \end{cases} \tag{19}$$

$$L_k(c,m) = \begin{cases} 1, & k = \arg\max_{k'} P(H_0|\{P(H_0|r_{c'm'}(\tau_{k'}(m')))\}) \text{ and} \\ & P(H_0|\{P(H_0|r_{c'm'}(\tau_k(m')))\}) > 0.5 \\ 0, & \text{else} \end{cases} \tag{20}$$

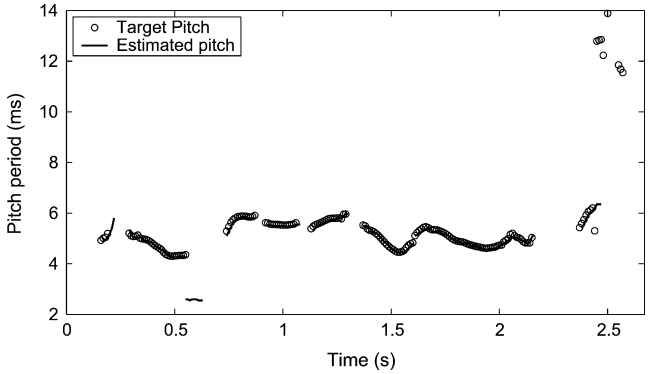


Fig. 4. Estimated pitch contours for the mixture of one female utterance and crowd noise.

active T-F units according to (20). If a T-segment fails to be labeled as the target, we still treat individual active T-F units as the target.

Fig. 4 shows the detected pitch contours for a mixture of the female utterance used in Fig. 1 and crowd noise at 0-dB SNR. The mixture is illustrated in Fig. 5, where Fig. 5(a) and (b) shows the cochleagram and the waveform of the female utterance and Fig. 5(c) and (d) the cochleagram and the waveform of the mixture. In Fig. 4, we use the pitch points detected by *Praat* from the clean utterance as the ground truth of the target pitch. As shown in the figure, our algorithm correctly estimates most of target pitch points. At the same time, it also yields one pitch contour for interference (the one overlapping with no target pitch point). Fig. 5(e) and (g) shows the obtained masks for the target utterance in the mixture without and with incorporating segmentation, respectively. Comparing the mask in Fig. 5(e) with the ideal binary mask shown in Fig. 5(i), we can see that our system is able to segregate most voiced portions of the target without including much interference. These two masks yield similar resynthesized targets in the voiced intervals, as shown in Fig. 5(f) and (j). By using T-segments, the tandem algorithm is able to recover even more target energy, but at the expense of adding a small amount of the interference, as shown in Fig. 5(g) and (h). Note that the output consists of several pitch contours and their associated masks. To determine whether a segregated sound is part of the target speech is the task of sequential grouping [6], [39], which is beyond the scope of this paper. The masks in Fig. 5(e) and (g) are obtained by assuming perfect sequential grouping.

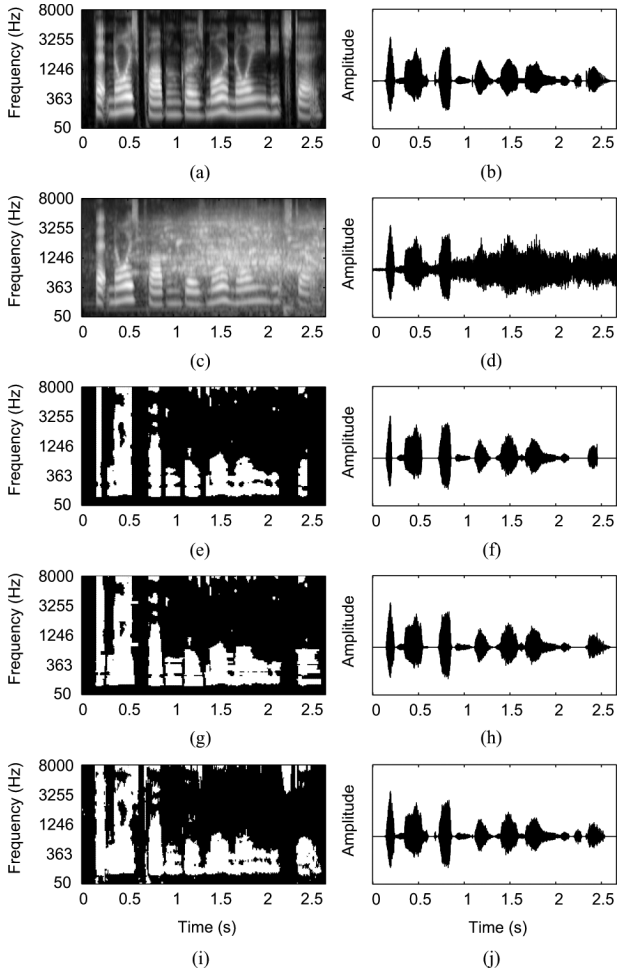


Fig. 5. Segregation illustration. (a) Cochleagram of a female utterance showing the energy of each T-F unit with brighter pixel indicating stronger energy. (b) Waveform of the utterance. (c) Cochleagram of the utterance mixed with a crowd noise. (d) Waveform of the mixture. (e) Mask of segregated voiced target where 1 is indicated by white and 0 by black. (f) Waveform of the target resynthesized with the mask in (e). (g) Mask of the target segregated after using T-segments. (h) Waveform of the target resynthesized with the mask in (g). (i) Ideal binary mask. (j) Waveform of the target resynthesized from the IBM in (i).

VI. Evaluation

As mentioned earlier, the tandem algorithm produces a set of pitch contours and their associated binary masks. This section separately evaluates pitch estimation and voiced speech segregation.

A. Pitch Estimation

We first evaluate the tandem algorithm on pitch determination with utterances from the FDA Evaluation Database [1]. This database was collected for evaluating pitch determination algorithms and provides accurate target pitch contours derived from laryngograph data. The database contains utterances from two speakers, one male and one female. We randomly select one sentence that is uttered by both speakers. These two utterances are mixed with a set of 20 intrusions at different SNR levels. These intrusions are: N1 – white noise, N2 – rock music, N3 – siren, N4 – telephone, N5 – electric fan, N6 – clock alarm, N7 – traffic noise, N8 – bird chirp with water flowing, N9 – wind, N10 – rain, N11 – cocktail party noise, N12 – crowd noise at a playground, N13 – crowd noise with music, N14 – crowd noise with clap, N15 – babble noise (16 speakers), $N16 \sim N20 - 5$ different utterances (see [15] for details). These intrusions have a considerable variety: some are noise-like (N9, N11) and some contain strong harmonic sounds (N3, N8). They form a reasonable corpus for testing the capacity of a CASA system in dealing with various types of interference.

Fig. 6(a) shows the average correct percentage of pitch determination with the tandem algorithm on these mixtures at different SNR levels. In calculating the correct detection percentage, we only consider estimated pitch contours that

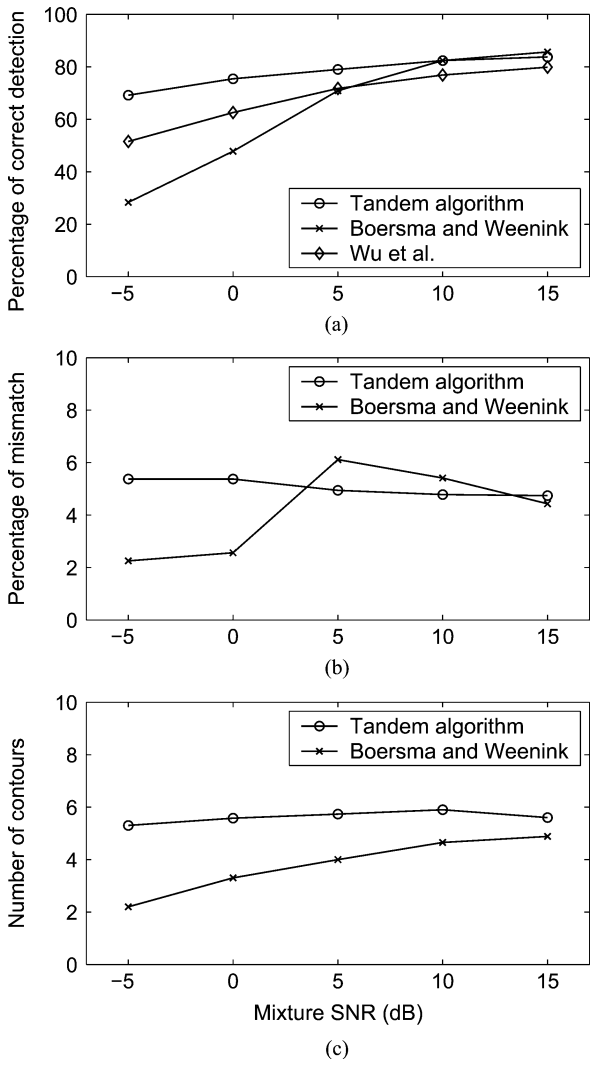


Fig. 6. Results of pitch determination for different algorithms. (a) Percentage of correct detection. (b) Percentage of mismatch. (c) Number of contours that match the target pitch.

match the target pitch: an estimated pitch contour matches the target pitch if at least half of its pitch points match the target pitch, i.e., the target is pitched at these corresponding frames and the estimated pitch periods differ from the true target pitch periods by less than 5%. As shown in the figure, the tandem algorithm is able to detect 69.1% of target pitch even at -5 dB SNR. The correct detection rate increases to about 83.8% as the SNR increases to 15 dB. In comparison, Fig. 6(a) also shows the results using *Praat* and from a multiple pitch tracking algorithm by Wu *et al.* [40], which produces competitive performance [23], [24]. Note that the Wu *et al.* algorithm does not yield continuous pitch contours. Therefore, the correct detection rate is computed by comparing estimated pitch with the ground truth frame by frame. As shown in the figure, the tandem algorithm performs consistently better than the Wu *et al.* algorithm at all SNR levels. The tandem algorithm is more robust to interference compared to *Praat*, whose performance is good at SNR levels above 10 dB, but drops quickly as SNR decreases.

TABLE II
PERFORMANCE OF THE TANDEM ALGORITHM WITH RESPECT TO THE NUMBER OF ITERATIONS

Iteration No.	0	1	2	3	4	Convergence
Percentage of detection	63.0	66.3	67.8	68.8	68.9	69.1
SNR (dB)	6.97	7.44	7.62	7.77	7.89	8.04

Besides the detection rate, we also need to measure how well the system separates pitch points of different sources. Fig. 6(b) shows the percentage of mismatch, which is the percentage of estimated pitch points that do not match the target pitch among the pitch contours matching the target pitch. An estimated pitch point is counted as mismatch if either target is not pitched at the corresponding frame or the difference between the estimated pitch period and the true period is more than 5%. As shown in the figure, the tandem algorithm yields a low percentage of mismatch, which is slightly lower than that of *Praat* when the SNR is above 5-dB SNR. In lower SNR levels, *Praat* has a lower percentage of mismatch because it detects fewer pitch points. Note that the Wu algorithm does not generate pitch contours, and the mismatch rate is 0. In addition, Fig. 6(c) shows the average number of estimated pitch contours that match target pitch contours. The actual average number of target pitch contours is 5. The tandem algorithm yields an average of 5.6 pitch contours for each mixture. This shows that the algorithm well separates target and interference pitch without dividing the former into many short contours. *Praat* yields almost the same numbers of contours as the actual ones at 15-dB SNR. However, it detects fewer contours when the mixture SNR drops. Overall, the tandem algorithm yields better performance than either *Praat* or the Wu *et al.* algorithm, especially at low SNR levels.

To illustrate the advantage of the iterative process for pitch estimation, we present the average percentage of correct detection for the above mixtures at -5 dB with respect to the number of iterations in the first row of Table II. Here 0 iteration corresponds to the result of initial estimation, and "convergence" corresponds to the final output of the algorithm. As shown in the table, the initial estimation already gives a good pitch estimate. The iterative procedure, however, is able to improve the detection rate, especially in the first iteration. Overall, the procedure increases the detection rate by 6.1 percentage points. It is worth pointing out that the improvement varies considerably among different mixtures, and the largest improvement is 22.1 percentage points.

B. Voiced Speech Segregation

The performance of the system on voiced speech segregation has been evaluated on a test corpus containing 20 target utterances from the test part of the TIMIT database mixed with the 20 intrusions described in the previous section. Note that the utterances in the test part of the TIMIT are produced by different speakers from those producing the utterances in the training part of the corpus.

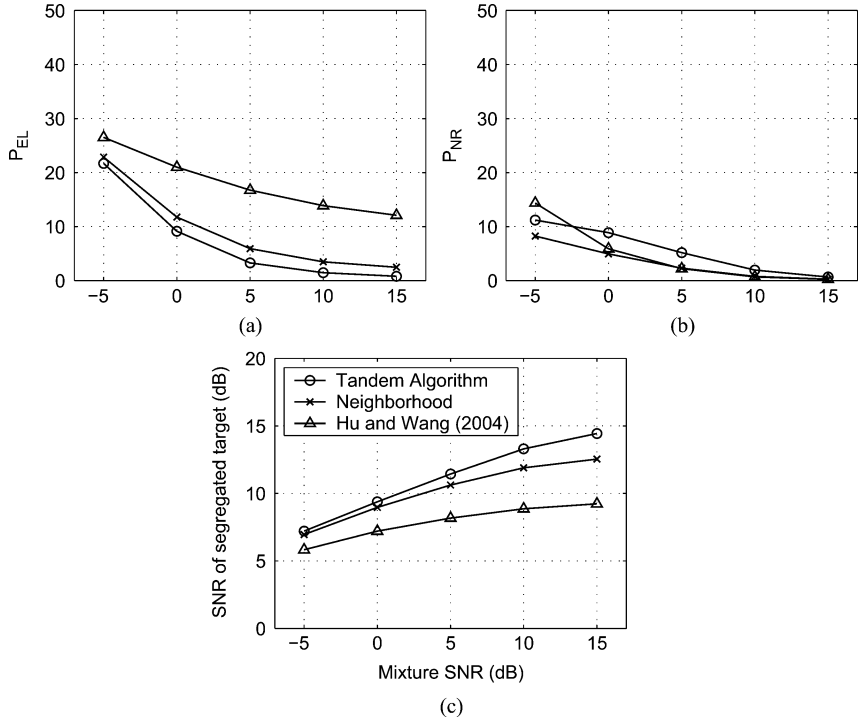


Fig. 7. Results of voiced speech segregation. (a) Percentage of energy loss on voiced target. (b) Percentage of noise residue. (c) SNR of segregated voiced target.

Estimated target masks are obtained by assuming perfect sequential grouping. Since our computational goal here is to estimate the IBM, we evaluate segregation performance by comparing the estimated mask to the IBM with two measures [16].

- The percentage of energy loss P_{EL} which measures the amount of energy in the active T-F units that are labeled as interference relative to the total energy in the active units.
- The percentage of noise residue P_{NR} which measures the amount of energy in the inactive T-F units that are labeled as the target relative to the total energy in the inactive units.

P_{EL} and P_{NR} provide complementary error measures of a segregation system and a successful system needs to achieve low errors in both measures.

In addition, to compare waveforms directly we measure the SNR of the segregated voiced target in decibels [16]

$$\text{SNR} = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n [s(n) - \hat{s}_V^2(n)]^2} \tag{21}$$

where $s(n)$ is the target signal resynthesized from the IBM and $\hat{s}_V(n)$ is the segregated voiced target.

The results from our system are shown in Fig. 7. Each point in the figure represents the average value of 400 mixtures in the test corpus at a particular SNR level. Fig. 7(a) and (b) shows the percentages of energy loss and noise residue. Note that since our goal here is to segregate voiced target, the P_{EL} values here are only for the target energy at the voiced frames of the target. However, the IBM used in (21) is constructed for the entire target, which contains both voiced and unvoiced speech, i.e., the lack of unvoiced speech segregation in this study is accounted for (or penalized) in the SNR measure.

As shown in the figure, our system segregates 78.3% of voiced target energy at −5-dB SNR and 99.2% at 15-dB SNR. At the same time, 11.2% of the segregated energy belongs to intrusion at −5 dB. This number drops to 0.6% at 15-dB SNR. Fig. 7(c) shows the SNR of the segregated target. Our system obtains an average 12.2-dB gain in SNR when the mixture SNR is −5 dB. This gain drops to 3.3 dB when the mixture SNR is 10 dB. Note that at 15 dB, our system does not improve the SNR because most unvoiced speech is not segregated. Fig. 7 also shows the result of the algorithm without using T-segments in the final estimation step ("Neighborhood"). As shown in the figure, the corresponding segregated target loses more target energy, but contains less interference. The SNR performance is a little better by incorporating T-segments.

Fig. 7 also shows the performance using our previous voiced speech segregation system [16], which is a representative CASA system. Because the previous system can only track one pitch contour of the target, in this implementation we provide target pitch estimated by applying *Praat* to clean utterances. As shown in the figure, the previous system yields a lower percentage of noise residue, but has a much higher percentage of energy loss. As clearly shown in the SNR measure of Fig. 7(c), even with provided target pitch, the previous system does not perform as well as the tandem algorithm, especially at higher input SNR levels.

To illustrate the effect of iterative estimation, we present the average SNR for the mixtures of two utterances and all the intrusions at −5-dB SNR in the second row of Table II. On average, the tandem algorithm improves the SNR by 1.07 dB. Again,

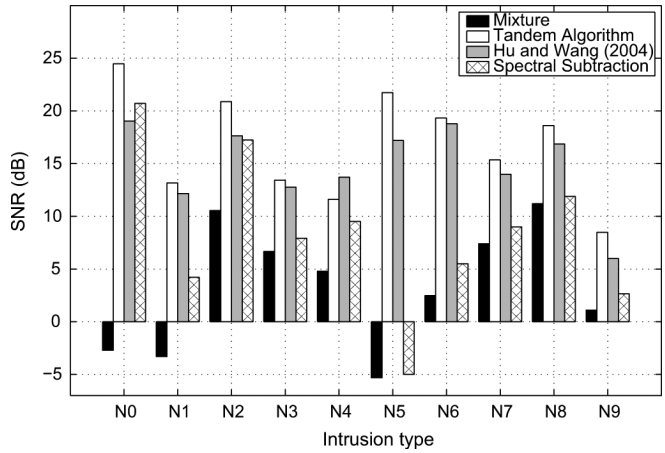


Fig. 8. SNR results for segregated speech and original mixtures for a corpus of voiced speech and various intrusions.

the SNR improvement varies considerably among different mixtures, and the largest improvement is 7.27 dB.

As an additional benchmark, we have evaluated the tandem algorithm on a corpus of 100 mixtures composed of ten target utterances mixed with ten intrusions [10]. Every target utterance in the corpus is totally voiced and has only one pitch contour. The intrusions have a considerable variety; specifically they are: N0 –1kHz 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. The average SNR of the entire corpus is 3.28 dB. This corpus is commonly used in CASA for evaluating voiced speech segregation [8], [16], [26]. The average SNR for each intrusion is shown for the tandem algorithm in Fig. 8, compared with those of the original mixtures, our previous system, and a spectral subtraction method. Note that here our previous system extracts pitch contours from mixtures instead of using pitch contours extracted from clean utterances with *Praat*. Spectral subtraction is a standard method for speech enhancement [21] (see also [16]). The tandem algorithm performs consistently better than spectral subtraction, and our previous system except for N4. On average, the tandem algorithm obtains a 13.4-dB SNR gain, which is about 1.9 dB better than our previous system and 8.3 dB better than spectral subtraction.

VII. Discussion

Classification, which is a form of supervised learning, plays an important role in the tandem algorithm, particularly for pitch-based unit labeling. As with any supervised approach, one wonders how well our algorithm generalizes to datasets not used in training. There are reasons to expect that the tandem algorithm has only weak dependency on the specific training corpus. First, pitch-based features used in IBM estimation capture general acoustic characteristics, not corpus-specific attributes. Second, in pitch determination, several parameter values, e.g., the 20% relative pitch changes in deciding pitch continuity, are chosen based on general observations such as temporal continuity. As described in Section VI, our evaluation results are obtained using test signals that are not used in training, and pitch tracking results in Fig. 6 and the benchmark segregation results in Fig. 8 are reported with no retraining using two datasets different from the TIMIT corpus used in training. A very recent study [20] has successfully applied the tandem algorithm to pitch tracking for segregating the IEEE sentences [22] without any retraining or change.

Although we have emphasized the iterative nature of the tandem algorithm, the algorithm also includes a specific method to jump start the iterative process, which gives an initial estimate of both pitch and mask with reasonable quality. In general, the performance of the algorithm depends on the initial estimate, and better initial estimates would lead to better performance. Even with a poor estimate, which is unavoidable in very low SNR conditions, our algorithm can still improve the initial estimate during the iterative process, e.g., through the pitch contour expansion described in Section V-B. The results in Section VI show that the tandem algorithm performs well even when the input SNR is -5 dB.

In terms of computational complexity, the main cost of the tandem algorithm arises from autocorrelation calculation and envelope extraction in the feature extraction stage and the iterative estimation of pitch contours and IBM consumes just a small fraction of the overall cost. We implemented both tasks in the frequency domain and their time complexity is $O(N \log N)$, where N is the number of samples in an input signal. Note that these operations need to be performed for each filter channel, and our system employs 128 channels. On the other hand, since feature extraction takes place in different filter channels independently, substantial speedup can be achieved through parallel computing.

This study concentrates on voiced speech, and does not deal with unvoiced speech. In a recent paper, we developed a model for separating unvoiced speech from nonspeech interference on the basis of auditory segmentation and feature-based classification [19]. This unvoiced segregation system operates on the output of voiced speech segregation, which was provided by Hu and Wang [17] assuming the availability of target pitch contours. The system in [17] is a simplified and slightly improved version of [16]. We have substituted the voiced segregation component of [19] by the tandem algorithm [15]. The combined system produces segregation results for both voiced and unvoiced speech that are as good as those reported in [19], but with detected pitch contours rather than ground-truth pitch contours (see [15] for details).

A natural speech utterance contains silent gaps and other intervals masked by interference. In practice, one needs to group the utterance across such time intervals. This is the problem of sequential grouping [6], [39]. This study does not address the problem of sequential grouping. The system in [19] handles the situation of nonspeech interference but not applicable to mixtures of multiple speakers. Sequentially grouping segments or masks could be achieved by using speech recognition in a top-down manner (also limited to nonspeech interference) [2] or by speaker recognition using trained speaker models [35]. Nevertheless, these studies are not yet mature, and substantial effort is needed in the future to fully address the problem of sequential grouping.

VIII. Conclusion

We have proposed an algorithm that estimates target pitch and segregates voiced target in tandem. This algorithm iteratively improves the estimation of both target pitch and voiced target. The tandem algorithm is novel not only for its iterative nature but also for the methods proposed for pitch-based labeling of T-F units and pitch estimation from a given binary mask. The tandem algorithm is robust to interference and produces good estimates of both pitch and voiced speech even in the presence of strong interference. Systematic evaluation shows that the tandem algorithm performs significantly better than previous CASA and speech enhancement systems. Together with our previous system for unvoiced speech segregation [19], we have a complete CASA system to segregate speech from various types of nonspeech interference.

Acknowledgment

The authors would like to thank Z. Jin for his assistance in manuscript formatting.

References

[1] P. C. Bagshaw, S. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in *Proc. Eurospeech*, 1993, pp. 1003–1006.

[2] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, pp. 5–25, 2005.

[3] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proc. IEEE*, vol. 80, pp. 520–538, 1992.

[4] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications," *Proc. IEEE*, vol. 80, no. 4, pp. 540–568, Apr. 1992.

[5] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," 2004.

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

[7] J. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, Architectures, and Applications*, F. Fogelman-Soulie and J. Herault, Eds. New York: Springer, 1989, pp. 227–236.

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

[9] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.

[10] M. Cooke, *Modelling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[11] A. de Cheveigne, "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley and IEEE Press, 2006, pp. 45–79.

[12] H. Dillon, *Hearing Aids*. New York: Thieme, 2001.

[13] J. Garofolo *et al.*, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," National Inst. of Standards and Technol., 1993, NISTIR 4930.

[14] H. Helmholtz, *On the Sensation of Tone*, 2nd ed. New York: Dover, 1863.

[15] G. Hu, "Monaural speech organization and segregation," Ph.D dissertation, Biophysics Program, Ohio State Univ., Columbus, 2006.

[16] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[17] G. Hu and D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, E. Hansler and G. Schmidt, Eds. Heidelberg, Germany: Springer, 2006, pp. 485–515.

[18] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[19] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[20] K. Hu and D. L. Wang, "Incorporating spectral subtraction and noise type for unvoiced speech segregation," in *Proc. IEEE ICASSP*, 2009, pp. 4425–4428.

[21] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language processing: A Guide to Theory, Algorithms, and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[22] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[23] A. Khurshid and S. L. Denham, "A temporal-analysis-based pitch estimation system for noisy speech with a comparative study of performance of recent systems," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1112–1124, Sep. 2004.

[24] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, and S. Sagayama, "Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1135–1145, May 2007.

[25] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, pp. 1673–1682, 2008.

[26] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2014–2023, Nov. 2006.

[27] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–16, 1997.

[28] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[29] S. G. Nooteboom, "The prosody of speech: Melody and rhythm," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Oxford, U.K.: Blackwell, 1997, pp. 640–673.

[30] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Applied Psychology Unit, 1988, 2341.

[31] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP J. Audio Speech Music Process.*, vol. 2007, p. 15, 2007, Article 84186.

[32] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[33] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Commun.*, vol. 21, pp. 191–207, 1997.

[34] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing*, D. E. Rumelhart and J. L. McClell, Eds. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[35] Y. Shao, "Sequential Organization in Computational Auditory Scene Analysis," Ph.D. dissertation, Dept. of Comput. Sci. Eng., Ohio State Univ., Columbus, OH, 2007.

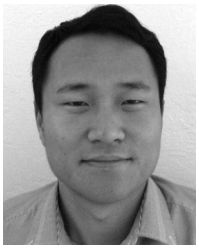
[36] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Commun.*, vol. 45, pp. 410–423, 2005.

[37] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

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

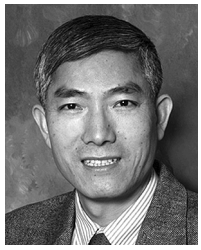
[39] *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley and IEEE Press, 2006.

[40] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.



Guoning Hu (M'05) received the B.S. and M.S. degrees in physics from Nanjing University, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree in biophysics from The Ohio State University, Columbus, in 2006.

He is currently with AOL Truveo Video Search, San Francisco, CA. His research interests include speech segregation, computational auditory scene analysis, and statistical machine learning.



DeLiang Wang (M'90–SM'01–F'04) received the B.S. and M.S. degrees from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1991, all in computer science.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology, Harvard University, Cambridge, MA. From October 2006 to June 2007, he was a Visiting Scholar at Oticon A/S, Denmark. His research interests include machine perception and neurodynamics.

Dr. Wang received the National Science Foundation Research Initiation Award in 1992, the Office of Naval Research Young Investigator Award in 1996, and the Helmholtz Award from the International Neural Network Society in 2008. He also received the 2005 Outstanding Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS.