Unvoiced Speech Segregation From Nonspeech Interference via CASA and Spectral Subtraction

Ke Hu, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—While a lot of effort has been made in computational auditory scene analysis to segregate voiced speech from monaural mixtures, unvoiced speech segregation has not received much attention. Unvoiced speech is highly susceptible to interference due to its relatively weak energy and lack of harmonic structure, and hence makes its segregation extremely difficult. This paper proposes a new approach to segregation of unvoiced speech from nonspeech interference. The proposed system first removes estimated voiced speech, and the periodic part of interference based on cross-channel correlation. The resultant interference becomes more stationary and we estimate the noise energy in unvoiced intervals using segregated speech in neighboring voiced intervals. Then unvoiced speech segregation occurs in two stages: segmentation and grouping. In segmentation, we apply spectral subtraction to generate time–frequency segments in unvoiced intervals. Unvoiced speech segments are subsequently grouped based on frequency characteristics of unvoiced speech using simple thresholding as well as Bayesian classification. The proposed algorithm is computationally efficient, and systematic evaluation and comparison show that our approach considerably improves the performance of unvoiced speech segregation.**

***Index Terms*—Bayesian classification, computational auditory scene analysis (CASA), nonspeech interference, spectral subtraction, unvoiced speech segregation.**

I. Introduction

SPEECH reaching our ears is almost never pure in the real world. Acoustic interference, such as fan noise, music, or another voice, poses a serious problem for many applications including automatic speech recognition [1] and hearing aid design [8]. While humans are remarkably adept in separating a particular sound from a mixture of many sources, such a task remains a major challenge for machines [36]. Monaural speech segregation refers to the task of separating speech from interference using a single microphone. This is a particularly difficult task because only one recording is available and one cannot explore the spatial information of sources present in multi-microphone situations. In a monaural case, one has to rely on the intrinsic properties of speech, such as harmonic structure and onset to perform segregation [4]. Research employing these features has made considerable advances in voiced speech segregation for anechoic [5], [12], [21] and reverberant conditions [18]. In contrast, the unvoiced speech segregation problem has not been much studied (see [13] for an exception) and remains a big challenge. In this paper, we study monaural segregation of unvoiced speech from nonspeech interference.

Speech enhancement methods have been proposed to enhance noisy speech based on a single recording [24]. Representative algorithms include spectral subtraction, Wiener filtering, minimum mean square error-based estimator, and subspace analysis. Such methods work with the whole noisy utterance and therefore have the potential to deal with unvoiced speech. However, speech enhancement methods often make assumptions about the statistical properties of interference, which limits their ability in dealing with general interference. For example, the assumption of stationary noise is often made, which is not true in typical real-world situations where interference can change abruptly over a short period of time. Another class of techniques, called model-based speech separation, focuses on modeling source patterns and formulates separation as an estimation problem in a probabilistic framework. By representing observations using source models, such a system either directly estimates individual speech utterances or derives a time–frequency (T-F) mask to segregate each source. For example, Radfar *et al.* [29] proposed a maximum-likelihood method to estimate vocal-tract-related filter responses, which are then combined with excitation signals to reconstruct individual speech signals based on a source-filter model. Along the same line, a composite source model in the form of Gaussian mixture is used in [28] to model individual speakers and a minimum mean square error estimator is used to segregate each source. Model-based techniques have the potential to segregate unvoiced speech, but the assumption that the mixture consists of only speech utterances limits the scope of their applications. It is also unclear that how the system performs when two speakers utter unvoiced speech simultaneously.

Computational auditory scene analysis (CASA) aims to achieve sound organization based on perceptual principles [36]. Segmentation and grouping are the two main stages of CASA. In segmentation, the input is decomposed to segments, each of which is a contiguous T-F region originating mainly from a single sound source. The grouping stage combines segments that likely arise from the same source into a stream. Ideal binary mask (IBM) has been suggested as a main goal of CASA [34]. The IBM is a binary T-F matrix where each T-F unit is labeled either as target dominant with a value of 1 or as interference dominant with a value of 0. IBM can

Manuscript received October 20, 2009; revised September 14, 2010; accepted November 02, 2010. Date of publication November 18, 2010; date of current version June 01, 2011. This work was supported in part by the National Science Foundation (NSF) under Grant IIS-0534707, in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155, and in part by the VA Biomedical Laboratory Research and Development Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

The authors are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: huk@cse.ohio-state.edu; dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2010.2093893

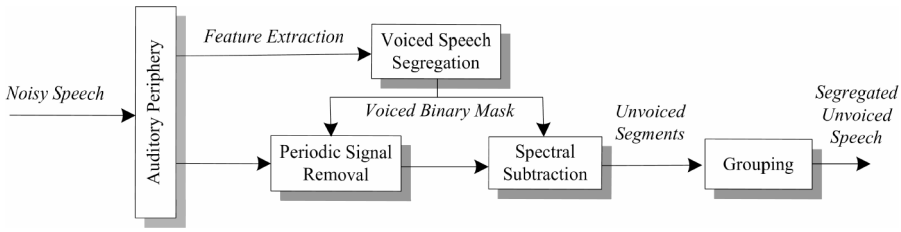


Fig. 1. Schematic diagram of the proposed unvoiced speech segregation system. The system first performs voiced speech segregation. The segregated voiced speech and periodic portions of interference are then removed in a periodic signal removal stage. Unvoiced speech segregation then occurs in two stages: segmentation and grouping. In segmentation, the system performs spectral subtraction on noise estimated using the voiced binary mask. Unvoiced speech segments are subsequently grouped to form an unvoiced speech stream.

be constructed by comparing the signal-to-noise ratio (SNR) within each T-F unit against a local criterion (LC). It is shown that IBM achieves optimal SNR gain under certain conditions [22]. Subject tests have shown that speech segregated by IBM leads to dramatic intelligibility improvements for both normal-hearing and hearing-impaired listeners [6], [20], [37].

As a subset of consonants, unvoiced speech consists of unvoiced fricatives, stops, and affricates [19], [32]. Recently, Hu and Wang studied unvoiced speech segregation and successfully extracted a majority of unvoiced speech from nonspeech interference [13]. They utilized onset and offset cues to extract candidate unvoiced speech segments. Acoustic–phonetic features are then used to separate unvoiced speech in a classification stage. In [15], we incorporated spectral subtraction and noise type in unvoiced speech segregation. The evaluation shows promising results but the grouping method involves a large amount of training and is designed for mixtures only at one SNR level.

In this paper, we extend the idea of spectral subtraction based segmentation in [15] and propose a simpler framework for unvoiced speech segregation. First, our system segregates voiced speech by using a tandem algorithm [14]. We then remove voiced speech as well as periodic components in interference based on cross-channel correlation. As periodic portions are removed, the interference is expected to become more stationary. Then unvoiced speech segregation occurs in two stages: segmentation and grouping. In segmentation, we first estimate interference energy in unvoiced intervals by averaging the mixture energy in inactive units (those labeled as 0) in neighboring voiced intervals. Estimated noise energy is then used by spectral subtraction to generate unvoiced T-F segments. In the grouping stage, unvoiced speech segments are extracted based on thresholding or classification.

The rest of the paper is organized as follows. The next section presents peripheral processing, feature extraction and voiced speech segregation. Unvoiced speech segregation is described in Section III. Systematic evaluation and comparison are provided in Section IV and we conclude the paper in Section V.

II. Background and Voiced Speech Segregation

Our system is shown in Fig. 1. Noisy speech is first analyzed by an auditory periphery model [36] and voiced speech is segregated using a tandem algorithm [14]. The segregated voiced speech is subsequently removed along with the periodic portions of interference from the mixture, and unvoiced speech segmentation and grouping are then carried out.

A. Peripheral Processing and Feature Extraction

To analyze noisy speech, the system first decomposes the signal in the frequency domain using a bank of 64 gammatone filters with center frequencies equally distributed on the equivalent rectangular bandwidth scale from 50 to 8000 Hz [27]. The gammatone filterbank is a standard model of cochlear filtering. The output of each channel is then transduced by the Meddis hair cell model [25]. Details of auditory peripheral processing can be found in [36]. In the time domain, channel outputs are decomposed to 20-ms time frames with a 10-ms frame shift. The resulting time–frequency representation is called a cochleagram [36].

Let $u_{c,m}$ denote a T-F unit at channel c and frame m, and $r(c,m)$ the corresponding hair cell output. We calculate a normalized correlogram by using the following autocorrelation function (ACF)

$$A(c,m,\tau) = \frac{\sum_{n=-N/2+1}^{N/2} r(c,mN/2+n)r(c,mN/2+n+\tau)}{\sqrt{\sum_{n=-N/2+1}^{N/2} r^2(c,mN/2+n) \sum_{n=-N/2+1}^{N/2} r^2(c,mN/2+n+\tau)}} \quad (1)$$

where τ denotes the time delay, and the frame length N is 320 corresponding to 20 ms with a sampling frequency of 16 kHz. Within each frame, the ACF carries periodicity information of the filter response and the delay corresponding to the global peak of the ACF indicates the dominant pitch period. In implementation, time delay τ varies between 0 ms and 12.5 ms, which includes the plausible pitch range of human speech.

Harmonics of voiced speech are resolved in the low-frequency range, but not at high frequencies. Each high-frequency filter responds to multiple harmonics so that the response is amplitude modulated and the envelope of the response fluctuates at the $F0$ (fundamental frequency) of the voiced speech [36]. Therefore, to encode unresolved harmonics, we extract the envelope of the response by half-wave rectification and bandpass filtering with the passband from 50 to 550 Hz [18]. The envelope ACF of $u_{c,m}$, $A_E(c,m,\tau)$, is then calculated similarly to (1).

Neighboring channels responding to the same harmonic or formant tend to have high cross-channel correlation [35]. We calculate the cross channel correlation between $u_{c,m}$ and $u_{c+1,m}$ by

$$C(c,m) = \frac{1}{L+1}\sum_{\tau=0}^{L} \hat{A}(c,m,\tau)\hat{A}(c+1,m,\tau) \quad (2)$$

where $\hat{A}(c,m,\tau)$ denotes the normalized ACF with zero mean and unity variance, and $L = 200$ corresponds to the maximum time delay of 12.5 ms. In addition, we calculate the cross-channel correlation of response envelope between $u_{c,m}$ and $u_{c+1,m}$, $C_E(c,m)$, similarly to (2).

B. Voiced Speech Segregation

After feature extraction, we use the tandem algorithm [10], [14] to estimate a voiced binary mask. The main purpose of estimating a voiced binary mask is to identify inactive T-F units in voiced intervals to estimate noise energy in unvoiced intervals.

Following [10], we extract a 6-dimensional feature vector for $u_{c,m}$

$$X_{c,m} = \begin{pmatrix} A(c,m,\tau_m) \\ \mathrm{int}\left(\overline{f}(c,m)\cdot\tau_m\right) \\ \left|\overline{f}(c,m)\cdot\tau_m - \mathrm{int}\left(\overline{f}(c,m)\cdot\tau_m\right)\right| \\ A_E(c,m,\tau_m) \\ \mathrm{int}\left(\overline{f}_E(c,m)\cdot\tau_m\right) \\ \left|\overline{f}_E(c,m)\cdot\tau_m - \mathrm{int}\left(\overline{f}_E(c,m)\cdot\tau_m\right)\right| \end{pmatrix}. \quad (3)$$

In (3), τ_m is the estimated pitch period at frame m. $A(c,m,\tau_m)$ measures periodicity similarity between the unit response and the estimated pitch at frame m. $\overline{f}(c,m)$ denotes the estimated average instantaneous frequency of the response within $u_{c,m}$, which is estimated using the zero-crossing rate of $A(c,m,\tau)$. The function $\mathrm{int}(x)$ returns the nearest integer. The product $\overline{f}(c,m)\cdot\tau_m$ provides another feature to determine the periodicity of a T-F unit, and its closest integer indicates a harmonic number. The third feature measures the deviation of the product from its nearest harmonic number. While the first three features in (3) are extracted from filter responses, the last three are extracted from response envelopes (indicated by the subscript E).

Given the pitch-based feature vector in (3), we train a multilayer perceptron (MLP) to label T-F units for each channel. The training samples are generated by mixing 100 utterances randomly selected from the training part of the TIMIT database [9] and 100 nonspeech interferences [11] at 0 dB. Feature extraction needs $F0$, which is extracted from clean speech utterances by Praat [3]. The IBM is generated with an LC of 0 dB and used to provide the desired output in training. All 64 MLPs have the same architecture of 6 input nodes, one hidden layer of 5 nodes and 1 output node according to [14]. The hyperbolic tangent activation function is used for both hidden and output layers. Since our system adopts a 64-channel filterbank in peripheral processing, we halve the frequency range in neighbor based unit labeling to 4 and retrain the MLP classifier. In addition, the thresholds in initial mask estimation are set to 0.945. In testing, the tandem algorithm performs pitch estimation and voiced speech segregation jointly.

III. Unvoiced Speech Segregation

The basic idea of our unvoiced speech segregation method is to capitalize on the segregated voiced speech to estimate interference energy. Since the estimated voiced binary mask contains inactive T-F units during voiced intervals, we utilize them to estimate noise energy and subtract it from the mixture during unvoiced intervals in order to form unvoiced segments. Before unvoiced segregation, we first remove periodic signals.

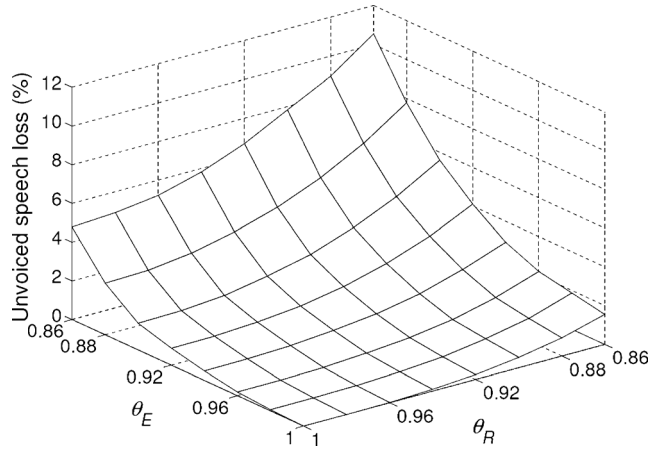


Fig. 2. Unvoiced speech energy loss as a function of thresholds for response and envelope cross-channel correlations. The horizontal axes represent two thresholds θ_R and θ_E, and the vertical axis represents the percent of unvoiced speech energy loss.

A. Periodic Signal Removal

Unvoiced speech is aperiodic in nature. Therefore, the T-F units that contain periodic signals do not originate from unvoiced speech and should be removed. Specifically, we consider unit $u_{c,m}$ to be dominated by a periodic signal if either of the following two conditions is satisfied: $u_{c,m}$ is included in the segregated voiced stream, or the unit has a high cross-channel correlation. The second condition stems from the observation that T-F units dominated by a periodic signal tend to have high cross-channel correlations [35]. The cross-channel correlation is deemed high if it is above a certain threshold

$$C(c,m) > \theta_R \quad \text{or} \quad C_E(c,m) > \theta_E. \quad (4)$$

Here, θ_R and θ_E are thresholds for the response and envelope cross-channel correlations, respectively. To maintain a balance between periodic signal removal and unvoiced speech preservation, the thresholds need to be carefully chosen. To find appropriate values, we vary both thresholds from 0.86 to 1 and calculate the percent of unvoiced speech energy loss. In this analysis, 100 speech sentences from the IEEE sentence database recorded by a single female speaker [17] are mixed with 15 nonspeech interferences (see Section IV for details) at 0 dB to generate mixtures. Different parts of an interfering signal are used in analysis and evaluation. Here, the first half of interference is mixed with speech for analysis or training, while in evaluation the second half is used. An interference is either cut or concatenated with itself to match the length of a corresponding speech signal. IBM is generated with an LC of 0 dB, and we use the portions in unvoiced intervals to represent ideally segregated unvoiced speech. To generate the unvoiced IBM, pitch contours are detected from clean speech using Praat. In addition, to exclude voiced speech which is not strongly periodic, we remove segments in the unvoiced IBM extending below 1 kHz. We calculate the percent of unvoiced speech lost with respect to total unvoiced speech in each noisy speech utterance and present the mean in Fig. 2. As shown in the figure, when both thresholds are set to 0.86, about 10% of unvoiced speech is wrongly removed. As the thresholds increase, less unvoiced speech is lost. To achieve a good compromise, we choose θ_R to be 0.9 and θ_E

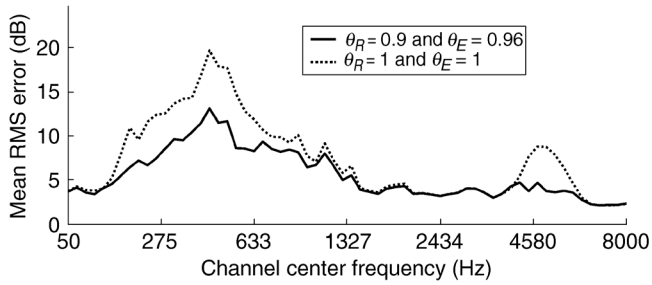


Fig. 3. Mean RMS errors of noise energy estimation over frequencies for bird chirp noise. The overall estimation performance with the chosen thresholds (solid line) is better than that without periodic signal removal (dotted line).

to be 0.96. As indicated by the figure, less than 2% of the unvoiced speech is lost in this case.

We have considered choosing different thresholds for different noise types. By analyzing the percentages of unvoiced speech loss for each noise type separately, we observe that, with the chosen thresholds, the loss percentages for different noises are all smaller than 6%. This indicates that the fixed thresholds perform well for individual noise types. As a result, we do not expect significant performance improvements by using different thresholds for different noise types. Of course, using fixed threshold values is desirable as it does not need detection of noise types, which would be required if thresholds need to be tuned based on noise type.

Based on the criterion in (4), we detect T-F units dominated by periodic signals and merge neighboring ones to form a mask. Together with the voiced binary mask obtained in Section II-B, we produce a periodic mask whereby active units are removed from the consideration of unvoiced speech grouping. Periodic signal removal serves two purposes. First, it reduces the possibility of false detection in unvoiced speech segregation. Second, the removal of periodic signal tends to make interference more stationary. Consequently, the noise estimated in voiced intervals is generalized to neighboring unvoiced intervals. To show how this process improves noise estimation, we calculate the root mean square (RMS) error of noise energy estimation for each channel with or without periodic signal removal. The RMS error is measured over unvoiced speech intervals, which are determined by the tandem algorithm. Here, 100 speech utterances different from those in the above analysis are randomly selected from the IEEE database and mixed with the bird chirp noise [11] at 0 dB for evaluation. Fig. 3 shows the mean RMS errors. The dotted line denotes the error with the cross-channel correlation thresholds set to 1, which amounts to no periodic signal removal. In contrast, the solid line represents the error with the chosen thresholds. The RMS error with periodic signal removal is uniformly smaller than that without the removal, especially at high frequencies where the energy of bird chirp noise is concentrated.

B. Unvoiced Speech Segmentation Based on Spectral Subtraction

After the removal of periodic signals, we deal with the mixture of only unvoiced speech and aperiodic interference. Obviously, the pitch-based feature vector in (3) cannot be used to segregate unvoiced speech. Our method first estimates the background noise and then removes it during unvoiced intervals. Without the periodic signals, we estimate the interference energy in an unvoiced interval by averaging the mixture energy within inactive T-F units in the two neighboring voiced intervals. For channel c, the interference energy (in dB) is estimated as shown in (5) at the bottom of the page, where $E_{dB}(c,i)$ denotes the energy within $u_{c,i}$ in dB, and $y(c,i)$ its estimated binary label. m_1 and m_2 are the indices of the first and last frames of the current unvoiced interval respectively, and l_1 and l_2 the frame lengths of the preceding and succeeding voiced intervals, respectively. For the unvoiced interval at the start or end of an utterance, estimation is only based on the succeeding or preceding voiced interval, respectively. In the situation where no inactive unit exists in the neighboring voiced intervals for certain channels, we search for the two further neighboring voiced intervals and continue this process until at least one of them contains inactive units. All detected inactive units are then used for estimation. If no inactive unit exists in this channel, the mixture energy of the first five frames is averaged to obtain the noise estimate. Besides averaging, we have tried linear interpolation and smoothing spline interpolation [7], but got no better performance.

Our segmentation method employs spectral subtraction, which is a widely used approach for enhancing signals corrupted by stationary noise [24]. Letting $X(c,m)$ be noisy speech energy and $\hat{N}(c,m)$ the estimated noise energy in $u_{c,m}$, we estimate the local SNR (in dB) in this unit as

$$\xi(c,m) = 10\log_{10}\left(\left[X(c,m) - \hat{N}(c,m)\right]^{+} \Big/ \hat{N}(c,m)\right) \tag{6}$$

where the function $[x]^{+} = x$ if $x \geq 0$ and $[x]^{+} = 0$ otherwise. Notice that $\hat{N}(c,m) = 10^{(\hat{N}_{dB}(c,m)/10)}$. A T-F unit is then labeled as 1 if $\xi(c,m)$ is greater than 0 dB, or 0 otherwise. Notice that estimating the local SNR using (6) is equivalent to performing power spectral subtraction [2], except that here we

$$\hat{N}_{\mathrm{dB}}(c,m) = \frac{\sum_{i=m_1-l_1}^{m_1-1} E_{\mathrm{dB}}(c,i)\cdot(1-y(c,i)) + \sum_{i=m_2+1}^{m_2+l_2} E_{\mathrm{dB}}(c,i)\cdot(1-y(c,i))}{\sum_{i=m_1-l_1}^{m_1-1}(1-y(c,i)) + \sum_{i=m_2+1}^{m_2+l_2}(1-y(c,i))}, \quad \text{for} \quad m \in [m_1, m_2] \tag{5}$$

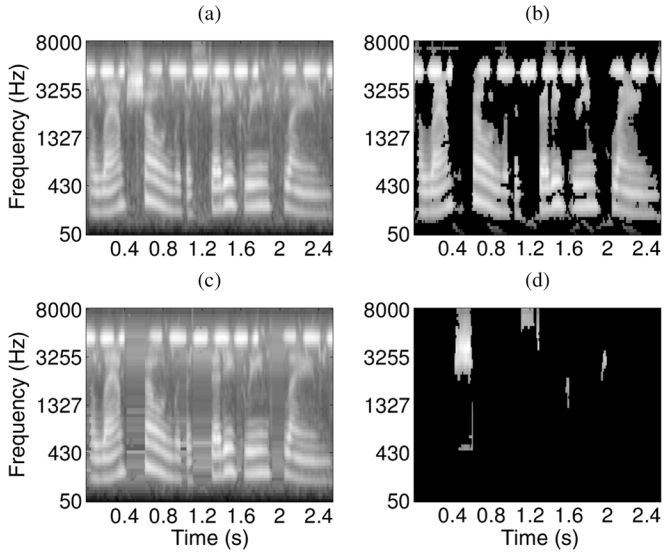


Fig. 4. Illustration of unvoiced speech segmentation via spectral subtraction. (a) Cochleagram of a female utterance, "The lamp shone with a steady green flame," mixed with the bird chirp noise at 0 dB. (b) Voiced speech as well as periodic portions of interference detected in the mixture. (c) The combination of (b) and estimated noise energy in voiced and unvoiced intervals. (d) Candidate unvoiced speech segments after spectral subtraction.

either keep or discard the mixture energy in $u_{c,m}$ depending on $\xi(c,m)$. We have investigated the over-subtraction technique proposed by Berouti *et al.* [2] to attenuate music noise, and found an over-subtraction factor of 2 to be a good tradeoff. Thus we double the noise estimate in (6) during labeling. Unvoiced speech segments are subsequently formed by merging neighboring active T-F units in the T-F domain.

As an illustration, Fig. 4(a) shows a T-F representation of the 0-dB mixture of a female utterance, "The lamp shone with a steady green flame," from the IEEE sentence database and the bird chirp noise, where a brighter unit indicates stronger energy. Fig. 4(b) shows the segregated voiced speech and the periodic portions of the interference detected using cross-channel correlation. Estimated noise in voiced and unvoiced intervals is shown in Fig. 4(c) together with detected periodic signals. Fig. 4(d) shows the extracted unvoiced speech segments based on the subtraction of Fig. 4(c) from Fig. 4(a) using (6).

C. Unvoiced Segment Grouping

Spectral subtraction based segmentation captures most of unvoiced speech, but some segments correspond to residual noise. To extract only unvoiced speech segments and remove residual noise is the task of grouping. Before grouping, let us analyze the characteristics of unvoiced speech. An unvoiced fricative is produced by forcing air through a constriction point in the vocal tract to generate turbulence noise [32]. In English, unvoiced fricatives consist of the labiodental (/f/), dental (/θ/), alveolar (/s/), and palatoalveolar /ʃ/. Except for the labiodental, the acoustic cavity of an unvoiced fricative is so small that resonance concentrates at high frequencies. For example, the alveolar fricative often has a spectral peak around 4.5 kHz, which corresponds to the natural frequency of its acoustic cavity. An unvoiced stop is generated by forming a complete closure in the vocal tract first and then releasing it abruptly [32]. At the stop release multiple acoustic events happen, including a transient, a burst of frication noise, and aspiration noise. As a result, the energy of an unvoiced stop usually concentrates in both middle (1.5–3 kHz) and high-frequency bands (3 kHz–8 kHz). The unvoiced affricate, /tʃ/, can be considered as a composite of a stop and a fricative. In summary, the energy of unvoiced speech often concentrates in the middle and high frequency ranges. This property, however, is not shared by nonspeech interference. To explore spectral characteristics of unvoiced speech and noise segments, we analyze their energy distributions with respect to frequency. Specifically, lower and upper frequency bounds of a segment are used to represent its frequency span. Notice that our task is to segregate only unvoiced speech; therefore, we consider voiced speech that is not strongly periodic as noise too. A statistical analysis is carried out using the 0-dB mixtures of 100 speech utterances and 15 interferences described in the first paragraph of Section III-A. Fig. 5(a) shows the normalized energy distribution of segments with respect to the segment lower bound and Fig. 5(b) the upper bound. In the plots, a white bar represents the aggregated energy of all unvoiced speech segments with a certain frequency bound and a black bar represents that of all interference segments. Energy bars are normalized to the sum of 1. For clear illustration, the bar with lower energy is displayed in front of the bar with higher energy for each frequency bound in the figure. The unvoiced IBM with an LC of 0 dB is used for ideal classification, i.e., segments with more than half of energy overlapping with the unvoiced IBM are considered as unvoiced speech and others as interference. We observe from the figure that unvoiced speech segments tend to reside at high frequencies while interference segments dominate at low frequencies. Interference is effectively removed at high frequencies probably because the corresponding noise estimate is relatively accurate due to weak voiced speech at these frequencies. Based on our analysis and acoustic–phonetic characteristics of unvoiced speech [32], we can simply select segments with a lower bound higher than 2 kHz or an upper bound higher than 6 kHz as unvoiced speech and remove others as noise. We call this grouping method thresholding.

We can also formulate grouping as a hypothesis test and perform Bayesian classification. Let S denote the segment to be classified. The two hypotheses are H_0: S is dominated by unvoiced speech, and H_1: S is dominated by interference. For classification, we construct three features for segment S

$$\mathbf{X}_S = \left(f_L^S, f_U^S, \|S\|\right) \tag{7}$$

where f_L^S and f_U^S denote the frequency lower and upper bounds of S, respectively. The third feature is the size (the number of T-F units) of segment S. We retain S as unvoiced speech if

$$P(H_0|\mathbf{X}_S) > P(H_1|\mathbf{X}_S). \tag{8}$$

As MLP directly estimates the *a posteriori* probability [26], we train an MLP to estimate $P(H_0|\mathbf{X}_S)$; note that $P(H_1|\mathbf{X}_S) = 1 - P(H_0|\mathbf{X}_S)$. Here, we adopt an SNR-based objective function in [18] for MLP training

$$J = \sum_S \left(d(S) - y(S)\right)^2 \cdot E(S) / \sum_S E(S) \tag{9}$$

where $E(S)$ denotes the energy in segment S, and $d(S)$ and $y(S)$ are the desired (binary) and actual MLP outputs, re-

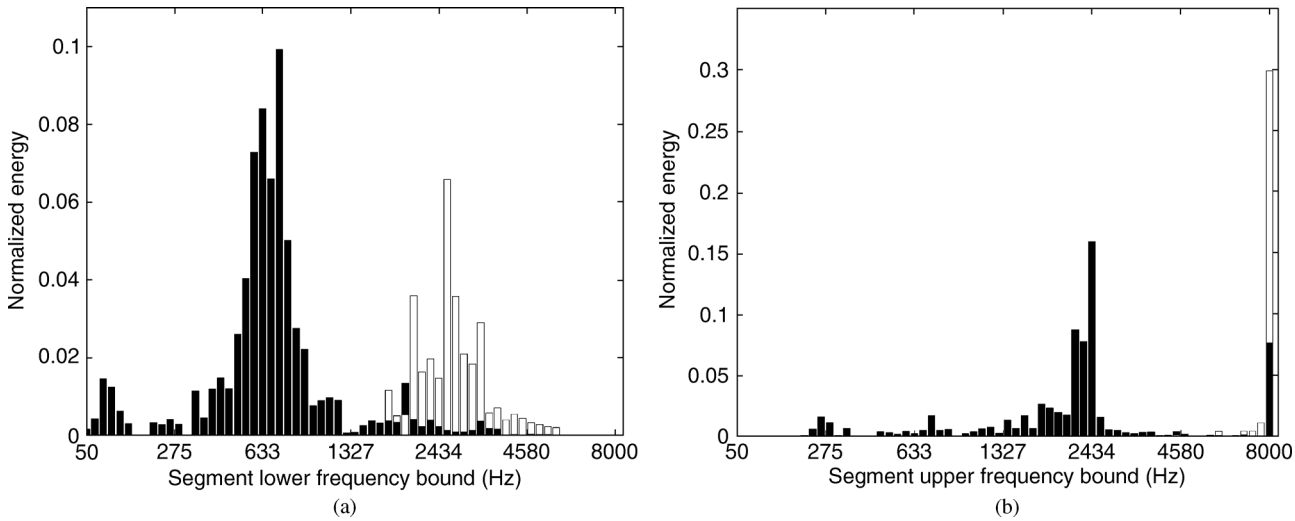


Fig. 5. Normalized energy distribution of unvoiced speech segments (white) and interference segments (black) over (a) segment lower bound and (b) segment upper bound.

spectively. This objective function penalizes labeling errors in segments with higher energy more than those with lower energy, hence maximizing the overall SNR. The configuration of the MLP is the same as that in Section II-B except that the hidden layer has three nodes as determined by ten-fold cross validation. The 0-dB mixtures described in the first paragraph of Section III-A are used for training and segments are compared with the unvoiced IBM to obtain desired labels. The performance of Bayesian classification is compared with that of simple thresholding in Section IV-A.

In addition, we have tried to incorporate the prior probability ratio in classification as in [13] but obtain no better performance. We have also considered using Bayesian classification of acoustic–phonetic features in [13] to group unvoiced segments. The performance did not improve maybe because of the assumption of independence among frames within a segment. Our features, on the other hand, are extracted from the whole segment. In terms of dimensionality, the acoustic phonetic feature used in [13] is 128-dimensional while ours is only 3-D. As a result, the MLP training for Bayesian classification using (7) is much faster.

IV. Evaluation and Comparison

We evaluate the proposed algorithm using a noisy speech corpus composed of 100 utterances and 15 nonspeech interferences. The 100 test sentences are randomly selected from those of the IEEE sentences not used in training (see Section III-C). All utterances are downsampled from 20 to 16 kHz and each is mixed with an individual interference at the SNR levels of -5, 0, 5, 10, and 15 dB. The interference set comprises electric fan (N1), white noise (N2), crowd noise at a playground (N3), crowd noise with clapping (N4), crowd noise with music (N5), rain (N6), babble noise (N7), rock music (N8), wind (N9), cocktail party noise (N10), clock alarm (N11), traffic noise (N12), siren (N13), bird chirp with water flowing (N14), and telephone ring (N15) [13]. They cover a wide variety of real-world noise types. As mentioned in Section III-A, the first half of an interference is mixed with speech to create mixtures in training or analysis, while in testing the second half is used.

The computational objective of our proposed system is to estimate the unvoiced IBM. Hence, we adopt the SNR measure in [14] and consider the resynthesized speech from the unvoiced IBM as the ground truth

$$\mathrm{SNR} = 10\log_{10}\left(\sum_n S_I^2[n] \Big/ \sum_n \left(S_I[n] - S_E[n]\right)^2\right) \quad (10)$$

where $S_I[n]$ and $S_E[n]$ are the signals resynthesized using the ideal and estimated unvoiced binary masks, respectively. The unvoiced IBM is determined by pitch contours extracted from clean speech signals using Praat. For estimation, pitch contours are detected from mixtures using the tandem algorithm. In both cases, an LC of 0 dB is used to generate the IBM for all SNR conditions. As mentioned earlier, to obtain only unvoiced IBM, segments extending below 1 kHz are removed unless they could correspond to unvoiced speech at high SNRs (above 10 dB) for some interferences.

A. SNR Performance

We evaluate the system performance based on simple thresholding described in Section III-C. To quantitatively evaluate the performance, an SNR gain is computed from the output SNR of segregated speech subtracted by the initial SNR of the mixture over unvoiced intervals. As mentioned earlier, a total of 100 mixtures are used for evaluation for each noise and input SNR condition. The SNR gains are shown in Table I. Our system achieves considerable SNR improvements for the large majority of noise and input SNR conditions, especially at low input SNRs. On average, the proposed system obtains an SNR gain of 18.5 dB when the input SNR is -5 dB. The SNR gain decreases gradually as the input SNR increases, and at 15-dB input SNR there is small degradation in a few noise conditions. Across all noise types and input SNR levels, the system generates an overall 10.8-dB SNR gain. It is worth noting that the performance of our system for nonstationary noises [e.g., cocktail party noise (N10) and siren (N13)] is not necessarily worse than for stationary noises, especially at relatively high-input SNR conditions. We have also evaluated

TABLE I
SNR GAIN (IN dB) AT DIFFERENT NOISE AND INPUT SNR CONDITIONS

Input SNR (dB)	Noise Types															Average
	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	
-5	22.7	15.9	20.4	21.9	18.2	16.1	9.2	17.3	20.1	19.4	17.3	21.7	20.9	18.5	18.1	18.5
0	19.3	12.6	17.7	19.6	14.1	13.3	11.1	16.1	18.6	17.2	13.4	19.0	16.2	14.8	14.3	15.8
5	12.8	8.6	13.8	14.3	10.7	10.0	10.6	12.6	9.9	13.8	9.2	15.3	9.5	9.0	7.1	11.1
10	4.8	4.5	9.3	9.4	7.6	6.4	8.3	8.3	5.4	9.2	3.4	10.4	5.3	4.9	5.0	6.8
15	-0.4	1.0	4.7	4.8	3.4	3.6	4.5	4.2	-0.2	3.6	-1.5	3.9	0.4	0.2	0.5	2.2

the system performance with different over-subtraction factors but got no improvement. In particular, when the factor is greater than 3, the overall SNR gain decreases gradually as the factor increases. It is probably because of the loss of unvoiced speech due to over-estimated noise.

In addition, we have evaluated the system performance using Bayesian classification and found that the classification method performs comparably with simple thresholding at all input SNR conditions. When averaged across different noises, the two methods perform almost equally. The lack of a significant improvement in classification is probably because the two frequency bounds chosen empirically are already very effective. Since simple thresholding does not require any training, this grouping method should be more desirable in real applications.

B. Comparisons

We compare our system (simple thresholding) with the unvoiced speech segregation system proposed by Hu and Wang in [13], the only previous system directly dealing with unvoiced speech segregation to our knowledge. In their system, segmentation is performed by multiscale onset–offset analysis and grouping is based on Bayesian classification as mentioned earlier. We retrain their MLP classifier using the 100 speech utterances mixed with 15 nonspeech interferences described in the first paragraph of Section III-A. The training and test conditions of the Hu and Wang system match exactly those of our system, i.e., the first half of each interference is used in training while the second half is for testing. In training, the unvoiced IBM provides the desired output. For both methods, the tandem algorithm is used for voiced speech segregation. The results are shown by solid curves in Fig. 6. Our proposed algorithm performs better than their system with an average of 1.6-dB SNR improvement over all input SNR levels. In terms of computational complexity, the proposed algorithm is much simpler than the Hu and Wang algorithm. First, spectral subtraction based segmentation is more efficient than the multiscale onset–offset analysis since the latter needs to analyze the signal in different scales. Second, grouping based on simple thresholding is computationally much simpler. It requires no training for MLP based segment removal and classification, which is time–consuming with 128-dimensional feature vectors in [13]. We have also tried a supervised learning algorithm [18] for voiced speech segregation. The supervised learning algorithm performs a little better than the tandem algorithm with training using the 100 speech utterances mixed with 15 nonspeech interferences described in the first paragraph of Section III-A. As a result, one might expect unvoiced segregation performance to improve slightly, but we observed that the system employing the supervised learning algorithm obtains almost the same results.

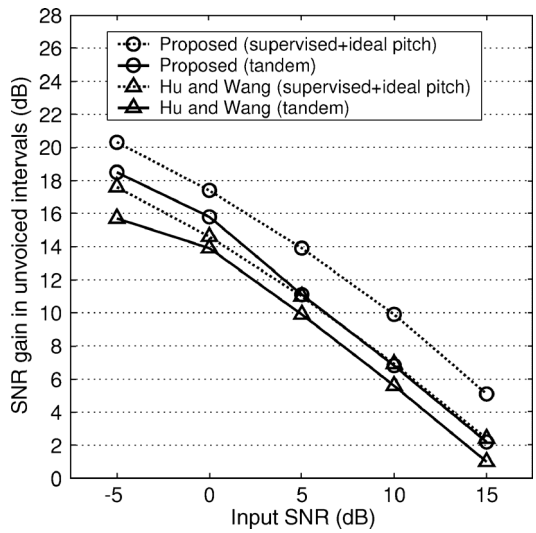


Fig. 6. Comparison in terms of SNR gain between the proposed algorithm and the Hu and Wang algorithm. Two kinds of pitch contours are used: 1) voiced speech and pitch contours detected using the tandem algorithm (solid line) and 2) voiced speech segregated using the supervised learning algorithm with ideal pitch contours (dotted line).

Errors in pitch tracking influence the determination of voiced and unvoiced intervals, hence likely degrading the unvoiced speech segregation performance. To evaluate how pitch tracking errors affect segregation performance, we perform unvoiced speech segregation using ideal pitch contours, which are extracted from clean speech utterances using Praat. As shown in Fig. 6, using ideal pitch contours in the supervised learning algorithm improves unvoiced speech segregation, and our system with simple thresholding obtains a larger SNR improvement over the Hu and Wang system: 2.8 dB on average.

The insensitivity to different voiced speech segregation methods with detected pitch suggests that our noise estimation is not very sensitive to voiced mask estimation. To further test how robust our system is, we have applied ideal voiced segregation. Specifically, the estimated binary mask is replaced by the IBM at voiced frames. As shown in Fig. 7, the system with ideal voiced mask information only performs slightly better. On average, it improves the SNR performance by only about 0.1 dB. With ideal pitch, the performance difference in terms of voiced mask is about 0.4 dB. This comparison shows that our system is not much affected by estimated voiced binary mask.

Since spectral subtraction plays a major role in the segmentation stage of our system, it is informative to compare our algorithm with speech enhancement methods. To isolate the ef-

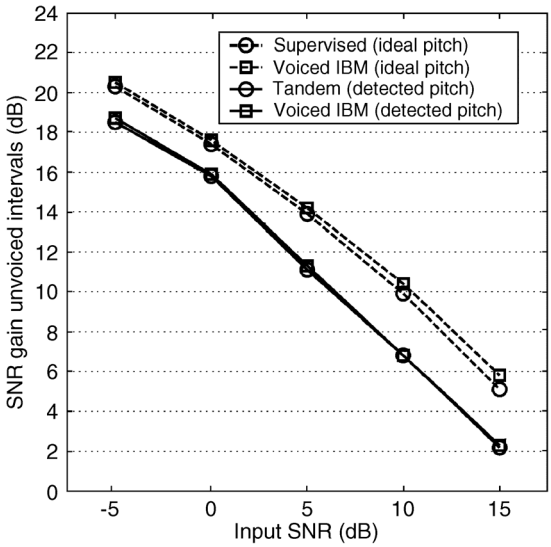


Fig. 7. SNR comparison between using estimated voiced binary mask and ideal voiced binary mask. Two pitch contours are used in voiced speech segregation: 1) pitch contours extracted by the tandem algorithm (solid line) and 2) ideal pitch contours extracted from clean speech utterance using Praat (dotted line).

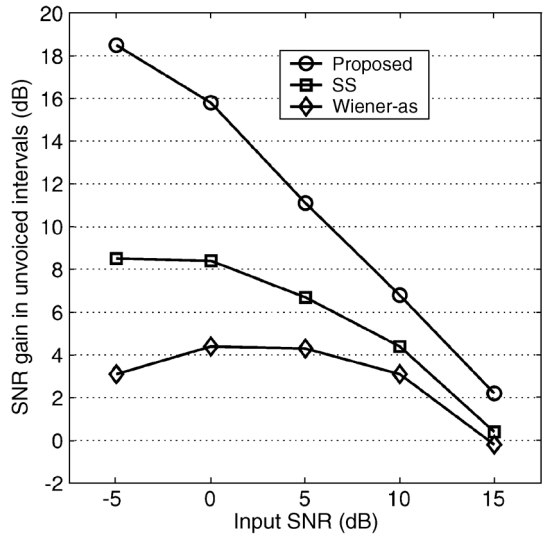


Fig. 8. Comparison with two speech enhancement methods at different SNR levels. The two representative methods are spectral subtraction (SS) and *a priori* SNR based Wiener algorithm (Wiener-as).

fects of the grouping stage of our CASA based system, we apply spectral subtraction alone to segregate unvoiced speech, i.e., the segments generated using spectral subtraction with an over-subtraction factor of 2 are directly combined to form an unvoiced stream. In addition, we also compare with a Wiener algorithm based on *a priori* SNR estimation (Wiener-as), which is reported as the best performing speech enhancement algorithm in speech intelligibility evaluations [16]. In this case, we binarize the amplitude gain in Wiener estimation with the threshold of 0.5 to generate segments and form a binary mask (see [20]). In both methods, noise is estimated in the same way as explained in Section III-B except that no periodic signal removal is carried out. As in our method of obtaining the unvoiced IBM, we remove the portions of the estimated unvoiced mask below 1 kHz to evaluate unvoiced speech segregation performance.

Fig. 8 shows the comparative results. As observed in the figure, the proposed algorithm performs much better than either of the two speech enhancement methods. In the case of using only spectral subtraction, the largest gap is about 10 dB when the input SNR is -5 dB and the gap is about 1.8 dB as the input SNR increases to 15 dB. The Wiener-as algorithm performs worse than spectral subtraction. We have also evaluated the SNR gains of the speech enhancement methods without binary masking, and only the Wiener-as method obtains about 1 dB improvement. Even in this case the performance gap from the proposed method is still large. It is worth noting that large gains at low input SNR levels are particularly useful for people with hearing loss [8]. Hence, the need to improve SNR in these conditions is more acute than at high-input SNRs.

Estimation and reduction methods have been proposed to deal with nonstationary noises in speech enhancement. For example, the algorithm in [31] trains codebooks for individual noises using *a priori* noise information and uses the codebooks to estimate speech and noise jointly. The system in [23] addresses noise tracking in highly nonstationary environments. Instead of building models using *a priori* noise information, this system relies on only noisy observations and utilizes harmonicity of voiced speech and unvoiced speech lengths to inform noise update. Since our system is designed specifically for separating unvoiced speech, direct comparisons with such speech enhancement methods are not appropriate. Nonetheless, we want to point out that our system deals with all interferences in a general way by first making them more stationary and then using general speech and noise characteristics for separation. As pointed out by the authors, the method in [23] may not work when noise exhibits harmonic properties. For a few common noises used (e.g., white and babble), our SNR gains are competitive although we should caution that test conditions and detailed SNR metrics are not the same.

TABLE II
AVERAGE PER-FRAME LABELING ERROR (%) IN IBM ESTIMATION

	Input SNR (dB)				
	-5	0	5	10	15
Overall	14.20	17.89	22.11	26.63	31.26
Miss	70.37	60.45	57.09	55.88	55.47
False alarm	2.62	3.56	4.56	5.61	6.08

Motivated by the relationship between intelligibility and labeling errors in IBM estimation [20], we have also evaluated our system performance in terms of error percentages in unit labeling. The overall percentage of mask error is calculated as the average error rate per frame for entire speech, counting flips from 0's to 1's and from 1's to 0's, relative to the IBM. These error rates are given in Table II. We have also examined two different types of error, misses and false alarms, which have been shown to have different impacts on speech intelligibility with false alarms to be particularly harmful [20]. Specifically, we compute the miss error as the per-frame average percentage of active units wrongly labeled as inactive ones, and the false alarm error as the per-frame average percentage of inactive units wrongly labeled as active ones. Results are also shown in Table II and indicate that miss errors are much more prevalent than false alarm errors in our system. In comparison with the

overall rates of the two representative speech enhancement algorithms examined in [20], our algorithm achieves considerably lower error rates.

V. Discussion

Unvoiced speech separation is a challenging task. Our proposed CASA system utilizes segregated voiced speech to assist unvoiced speech segregation. Specifically, the system first removes periodic signals from the noisy input and then estimates interference energy by averaging mixture energy within inactive T-F units in neighboring voiced intervals. The estimated interference is used by spectral subtraction to extract unvoiced segments, which are then grouped by either simple thresholding or Bayesian classification. A systematic comparison shows the proposed system outperforms a recent system in [13] over a wide range of input SNR levels. In addition, segmentation based on spectral subtraction is simpler and faster than multiscale onset-offset analysis, and grouping based on simple thresholding does not need MLP training. Our CASA based approach also performs substantially better than speech enhancement methods, indicating the effectiveness of a grouping stage.

In our study, the segregation performance is measured in terms of SNR gain in unvoiced intervals. Since unvoiced speech is generally much weaker than voiced speech in an utterance, high unvoiced SNR gains we have obtained will not directly translate to comparable improvements when measuring over whole utterances. However, unvoiced speech accounts for a significant portion of total speech and is important for speech intelligibility [13]. The lack of separate treatment of unvoiced speech could be a main reason for the well-known lack of speech intelligibility improvement of speech enhancement methods [16].

We use a 64-channel gammatone filterbank in T-F analysis. Compared with systems employing 128-channel filterbanks [13], [14], [18], the use of a 64-channel filterbank halves the computing time. In terms of segregation performance, we have observed comparable performance to that using a 128-channel filterbank. We have also reduced the number of channels in other algorithms used in our system, such as the tandem algorithm and supervised learning algorithm, to 64 and found similar performance. Those comparisons indicate that a 64-channel filterbank may be sufficient for T-F analysis in CASA systems, as in perceptual studies [37].

Speech interference, which often occurs in a meeting or a daily conversation, is not considered in this study. To tackle this problem in our framework, a multipitch tracker would be needed and the system has to address the sequential grouping problem [30]. In [33], voiced-voiced separation and unvoiced-voiced (or voiced-unvoiced) separation have been studied, but not unvoiced-unvoiced separation. Our future research will address multi-talker separation problem.

Acknowledgment

The authors would like to thank Z. Jin and G. Hu for providing their programs for this work.

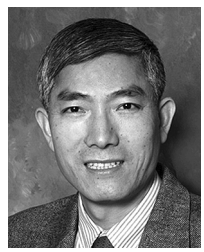
References

[1] J. B. Allen, *Articulation and Intelligibility*. San Rafael, CA: Morgan & Claypool, 2005.
[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, 1979, pp. 208–211.
[3] P. Boersma and D. Weenink, December 27, 2007, Praat: Doing Phonetics by Computer, ver. 5.0.02 [Online]. Available: http://www.fon.hum.uva.nl/praat
[4] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[5] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.
[6] 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.
[7] C. de Boor, *A Practical Guide to Splines*. New York: Springer-Verlag, 1978.
[8] H. Dillon, *Hearing Aids*. New York: Thieme Medical Publishers, 2001.
[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, 1993, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html
[10] G. Hu, "Monaural Speech Organization and Segregation," Ph.D. dissertation, Biophys. Program, Ohio State Univ., Columbus, 2006.
[11] G. Hu, 2006, 100 Nonspeech Sounds Online. [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html
[12] 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, Sep. 2004.
[13] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.
[14] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.
[15] K. Hu and D. L. Wang, "Incorporating spectral subtraction and noise type for unvoiced speech segregation," in *Proc. IEEE ICASSP*, 2009, pp. 4425–4428.
[16] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.
[17] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, pp. 225–246, 1969.
[18] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.
[19] P. Ladefoged, *Vowels and Consonants: An Introduction to the Sounds of Languages*. Oxford, U.K.: Blackwell, 2001.
[20] 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.
[21] 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.
[22] Y. Li and D. L. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2009.
[23] Z. Lin, R. A. Goubran, and R. M. Dansereau, "Noise estimation using speech/non-speech frame decision and subband spectral tracking," *Speech Commun.*, vol. 49, pp. 542–557, 2007.
[24] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.
[25] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.
[26] H. Ney, "On the probabilistic interpretation of neural network classifiers and discriminative training criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 2, pp. 107–119, Feb. 1995.
[27] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Appl. Psychol. Unit*, 1988, Cambridge, U.K., APU Rep. 2341.

[28] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft masking filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.
[29] 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, 2007, 10.1155/2007/84186, Article ID 84186, 15 pages.
[30] Y. Shao, "Sequential Organization in Computational Auditory Scene Analysis," Ph.D. dissertation, Dept. of Comput. Sci. and Eng., Ohio State Univ., Columbus, 2007.
[31] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.
[32] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.
[33] S. Vishnubhotla and C. Y. Espy-Wilson, "An algorithm for speech segregation of co-channel speech," in *Proc. IEEE ICASSP*, 2009, pp. 109–112.
[34] 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.
[35] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.
[36] *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley-IEEE Press, 2006.
[37] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time–frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.

Ke Hu (S'09) received the B.E. and M.E. degrees in automation from the University of Science and Technology of China, Hefei, in 2003 and 2006, respectively, and the M.S. degree in computer science and engineering from The Ohio State University, Columbus, in 2010, where he is currently pursuing the Ph.D. degree.

His research interests include computational auditory scene analysis, speech processing, 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 and 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.