Segregation of unvoiced speech from nonspeech interference

Guoning Hu[a)]
Biophysics Program, The Ohio State University, Columbus, Ohio 43210

DeLiang Wang[b)]
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 5 September 2007; revised 7 May 2008; accepted 8 May 2008)

Monaural speech segregation has proven to be extremely challenging. While efforts in computational auditory scene analysis have led to considerable progress in voiced speech segregation, little attention has been given to unvoiced speech, which lacks harmonic structure and has weaker energy, hence more susceptible to interference. This study proposes a new approach to the problem of segregating unvoiced speech from nonspeech interference. The study first addresses the question of how much speech is unvoiced. The segregation process occurs in two stages: Segmentation and grouping. In segmentation, the proposed model decomposes an input mixture into contiguous time-frequency segments by a multiscale analysis of event onsets and offsets. Grouping of unvoiced segments is based on Bayesian classification of acoustic-phonetic features. The proposed model for unvoiced speech segregation joins an existing model for voiced speech segregation to produce an overall system that can deal with both voiced and unvoiced speech. Systematic evaluation shows that the proposed system extracts a majority of unvoiced speech without including much interference, and it performs substantially better than spectral subtraction. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2939132]

PACS number(s): 43.72.Dv [DOS] Pages: 1306–1319

I. INTRODUCTION

In a daily environment, target speech is often corrupted by various types of acoustic interference, such as crowd noise, music, and other voices. Acoustic interference poses a serious problem for many applications including hearing aid design, automatic speech recognition (ASR), telecommunication, and audio information retrieval. In the hearing aid application, for example, it is well known that listeners with hearing loss have substantially greater difficulty in understanding speech in a noisy background (Moore, 2007). Hearing aids improve the audibility of noisy speech by means of amplification. However, their ability to improve the intelligibility of noisy speech is very limited, and how to remove or attenuate background noise is considered one of the biggest challenges facing hearing aid design (Dillon, 2001). Applications like this often require speech segregation. In addition, in many practical situations, monaural segregation is either necessary or desirable. Monaural speech segregation is especially difficult because one cannot utilize spatial filtering afforded by a microphone array to separate sounds from different directions. For monaural segregation, one has to consider the intrinsic properties of target speech and interference in order to disentangle them. Various methods have been proposed for monaural speech enhancement (Benesty *et al.*, 2005), and they usually assume stationary and quasistationary interference and achieve speech enhancement based on certain assumptions or models of speech and interference. These methods tend to lack the capacity to deal with general interference as the variety of interference makes it very difficult to model and predict.

While monaural speech segregation by machines remains a great challenge, the human auditory system shows a remarkable ability for this task. The perceptual segregation process is called auditory scene analysis (ASA) by Bregman (1990), who considers ASA to take place in two conceptual stages. The first stage, called segmentation (Wang and Brown, 1999), decomposes the auditory scene into sensory elements (or segments), each of which should primarily originate from a single sound source. The second stage, called grouping, aggregates the segments that likely arise from the same source. Segmentation and grouping are governed by perceptual principles, or ASA cues, which reflect intrinsic sound properties, including harmonicity, onset and offset, location, and prior knowledge of specific sounds (Bregman, 1990; Darwin, 1997).

Research in ASA has inspired considerable work in computational ASA (CASA) [for a recent, extensive review see Wang and Brown (2006)]. Many CASA studies have focused on monaural segregation and have performed the task without making strong assumptions about interference. Mirroring the two-stage model of ASA, a typical CASA system includes separate stages of segmentation and grouping that operate on a two-dimensional time-frequency (T-F) representation of the auditory scene (see Wang and Brown, 2006, Chap. 1). The T-F representation is typically created by an auditory peripheral model that analyzes an acoustic input by an auditory filterbank and decomposes each filter output

a)Electronic mail: hu.117@osu.edu
b)Electronic mail: dwang@cse.ohio-state.edu

 0001-4966/2008/124(2)/1306/14/$23.00

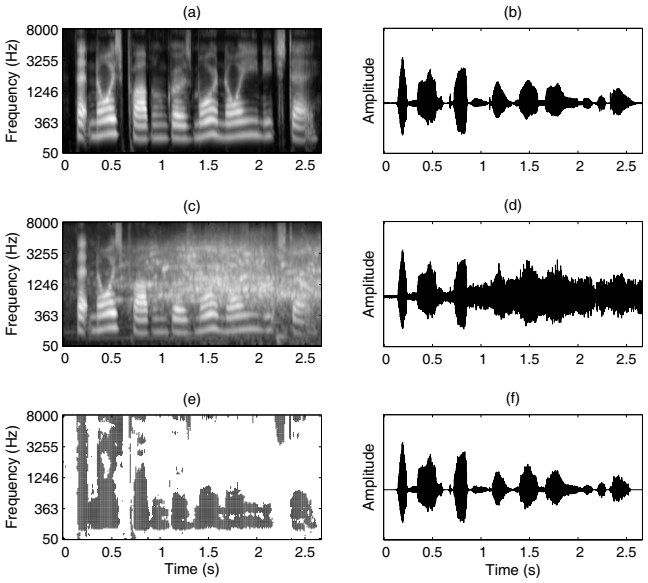

FIG. 1. CASA illustration. (a) T-F decomposion of a female utterance, "That noise problem grows more annoying each day." (b) Waveform of the utterance. (c) T-F decomposition of the utterance mixed with a crowd noise. (d) Waveform of the mixture. (e) Target stream composed of all the T-F units (black regions) dominated by the target (ideal binary mask). (f) Waveform resynthesized from the target stream.

into time frames. The basic element of the representation is called a T-F unit, corresponding to a filter channel and a time frame.

We have suggested that a reasonable goal of CASA is to retain the mixture signals within the T-F units where target speech is more intense than interference and to remove others (Hu and Wang, 2001; 2004). In other words, the goal is to compute a binary T-F mask, referred to as an ideal binary mask, where 1 indicates that the target is stronger than interference in the corresponding T-F unit and 0 otherwise. See Wang (2005) and Brungart *et al.* (2006) for more discussions on the notion of the ideal binary mask and its psychoacoustical support.

As an illustration, Fig. 1(a) shows a T-F representation of the waveform signal in Fig. 1(b). The signal is a female utterance, "That noise problem grows more annoying each day," from the TIMIT database (Garofolo *et al.*, 1993). The peripheral processing is carried out by a 128-channel gammatone filterbank with 20-ms time frames and a 10-ms frame shift (see Sec. III A for details). Figures 1(c) and 1(d) show the corresponding representations of a mixture of this utterance and crowd noise, where the signal-to-noise ratio (SNR) is 0 dB. In Figs. 1(a) and 1(c) a brighter unit indicates stronger energy. Figure 1(e) illustrates the ideal binary mask for the mixture in Fig. 1(d). With this mask, target speech can then be synthesized by retaining the filter responses of the T-F units having the value of 1 and eliminating the filter responses of the units of the value of 0. Figure 1(f) shows the synthesized waveform signal, which is close to the clean utterance in Fig. 1(b).

Natural speech contains both voiced and unvoiced portions (Stevens, 1998; Ladefoged, 2001). Voiced speech consists of portions that are mainly periodic (harmonic) or quasiperiodic. Previous CASA and related separation studies have focused on segregating voiced speech based on harmonicity (Parsons, 1976; Weintraub, 1985; Brown and Cooke, 1994; Hu and Wang, 2004). Although substantial advances have been made on voiced speech segregation, unvoiced speech segregation has not been seriously addressed and remains a major challenge. A recent system by Radfar *et al.* (2007) exploits vocal-tract filter characteristics (spectral envelopes) to separate two voices, which have the potential to deal with unvoiced speech. However, it is not clear how well their system performs when both speakers utter unvoiced speech and the assumption of two-speaker mixtures limits the scope of application.

Compared with voiced speech segregation, unvoiced speech segregation is clearly more difficult for two reasons. First, unvoiced speech lacks harmonic structure and is often acoustically noiselike. Second, the energy of unvoiced speech is usually much weaker than that of voiced speech; as a result, unvoiced speech is more susceptible to interference. Nevertheless, both voiced and unvoiced speech carry crucial information for speech understanding, and both need to be segregated.

In this paper, we propose a CASA system to segregate unvoiced speech from nonspeech interference. For auditory segmentation, we apply a multiscale analysis of event onsets and offsets (Hu and Wang, 2007), which has the important property that segments thus formed correspond to both voiced and unvoiced speech. By limiting interference to nonspeech signals, we propose to identify and group segments corresponding to unvoiced speech by a Bayesian classifier that decides whether segments are dominated by unvoiced speech on the basis of acoustic-phonetic features derived from these segments. The proposed algorithm, together with our previous system for voiced speech segregation (Hu and Wang, 2004; 2006), leads to a CASA system that segregates both unvoiced and voiced speech from nonspeech interference.

Before tackling unvoiced speech segregation, we first address the question of how much speech is unvoiced. This is the topic of the next section. Section III describes early stages of the proposed system, and Sec. IV details the grouping of unvoiced speech. Section V presents systematic evaluation results. Further discussions are given in Sec. VI.

II. HOW MUCH SPEECH IS UNVOICED?

Voiced speech refers to the part of speech signal that is periodic (harmonic) or quasiperiodic. In English, voiced speech includes all vowels, approximants, nasals, and certain stops, fricatives, and affricates (Stevens, 1998; Ladefoged, 2001). It comprises a majority of spoken English. Unvoiced speech refers to the part that is mainly aperiodic. In English, unvoiced speech comprises a subset of stops, fricatives, and affricates. These three consonant categories contain the following phonemes:

(1) Stops: /t/, /d/, /p/, /b/, /k/, and /g/.
(2) Fricatives: /s/, /z/, /f/, /v/, /ʃ/, /ʒ/, /θ/, /ð/, and /h/.
(3) Affricates: /tʃ/ and /dʒ/.

TABLE I. Occurrence percentages of six consonant categories.

Phoneme type	Conversational	Written	TIMIT
Voiced stop	6.7	6.9	7.9
Unvoiced stop	15.1	11.9	12.8
Voiced fricative	7.5	9.5	7.7
Unvoiced fricative	8.6	8.6	9.8
Voiced affricate	0.3	0.4	0.6
Unvoiced affricate	0.3	0.5	0.5
Total	38.5	37.8	39.3

TABLE II. Duration percentages of six consonant categories.

Phoneme type	Conversational	TIMIT
Voiced stop	5.6	5.2
Unvoiced stop	16.2	12.9
Voiced fricative	5.3	5.8
Unvoiced fricative	9.6	12.0
Voiced affricate	0.3	0.6
Unvoiced affricate	0.4	0.7
Total	37.4	37.2

In phonetics, all these phonemes, except /h/, are called obstruents. To simplify notations, we refer to the above phonemes as expanded obstruents. Eight of the expanded obstruents, /t/, /p/, /k/, /s/, /f/, /ʃ/, /θ/, and /tʃ/, are categorically unvoiced. In addition, /h/ may be pronounced either in the voiced or the unvoiced manner. The other phonemes are categorized as voiced, although in articulation they often contain unvoiced portions. Note that an affricate can be treated as a composite phoneme, with a stop followed by a fricative.

Dewey (1923) conducted an extensive analysis of the relative frequencies of individual phonemes in written English, and this analysis concludes that unvoiced phonemes account for 21.0% of the total phoneme usage. For spoken English, French *et al.* (1930) [see also Fletcher (1953)] conducted a similar analysis on 500 telephone conversations containing a total of about 80 000 words and concluded that unvoiced phonemes account for about 24.0%. Another extensive phonetically labeled corpus is the TIMIT database, which contains 6300 sentences read by 630 different speakers from various dialect regions in America (Garofolo *et al.*, 1993). Note that the TIMIT database is constructed to be phonetically balanced. Many of the same sentences are read by multiple speakers, and there are a total of 2342 different sentences. We have performed an analysis of relative phoneme frequencies for distinct sentences in the TIMIT corpus, and found that unvoiced phonemes account for 23.1% of the total phonemes.

Table I shows the occurrence percentages of six phoneme categories from these studies. Several observations may be made from the table. First, unvoiced stops occur much more frequently than voiced stops, particularly in conversations where they occur more than twice as often as their voiced counterparts. Second, affricates are used only occasionally. It is remarkable that the percentages of the six consonant categories are comparable despite the fact that written, read, and conversational speech are different in many ways. In particular, the total percentages of these consonants are almost the same for the three different kinds of speech.

What about the relative durations of unvoiced speech in spoken English? Unfortunately, the data reported on the telephone conversations (French *et al.*, 1930) do not contain durational information. To get an estimate, we use the durations obtained from a phonetically transcribed subset of the switchboard corpus (Greenberg *et al.*, 1996), which also consists of conversations over the telephone. The amount of labeled data in the Switchboard corpus, i.e., 72 min of conversation, is much smaller than that in the telephone conservations analyzed by French *et al.* (1930). Hence we do not use the labeled Switchboard corpus to obtain phoneme frequencies; instead we assign the median durations from the transcription to the occurrence frequencies in the telephone conservations in order to estimate the relative durations of unvoiced sounds. Table II lists the resulting duration percentages of six phoneme categories. Also listed in the table are the corresponding data from the TIMIT corpus. The table shows that for stops and fricatives, unvoiced sounds last much longer than their voiced counterparts. In addition, affricates have a minor contribution in terms of duration, similar to that in terms of occurrence frequency. Once again, the percentages from conversational speech are comparable to those from read speech. In terms of overall time duration, unvoiced speech accounts for 26.2% in telephone conversations and 25.6% in the read speech of the TIMIT corpus. These duration percentages are a little higher than the corresponding frequency percentages.

Tables I and II show that unvoiced sounds account for more than 20% of spoken English in terms of both occurrence frequency and time duration. In addition, since voiced obstruents are often not entirely voiced, unvoiced speech may occur more than suggested by the above estimates.

III. EARLY PROCESSING STAGES

Our proposed system for unvoiced speech segregation has the following stages of computation: Peripheral analysis, feature extraction, auditory segmentation, and grouping. In this section, we describe the first three stages. The stage of grouping is described in the next section. A list of all the symbols used in system description is given in the Nomenclature.

A. Auditory peripheral analysis

This stage derives a T-F representation of an input scene by performing a frequency analysis using a gammatone filterbank (Patterson *et al.*, 1988), which models human cochlear filtering. Specifically, we employ a bank of 128 gammatone filters, whose center frequencies range from 50 to 8000 Hz; this frequency range is adequate for speech understanding (Fletcher, 1953; Pavlovic, 1987). The impulse response of a gammatone filter centered at frequency f is

$$g(f,t) = \begin{cases} b^a t^{a-1} e^{-2\pi bt} \cos(2\pi ft), & t \geqslant 0 \\ 0 & \text{otherwise}, \end{cases} \quad (1)$$

where $a=4$ is the order of the filter and b is the equivalent rectangular bandwidth (Glasberg and Moore, 1990), which increases as the center frequency f increases.

Let $x(t)$ be the input signal. The response from a filter channel c, $x(c,t)$, is given by

$$x(c,t) = x(t) * g(f_c,t), \quad (2)$$

where "*" denotes convolution, and f_c the center frequency of filter channel c. In each filter channel, the output is further divided into 20-ms time frames with a 10-ms shift between consecutive frames.

B. Feature extraction

Previous studies suggest that in a T-F region dominated by a periodic signal, T-F units in adjacent channels tend to have highly correlated filter responses (Wang and Brown, 1999) or response envelopes (Hu and Wang, 2004). In this stage, we calculate such cross-channel correlations. These correlations will be used to determine T-F units dominated by unvoiced speech in the grouping stage.

Cross-channel correlation of filter responses measures the similarity between the responses of two adjacent filter channels. Since these responses have channel-dependent phases, we perform phase alignment before measuring their correlation. Specifically, we first compute their autocorrelation functions (Licklider, 1951; Lyon, 1984; Slaney and Lyons, 1990) and then use their autocorrelation responses to calculate cross-channel correlation.

Let u_{cm} denote a T-F unit for frequency channel c and time frame m, the corresponding autocorrelation of the filter response is given by

$$A(c,m,\tau) = \sum_n x(c,mT_m - nT_n)x(c,mT_m - nT_n - \tau T_n). \quad (3)$$

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. The above summation is over 20 ms, the length of a time frame. The cross-channel correlation between u_{cm} and $u_{c+1,m}$ is given by

$$C(c,m) = \frac{\Sigma_\tau [A(c,m,\tau) - \overline{A(c,m)}][A(c+1,m,\tau) - \overline{A(c+1,m)}]}{\sqrt{\Sigma_\tau [A(c,m,\tau) - \overline{A(c,m)}]^2 \Sigma_\tau [A(c+1,m,\tau) - \overline{A(c+1,m)}]^2}}, \quad (4)$$

where $\bar{A}$ denotes the average value of A.

When the input contains a periodic signal, auditory filters with high center frequencies respond to multiple harmonics. Such a filter response is amplitude modulated, and the response envelope fluctuates at the F0 of the periodic signal (Helmholtz, 1863). As a result, adjacent channels in the high-frequency range tend to have highly correlated response envelopes. To extract these correlations, we calculate response envelope through half-wave rectification and bandpass filtering, where the passband corresponds to the plausible F0 range of target speech, i.e., [70 Hz, 400 Hz], the typical pitch range for adults (Nooteboom, 1997). The resulting bandpassed envelope in channel c is denoted by $x_E(c,t)$.

Similar to Eqs. (3) and (4), we compute envelope autocorrelation as

$$A_E(c,m,\tau) = \sum_n x_E(c,mT_m - nT_n)x_E(c,mT_m - nT_n - \tau T_n) \quad (5)$$

and then obtain cross-channel correlation of response envelopes as

$$C_E(c,m) = \frac{\Sigma_\tau [A_E(c,m,\tau) - \overline{A_E(c,m)}][A_E(c+1,m,\tau) - \overline{A_E(c+1,m)}]}{\sqrt{\Sigma_\tau [A_E(c,m,\tau) - \overline{A_E(c,m)}]^2 \Sigma_\tau [A_E(c+1,m,\tau) - \overline{A_E(c+1,m)}]^2}}. \quad (6)$$

C. Auditory segmentation

Previous CASA systems perform auditory segmentation by analyzing common periodicity (Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004), and thus cannot handle unvoiced speech. In this study, we apply a segmentation algorithm based on a multiscale analysis of event onsets and offsets (Hu and Wang, 2007). Onsets and offsets are important ASA cues (Bregman, 1990) because different sound sources in an acoustic environment seldom start and end at the same time. In the time domain, boundaries between different sound sources tend to produce onsets and offsets. Common onsets and offsets also provide natural cues

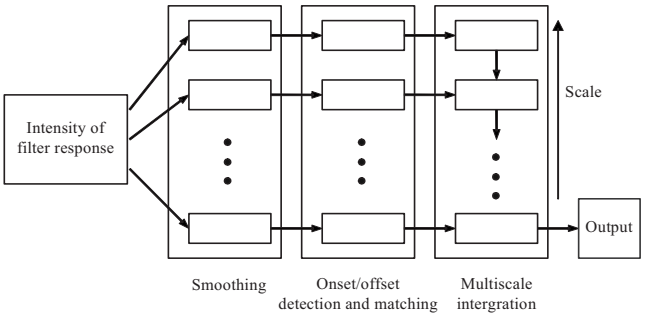

FIG. 2. Diagram of the segmentation stage. In each processing step, a rectangle represents a particular scale, which increases from bottom to top.

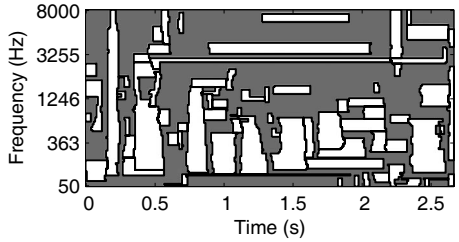

FIG. 3. Bounding contours of estimated segments. The input is the mixture shown in Fig. 1(d). The background is represented by gray.

to integrate sounds from the same source across frequency. Because onset and offset are cues common to all the sounds, this algorithm is applicable to both voiced and unvoiced speech. Figure 2 shows the diagram of the segmentation stage. It has three steps: Smoothing, onset/offset detection, and multiscale integration.

Onsets and offsets correspond to sudden intensity increases and decreases, respectively. A standard way to identify such intensity changes is to find the peaks and valleys of the time derivative of signal intensity (Wang and Brown, 2006, Chap. 3). We calculate the intensity of a filter response as the square of the response envelope, which is extracted using half-wave rectification and low-pass filtering. Because of the intensity fluctuation within individual events, many peaks and valleys of the derivative do not correspond to real onsets and offsets. Therefore, in the first step of segmentation, we smooth the intensity over time to reduce such fluctuations. Since an acoustic event tends to have synchronized onset and offset across frequency, we additionally perform smoothing over frequency, which helps to enhance such coincidences in neighboring frequency channels. This procedure is similar to the standard Canny edge detector in image processing (Canny, 1986). The degree of smoothing over time and frequency is referred to as the two-dimensional scale. The larger the scale is, the smoother the intensity is. The smoothed intensities at different scales form the so-called scale space (Romeny *et al.*, 1997).

In the second step of segmentation, our system detects onsets and offsets in each filter channel. Onset and offset candidates are detected by marking peaks and valleys of the time derivative of the smoothed intensity. The system then merges simultaneous onsets and offsets in adjacent channels into onset and offset fronts, which are contours connecting onset and offset candidates across frequency. Segments are obtained by matching individual onset and offset fronts.

As a result of smoothing, event onsets and offsets of small T-F regions may be blurred at a larger (coarser) scale. Consequently, we may miss some true onsets and offsets. On the other hand, at a smaller (finer) scale, the detection may be sensitive to insignificant intensity fluctuations within individual events. Consequently, false onsets and offsets may be generated and some true segments may be oversegmented. We find it generally difficult to obtain satisfactory segmentation with a single scale. In the last step of segmentation, we deal with this issue by performing multiscale integration from the largest scale to the smallest scale in an orderly manner. More specifically, at each scale, our system first locates more accurate boundaries for the segments obtained at a larger scale. Then it creates new segments outside the existing ones. The details of the segmentation stage are given in Hu and Wang (2007); see also Hu (2006).

As an illustration, Fig. 3 shows the bounding contours of obtained segments for the mixture in Fig. 1(d). The background is represented in gray. Compared with the ideal binary mask in Fig. 1(e), the obtained segments capture a majority of the target speech. Some segments for the interference are also formed. Note that the system does not, in this stage, distinguish between target and interference for each segment, which is the task of grouping described below.

IV. GROUPING

Our general strategy for grouping is to first segregate voiced speech and then deal with unvoiced speech. This strategy is motivated by the consideration that voiced speech segregation has been well studied and can be applied separately, and segregated voiced speech can be useful in subsequent unvoiced speech segregation.

To segregate the voiced portions of a target utterance, we apply our previous system for voiced speech segregation (Hu and Wang, 2006), which is slightly extended from an earlier version (Hu and Wang, 2004) and produces good segregation results. Target pitch contours needed for segregation are obtained from a clean target by PRAAT, a standard pitch determination algorithm for clean speech (Boersma and Weenink, 2004). This way, we avoid pitch tracking errors which could adversely influence the performance of unvoiced speech segregation—the focus of this study. We refer to the resulting stream of voiced target as S_T^1.

The task of grouping unvoiced target amounts to labeling segments already obtained in the segmentation stage. A segment may be dominated by voiced target, unvoiced target, or interference, and we want to group segments dominated by unvoiced target while rejecting segments dominated by interference. Since an unvoiced phoneme is often strongly coarticulated with a neighboring voiced phoneme, some unvoiced target is included in segments dominated by voiced target (Hu, 2006; Hu and Wang, 2007). So we need to group segments dominated by voiced target to recover this part of unvoiced speech.

Our system first groups segments dominated by voiced target. Then among the remaining segments, we label those dominated by unvoiced target in two steps: Segment removal and segment classification.

A. Grouping segments dominated by voiced target

A segment dominated by voiced target should have a significant overlap with the segregated voiced target, S_T^1. Hence we label a segment as dominated by voiced target if

(1) more than half of its total energy is included in the voiced time frames of target, and
(2) more than half of its energy in the voiced frames is included in the T-F units belonging to S_T^1.

All the segments labeled as dominated by voiced target are grouped into the segregated target stream.

By grouping segments dominated by voiced target, we recover more target-dominant T-F units than S_T^1. However, some interference-dominant T-F units are also included due to the mismatch error in segmentation, i.e., the error of putting both target- and interference-dominant units into one segment (Hu and Wang, 2007). We found that a significant amount of the mismatch error in segmentation stems from merging T-F areas in adjacent channels into one segment (Hu, 2006). To minimize the amount of interference-dominant T-F units being wrongly grouped into the target stream, we consider estimated segments in individual channels, referred to as T-segments, instead of whole T-F segments. Specifically, if a T-segment is dominated by a voiced target based on the above two criteria, all the T-F units within the T-segment are grouped into the voiced target. The resulting stream is referred to as S_T^2.

B. Acoustic-phonetic features for segment classification

The next task is to label or classify segments dominated by unvoiced speech. Since the signal within a segment is mainly from one source, it is expected to have similar acoustic-phonetic properties to that source. Therefore, we identify segments dominated by unvoiced speech using acoustic-phonetic features.

A basic speech sound is characterized by the following acoustic-phonetic properties: Short-term spectrum, formant transition, voicing, and phoneme duration (Stevens, 1998; Ladefoged, 2001). These features have proven to be useful in speech recognition, e.g., to distinguish different phonemes or words (Rabiner and Juang, 1993; Ali and Van der Spiegel, 2001b, 2001a). These properties may also be useful in distinguishing speech from nonspeech interference. However, it is important to treat these properties, appropriately considering that we are dealing with noisy speech. In particular, we give the following considerations.

(1) *Spectrum*. The short-term spectrum of an acoustic mixture at a particular time may be quite different from that of the target utterance or that of the interference in the mixture. Therefore, features representing the overall shape of a short-term spectrum may not be appropriate for our task. On the other hand, the short-term spectra in the T-F regions dominated by speech are expected to be similar to those of clean utterances, while the short-term spectra of other T-F regions tend to be different. Therefore, we use the short-term spectrum within a T-F region as a feature to decide whether this region is dominated by speech or interference. More specifically, we use the energy within individual T-F units as the feature to represent the short-term spectrum.
(2) *Formant transition*. It is difficult to estimate the formant frequency of a target utterance in the presence of strong interference. In addition, formant transition is embodied in the corresponding short-term spectrum. Therefore, we do not explicitly use formant transition in this study.
(3) *Voicing*. Voicing information of a target utterance is not utilized since we are handling unvoiced speech.
(4) *Duration*. While the duration of an interfering sound is unpredictable, for speech each phoneme lasts for a range of durations. However, we may not be able to detect the boundaries of phonemes that are strongly coarticulated. Therefore it is difficult to find the accurate durations of individual phonemes from an acoustic mixture, and the durations of individual phonemes are not utilized in this study.

In summary, we use the signal energy within individual T-F segments to derive the acoustic-phonetic features for distinguishing speech and nonspeech interference.

C. Segment removal

Since our task is to group segments for unvoiced speech, segments that mainly contain periodic or quasiperiodic signals unlikely originate from unvoiced speech and should be removed. A segment is removed if more than half of its total energy is included in the T-F units dominated by a periodic signal. We consider unit u_{cm} dominated by a periodic signal if it is included in the segregated voiced stream or has a high cross-channel correlation, the latter indicating that two neighboring channels respond to the same harmonic or formant (Wang and Brown, 1999). Specifically, a cross-channel correlation is considered high if $C(c,m)>0.985$ or $C_E(c,m)>0.985$.

Among the remaining segments, a segment dominated by unvoiced target is unlikely located at time frames corresponding to voiced phonemes other than expanded obstruents. This property is, however, not shared by some interference-dominant segments that can have significant energy in such voiced frames. We remove these segments as follows.

We first label the voiced frames of a target utterance that unlikely contain an expanded obstruent, according to the segregated voiced target. Let $H_0(m_1,m_2)$ be the hypothesis that a T-F region between frame m_1 and frame m_2 is dominated by speech and $H_1(m_1,m_2)$ the hypothesis that the region is dominated by interference. In addition, let $H_{0,a}(m_1,m_2)$ be the hypothesis that this region is dominated by an expanded obstruent and $H_{0,b}(m_1,m_2)$ by any other phoneme.

Let $X(c,m)$ be the energy in u_{cm} and $X(m)=\{X(c,m),\forall c\}$ the vector of the energy in all the T-F units at

time frame m. $X(m)$ is referred to as the *cochleagram* at frame m (Wang and Brown, 2006). Let $X_T(m) = \{X_T(c,m), \forall c\}$ be the cochleagram of the segregated target at frame m, that is,

$$X_T(c,m) = \begin{cases} X(c,m) & \text{if } u_{cm} \in S_T^2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

A voiced frame m is labeled as obstruent dominant if

$$P(H_{0,a}(m)|X_T(m)) > P(H_{0,b}(m)|X_T(m)). \tag{8}$$

We assume that, given $X_T(m)$, these posterior probabilities do not depend on a particular frame index. In other words, for any two frames m_1 and m_2,

$$P(H(m_1)|X_T(m_1)) = P(H(m_2)|X_T(m_2)) \quad \text{if } X_T(c,m_1) = X_T(c,m_2), \forall c. \tag{9}$$

To simplify calculations, we further assume that the prior probabilities of $H_{0,a}(m)$, $H_{0,b}(m)$, and $H_1(m)$ are constant for individual frames within a given T-F region. A frame index can then be dropped from these frame-level hypotheses. In the following, we use a hypothesis without a frame index to refer to that hypothesis for a single frame of a T-F segment. Then Eq. (8) becomes

$$P(H_{0,a}|X_T(m)) > P(H_{0,b}|X_T(m)). \tag{10}$$

Given that $X_T(m)$ corresponds to the voiced target, we have $P(H_{0,b}|X_T(m)) = 1 - P(H_{0,a}|X_T(m))$. Therefore, we have

$$P(H_{0,a}|X_T(m)) > 0.5. \tag{11}$$

We construct a multilayer perceptron (MLP) to compute $P(H_{0,a}|X_T(m))$. The MLP uses sigmoidal activation functions and has one hidden player. The input to the MLP is $X_T(m)$, the cochleagram of a target utterance at a voiced frame. Consequently, this MLP has 128 units in the input layer. It has one unit in the output layer. The desired output of this unit is 1 if the corresponding frame is dominated by an expanded obstruent and 0 otherwise. Note that when there are sufficient training samples, the trained MLP yields a good estimate of the probability (Bridle, 1989). The MLP is trained with a corpus that includes all the utterances from the training part of the TIMIT database 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. We use PRAAT to label voiced frames. The cochleagram of the target at voiced frames is determined using the ideal binary mask of each mixture. The number of units in the hidden layer of the MLP is determined using cross validation. Specifically, we divide the training samples into two equal sets, one for training and the other for validation. The resulting MLP has 20 units in the hidden layer.

We label every voiced frame based on Eq. (11). A segment is removed if more than 50% of its energy is included in the voiced frames that are not dominated by an expanded obstruent. As a result of segment removal, many segments dominated by interference are removed. We find that this step increases the robustness of the system and greatly reduces the computational burden for the following segment classification.

D. Segment classification

In this step, we classify the remaining segments as dominated by either unvoiced speech or interference. Let s be a remaining segment lasting from frame m_1 to m_2, and $X_s(m) = \{X_s(c,m), \forall c\}$ be the corresponding cochleagram at frame m. That is,

$$X_s(c,m) = \begin{cases} X(c,m) & \text{if } u_{cm} \in s \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Let $\mathbf{X}_s = [X_s(m_1), X_s(m_1+1), \ldots, X_s(m_2)]$. s is classified as dominated by unvoiced speech if

$$P(H_{0,a}(m_1,m_2)|\mathbf{X}_s) > P(H_1(m_1,m_2)|\mathbf{X}_s). \tag{13}$$

Because segments have varied durations, directly evaluating $P(H_{0,a}(m_1,m_2)|\mathbf{X}_s)$ and $P(H_1(m_1,m_2)|\mathbf{X}_s)$ for each possible duration is not computationally feasible. Therefore, we consider a simplifying approximation that each time frame is statistically independent (more discussion on this approximation will be given later in this section). Since

$$P(H_{0,a}(m_1,m_2)|\mathbf{X}_s) = P(H_{0,a}(m_1), H_{0,a}(m_1+1), \ldots, H_{0,a}(m_2)|\mathbf{X}_s) \tag{14}$$

Applying the chain rule,

$$\begin{aligned} P(H_{0,a}(m_1,m_2)|\mathbf{X}_s) &= P(H_{0,a}(m_1)|\mathbf{X}_s) \\ &\times P(H_{0,a}(m_1+1)|H_{0,a}(m_1),\mathbf{X}_s)\cdots \\ &\times P(H_{0,a}(m_2)|H_{0,a}(m_1), H_{0,a}(m_1+1), \ldots, H_{0,a}(m_2 - 1),\mathbf{X}_s). \end{aligned} \tag{15}$$

From the independence assumption, we have

$$\begin{aligned} &P(H_{0,a}(m_1+k)|H_{0,a}(m_1), H_{0,a}(m_1+1), \\ &\quad \ldots, H_{0,a}(m_2+k-1), \mathbf{X}_s) \\ &= P(H_{0,a}(m_1+k)|\mathbf{X}_s) = P(H_{0,a}(m_1+k)|X_s(m_1+k)). \end{aligned} \tag{16}$$

Therefore,

$$P(H_{0,a}(m_1,m_2)|\mathbf{X}_s) = \prod_{m=m_1}^{m_2} P(H_{0,a}(m)|X_s(m)), \tag{17}$$

and the same calculation can be done for $P(H_1(m_1,m_2)|\mathbf{X}_s)$. Now Eq. (13) becomes

$$\prod_{m=m_1}^{m_2} P(H_{0,a}(m)|X_s(m)) > \prod_{m=m_1}^{m_2} P(H_1(m)|X_s(m)). \tag{18}$$

By applying the Bayesian rule and the assumption made in Sec. IV C that the prior and the posterior probabilities do not depend on a frame index within a given segment, the above inequality becomes

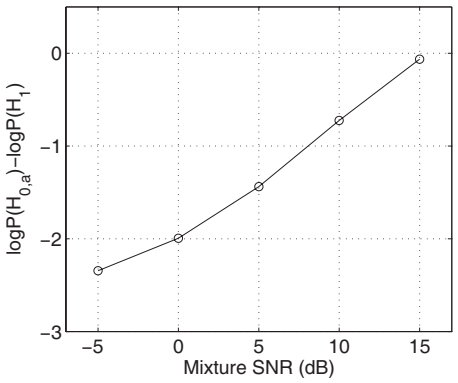

FIG. 4. Ratio of the prior probability of the target to that of interference as a function of mixture SNR.

$$\left[\frac{P(H_{0,a})}{P(H_1)}\right]^{m_2-m_1+1}\prod_{m=m_1}^{m_2}\frac{p(X_s(m)|H_{0,a})}{p(X_s(m)|H_1)}>1. \tag{19}$$

The prior probabilities $P(H_{0,a})$ and $P(H_1)$ depend on the SNR of acoustic mixtures. Figure 4 shows the observed logarithmic ratios between $P(H_{0,a})$ and $P(H_1)$ from the training data at different mixture SNR levels. We approximate the relationship shown in the figure by a linear function,

$$\log\frac{P(H_{0,a})}{P(H_1)}=0.1166\ \text{SNR}-1.8962. \tag{20}$$

If we can estimate the mixture SNR, we will be able to estimate the log ratio of $P(H_{0,a})$ and $P(H_1)$ and use it in Eq. (19). This allows us to be more stringent in labeling a segment as speech dominant when the mixture SNR is low.

We propose to estimate the SNR of an acoustic mixture by capitalizing on the voiced target that has already been segregated from the mixture. Let E_1 be the total energy included in the T-F units labeled 1 at the voiced frames of the target. One may use E_1 to approximate the target energy at voiced frames and estimate the total target energy as αE_1 that includes unvoiced target speech. By analyzing the training part of the TIMIT database, we find that parameter α—the ratio between the total energy of a speech utterance and the total energy at the voiced frames of the utterance—varies substantially across individual utterances. In this study, we set α to 1.09, the average value of all the utterances in the training part of the TIMIT database. Similarly, let E_2 be the total energy included in the T-F units labeled 0 at the voiced frames of the target, N_1 the total number of these voiced frames, and N_2 the total number of other frames. Hence, E_2 approximates the interference energy at voiced frames, and the average interference energy per voiced frame is then E_2/N_1. Assuming that interference is relatively steady, we can use E_2/N_1 to approximate the interference energy per frame and estimate the total interference energy as $E_2(N_1+N_2)/N_1$. Consequently, the estimated mixture SNR is

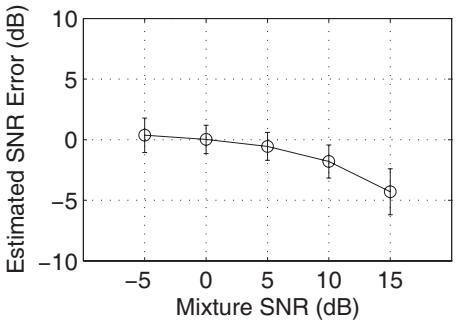

FIG. 5. Mean and standard deviation of estimated mixture SNRs in the test corpus.

$$\overline{\text{SNR}}=10\log_{10}\frac{\alpha N_1E_1}{(N_1+N_2)E_2}=10\log_{10}\frac{E_1}{E_2}+10\log_{10}\alpha+10\log_{10}\frac{N_1}{N_1+N_2}. \tag{21}$$

With $\alpha=1.09$, $10\log_{10}\alpha=0.37$ dB. We have applied this SNR estimation to the test corpus. Figure 5 shows the mean and the standard deviation of the estimation error at each SNR level of the original mixtures; the estimation error equals the estimated SNR subtracted by the true SNR. As shown in the figure, the system yields a reasonable estimate when the mixture SNR is lower than 10 dB. When the mixture SNR is greater than or equal to 10 dB, Eq. (21) tends to underestimate the true SNR. As discussed in Sec. II, some voiced frames of the target, such as those corresponding to expanded obstruents, may contain unvoiced target energy that fails to be included in E_1 but ends up in E_2. When the mixture SNR is low, this part of unvoiced energy is much lower than the interference energy. Therefore, it is negligible and Eq. (21) provides a good estimate. When the mixture SNR is high, this unvoiced target energy can be comparable to interference energy, and as a result the estimated SNR tends to be systematically lower than the true SNR.

Alternatively, one can also estimate the mixture SNR at the unvoiced frames of the target or estimate the target energy at the unvoiced frames based on the average frame-level energy ratio of unvoiced speech to voiced speech. These alternatives have been evaluated in Hu (2006), and they do not yield more accurate estimates. Of course, for the TIMIT corpus we can simply correct the systematic bias shown in Fig. 5. We choose not to do so for the sake of generality.

To label a segment as either expanded obstruent or interference according to Eq. (19), we need to estimate the likelihood ratio between $p(X_s(m)|H_{0,a})$ and $p(X_s(m)|H_1)$. When $P(H_{0,a})$ and $P(H_1)$ are equal, we have by the Bayesian rule

$$\frac{p(X_s(m)|H_{0,a})}{p(X_s(m)|H_1)}=\frac{P(H_{0,a}|X_s(m))}{P(H_1|X_s(m))}. \tag{22}$$

We train an MLP to estimate $P(H_{0,a}|X_s(m))$ when $P(H_{0,a})$ and $P(H_1)$ are equal. The MLP has the same structure as the one described in Sec. IV C. The desired output of this MLP is 1 if a frame of a segment is dominated by an expanded obstruent and 0 if it is dominated by nonspeech interference.

The MLP is trained with the cochleagrams of target utterances at time frames corresponding to expanded obstruents and those of nonspeech intrusions from the same training set described in Sec. IV C. Since $P(H_1|X_s(m))=1-P(H_{0,a}|X_s(m))$, given that frame m corresponds to an expanded obstruent, we are able to calculate the likelihood ratio of $p(X_s(m)|H_{0,a})$ and $p(X_s(m)|H_1)$ using the output from the trained MLP.

Using the above estimate of the likelihood ratio and the estimated mixture SNR to calculate the prior probability ratio of $P(H_{0,a})$ and $P(H_1)$, we label a segment as either expanded obstruent or interference according to Eq. (19). All the segments labeled as unvoiced speech are added to the segregated voiced stream, S_T^2, yielding the final segregated stream, referred to as S_T^3.

This method for segregating unvoiced speech is very similar to a previous version (Wang and Hu, 2006) where we used fixed prior probabilities for all SNR levels. We find that using SNR-dependent prior probabilities gives better performance, especially when the mixture SNR is high. In an earlier study (Hu and Wang, 2005), we used Gaussian mixture models (GMM) to model both speech and interference and then classify a segment using the obtained models. The performance in that study is not as good as the present method. The main reason, we believe, is that although GMM is trained to represent the distributions of speech and interference accurately, MLP is trained to distinguish speech and interference and therefore has more discriminative power. We have also considered the dependence between consecutive frames, instead of treating individual frames as independent. The obtained result is comparable to that obtained with the independence assumption, probably due to the fact that the signal within a segment is usually quite stable across time so that considering the dynamics within a segment does not provide much additional information for classification.

As an example, Figs. 6(e) and 6(f) show the final segregated target and the corresponding synthesized waveform for the mixture in Fig. 1(d). Compared with the ideal mask in Fig. 1(e) and the corresponding synthesized wave form in Fig. 1(f), our system segregates most of target energy and rejects most of interfering energy. In addition, Figs. 6(a) and 6(b) show the mask and the waveform of the segregated voiced target, i.e., S_T^1. Figures 6(c) and 6(d) show the mask and the waveform of the resulting stream after grouping T-segments dominated by voiced speech, i.e., S_T^2. The target utterance, "That noise problem grows more annoying each day," includes five stops (/t/ in "that," /p/ and /b/ in "problem," /g/ in "grows," and /d/ in "day"), three fricatives (/ð/ in "that," /z/ in "noise," and /z/ in "grows"), and one affricate (/tʃ/ in "each"). The unvoiced parts of some consonants with strong coarticulation with the voiced speech, such as /ð/ in "that" and /d/ in "day," are segregated by using T-segments. The unvoiced part of /z/ in "noise" and /tʃ/ in "each" are segregated by grouping the corresponding segments. Except for a significant loss of energy for /p/ in "problem" and some energy loss for /t/ in "that," our system segregates most of the energy of the above consonants.

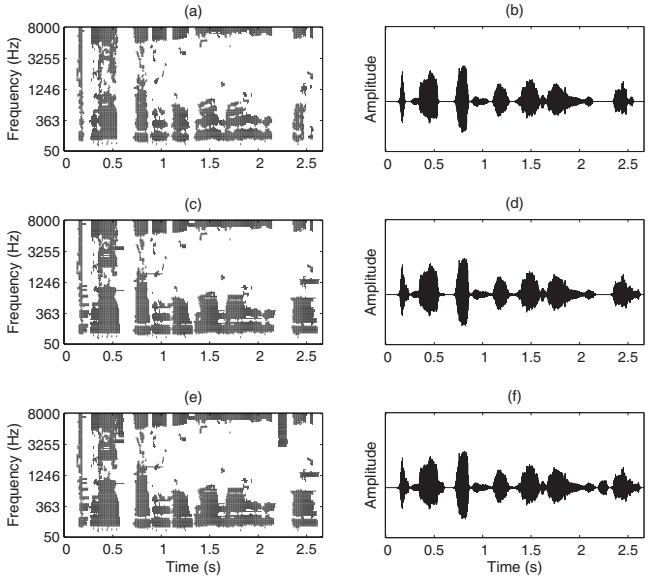

FIG. 6. Segregated target of the mixture shown in Fig. 1(d). (a) Mask of segregated voiced target (black regions). (b) Waveform resynthesized from the mask in (a). (c) Mask of the resulting target stream after grouping estimated T-segments (black regions). (d) Waveform resynthesized from the mask in (c). (e) Mask of the final segregated target (black regions). (f) Waveform resynthesized from the mask in (e).

V. EVALUATION

We now systematically evaluate the performance of our system. Here, we use a test corpus containing 20 target utterances randomly selected from the test part of the TIMIT database mixed with 15 nonspeech intrusions including five with crowd noise. Table III lists the 20 target utterances. The intrusions are as follows: N1—white noise, N2—rock music,

TABLE III. Target utterances in the test corpus.

Target	Content
S1	Put the butcher block table in the garage.
S2	Alice's ability to work without supervision is noteworthy.
S3	Barb burned paper and leaves in a big bonfire.
S4	Swing your arm as high as you can.
S5	Shaving cream is a popular item on Halloween.
S6	He then offered his own estimate of the weather, which was unenthusiastic.
S7	The morning dew on the spider web glistened in the sun.
S8	Her right hand aches whenever the barometric pressure changes.
S9	Why yell or worry over silly items.
S10	Aluminum silverware can often be flimsy.
S11	Guess the question from the answer.
S12	Medieval society was based on hierarchies.
S13	That noise problem grows more annoying each day.
S14	Don't ask me to carry an oily rag like that.
S15	Each untimely income loss coincided with the breakdown of a heating system part.
S16	Combine all the ingredients in a large bowl.
S17	Fuss, fuss, old man.
S18	Don't ask me to carry an oily rag like that.
S19	The fish began to leap frantically on the surface of the small lake.
S20	The redcoats ran like rabbits.

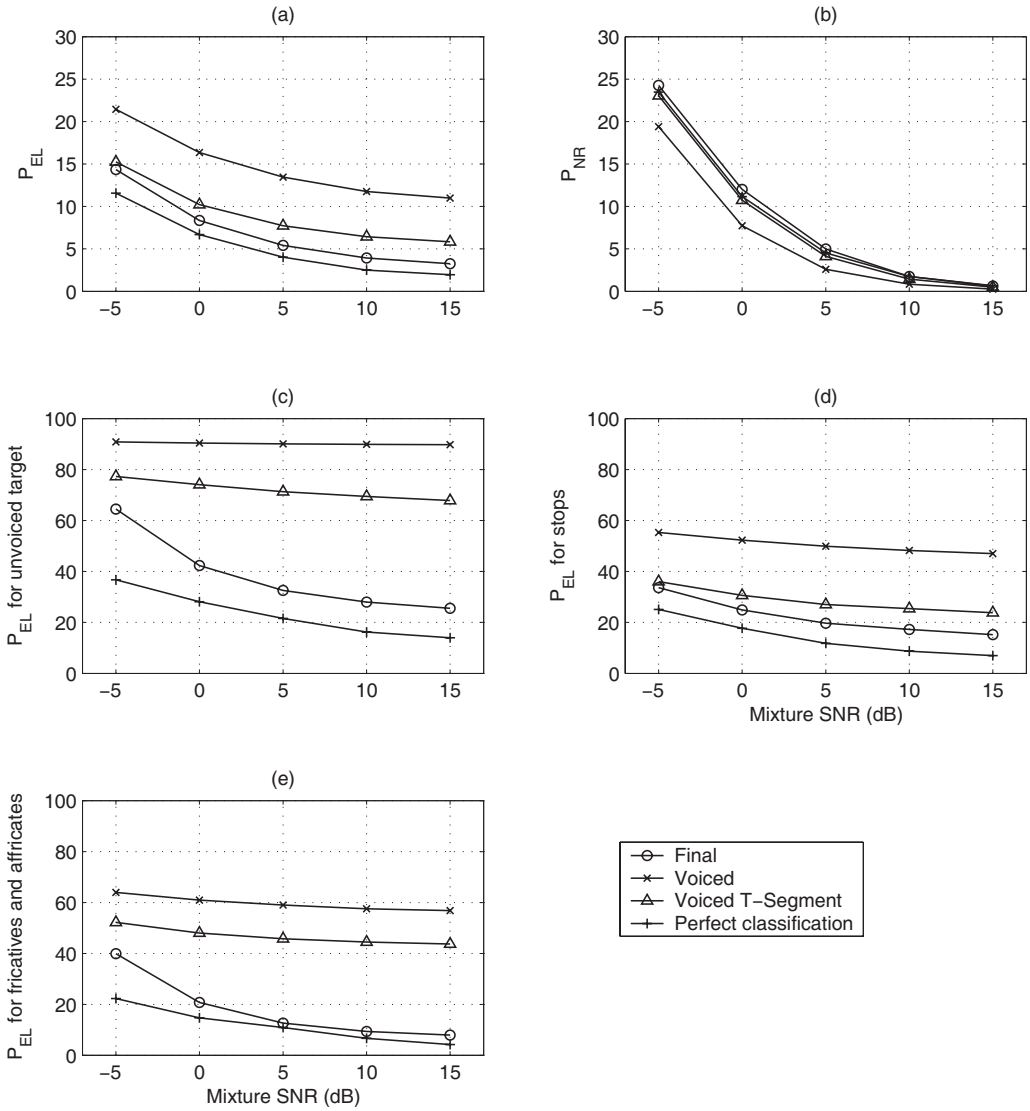

FIG. 7. System performance. In this figure, "Final" refers to the final segregated target, "Voiced" the segregated voiced target, "Voiced T-segment" the segregated target after grouping T-segments dominated by voiced target, and "Perfect classification" segregated target with perfect segment classification. (a) Average percentage of energy loss. (b) Average percentage of noise residue. (c) Average percentage of energy loss for unvoiced speech. (d) Average percentage of energy loss for stop consonants. (e) Average percentage of energy loss for fricatives and affricates.

N3—siren, N4—telephone ring, N5—electric fan, N6—clock alarm, N7—traffic noise, N8—bird chirp with water flowing, N9—wind, and N10—rain, N11—cocktail party noise, N12—crowd noise at a playground, N13—crowd noise with music, N14—crowd noise with clap, and N15—babble noise (16 speakers). This set of intrusions is not used during training, and represents a broad range of nonspeech sounds encountered in typical acoustic environments. Each target utterance is mixed with individual intrusions at −5-, 0-, 5-, 10-, and 15-dB SNR levels. The test corpus has 300 mixtures at each SNR level and 1500 mixtures altogether.

We evaluate our system by comparing the segregated target with the ideal binary mask—the stated computational goal. The performance of segregation is given by comparing the estimated mask and the ideal binary mask with two measures (Hu and Wang, 2004):

(1) the percentage of energy loss, P_{EL}, which measures the amount of energy in the target-dominant T-F units that are labeled as interference (hence removed) relative to the total energy in target-dominant T-F units and
(2) the percentage of noise residue, P_{NR}, which measures the amount of energy in the interference-dominant T-F units that are labeled as target (hence retained) relative to the total energy in T-F units estimated as target dominant.

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.

The P_{EL} and P_{NR} values for S_T^3 at different input SNR levels are shown in Figs. 7(a) and 7(b). Each value in the figure is the average over the 300 mixtures of individual targets and intrusions N1–N15. As shown in the figure, for the final segregation, our system captures an average of 85.7% of target energy at −5-dB SNR. This value increases to 96.7% when the mixture SNR increases to 15 dB. On average 24.3% of the segregated target belongs to interfer-

ence at −5 dB. This value decreases to 0.6% when the mixture SNR increases to 15 dB. In summary, our system captures a majority of target without including much interference.

To see the performance of our system on unvoiced speech in detail, we measure P_{EL} for target speech in the unvoiced frames. The average of these P_{EL} values at different SNR levels are shown in Fig. 7(c). Note that since some voiced frames contain unvoiced target, these are not exactly the P_{EL} values of unvoiced speech. Nevertheless, they are close to the real values. As shown in the figure, our system captures 35.5% of the target energy at the unvoiced frames when the mixture SNR is −5 dB and 74.4% when the mixture SNR is 15 dB. Overall, our system is able to capture more than 50% of target energy at the unvoiced frames when the mixture SNR is 0 dB or higher.

As discussed in Sec. II, expanded obstruents often contain voiced and unvoiced signals at the same time. Therefore, we measure P_{EL} for these phonemes separately in order to gain more insight into system performance. Because affricates do not occur often and they are similar to fricatives, we measure P_{EL} for fricatives and affricates together. The averages of these P_{EL} values at different SNR levels are shown in Figs. 7(d) and 7(e). As shown in the figure, our system performs somewhat better for fricatives and affricates when the mixture SNR is 0 dB or higher. On average, the system captures about 65% of these phonemes when the mixture SNR is −5 dB and about 90% when the mixture SNR is 15 dB.

For comparison, Fig. 7 also shows the P_{EL} and P_{NR} values for segregated voiced target, i.e., S_T^1 (labeled as "Voiced"), and the resulting stream after grouping T-segments dominated by voiced target, S_T^2 (labeled as "Voiced T-segments"). As shown in the figure, S_T^1 only includes about 10% of target energy in unvoiced frames, while S_T^2 includes about 17% more on average [Fig. 7(c)]. This additional 17% mainly corresponds to unvoiced phonemes that have strong coarticulation with neighboring voiced phonemes. By comparing these P_{EL} and P_{NR} values with those of the final segregated target, we can see that grouping segments dominated by unvoiced speech helps to recover a large amount of unvoiced speech. It also includes a small amount of additional interference energy, especially when the mixture SNR is low [Fig. 7(b)].

In addition, Fig. 7 shows the P_{EL} and P_{NR} values for segregated target obtained with perfect segment classification. As shown in the figure, there is a performance gap that can be narrowed with better classification, especially when the mixture SNR is low.

We also measure the system performance in terms of SNR by treating the target synthesized from the corresponding ideal binary mask as signal (Hu and Wang, 2004, 2006). Figures 8(a) and 8(b) show the overall average SNR values of segregated targets at different levels of mixture SNR and the corresponding SNR gain. Figures 8(c) and 8(d) show the corresponding values at unvoiced frames. Our system improves SNR in all input conditions.

To put our performance in perspective, we have compared with spectral subtraction, a standard method for speech enhancement (Huang *et al.*, 2001), with the above SNR measures. The spectral subtraction method is applied as follows. For each acoustic mixture, we assume that the silent portions of a target utterance are known and use the short-term spectra of interference in these portions as the estimates of interference. Interference is attenuated by subtracting the most recent interference estimate from the mixture spectrum at every time frame. The resulting SNR measures of the spectral subtraction method are also shown in Fig. 8. As shown in the figure, our system performs substantially better for both voiced and unvoiced speech than the spectral subtraction method even when it is applied with perfect speech pause detection; the only exception occurs for unvoiced speech at the input SNR of 15 dB. The improvement is more pronounced when the mixture SNR is low. At 15-dB SNR, the error in SNR estimation becomes significant (see Fig. 5), and the unvoiced speech energy that fails to be grouped becomes relatively large in comparison with interference energy. The spectral subtraction method is based on the estimation of interference and is less sensitive to input SNR.

We should point out that our system has significantly higher computational complexity than the spectral subtraction method. Note, however, that the spectral subtraction method implemented in our comparison assumes the prior knowledge of silent intervals of target speech, which greatly simplifies noise estimation—a nontrivial task that can itself be computationally intense. The major computational load of our system stems from calculating autocorrelations in the stage of feature extraction (Sec. III B) and extracting response envelopes in the stages of feature extraction and segmentation (Sec. III C). Also, our system needs to perform these computations in 128 frequency channels. As the calculations of the autocorrelations and response envelopes can be carried out in different channels independently, one can substantially improve computational efficiency by utilizing parallel computing hardware.

VI. DISCUSSION

Several insights have emerged from this study. The first is that the temporal properties of acoustic signals are crucial for speech segregation. Our system makes extensive use of temporal properties. In particular, we group target sound in consecutive frames based on the temporal continuity of speech signal. Furthermore, our system generates segments by analyzing sound intensity across time, i.e., onset and offset detection. The importance of temporal properties of speech for human speech recognition has been convincingly demonstrated by Shannon *et al.* (1995). In addition, studies in ASR suggest that long-term temporal information helps to improve recognition rate (see, e.g., Hermansky and Sharma, 1999). All these observations show that temporal information plays a critical role in sound organization and recognition.

Second, we find it advantageous to segregate voiced speech first and then use the segregated voiced speech to aid the segregation of unvoiced speech. As discussed before, unvoiced speech is more vulnerable to interference and more difficult to segregate. Segregation of voiced speech is more reliable and can be used to assist in the segregation of unvoiced speech. Our study shows that the unvoiced speech

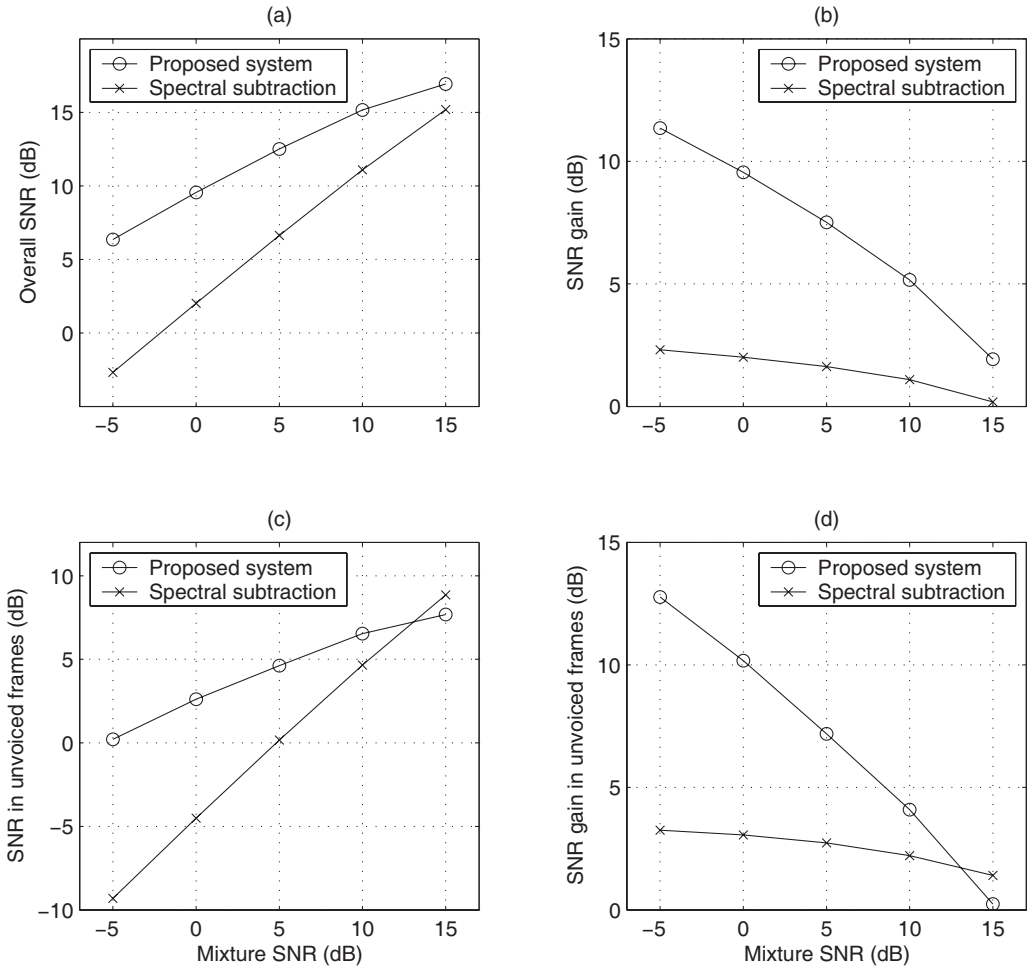

FIG. 8. SNR performances of the proposed system and spectral subtraction. (a) SNR of segregated target. (b) SNR gain of segregated target. (c) SNR of segregated target at unvoiced frames. (d) SNR gain of segregated target at unvoiced frames.

with strong coarticulation with voiced speech can be segregated using segregated voiced speech and estimated T-segments. Segregated voiced speech is also used to delineate the possible T-F locations of unvoiced speech. As a result, our system need not search the entire T-F domain for segments dominated by unvoiced speech and less likely identifies an interference-dominant segment as target. In addition, we have proposed an estimate of the mixture SNR from segregated voiced speech, which helps the system to adapt the prior probabilities in segment classification.

In addition, auditory segmentation is important for unvoiced speech segregation. In our system, the segmentation stage provides T-segments that help to segregate unvoiced speech that has strong coarticulation with voiced speech. As shown by Cole *et al.* (1996), such portions of speech are important for speech intelligibility. More importantly, segments are the basic units for classification, which enables the grouping of unvoiced speech.

A natural speech utterance contains silent gaps and other sections masked by interference. In practice, one needs to group the utterance across such time intervals. This is the problem of sequential grouping (Bregman, 1990; Wang and Brown, 2006). In this study, we handle this problem in a limited way by applying feature-based classification, assuming nonspeech interference. Systematic evaluation shows that, although our system yields good performance, it can be further improved with better sequential grouping. The assumption of nonspeech interference is obviously not applicable to mixtures of multiple speakers. Alternatively, grouping T-F segments sequentially may be achieved by using speech recognition (Barker *et al.*, 2005) or speaker recognition (Shao and Wang, 2006) in a top-down manner. Although these model-based studies on sequential grouping show promising results, the need for training with a specific lexicon or speaker set limits their scope of application. Substantial effort is needed to develop a general approach to sequential grouping.

VII. CONCLUSION

We have proposed a monaural CASA system that segregates unvoiced speech by performing onset-offset-based segmentation and feature-based classification. This system, together with our previous model for voiced speech segregation, yields a complete system that segregates both voiced and unvoiced speech from nonspeech interference. To our knowledge, this is the first systematic study on unvoiced speech segregation. Quantitative evaluation shows that our system captures most of unvoiced speech without including much interference.

ACKNOWLEDGMENT

This research was supported in part by an AFOSR grant (No. FA9550-04-01-0117), an AFRL grant (No. FA8750-04-1-0093), and an NSF grant (No. IIS-0534707).

NOMENCLATURE

a = Order of a gammatone filter
$A(c,m,\tau)$ = Autocorrelation function of filter response at delay τ in channel c and frame m
$\overline{A(c,m)}$ = Average of $A(c,m,\tau)$ over τ
$A_E(c,m,\tau)$ = Autocorrelation function of response envelope at delay τ in channel c and frame m
$\overline{A_E(c,m)}$ = Average of $A_E(c,m,\tau)$ over τ
α = Ratio of total speech energy to total voiced speech energy
b = Equivalent rectangular bandwidth of a gammatone filter
c = Filter channel index
$C(c,m)$ = Cross-channel correlation of filter responses between channels c and $c+1$ at frame m
$C_E(c,m)$ = Cross-channel correlation of response envelopes between channels c and $c+1$ at frame m
E_1 = Total target energy in voiced speech frames
E_2 = Total interference energy in voiced speech frames
f = Center frequency of a gammatone filter
f_c = Center frequency of filter channel c
$g(f,t)$ = Impulse response of a gammatone filter centered at frequency f
H = Hypothesis
H_0 = Hypothesis that a T-F region is target dominant
$H_{0,a}$ = Hypothesis that a T-F region is dominated by an expanded obstruent
$H_{0,b}$ = Hypothesis that a T-F region is dominated by any phoneme other than an expanded obstruent
H_1 = Hypothesis that a T-F region is interference dominant
m = Time frame index
n = Discrete time
N_1 = Total number of voiced speech frames
N_2 = Total number of frames other than voiced speech frames
p = Probability density
P = Probability
P_{EL} = Percentage of energy loss
P_{NR} = Percentage of noise residue
s = Segment
S_T^1 = Segregated stream for voiced speech
S_T^2 = Segregated stream for voiced speech and voiced T-segments
S_T^3 = Segregated stream for both voiced and unvoiced speech
t = Continuous time
T_m = Frame shift
T_n = Discrete sampling time
τ = Time delay
u_{cm} = Time-frequency unit of channel c and frame m
$x(t)$ = Input signal
$x(c,t)$ = Response of filter channel c to input signal
$x_E(c,t)$ = Response envelope of filter channel c
$X(c,m)$ = Cochleagram value in channel c and frame m
$X(m)$ = Cochleagram at frame m
$\mathbf{X}_s$ = Cochleagram of segment s
$X_s(c,m)$ = Cochleagram value of segment s in channel c and frame m
$X_s(m)$ = Cochleagram of segment s at frame m
$X_T(c,m)$ = Cochleagram value of segregated voiced target in channel c and frame m
$X_T(m)$ = Cochleagram of segregated voiced target at frame m

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

Ali, A. M. A., and Van der Spiegel, J. (**2001a**). "Acoustic-phonetic features for the automatic classification of fricatives," J. Acoust. Soc. Am. **109**, 2217–2235.
Ali, A. M. A., and Van der Spiegel, J. (**2001b**). "Acoustic-phonetic features for the automatic classification of stop consonants," IEEE Trans. Audio, Speech, Lang. Process. **9**, 833–841.
Barker, J., Cooke, M., and Ellis, D. (**2005**). "Decoding speech in the presence of other sources," Speech Commun. **45**, 5–25.
Benesty, J., Makino, S., and Chen, J., Ed. (**2005**). *Speech Enhancement* (Springer, New York).
Boersma, P., and Weenink, D. (**2004**). PRAAT, doing phonetics by computer, Version 4.2.31, (http://www.fon.hum.uva.nl/praat/) Last viewed 6/20/08.
Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).
Bridle, J. (**1989**). "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, Architectures, and Applications*, edited by F. Fogelman-Soulie and J. Herault, (Springer, New York), pp. 227–236.
Brown, G. J., and Cooke, M. (**1994**). "Computational Auditory Scene Analysis," Comput. Speech Lang. **8**, 297–336.
Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Canny, J. (**1986**). "A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell. **8**, 679–698.
Cole, R. A., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," in Proceedings of IEEE ICASSP, pp. 853–856.
Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.
Dewey, G. (**1923**). *Relative Frequency of English Speech Sounds* (Harvard University Press, Cambridge, MA).
Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York).
Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, New York).
French, N. R., Carter, C. W., and Koenig, W. (**1930**). "The words and sounds of telephone conversations," Bell Syst. Tech. J. **9**, 290–324.
Garofolo, J. *et al.* (**1993**). "DARPA TIMIT acoustic-phonetic continuous speech corpus," Technical Report No. NISTIR 4930, National Institute of Standards and Technology.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Greenberg, S., Hollenback, J., and Ellis, D. (**1996**). "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in Proceedings of ICSLP, pp. 24–27.
Helmholtz, H. (**1863**). *On the Sensation of Tone*, 2nd English ed., translated by A. J. Ellis (Dover, New York).
Hermansky, H., and Sharma, S. (**1999**). "Temporal patterns (TRAPs) in ASR of noisy speech," in Proceedings of IEEE ICASSP, pp. 289–292.

Hu, G. (**2006**). "Monaural speech organization and segregation," Ph.D. thesis, The Ohio State University Biophysics Program.
Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 79–82.
Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.
Hu, G., and Wang, D. L. (**2005**). "Separation of fricatives and affricates," in Proceedings of IEEE ICASSP, pp. 749–752.
Hu, G., and Wang, D. L. (**2006**). "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, edited by E. Hansler and G. Schmidt (Springer, Heidelberg, Germany), pp. 485–515.
Hu, G., and Wang, D. L. (**2007**). "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio, Speech, Lang. Process. **15**, 396–405.
Huang, X., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing: A Guide to Theory, Algorithms, and System Development* (Prentice-Hall, Upper Saddle River, NJ).
Ladefoged, P. (**2001**). *Vowels and Consonants* (Blackwell, Oxford, UK).
Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–134.
Lyon, R. F. (**1984**). "Computational models of neural auditory processing," in Proceedings of IEEE ICASSP, pp. 41–44.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK).
Nooteboom, S. G. (**1997**). "The prosody of speech: Melody and rhythm," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle, and J. Laver (Blackwell, Oxford, UK), pp. 640–673.
Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. **60**, 911–918.
Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS final report, Part B: Implementing a gammatone filterbank," Report No. 2341, MRC Applied Psychology Unit.
Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413–422.
Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Radfar, M. H., Dansereau, R. M., and Sayadiyan, A. (**2007**). "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," EURASIP J. Audio Speech Music Process., Paper No. 84186.
Romeny, B. H., Florack, L., Koenderink, J., and Viergever, M., Ed. (**1997**). *Scale-Space Theory in Computer Vision* (Springer, New York).
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shao, Y., and Wang, D. L. (**2006**). "Model-based sequential organization in cochannel speech," IEEE Trans. Audio, Speech, Lang. Process. **14**, 289–298.
Slaney, M., and Lyons, R. F. (**1990**). "A perceptual pitch detector," in Proceedings of IEEE ICASSP, pp. 357–360.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J. (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Netw. **10**, 684–697.
Wang, D. L., and Brown, G. J., Ed. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ).
Wang, D. L., and Hu, G. (**2006**). "Unvoiced speech segregation," in Proceedings of IEEE ICASSP, pp. 953–956.
Weintraub, M. (**1985**). "A theory and computational model of auditory monaural sound separation," Ph.D. thesis, Stanford University Department of Electrical Engineering.