Auditory Segmentation Based on Onset and Offset Analysis

Guoning Hu and DeLiang Wang, *Fellow, IEEE*

***Abstract*—A typical auditory scene in a natural environment contains multiple sources. Auditory scene analysis (ASA) is the process in which the auditory system segregates a scene into streams corresponding to different sources. Segmentation is a major stage of ASA by which an auditory scene is decomposed into segments, each containing signal mainly from one source. We propose a system for auditory segmentation by analyzing onsets and offsets of auditory events. The proposed system first detects onsets and offsets, and then generates segments by matching corresponding onset and offset fronts. This is achieved through a multiscale approach. A quantitative measure is suggested for segmentation evaluation. Systematic evaluation shows that most of target speech, including unvoiced speech, is correctly segmented, and target speech and interference are well separated into different segments.**

***Index Terms*—Auditory segmentation, event detection, multiscale analysis, onset and offset.**

I. Introduction

In a natural environment, multiple sounds from different sources form a typical auditory scene. An effective system that segregates target speech in a complex acoustic environment is required for many applications, such as robust speech recognition in noise and hearing aids design. In these applications, a monaural (one microphone) solution of speech segregation is often desirable. Many techniques have been developed to enhance speech monaurally, such as spectral subtraction [20] and hidden Markov models [30]. Such techniques tend to assume *a priori* knowledge or statistical properties of interference, and these assumptions are often too strong in realistic situations. Other approaches, including sinusoidal modeling [21] and comb filtering [11], attempt to extract speech by exploiting the harmonicity of voiced speech. Obviously, their approaches cannot handle unvoiced speech. Monaural speech segregation remains a very challenging task.

On the other hand, the auditory system shows a remarkable capacity in monaural segregation of sound sources. This perceptual process is referred to as auditory scene analysis (ASA) [4]. According to Bregman, ASA takes place in the brain in two stages: The first stage decomposes an auditory scene into segments (or sensory elements) and the second stage groups segments into streams. Considerable research has been carried out to develop *computational auditory scene analysis* (CASA) systems for sound separation and has obtained success in separating voiced speech [5], [8], [15], [18], [33], [34] (see [6], [12] for recent reviews). A typical CASA system decomposes an auditory scene into a matrix of time-frequency (T-F) units via bandpass filtering and time windowing. Then, the system separates sounds from different sources in two stages, *segmentation* and *grouping*. In segmentation, neighboring T-F units responding to the same source are merged into segments. In grouping, segments likely belonging to the same source are grouped together.

We should clarify that the term *segmentation* used in CASA has a different meaning than that in speech segmentation used in speech processing, which refers to identifying temporal boundaries between speech units (e.g., phonemes or syllables) of clean speech. Auditory segmentation here occurs on a two-dimensional (2-D) time-frequency representation of the input scene. In addition, the scene as a rule contains multiple sound sources. Segmentation in CASA has a similar meaning as segmentation in visual analysis (more discussion below).

In addition to the conceptual importance of segmentation for ASA, a segment as a region of T-F units contains global information of the source that is missing from individual T-F units, such as spectral and temporal envelope. This information could be key for distinguishing sounds from different sources. As shown in [18], grouping segments instead of individual T-F units is more robust for segregating voiced speech. A recent model of robust automatic speech recognition operates directly on auditory segments [2]. In our view, effective segmentation provides a foundation for grouping and is essential for successful CASA.

Previous CASA systems generally form segments according to two assumptions [5], [8], [18], [33]. First, signal from the same source likely generates responses with similar temporal or periodic structure in neighboring auditory filters. Second, signals with good continuity in time likely originate from the same source. The first assumption works well for harmonic sounds, but not for noise-like signals, such as unvoiced speech. The second assumption is problematic when target and interference have significant overlap in time.

From a computational standpoint, auditory segmentation corresponds to image segmentation, which has been extensively studied in computer vision. In image segmentation, the main task is to find bounding contours of visual objects. These contours usually correspond to sudden changes of certain local image properties, such as luminance and color. In auditory

Manuscript received September 9, 2005; revised January 19, 2006. This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-04-01-0117, by the Air Force Research Laboratory under Grant FA8750-04-1-0093, and by the National Science Foundation under Grant IIS-0081058. An earlier version of this paper was presented at the 2004 ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hong-Goo Kang.

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

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

Digital Object Identifier 10.1109/TASL.2006.881700

segmentation, the corresponding task is to find onsets and offsets of individual auditory events, which correspond to sudden changes of acoustic energy. In this paper, we propose a system for auditory segmentation based on onset and offset analysis of auditory events. Onsets and offsets are important ASA cues [4] for the reason that different sound sources in an environment seldom start and end at the same time. In addition, there is strong evidence for onset detection by auditory neurons [27]. There are several advantages for applying onset and offset analysis to auditory segmentation. In the time domain, onsets and offsets form boundaries between sounds from different sources. Common onsets and offsets provide natural cues to integrate sounds from the same source across frequency. In addition, since onset and offset cues are common to all types of sounds, the proposed system can in principle deal with both voiced and unvoiced speech.

Specifically, we apply a multiscale analysis, motivated by scale-space theory widely used in image segmentation [29], to onset and offset analysis for auditory segmentation. The advantage of using a multiscale analysis is to provide different levels of detail for an auditory scene so that one can detect and localize auditory events at appropriate scales. Our multiscale segmentation takes place in three stages. First, an auditory scene is smoothed to different degrees (scales). Second, the system detects onsets and offsets at certain scales, and forms segments by matching individual onset and offset fronts. Third, the system generates a final set of segments by integrating analysis at different scales. Scale-space analysis for speech segmentation as in speech processing (see earlier discussion) has been studied before [24].

This paper is organized as follows. In Section II, we propose a working definition for an auditory event to clarify the computational goal of segmentation. In Section III, we first give a brief description of the system and then present the details of each stage. We propose a quantitative measure to evaluate the performance of auditory segmentation in Section IV. The results of the system on segmenting target speech in noise are reported in Section V. This paper concludes with a discussion in Section VI.

II. What is an Auditory Event?

Because at any time there are infinite acoustic events taking place simultaneously in the world, one must limit the focus of CASA to an acoustic environment relative to a listener; in other words, only events audible to a listener should be considered. To determine the audibility of a sound, two perceptual effects need to be considered. First, a sound must be audible on its own, i.e., its intensity must exceed a certain level, referred to as the absolute threshold in a frequency band [25]. Second, when there are multiple sounds in the same environment, a weaker sound tends to be masked by a stronger one [25]. Hence, we consider a sound to be audible in a local T-F region if it satisfies the following two criteria.

- Its intensity is above the absolute threshold.
- Its intensity is higher than the summated intensity of all other signals in that region.

The absolute threshold of a sound depends on frequency and is different among listeners [25]. For young adults with normal hearing, the absolute threshold is about 15 dB sound pressure level (SPL) within the frequency range of [300 Hz, 10 kHz] [22]. Therefore, we take 15-dB SPL as a constant absolute threshold for the sake of simplicity. Based on the above criteria, we define an auditory event as the collection of all the audible T-F regions for an acoustic event. Thus, the computational goal of auditory segmentation is to generate segments for contiguous T-F regions from the same auditory event. This goal is consistent with the ASA principle of exclusive allocation, that is, a T-F region should be attributed to only one event [4]. We note that the exclusive allocation principle seems to contradict the fact that acoustic signals tend to add linearly (see, e.g., [1]). Besides the aforementioned auditory masking phenomenon, there is considerable evidence supporting this principle from both human speech intelligibility [7], [28] and automatic speech recognition [9], [28] studies (for an extensive discussion, see [32]).

To make this goal of auditory segmentation concrete requires a T-F representation of an acoustic input. Here, we employ a cochleagram representation of an acoustic signal, which refers to analyzing the signal in frequency by cochlear filtering (e.g., by a gammatone filterbank) followed by some form of nonlinear rectification corresponding to hair cell transduction, and in time through some form of windowing [23]. Specifically, we use a filterbank with 128 gammatone filters centered from 50 Hz to 8 kHz [26], and decompose filter responses into consecutive 20-ms windows with 10-ms window shifts. [18], [33]. Fig. 1(a) shows such a cochleagram for a mixture of a target female utterance and crowd noise with music, with the overall signal-to-noise ratio (SNR) of 0 dB. Here, the nonlinear rectification is simply the response energy within each T-F unit. With this T-F representation, we obtain the ideal segments of an event in an acoustic mixture as follows. First, we mark the audible T-F units of the event according to the premixing target and interference. Then we merge all marked units into spatially contiguous regions; each region then corresponds to a segment. Fig. 1(b) shows the resulting bounding contours (black line) of the target segments in the mixture. Gray regions form the background corresponding to the entire interference. Because the passbands of gammatone filters are relatively wide, particularly in the high-frequency range, adjacent harmonics may activate a number of adjacent filters. As a result, an ideal segment can combine several harmonics, as shown in Fig. 1(b).

As a working definition, we consider a phoneme, a basic phonetic unit of speech, as an acoustic event. There are two issues for treating individual phonemes as events. First, two types of phonemes, stops and affricates, have clear boundaries between a closure and a subsequent release in the middle of these phonemes. Therefore, we treat a closure in a stop or an affricate as an event on its own. This way, the acoustic signal within each event is generally stable. The second issue is that neighboring phonemes can be coarticulated. As a result, coarticulation may lead to unnatural boundaries between some consecutive ideal segments. These ideal segments may be put together by a real segmentation system, creating a case of under-segmentation. Alternatively, one may define a syllable, a word, or even a whole utterance from the same speaker as an acoustic event. However, in such a definition, many valid acoustic boundaries between phonemes are not taken into account. Consequently, some ideal segments are likely to be divided by a segmentation

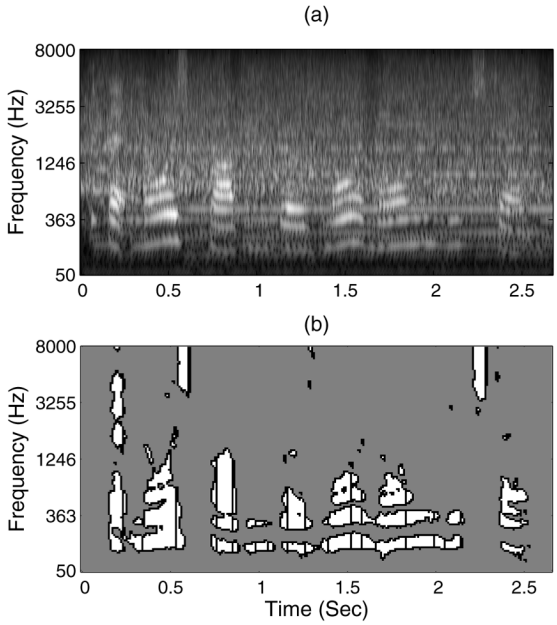


Fig. 1. Sound mixture and its ideal speech segments. (a) Cochleagram of a female utterance, "That noise problem grows more annoying each day," mixed with a crowd noise with music. (b) Bounding contours (black line) of the ideal segments of the utterance. The total number of ideal segments is 96.

system into smaller segments, creating a case of over-segmentation. We will come back to this issue in the evaluation and discussion sections.

III. System Description

Our system estimates ideal segments of auditory events via an analysis of signal onsets and offsets. Onsets and offsets, corresponding to sudden intensity changes, tend to delineate auditory events. In addition, onset/offset times of a segment, which is a part of an event, usually vary smoothly across frequency. Such smooth variation is partly due to the fact that certain speech events, such as stops and fricatives, exhibit smooth-varying onset and offset boundaries in certain ranges of frequency. Also, the passbands of neighboring frequency channels have significant overlap. Temporal alignment is an effective cue to group neighboring frequency channels. As shown in Fig. 1(b), even with strong interference, boundaries of most segments are reasonably smooth across frequency.

Fig. 2 gives the diagram of our system. An acoustic mixture is first normalized so that the average intensity is 60-dB SPL. Then it is passed through a bank of gammatone filters [26] (see Section II). To extract its temporal envelope, the output from each filter channel is half-wave rectified, low-pass filtered (a filter with a 74.5-ms Kaiser window and a transition band from 30 to 60 Hz) and downsampled to 400 Hz. The temporal envelope, indicating the intensity of a filter output, is used for onset and offset analysis. Note that, unlike the cochleagram representation, we do not divide the temporal envelope into consecutive frames in this analysis.

Onsets and offsets correspond to the peaks and valleys of the time derivative of the intensity. However, because of the intensity fluctuation within individual events, many peaks and valleys of the derivative do not correspond to real onsets and offsets.

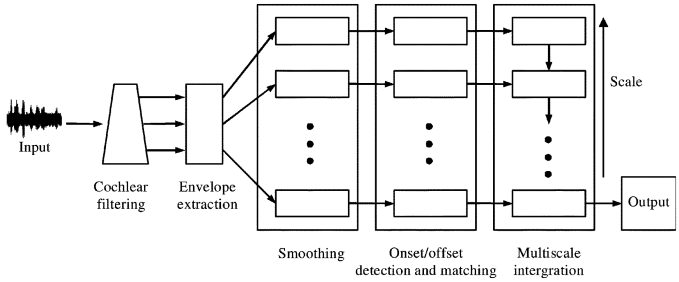


Fig. 2. Diagram of the system. Note that the scale increases from bottom to top.

Therefore, the intensity is smoothed over time to reduce the fluctuations in the smoothing stage. The system further smoothes the intensity over frequency to enhance the alignment of onsets and offsets. The degree of smoothing is called the scale—the larger the scale is, the smoother the intensity becomes.

In the stage of onset/offset detection and matching, the system detects onsets and offsets in each filter channel and merges detected onsets and offsets into onset and offset fronts if they occur at close times. It then matches individual onset and offset fronts to form segments.

As a result of smoothing, event onsets and offsets of small T-F regions may be blurred at a larger (coarser) scale. Consequently, the system may miss small events or generate segments combining different events, a case of under-segmentation. On the other hand, at a smaller (finer) scale, the system may be sensitive to insignificant intensity fluctuations within individual events. Consequently, the system tends to separate a continuous event into several segments, a case of over-segmentation. Therefore, it is difficult to obtain satisfactory segmentation with a single scale. Our system handles this issue by integrating onset/offset information across different scales in an orderly manner in the stage of multiscale integration, which yields the final set of segments. The detailed description of the last three stages is given below.

A. Smoothing

Smoothing corresponds to low-pass filtering. Our system first smoothes the intensity over time with a low-pass filter and then smoothes the intensity over frequency with a Gaussian kernel. Let $v(c,t,0,0)$ denote the initial intensity—logarithmic temporal envelope—at time t in filter channel c. We have

$$v(c,t,0,s_t) = v(c,t,0,0) * h(s_t) \quad (1)$$

$$v(c,t,s_c,s_t) = v(c,t,0,s_t) * g(0,s_c) \quad (2)$$

where $h(s_t)$ is a low-pass filter with passband $[0, 1/s_t]$ in hertz, and $g(0,s_c)$ is a Gaussian function with zero mean and standard deviation s_c. "$*$" denotes convolution. The parameter pair (s_c,s_t) indicates the degree of smoothing. The larger (s_c,s_t) is, the smoother $v(c,t,s_c,s_t)$ is. We refer to (s_c,s_t) as the (2-D) scale, and the smoothed intensities at different scales form the so-called scale space [29].

Here we apply low-pass filtering instead of generic diffusion [29] for smoothing over time because this way it is more intuitive to decide the appropriate scales for segmentation according to the acoustic and perceptual properties of the target we are interested in (see Section III-C). In an earlier study, we applied anisotropic diffusion and obtained similar results [19].

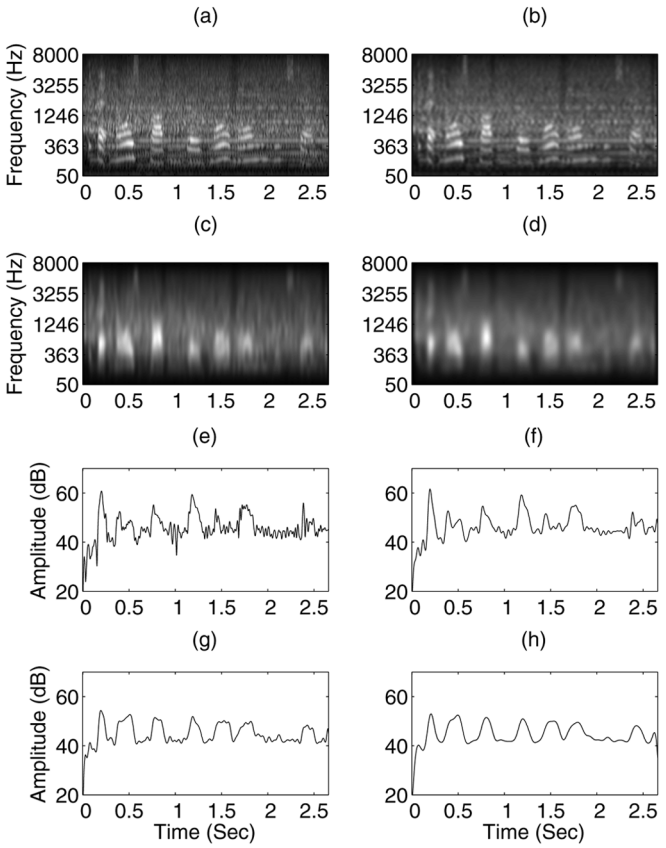


Fig. 3. Smoothed intensity values at different scales. (a) Initial intensity for all the channels. (b) Smoothed intensity at the scale (1/2, 1/14). (c) Smoothed intensity at the scale (6, 1/14). (d) Smoothed intensity at the scale (6, 1/4). (e) Initial intensity in a channel centered at 560 Hz. (f) Smoothed intensity in the channel at the scale (1/2, 1/14). (g) Smoothed intensity in the channel at the scale (6, 1/14). (h) Smoothed intensity in the channel at the scale (6, 1/4). The input is the same as shown in Fig. 1(a).

As an example, Fig. 3 shows the initial and smoothed intensities for the input mixture shown in Fig. 1(a). Fig. 3(a) shows the initial intensity. The smoothed intensities at three scales, (1/2, 1/14), (6, 1/14), and (6, 1/4) are shown in Fig. 3(b)–(d), respectively. To display more details, Fig. 3(e)–(h) shows the initial and smoothed intensities at these three scales in a single frequency channel centered at 560 Hz, respectively (see Section III-C for the implementation details of the low-pass filter). As shown in the figure, the smoothing process gradually reduces the intensity fluctuations. Local details of onsets and offsets also become blurred, but the major intensity changes corresponding to onsets and offsets are preserved.

B. Onset/Offset Detection and Matching

At a certain scale (s_c, s_t), onset and offset candidates are detected by marking peaks and valleys of the time derivative of the smoothed intensity

$$\begin{aligned}\frac{d}{dt}v(c,t,s_c,s_t) &= \frac{d}{dt}[v(c,t,0,0) * h(s_t) * g(0,s_c)] \\ &= \frac{d}{dt}[v(c,t,0,0) * h(s_t)] * g(0,s_c) \\ &= v(c,t,0,0) * \left[\frac{d}{dt}h(s_t)\right] * g(0,s_c). \quad (3)\end{aligned}$$

An onset candidate is removed if the corresponding peak is smaller than a threshold θ_{ON}, which suggests that the candidate is likely an insignificant intensity fluctuation. Since the peaks corresponding to true onsets are usually significantly higher than other peaks, we use the threshold $\theta_{\mathrm{ON}}(s_c, s_t) = \mu(s_c, s_t) + \sigma(s_c, s_t)$, where $\mu(s_c, s_t)$ and $\sigma(s_c, s_t)$ are the mean and standard deviation of all the derivative values, respectively. We have also tested $\mu(s_c, s_t) + 0.5\sigma(s_c, s_t)$, $\mu(s_c, s_t) + 1.5\sigma(s_c, s_t)$, and $\mu(s_c, s_t) + 2\sigma(s_c, s_t)$ as the threshold, but the performance is not as good.

Then in each filter channel, the system determines the offset time for each onset candidate. Let $t_{\mathrm{ON}}[c, i]$ represent the time of the ith onset candidate in channel c. The corresponding offset time, denoted as $t_{\mathrm{OFF}}[c, i]$, is chosen among the offset candidates located between $t_{\mathrm{ON}}[c, i]$ and $t_{\mathrm{ON}}[c, i + 1]$. The decision is simple if there is only one offset candidate in this range. When there are multiple offset candidates, we choose the one with the largest intensity decrease, i.e., the smallest dv/dt. Note that there is at least one offset candidate between two onset candidates since there is at least one local minimum between two local maxima.

Since frequency components with close onset or offset times likely arise from the same source, our system connects common onsets and offsets into onset and offset fronts. There is usually some onset time shifts in adjacent channels in response to the same event. This is because the onset times of the components of an acoustic event may vary across frequency. Masking by interference may further shift detected onset and offset times. Also, each gammatone filter introduces a small, frequency-dependent delay in its response. Based on these considerations, we allow a tolerance interval when connecting onset/offset candidates in neighboring frequency channels. Specifically, we connect an onset candidate with the closest onset candidate in an adjacent channel if their distance in time is less than a certain threshold; the same applies to offset candidates. This threshold value should not be too small; otherwise onsets (or offsets) from the same event will be prevented from joining together. On the other hand, a threshold value that is too big will connect some onsets from different events together. As found in [10], [31], human listeners start to segregate two sounds when their onset times differ by 20–30 ms. Therefore, we choose 20 ms as the threshold. If an onset front thus formed occupies less than three channels, we do not further process it because the front is likely insignificant. Onset and offset fronts are vertical contours across frequency in a cochleagram.

The next step is to match individual onset and offset fronts to form segments. Let $(t_{\mathrm{ON}}[c, i_1], t_{\mathrm{ON}}[c + 1, i_2], \ldots, t_{\mathrm{ON}}[c + m - 1, i_m])$ denote an onset front with m consecutive channels, and $(t_{\mathrm{OFF}}[c, i_1], t_{\mathrm{OFF}}[c+1, i_2], \ldots, t_{\mathrm{OFF}}[c+m-1, i_m])$ the corresponding offset times as described earlier. The system first selects all the offset fronts that cross at least one of these offset times. Among them, the one that crosses the most of the these offset times is chosen as the matching offset front, and all the channels from c to $c + m - 1$ occupied by the matching offset front are labeled as "matched." The offset times in these matched channels are updated to those of the matching offset front. If all the channels from c to $c + m - 1$ are labeled as matched, the matching procedure is finished. Otherwise, the process repeats for the remaining unmatched chan-

nels. In the end, the T-F region between $(t_{\rm ON}[c, i_1], t_{\rm ON}[c+1, i_2], \ldots, t_{\rm ON}[c+m-1, i_m])$ and the updated offset times $(t_{\rm OFF}[c, i_1], t_{\rm OFF}[c+1, i_2], \ldots, t_{\rm OFF}[c+m-1, i_m])$ yields a segment.

In the aforementioned segmentation, we assume that onset candidates in adjacent channels correspond to the same event if they are sufficiently close in time. This assumption may not always hold. To reduce the error of merging different sounds with similar onsets, we further require the corresponding temporal envelopes to be similar since sounds from the same source usually produce similar temporal envelopes. More specifically, for an onset candidate $t_{\rm ON}[c, i_1]$, let $t_{\rm ON}[c+1, i_2]$ be the closest onset candidate in an adjacent channel, and (t_1, t_2) be the overlapping duration between $(t_{\rm ON}[c, i_1], t_{\rm OFF}[c, i_1])$ and $(t_{\rm ON}[c+1, i_2], t_{\rm OFF}[c+1, i_2])$. The similarity between the temporal envelopes from these two channels in this duration is measured by their correlation (see [33])

$$C(c, i_1, i_2, s_c, s_t) = \sum_{t=t_1}^{t_2} \hat{v}(c, t, s_c, s_t)\hat{v}(c+1, t, s_c, s_t) \quad (4)$$

where $\hat{v}$ indicates the normalized v with zero mean and unity variance within (t_1, t_2). Then in forming onset fronts, we further require temporal envelope correlation to be higher than a threshold θ_C. By including this requirement, our system reduces the errors of accidentally merging sounds from different sources into one segment.

C. Multiscale Integration

Our system integrates analysis at different scales to form segments. It starts at a coarse scale, i.e., generating segments as described in Section III-B. Then, at a finer (smaller) scale, it locates more accurate onset and offset positions for segments, and new segments can be created within the current background. Segments are also expanded along the formed onset and offset fronts as follows. Let $(t_{\rm ON}[c, i_1], t_{\rm ON}[c+1, i_2], \ldots, t_{\rm ON}[c+m-1, i_m])$ and $(t_{\rm OFF}[c, i_1], t_{\rm OFF}[c+1, i_2], \ldots, t_{\rm OFF}[c+m-1, i_m])$ be the onset times and offset times of a segment occupying m consecutive channels. Note that lower-frequency channels are at lower positions in our cochleagram representation [see Fig. 1(a)]. The expansion works by considering the onset front at the current scale crossing $t_{\rm ON}[c+m-1, i_m]$ and the offset front crossing $t_{\rm OFF}[c+m-1, i_m]$. If both of these fronts extend beyond the segment, i.e., occupying channels above $c+m-1$, or channels with higher center frequencies, the segment will expand to include the channels that are crossed by both the onset and the offset fronts. Similarly, the expansion considers the channels below c, or the channels with lower center frequencies. At the end of expansion, segments with the same onset times in at least one channel are merged.

One could also start from a fine scale and then move to coarser scales. However, in this case, the chances of over-segmenting an input mixture are much higher, which is less desirable than under-segmentation since in subsequent grouping larger segments are preferred (see Section IV).

In this study, we are interested in estimating T-F segments of speech. Since temporal envelope variations down to 4 Hz are essential for speech intelligibility [13], [14], the system starts segmentation at the time scale $s_t = 1/4$. In addition, the system starts at the frequency scale $s_c = 6$. We have also considered starting at $s_c = 8$ and $s_c = 4$. In both situations, the system performed slightly worse. In the results reported here, the system forms segments in three scales from coarse to fine: $(s_c, s_t) =$ (6, 1/4), (6,1/14), and (1/2, 1/14). At the finest scale, i.e., (1/2, 1/14), the system does not form new segments since these segments tend to occupy insignificant T-F regions. The threshold θ_C is 0.95, 0.95, and 0.85, respectively; a larger θ_C is used in the first two scales because smoothing over frequency increases the similarity of temporal envelopes in adjacent channels. At each scale, a low-pass filter with a 182.5-ms Kaiser window and a 10-Hz transition band is applied for smoothing over time. Note that the passband of the filter corresponds to the time scale. We have also considered segmentation using more scales and with different types and parameters for the low-pass filter, and obtained similar results.

Fig. 4 shows the bounding contours of segments at different scales for the mixture in Fig. 1(a), where Fig. 4(a) shows the segments formed at the starting scale (6, 1/4), and Fig. 4(b) and (c) those from the multiscale integration of 2 and 3 scales, respectively. The background is represented by gray. Compared with the ideal segments in Fig. 1(b), the system captures a majority of speech events at the largest scale, but misses some small segments. As the system integrates analysis at smaller scales, more speech segments are formed; at the same time, more segments from interference also appear. Note that the system does not specify the sound source for each segment, which is the task of grouping not addressed here.

IV. Evaluation Metrics

Only a few previous models have explicitly addressed the problem of auditory segmentation [5], [8], [18], [33], but none have separately evaluated the segmentation performance. How to quantitatively evaluate segmentation results is a complex issue, since one has to consider various types of mismatch between a collection of ideal segments and that of estimated segments. On the other hand, similar issues occur also in image segmentation, which has been extensively studied in computer vision and image analysis. So we have adapted region-based metrics by Hoover *et al.* [16], which have been widely used for evaluating image segmentation systems.

Our region-based evaluation compares estimated segments with ideal segments of a target source since in many situations one is interested in only target extraction. In other words, how the system segments interference will not be considered in evaluation. Hence, we treat all the T-F regions dominated by interference as the ideal background. Note that this can be extended to situations where one is interested in evaluating segmentation of multiple sources, say, when interference is a competing talker. For example, one may evaluate how the system segments each source separately.

The general idea is to examine the overlap between ideal segments and estimated segments. Based on the degree of overlapping, we label a T-F region as correct, under-segmented, over-segmented, missing, or mismatch. Fig. 5(a) illustrates these cases, where ovals represent ideal target segments (numbered with Arabic numerals) and rectangles estimated segments

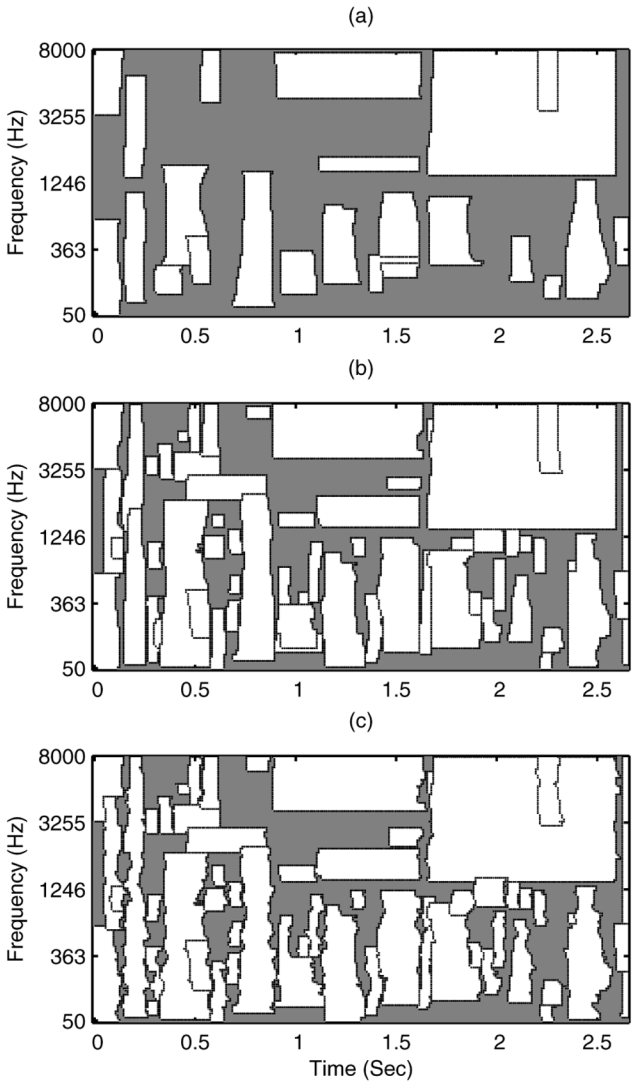


Fig. 4. Bounding contours of estimated segments from multiscale analysis. (a) One-scale analysis at the scale of (6, 1/4). (b) Two-scale analysis at the scales of (6, 1/4) and (6, 1/14). (c) Three-scale analysis at the scales of (6, 1/4), (6, 1/14), and (1/2, 1/14). The input is the same as shown in Fig. 1(a). The background is represented by gray.

(numbered with Roman numerals). As shown in Fig. 5(a), estimated segment I well covers ideal segment 1, and we label the overlapping region as correct. So is the overlap between segment 7 and VII. Segment III well covers two ideal segments, 3 and 4, and the overlapping regions are labeled as under-segmented. Segment IV and V are both well covered by segment 5, and the overlapping regions are labeled as over-segmented. All the remaining regions from ideal segments—segments 2 and 6 and the parts of segments 5 and 7 marked by diagonal lines—are labeled as missing. The black region in segment I belongs to the ideal background, but since it is merged with ideal segment 1 into an estimated segment we label this black region as mismatch, as well as the black region in segment III. Note the major differences among under-segmentation, missing, and mismatch. Under-segmentation denotes the error of combining multiple T-F regions belonging to different segments of the same source, whereas missing and mismatch denote the error

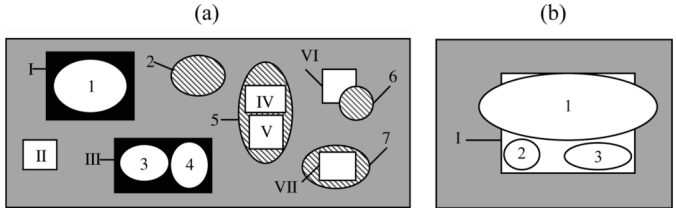


Fig. 5. Illustration of different matching situations between ideal and estimated segments. (a) Correct segmentation, under-segmentation, over-segmentation, missing, and mismatch. (b) Multiple labels for one overlapping region. Here, an oval indicates an ideal segment and a rectangle an estimated one. The background is represented by gray.

of mixing T-F regions from different sources. Therefore, if an estimated segment combines T-F regions belonging to different speakers, it is not under-segmentation, but missing or mismatch depending on the degree of overlapping. Segment II is well covered by the ideal background, which is not considered in the evaluation. Much of segment VI is covered by the ideal background and, therefore, we treat the white region of the segment the same as segment II (Note the difference between I and VI).

Quantitatively, let $\{r_I[k]\}, k = 0, 1, \ldots, K$ be the set of ideal segments, where $r_I[0]$ indicates the ideal background and others the ideal segments of target. Let $\{r_S[l]\}, l = 0, 1, \ldots, L$, be the estimated segments produced by the system, where $r_S[l], l > 0$, corresponds to an estimated segment and $r_S[0]$ the estimated background. Let $r[k, l]$ be the overlapping region between $r_I[k]$ and $r_S[l]$. Furthermore, let $E[k, l]$, $E_I[k]$, and $E_S[l]$ denote the corresponding energy in these regions. Given a threshold, we define that an ideal segment $r_I[k]$ is well-covered by an estimated segment $r_S[l]$ if $r[k, l]$ includes most of the energy of $r_I[k]$. That is

$$E[k, l] > \theta_E \cdot E_I[k]. \tag{5}$$

Similarly, $r_S[l]$ is well-covered by $r_I[k]$ if

$$E[k, l] > \theta_E \cdot E_S[l]. \tag{6}$$

For any $\theta_E \in [0.5, 1)$, the above definition of well-coveredness ensures that an ideal segment is well covered by at most one estimated segment, and vice versa.

Then we label a nonempty overlapping region as follows.

- A region $r[k, l]$, $k > 0$ and $l > 0$ is labeled as correct if $r_I[k]$ and $r_S[l]$ are mutually well-covered.
- Let $\{r_I[k']\}, k' = k_1, k_2, \ldots, k'_K$, and $K' > 1$ be all the ideal target segments that are well-covered by one estimated segment, $r_S[l]$, $l > 0$. The corresponding overlapping regions, $\{r[k', l]\}, k' = k_1, k_2, \ldots, k'_K$, are labeled as under-segmented if these regions combined include most of the energy of $r_S[l]$, that is

$$\sum_{k'} E[k', l] > \theta_E \cdot E_S[l], \qquad k' = k_1, k_2, \ldots, k_{K'}. \tag{7}$$

- Let $\{r_S[l']\}, l' = l_1, l_2, \ldots, l'_L$, and $L' > 1$ be all the estimated segments that are well-covered by one ideal segment, $r_I[k]$, $k > 0$. The corresponding overlapping regions, $\{r[k, l']\}$, $l' = l_1, l_2, \ldots, l'_L$, are labeled as over-

segmented if these regions combined include most of the energy of $r_I[k]$, that is

$$\sum_{l'} E[k,l'] > \theta_E \cdot E_I[k], \qquad l' = l_1, l_2, \ldots, l_{L'}. \quad (8)$$

- If a region $r[k,l]$ is part of an ideal segment of target speech, i.e., $k > 0$, but cannot be labeled as correct, under-segmented, or over-segmented, it is labeled as missing.
- For a region $r[0,l]$, the overlap between the ideal background $r_I[0]$ and an estimated segment $r_S[l]$, it is labeled as mismatch if $r_S[l]$ is not well-covered by the ideal background.

According to the above definitions, some regions may be labeled as either correct or under-segmented. Fig. 5(b) illustrates this situation, where estimated segment I and ideal segment 1 are mutually well-covered. Hence, $r[1,I]$ is labeled as correct. On the other hand, segment I also well covers ideal segments 2 and 3, and obviously ideal segments 1-3 together well cover segment I. According to the definition of under-segmentation, $r[1,I]$, $r[2,I]$, and $r[3,I]$ should all be labeled as under-segmented. Therefore, $r[1,I]$ can be labeled as either correct or under-segmented. Similarly, some regions may be labeled as either correct or over-segmented. To avoid labeling a region more than once, we consider a region to be correctly labeled as long as it satisfies the definition of correctness.

Let E_C, E_U, E_O, E_M, and E_N be the summated energy in all the regions labeled as correct, under-segmented, over-segmented, missing, and mismatch, respectively. Further, let E_I be the total energy of all ideal segments of target, and E_S that of all estimated segments, except for the estimated background. We use the following metrics for evaluation.

- The correct percentage: $P_C = E_C/E_I \times 100\%$.
- The percentage of under-segmentation: $P_U = E_U/E_I \times 100\%$.
- The percentage of over-segmentation: $P_O = E_O/E_I \times 100\%$.
- The percentage of missing: $P_M = E_M/E_I \times 100\%$.
- The percentage of mismatch: $P_N = E_N/E_S \times 100\%$.

Since $E_C + E_U + E_O + E_M = E_I$, or $P_C + P_U + P_O + P_M = 100\%$, only three out of these four percentages need to be measured.

The advantage of evaluation according to each category is that it clearly shows different types of error. In the context of speech segregation, under-segmentation is not really an error since it basically produces larger segments for target speech, which is good for subsequent grouping. In image segmentation, the region size corresponding to each segment is used for evaluation literally. Here, we use the energy of each segment because for acoustic signal, T-F regions with strong energy are much more important to segment than those with weak energy.

V. Evaluation Results

To systematically evaluate the performance of the proposed system, we have applied it to a mixture corpus created by mixing 20 speech utterances and ten intrusions. We consider as target the utterances that are randomly selected from the TIMIT database. The phonetically-labeled TIMIT database provides phoneme boundaries, which are used to generate ideal segments. There are altogether 696 phonemes occurring in these sentences. The intrusions are: white noise, electrical fan, rooster crow and clock alarm, traffic noise, crowd noise in a playground, crowd noise with music (used earlier), crowd noise with clapping, bird chirp with waterflow, wind, and rain. This set of intrusions represents a broad range of real sounds encountered in typical acoustic environments. As described in Section II, we consider each phoneme as an acoustic event of speech and obtain ideal target segments from target speech and interference before mixing. Because estimated segments and ideal segments have different T-F representations (see Section III), we convert the estimated segments into the T-F representation of ideal segments before evaluation.

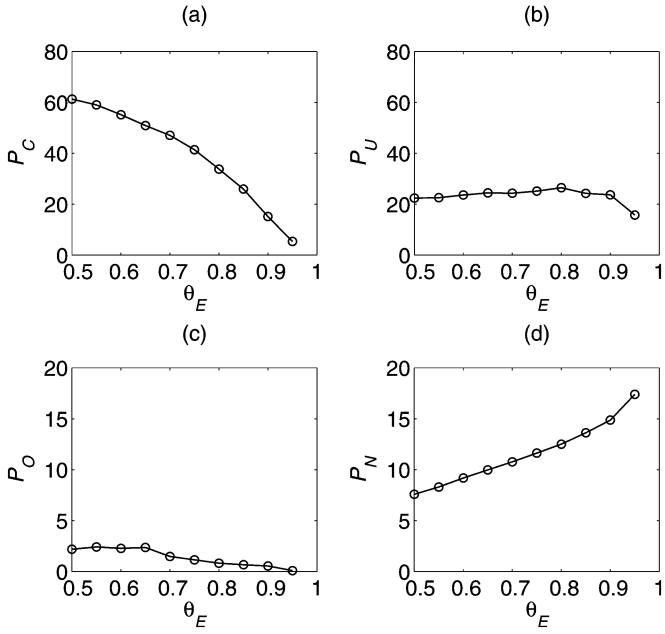


Fig. 6. Results of auditory segmentation. Target and interference are mixed at 0 dB SNR. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation. (d) Average percentage of mismatch.

Fig. 6 shows the average P_C, P_U, P_O, and P_N for different θ_E values. The evaluation is more stringent for higher θ_E. Note that we limit θ_E to be no smaller than 0.5 so that an ideal segment is well covered by at most one estimated segment, and vice versa (see Section IV). Speech and interference are mixed at 0 dB SNR. The total number of target events in these 200 mixtures is 24 753. As shown in the figure, the correct percentage is 61.3% when θ_E is 0.5, and it decreases to 5.4% as θ_E increases to 0.95. A significant amount of speech is under-segmented, which is due mainly to coarticulation of phonemes. As we have discussed in Section IV, under-segmentation is not really an error. By combining P_C and P_U together, the system correctly segments 83.8% of target speech when θ_E is 0.5. Even when θ_E increases to 0.85, more than 50% of speech is correctly segmented. In addition, we can see from the figure that over-segmentation is negligible. The main error comes from missing, which indicates that portions of target speech are buried in the background. The percentage of mismatch is 7.6% when θ_E is 0.5, and increases to 17.4% when θ_E increases to 0.95. Considering the overall SNR of 0 dB, the percentage of mismatch is

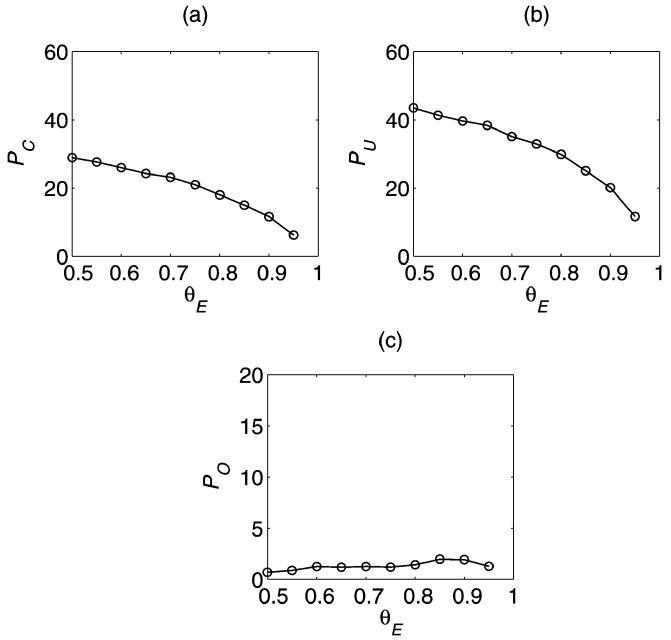


Fig. 7. Results of auditory segmentation for stops, fricatives, and affricates. Target and interference are mixed at 0 dB SNR. (a) Average correct percentage. (b) Average percentage of under-segmentation. (c) Average percentage of over-segmentation.

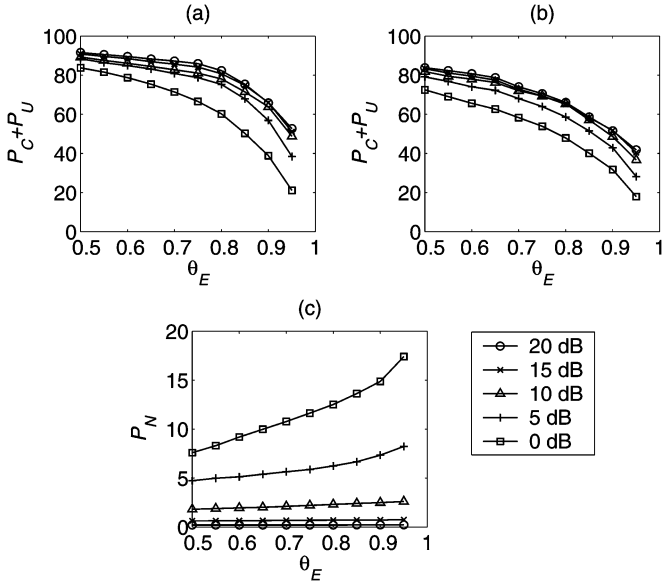


Fig. 8. Results of auditory segmentation at different SNR levels. (a) Average correct percentage plus the average percentage of under-segmentation for all the phonemes. (b) Average correct percentage plus the average percentage of under-segmentation for stops, fricatives, and affricates. (c) Average percentage of mismatch.

not significant. This shows that the interference and the target speech are well separated in the estimated segments.

Since voiced speech is generally much stronger than unvoiced speech, the above result mainly reflects the performance of the system on voiced speech. To see how the system performs on unvoiced speech, Fig. 7 shows the average P_C, P_U, and P_O for stops, fricatives, and affricates, which are the three main consonant categories that contain unvoiced speech energy. As shown in Fig. 7, much energy of these phonemes is under-segmented. As expected, the overall performance on these phoneme categories is not as good as that for other phonemes since unvoiced speech is weaker and more prone to interference. The average $P_C + P_U$ in the figure is 72.5% when θ_E is 0.5, and it drops below 50% when θ_E is larger than 0.75.

Fig. 8 shows the performance of the system at different SNR levels, where Fig. 8(a) shows the average $P_C + P_U$ for all the phonemes, Fig. 8(b) the average $P_C + P_U$ for stops, fricatives, and affricates, and Fig. 8(c) the average P_N. When SNR is 10 dB or higher, the interference has relatively insignificant influence on the system performance, and the $P_C + P_U$ scores are similar. The performance drops as SNR decreases beyond 10 dB, and the drop is most pronounced from 5 to 0 dB.

Because the low-frequency portion of speech is usually more intense than the high-frequency portion, the above energy-based evaluation may be dominated by the low-frequency range. To present a more balanced picture, we apply a first-order high-pass filter with the coefficient 0.95 to the input mixture to pre-emphasize its high-frequency portion, which approximately equalizes the average energy of speech in each filter channel. Then energy of each segment after pre-emphasis is used for evaluation. Fig. 9 presents a comparison with and without pre-emphasis for mixtures at 0 dB SNR. Fig. 9(a) and (b) shows the resulting average P_C and P_U for all the phonemes. With pre-emphasis the P_C scores are slightly higher than those without pre-emphasis, whereas the P_U scores are about 10% lower. This suggests that more voiced speech is under-segmented in the low-frequency range. Fig. 9(c) and (d) shows the average P_C and P_U for stops, fricatives, and affricates. With pre-emphasis, the P_C scores for these phonemes are much higher, whereas the P_U scores are much lower. The $P_C + P_U$ scores together are slightly higher with pre-emphasis. This suggests that our system under-segments most of the energy of stops, fricatives, and affricates in the low-frequency range, which is mainly voiced. On the other hand, it correctly separates most of the energy of stops, fricatives, and affricates in the high-frequency range, where the energy of unvoiced speech is more distributed, from neighboring phonemes as well as from interference. Fig. 9(e) shows the average P_N, which is reduced with pre-emphasis, showing less mismatch in the high-frequency range.

To put the system performance in perspective, we now compare it with the segmentation algorithm described by Brown and Cooke [5]. Their algorithm first produces spectral peak tracks on a frequency transition map and then extends each track in frequency by clustering cross-channel correlation values (a simplified algorithm [33] is compared in [19]). Fig. 10 shows the comparative results for mixtures at 0 dB SNR without pre-emphasis. Fig. 10(a) shows the average $P_C + P_U$ scores for all the phonemes. The Brown and Cooke algorithm yields much lower $P_C + P_U$ scores. The primary reason is that their algorithm is based on cross-channel correlation, which often fails to merge target speech across frequency because target speech may yield different responses in neighboring filter channels. Since their algorithm was mainly intended for segmenting voiced sound, a further comparison for only voiced speech in terms of $P_C + P_U$ is given in Fig. 10(b). In this case, the voiced portions of each utterance are determined using Praat, which has a standard pitch

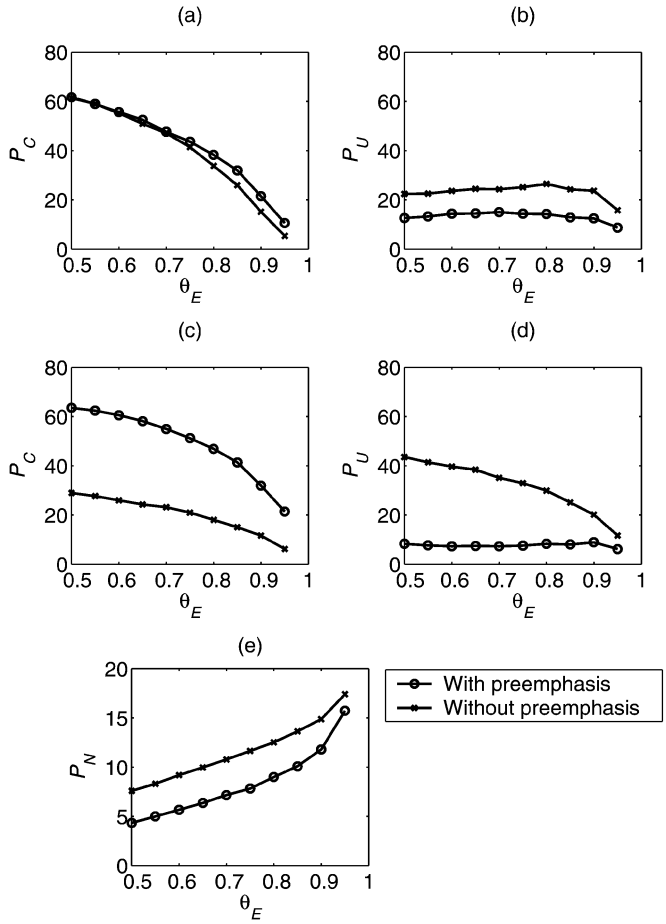


Fig. 9. Results of auditory segmentation with and without pre-emphasis. Target and interference are mixed at 0 dB SNR. (a) Average correct percentage for all the phonemes. (b) Average percentage of under-segmentation for all the phonemes. (c) Average correct percentage for stops, fricatives, and affricates. (d) Average percentage of under-segmentation for stops, fricatives, and affricates. (e) Average percentage of mismatch.

determination algorithm for clean speech [3]. The performance gap in Fig. 10(b) is not much different from that in Fig. 10(a). Fig. 10(c) shows the average P_N. Their algorithm produces lower P_N errors, because segmentation exploits harmonic structure and most intrusions in the evaluation corpus are noise-like. Taken together, our method performs much better than their algorithm for auditory segmentation.

VI. Discussion

To determine ideal segments of target speech, we need to decide what constitutes acoustic events of a speech utterance (see Section II). Here we treat a phoneme as an acoustic event. As we discussed in Section II, coarticulation between neighboring phonemes may create unnatural boundaries in ideal segments, a case of under-segmentation. This problem is partly taken care of in our evaluation which does not consider under-segmentation as an error. To avoid the problem of coarticulation, one could define a larger unit (e.g., a syllable or a word) as an acoustic event. As discussed earlier, over-segmentation becomes an issue in such a definition. Because it is not clear whether an instance of over-segmentation is caused by a true boundary between two phonemes or a genuine error, over-segmentation is a more thorny issue. This consideration has led us to choose phonemes as event units.

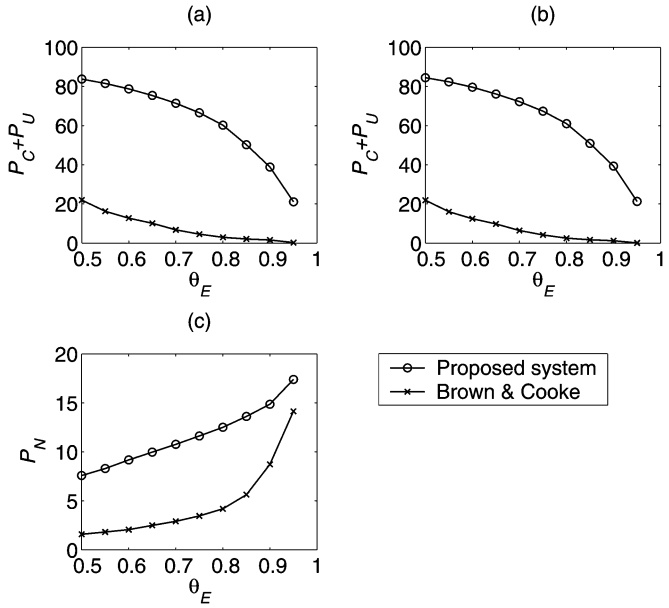


Fig. 10. Results of auditory segmentation for the proposed system and those from the Brown and Cooke algorithm. Target and interference are mixed at 0 dB SNR. (a) Average correct percentage plus the average percentage of under-segmentation for all the phonemes. (b) Average correct percentage plus the average percentage of under-segmentation for voiced portions of utterance. (c) Average percentage of mismatch.

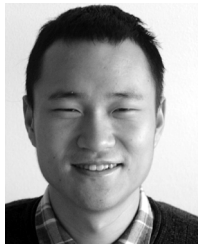
Our system employs two steps to integrate sounds from the same source across frequency based on common onset/offset and cross-channel correlation. The latter step helps to reduce the errors of merging different sounds with similar onsets. In our evaluation, the improvement from this step is not significant. This is mainly due to the fact that common onset and offset are already quite effective for our test corpus. However, under reverberant conditions, onset and offset information is likely to be more corrupted than that of temporal envelope. We expect that cross-channel correlation of temporal envelope will play a more significant role for segmentation in reverberant conditions.

In summary, our study on auditory segmentation makes a number of novel contributions. First, it provides a general framework for auditory segmentation. Second, it performs segmentation for general auditory events based on onset and offset analysis. Although it is well known that onset and offset are important ASA cues, few computational studies have explored their use. Brown and Cooke incorporated common onset and common offset as grouping cues but did not find significant performance improvements [5]. In a previous study, we demonstrated the utility of the onset cue for segregating stop consonants [17]. The present study further shows that event onsets and offsets may play a fundamental role in sound organization. Finally, our system generates segments for both unvoiced and voiced speech. Little previous research has been conducted on organization of unvoiced speech, and yet monaural speech segregation must address unvoiced speech.

References

[1] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Proc. ICASSP*, 2004, vol. 3, pp. 881–884.

[2] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," in *Speech Commun.*, 2005, vol. 45, pp. 5–25.
[3] P. Boersma and D. Weenink, Praat: Doing phonetics by computer, Version 4.2.31 2004 [Online]. Available: http://www.fon.hum.uva.nl/praat/
[4] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[5] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.
[6] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York: Springer, 2005, pp. 371–402.
[7] P. S. Chang, "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," M.S. thesis, Dept. Comput. Sci. Eng., The Ohio State Univ., Columbus, 2004.
[8] M. P. Cooke, *Modelling Auditory Processing and Organisation*. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[9] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Commun.*, 2001, vol. 34, pp. 267–285.
[10] C. J. Darwin, "Perceiving vowels in the presence of another sound: Constraints on formant perception," *J. Acoust. Soc. Amer.*, vol. 76, pp. 1636–1647, 1984.
[11] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.
[12] P. Divenyi, Ed., *Speech Separation by Humans and Machines*. Norwell, MA: Kluwer, 2005.
[13] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1053–1064, 1994.
[14] ——, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 2670–2680, 1994.
[15] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. Elec. Eng. and Comput. Sci., Mass. Inst. Technol., Cambridge, 1996.
[16] A. Hoover *et al.*, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 673–689, Jul. 1996.
[17] G. Hu and D. L. Wang, "Separation of stop consonants," in *Proc. ICASSP*, 2003, vol. 2, pp. 749–752.
[18] ——, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.
[19] ——, "Auditory segmentation based on event detection," in *Proc. ISCA Tutorial and Research Workshop on Stat. Percept. Audio Process.*, 2004.
[20] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithms, and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.
[21] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 731–740, Oct. 2001.
[22] M. C. Killion, "Revised estimate of minimal audible pressure: Where is the "missing 6 dB"?," *J. Acoust. Soc. Amer.*, vol. 63, pp. 1501–1510, 1978.
[23] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. ICASSP*, 1982, vol. 2, pp. 1282–1285.
[24] ——, "Speech recognition in scale space," in *Proc. ICASSP*, 1987, vol. 12, pp. 1265–1268.
[25] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. San Diego, CA: Academic, 2003.
[26] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *MRC Appl. Psychol. Unit.*, 1988.
[27] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 2nd ed. London, U.K.: Academic, 1988.
[28] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2236–2252, 2003.
[29] B. Romeny, L. Florack, J. Koenderink, and M. Viergever, Eds., *Scale-Space Theory in Computer Vision*. New York: Springer, 1997.
[30] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.
[31] M. Turgeon, A. S. Bregman, and P. A. Ahad, "Rhythmic masking release: Contribution of cues for perceptual organization to the cross-spectral fusion of concurrent narrow-band noises," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1819–1831, 2002.
[32] D. L. Wang, P. Divenyi, Ed., "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, 2005, pp. 181–197.
[33] 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.
[34] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1985.



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

His research interests include speech segregation, computational auditory scene analysis, and statistical machine learning.

DeLiang Wang (M'90–SM'01–F'04) received the B.S. and M.S. degrees from Peking 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.

Dr. Wang's research interests include machine perception and neurodynamics. He is the President of the International Neural Network Society (for 2006). He was a recipient of the 1996 U.S. Office of Naval Research Young Investigator Award.