Technical Report: OSU-CISRC-1/05-TR04
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2005**
File in pdf format: **TR4.pdf**
Web site: **http://www.cse.ohio-state.edu/research/tech-report.html**

Auditory Segmentation Based on Onset and Offset Analysis

Guoning Hu [a] and DeLiang Wang [b]

[a] *Biophysics Program*
The Ohio State University
Columbus, OH 43210
hu.117@osu.edu

[b] *Department of Computer Science and Engineering & Center for Cognitive Science*
The Ohio State University
Columbus, OH 43210
dwang@cse.ohio-state.edu

***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 an auditory 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 based on 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 based on scale-space theory. 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. Our approach performs much better than a cross-channel correlation method.**

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

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 [15] and hidden Markov models [23]. Such techniques tend to assume *a priori* knowledge or certain statistical properties of interference, and these assumptions are often too strong in realistic situations. Other approaches, including sinusoidal modeling [16] and comb filtering [8], attempt to extract speech by exploiting the harmonicity of voiced speech. Obviously such 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) [3]. 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 [3]. 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 [26] [7] [4] [10] [24] [14] (see [22] [5] 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.

In addition to the conceptual importance of segmentation for auditory scene analysis, 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 [14], 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 [1]. 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 [7] [4] [24] [14]. First, signals from the same source are likely to generate responses with similar temporal or periodic structure in neighboring frequency channels. 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 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 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 [21]. 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 onsets and offsets are common cues of all types of sounds, the proposed system can in principle deal with both voiced and unvoiced speech.

Specifically, we apply scale-space theory, a multiscale analysis widely used in image segmentation [25], 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. The smoothed scenes at different scales compose a scale space. 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 segments at different scales.

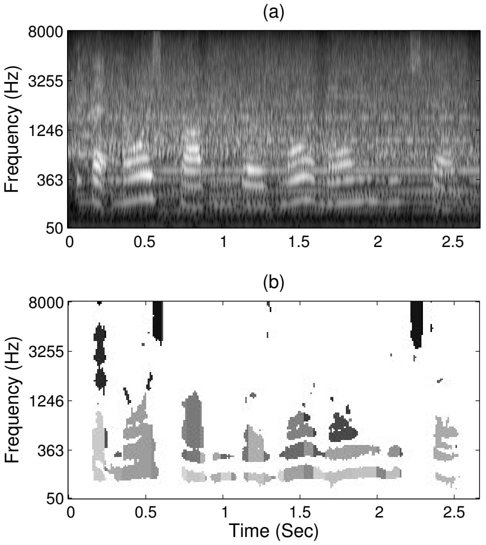

Fig. 1. A sound mixture and its ideal speech segments. (a) Cochleogram representation of a female utterance, “That noise problem grows more annoying each day,” mixed with a crowd noise with music. (b) The ideal segments of the utterance. The total number of ideal segments is 96.

This paper is organized as follows. In Sect. II, we propose a working definition for an auditory event in order to clarify the computational goal of segmentation. Details of the system are given in Sect. III. In Sect. IV, we propose a quantitative measure to evaluate the performance of auditory segmentation. The results of the system on speech segmentation are reported in Sect. V. The paper is ended with a discussion in Sect. VI.

II. What Is An Auditory Event?

Consider the signal from one source as containing a series of acoustic events separate in time. One may define the computational goal of segmentation as identifying the onsets and offsets of these events. However, at any time there are infinite acoustic events taking place simultaneously in the world, and one must limit the definition 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 [19]. Second, when there are multiple sounds in the same environment, a weaker sound tends to be masked by a stronger one [19]. 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 for different listeners [19]. For simplicity, we take as the absolute threshold a constant value, 15 dB sound pressure level (SPL), which is approximately the average absolute threshold from 200 Hz to 10 kHz for young adults with normal hearing [17].

Based on the above criteria, we define an auditory event as the collection of all the audible T-F regions for an acoustic event. This definition is consistent with the ASA principle of exclusive allocation, that is, a T-F region should be attributed to only one event [3]. Thus the computational goal of auditory segmentation is to generate segments for contiguous T-F regions from the same auditory event. To make this goal concrete requires a T-F representation of an acoustic input. Here we employ a cochleogram 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 [18]. Specifically, we use a filterbank with 128 gammatone filters centered from 50 Hz to 8 kHz [20], and decompose filter responses into consecutive 20-ms windows with 10-ms window shifts. Fig. 1(a) shows such a cochleogram for a mixture of a target female utterance and crowd noise with music, with the overall mixture signal-to-noise ratio (SNR) of 0 dB. Here, the nonlinear rectification is simply the response energy within each T-F unit.

As a working definition, we consider each phoneme of the target utterance as an acoustic event (see Sect. VI for more discussion on this working definition). Fig. 1(b) shows the ideal segments – the segments produced from premixing target and interference – of the target utterance in the mixture. Segments are represented by regions with different gray levels between neighboring regions, except for white regions, which form the background corresponding to the entire interference.

III. System Description

The proposed system contains three stages: smoothing, onset/offset detection and matching, along with multiscale integration. An acoustic mixture is first normalized at 60 dB SPL. Then it is passed through a bank of gammatone filters – a standard model of cochlear filtering [20]. 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 Hz to 60 Hz) and then downsampled to 400 Hz, which yield the temporal envelope of each filter output. The logarithm of the temporal envelope, referred to as the intensity of filter output across time, is used for onset and offset analysis.

A. Smoothing

Onsets and offsets correspond to sudden intensity increases and decreases. To find these sudden intensity changes, we take the derivative of the intensity with respect to time and then identify the peaks and valleys of the derivative. However, because of the intensity fluctuation within individual events, many peaks and valleys of the derivative do not correspond to real onsets and offsets. Therefore, the temporal envelope is smoothed over time to reduce the intensity fluctuation. Since an event usually has synchronized onsets and offsets across frequency, the temporal envelope is further smoothed over frequency (or filter channels) to enhance common onsets and offsets in adjacent channels. One way to perform the smoothing is to use a diffusion process [25], which is often applied for smoothing in image segmentation. A one-dimensional diffusion of a quantity v across a physical dimension x can be described by the following partial differential equation:

$$\frac{\partial v}{\partial t} = \frac{\partial}{\partial x}\left(D(v) \cdot \frac{\partial v}{\partial x}\right), \tag{1}$$

where t is the diffusion time, and D is a function controlling the diffusion process. Eq. (1) describes a process in which the change of v is determined by its gradient across x. When D satisfies certain conditions, v will change in a manner so that its gradient across x gradually approaches a constant, i.e., v is gradually smoothed over x [25]. The longer is t, the smoother is v. The diffusion time t is referred to as the scale parameter. The smoothed v values at different scales compose a scale space.

As an example, we consider a simple case where $D = 1$. Eq. (1) becomes

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} \, . \tag{2}$$

According to Eq. (2), the v value at a local minimum increases as t increases since $\partial^2 v/\partial x^2 > 0$ at such a point. Similarly, the v value at a local maximum will decrease as t increases since $\partial^2 v/\partial x^2 < 0$. As local minima of v gradually increase and local maxima of v gradually decrease, v becomes smoother over x during the diffusion process. In fact, (2) is equivalent to Gaussian smoothing [25]:

$$v(x,t) = v(x,0) * G(0,2t), \tag{3}$$

where $G(0, 2t)$ is a Gaussian function with mean 0 and variance $2t$, and "$*$" denotes convolution.

To perform smoothing, we let the intensity or logarithmic temporal envelope of each filter output be the initial value of v, and let v diffuse across time and frequency. That is,

$$v(c,n,0,0) = x(c,n), \tag{4}$$

$$\frac{\partial v}{\partial t_n} = \frac{\partial}{\partial n}\left(D_n(v) \cdot \frac{\partial v}{\partial n}\right), \tag{5}$$

$$\frac{\partial v}{\partial t_c} = \frac{\partial}{\partial c}\left(D_c(v) \cdot \frac{\partial v}{\partial c}\right), \tag{6}$$

where $x(c, n)$ is the intensity at time step n in channel c. t_c is the scale, or diffusion time, for the diffusion across frequency, and t_n the scale for the diffusion across time characterizing the input. Note the difference between the diffusion time,

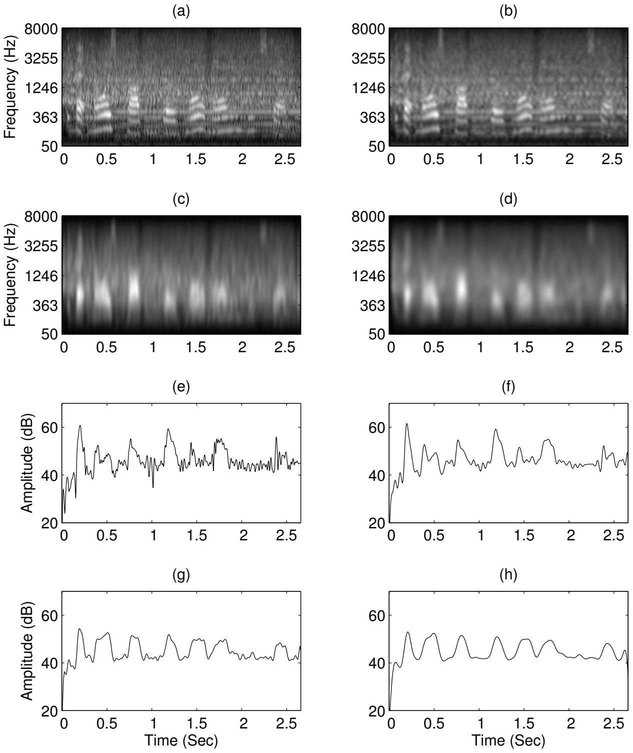

Fig. 2. Smoothed intensity values at different scales. (a) Initial intensity for all the channels. (b) Smoothed intensity at the scale (0.125, 1/14). (c) Smoothed intensity at the scale (18, 1/14). (d) Smoothed intensity at the scale (18, 1/4). (e) Initial intensity in a channel centered at 560 Hz. (f) Smoothed intensity in the channel at the scale (0.125, 1/14). (g) Smoothed intensity in the channel at the scale (18, 1/14). (h) Smoothed intensity in the channel at the scale (18, 1/4). The input is the same as shown in Fig. 1(a).

represented by t, and the time domain characterizing acoustic signal, represented by n. To avoid confusion, in the following text, we use "time" exclusively to refer to the time dimension of the input signal, and "scale" to refer to the diffusion time. With appropriate D_c and D_n, the output of the diffusion process at each scale, $v = v(c, n, t_c, t_n)$, will be a smoothed version of $x(c, n)$. Unlike the horizontal and the vertical dimension of a visual image, time and frequency are very different physical dimensions and shall undergo the diffusion process separately. More specifically, to obtain $v(c, n, t_c, t_n)$, the intensity first diffuses across time to yield $v(c, n, 0, t_n)$. Then it diffuses across frequency to yield $v(c, n, t_c, t_n)$.

We apply Gaussian smoothing for the diffusion across frequency, i.e., $D_c = 1$. For the diffusion across time, we have considered an isotropic diffusion as well as Gaussian smoothing in a preliminary study [13]. A critical issue of both diffusion processes is how to determine the scale, or when the diffusion process stops (see [6] for further discussion). The diffusion process needs to stop at a certain scale to preserve sharp intensity changes signaling onsets and offsets; otherwise, this important information will eventually be lost. This scale is task-dependent and there is no general rule to determine it. Given that smoothing is in fact lowpass-filtering, we use a series of lowpass filters to smooth the intensity instead. The cutoff frequency of each lowpass filter corresponds to a particular smoothing scale, and a smaller cutoff frequency corresponds to a larger smoothing scale. The smallest cutoff frequency, which corresponds to the scale when the diffusion stops, can be determined according to the acoustic and perceptual properties of the target. For speech, 4 Hz may be used as the smallest cutoff frequency since temporal envelope variations down to 4 Hz is essential for speech intelligibility [9]. Consequently we represent the smoothing scale as (t_c, t_n), where $2t_c$ is the variance of the Gaussian function for the smoothing over frequency and t_n is the reciprocal of the cutoff frequency of the lowpass filter for the smoothing over time.

As an example, Figure 2 shows the initial and smoothed intensities at three scales, (0.125, 1/14), (18, 1/14) and (18, 1/4), for the input mixture shown in Fig. 1(a). Fig. 2(a) shows the initial intensity. The corresponding smoothed intensities at the three scales are shown in Fig. 2(b), 2(c) and 2(d), respectively. A lowpass filter with a 112.5-ms Kaiser window and a 10-Hz transition band is used for the smoothing across time. As we can see from 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. To display more details, Fig. 2(e) shows the initial intensity of the output from a single frequency channel centered at 560 Hz. The corresponding smoothed intensities at three scales are shown in Fig. 2(f), 2(g) and 2(h), respectively.

B. *Onset/offset Detection and Matching*

At a certain scale (t_c, t_n), onset and offset candidates are detected by marking peaks and valleys of the time derivative of the smoothed intensity, $\partial v(c, n, t_c, t_n)/\partial n$. The derivative is calculated by taking the difference between consecutive samples. An onset candidate is removed if the corresponding difference is smaller than a threshold θ_{ON}, which suggests that the candidate is likely an insignificant intensity fluctuation. We choose $\theta_{ON}(t_c, t_n) = \mu(t_c, t_n) + \sigma(t_c, t_n)$, where $\mu(t_c, t_n)$ and $\sigma(t_c, t_n)$ are the mean and standard deviation of all the samples of $\partial v(c, n, t_c, t_n)/\partial n$, respectively.

To perform onset and offset matching, the system first determines in each channel the offset time for each onset candidate. Let $n_{ON}[c, i]$ represent the time of the ith onset candidate in channel c. The system identifies the corresponding offset time, denoted as $n_{OFF}[c, i]$, among the offset candidates located between $n_{ON}[c, i]$ and $n_{ON}[c, i+1]$. 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., with the smallest $\partial v/\partial n$. We have also considered choosing either the first or the last offset candidate, but their performance is not as good. 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.

In order to merge adjacent channels from the same event, the system first merges common onsets and offsets into onset and offset fronts since an event usually has synchronized onsets and offsets. More specifically, an onset candidate is merged with the closest onset candidate in an adjacent channel if their distance in time is less than 20 ms in our implementation; the same applies to offset candidates. If an onset front thus formed occupies less than three channels, we do not further process it because it is likely insignificant. Onset and offset fronts are vertical contours in the 2-D time-frequency representation.

The next step is to match individual onset and offset fronts to form segments. Let ($n_{ON}[c, i_1]$, $n_{ON}[c+1, i_2]$, …, $n_{ON}[c+m-1, i_m]$) denote an onset front with m consecutive channels starting from c, and ($n_{OFF}[c, i_1]$, $n_{OFF}[c+1, i_2]$, …, $n_{OFF}[c+m-1, i_m]$) 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 channels. In the end, the T-F region between ($n_{ON}[c, i_1]$, $n_{ON}[c+1, i_2]$, …, $n_{ON}[c+m-1, i_m]$) and the updated offset times ($n_{OFF}[c, i_1]$, $n_{OFF}[c+1, i_2]$, …, $n_{OFF}[c+m-1, i_m]$) yields a segment.

In 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 $n_{ON}[c, i_1]$, let $n_{ON}[c+1, i_2]$ be the closest onset candidate in an adjacent channel $c+1$. Let (n_1, n_2) be the overlapping duration between ($n_{ON}[c, i_1]$, $n_{OFF}[c, i_1]$) and ($n_{ON}[c+1, i_2]$, $n_{OFF}[c+1, i_2]$), where n_{OFF} in a channel is the corresponding offset time of n_{ON} as described earlier. The similarity between the temporal envelopes from these two channels in this duration is measured by their correlation (see [24]):

$$C(c,i_1,i_2,t_c,t_n) = \sum\nolimits_{n=n_1}^{n_2} \hat{v}(c,n,t_c,t_n)\hat{v}(c+1,n,t_c,t_n)\,, \qquad (7)$$

where $\hat{v}$ indicates the normalized v with zero mean and unity variance within (n_1, n_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

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 an 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 segments generated across different scales in an orderly manner. It starts to segment at a larger scale. Then, at a 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 ($n_{ON}[c, i_1]$, $n_{ON}[c+1, i_2]$, …, $n_{ON}[c+m-1, i_m]$) and ($n_{OFF}[c, i_1]$, $n_{OFF}[c+1, i_2]$, …, $n_{OFF}[c+m-1, i_m]$) be the onset times and offset times of a segment occupying m consecutive channels starting from c. Note that lower-frequency channels are at lower positions in our cochleogram representation (see Fig. 1(a)). The expansion works by considering the onset front at the current scale crossing $n_{ON}[c+m-1, i_m]$ and the offset front crossing $n_{OFF}[c+m-1, i_m]$. 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.

Since we let the temporal envelope diffuse across time and frequency separately, it is possible to move from a coarser scale to two finer scales so that one has a smaller t_c and the other has a smaller t_n. In this situation, how to order the two scales becomes ambiguous in multiscale integration. To avoid this situation, we only consider scales that are unambiguously ordered. In other words, among the scales considered, t_c and t_n of a coarser one are always not smaller than those of a finer one.

In our implementation, the system forms segments in three scales from coarse to fine: (t_c, t_n) = (18, 1/4), (18, 1/14) and (0.125, 1/14). At the finest scale, i.e. (0.125, 1/14), we do 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; the larger θ_C is used in the coarser scales because smoothing over frequency increases the similarity of temporal envelopes in adjacent channels. At each scale, a lowpass filter with a 112.5-ms Kaiser window and a 10-Hz transition band is applied for the smoothing over time. We have also considered segmentation using more scales, but results are not significantly better.

Fig. 3 shows the bounding contours of segments at different scales for the mixture in Fig. 1(a), where Fig. 3(a) shows the segments formed at scale (18, 1/4), Fig. 3(b) those from the multiscale integration of two scales (18, 1/4) and (18, 1/14), and Fig 3(c) those from the integration of all three scales. Comparing these contours with Fig. 1(b), we can see that at the largest scale, the system captures a majority of speech events,

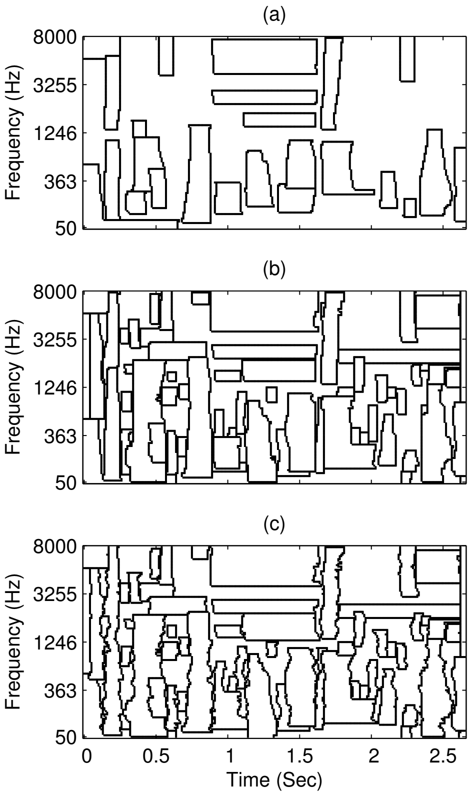

Fig. 3. The bounding contours of estimated segments from multiscale analysis. (a) One scale analysis at the scale of (18, 1/4). (b) Two-scale analysis at the scales of (18, 1/4) and (18, 1/14). (c) Three-scale analysis at the scales of (18, 1/4), (18, 1/14), and (0.125, 1/14). The input is the same as shown in Fig. 1(a).

but misses some small segments. As the system integrates segments generated at smaller scales, more speech segments are formed; at the same time, some segments from interference also appear.

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 Sect. IV).

IV. Evaluation Metrics

Only a few previous models have explicitly addressed the problem of auditory segmentation [7] [4] [24] [14] 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 decided to adapt region-based metrics by Hoover *et al.* [11], which have been widely used for evaluating image segmentation systems. Our evaluation is focused on comparing estimated segments with ideal segments for target, since it is sometimes hard to determine the ideal segments of interference and in many situations one is interested in only extracting target speech. Hence we will treat all the T-F regions where interference dominates as the background. Furthermore, the evaluation scheme discussed below can be easily extended to situations where one aims to evaluate segments from interference, say, when interference is a competing talker.

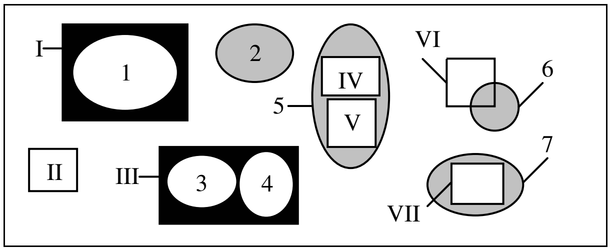

Fig. 4. Illustration of correct segmentation, under-segmentation, over-segmentation, missing, and mismatch. Here an oval indicates an ideal segment and a rectangle an estimated one.

The general idea of the region-based evaluation 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. 4 illustrates these cases, where ovals represent ideal target segments (numbered with Arabic numerals) and rectangles estimated segments (numbered with Roman numerals). As shown in Fig. 4, 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 — segment 2 and 6 and the gray parts of segments 5 and 7 — are labeled as missing. The black region in segment I belongs to the ideal background, but 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 difference between under-segmentation and mismatch. The former occurs when multiple segments from the same source are merged. The latter occurs when segments from different sources are merged. 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 $\theta_E \in [0.5, 1)$, we define that an ideal segment $r_I[k]$ is

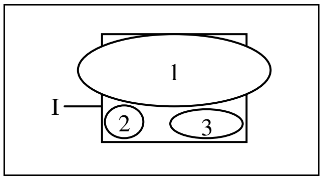

Fig. 5. Illustration of multiple labels for one overlapping region. Here an oval indicates an ideal segment and a rectangle an estimated one.

well-covered by an estimated segment $r_S[l]$ if $r[k, l]$ includes most of the energy of $r_I[k]$. That is,

$$E[k,l] > \theta_E \cdot E_I[k]\,. \tag{8}$$

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

$$E[k,l] > \theta_E \cdot E_S[l]\,. \tag{9}$$

The 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 non-empty 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:

$$\sum_{k'} E[k',l] > \theta_E \cdot E_S[l],\quad k' = k_1, k_2, \ldots, k_{K'} \tag{10}$$

- 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:

$$\sum_{l'} E[k,l'] > \theta_E \cdot E_I[k],\quad l' = l_1, l_2, \cdots, l_{L'} \tag{11}$$

- If a region $r[k, l]$ is part of an ideal segment of target speech, i.e., $k>0$, but cannot be labeled as either 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. Figure 5 illustrates this situation, where estimated segment I and ideal segment 1 are mutually well-covered. Hence, $r[1, \text{I}]$ 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, $r[1, \text{I}]$, $r[2, \text{I}]$, and $r[3, \text{I}]$ should all be labeled as under-segmented. Therefore, $r[1, \text{I}]$ 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 is the percentage of the total energy of correctly segmented target to the total energy of ideal segments of target, i.e., $P_C = E_C/E_I \times 100\%$.
- The percentage of under-segmentation is the percentage of the total energy of under-segmented target to the total energy of ideal segments of target, i.e., $P_U = E_U/E_I \times 100\%$.
- The percentage of over-segmentation is the percentage of the total energy of over-segmented target to the total energy of ideal segments of target, i.e., $P_O = E_O/E_I \times 100\%$.
- The percentage of missing is the percentage of the total energy of target missing from the estimated segments to the total energy of ideal segments of target, i.e., $P_M = E_M/E_I \times 100\%$.
- The percentage of mismatch is the percentage of total interference energy captured in estimated target segments to the total energy of estimated segments, i.e., $P_N = E_N/E_S \times 100\%$.

Since $E_C + E_U + E_O + E_M = E_I$, or $P_C + P_U + P_O + P_M = 1$, 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 10 intrusions. We consider the utterances, which are randomly selected from the TIMIT database, as target. 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

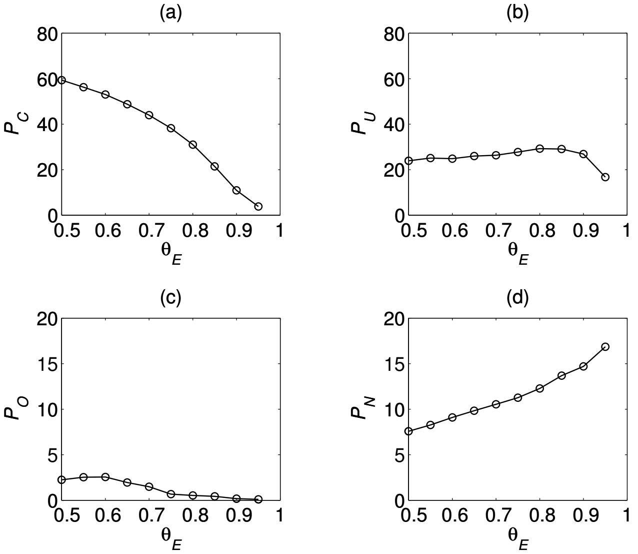

Fig. 6. The results of auditory segmentation. Target and interference are mixed at 0 dB SNR. (a) The average correct percentage. (b) The average percentage of under-segmentation. (c) The average percentage of over-segmentation. (d) The average percentage of mismatch.

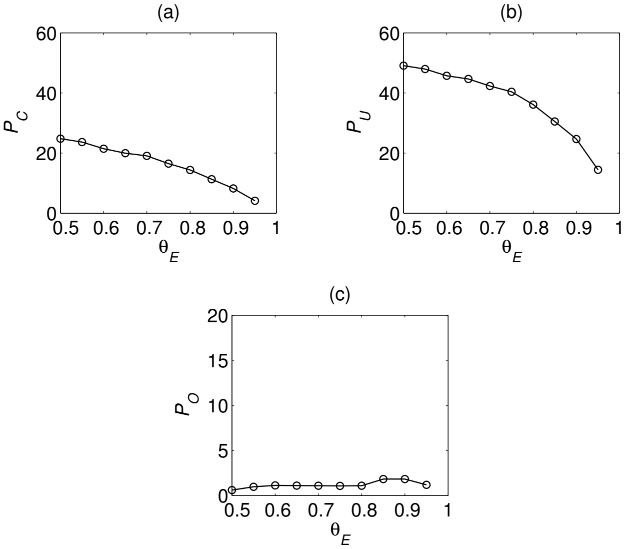

Fig. 7. The results of auditory segmentation for stops, fricatives, and affricates. Target and interference are mixed at 0 dB SNR. (a) The average correct percentage. (b) The average percentage of under-segmentation. (c) The average percentage of over-segmentation.

Sect. II, we consider each phoneme as an acoustic event of speech and obtain ideal target segments from target speech and interference before mixing.

Fig. 6 shows the average P_C, P_U, P_O, and P_N for different θ_E values. Note that the evaluation is more stringent for higher θ_E. Speech and interference are mixed at 0 dB SNR. As shown in the figure, the correct percentage is 59.4% when θ_E is 0.5, and it decreases to 3.8% 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 Sect. IV, under-segmentation is not really an error. By combining P_C and P_U together, the system correctly segments 83.3% 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 16.9% when θ_E increases to 0.95. Considering the overall SNR of 0 dB, the percentage of mismatch is not significant. This shows that the interference and the target speech are well separated in the estimated segments.

Since the 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. Here we compute P_C as the percentage of the total energy of correctly segmented stops, fricatives, and affricates to the total energy of these phonemes in the ideal segments. P_U and P_O are computed similarly. 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 73.9% when θ_E is 0.5, and it drops below 50% when θ_E is larger than 0.8.

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 dB to 0 dB.

Because the low-frequency portion of speech is usually much 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 highpass 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. Figure 9 presents a comparison with and without pre-emphasis for mixtures at 0 dB SNR. Figs. 9(a) and 9(b) show 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. Figs. 9(c) and 9(d) show 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-

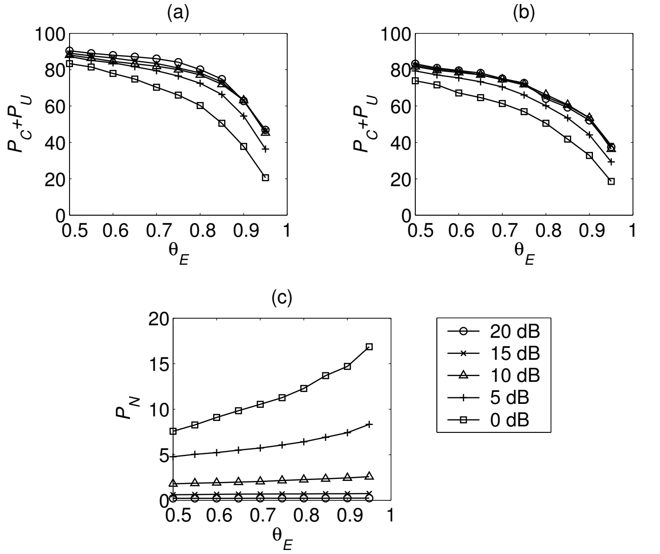

Fig. 8 The results of auditory segmentation at different SNR levels. (a) The average correct percentage plus the average percentage of under-segmentation for all the phonemes. (b) The average correct percentage plus the average percentage of under-segmentation for stops, fricatives, and affricates. (c)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

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 cross-channel correlation method for segmentation described in [14]; a more complex method of cross-channel correlation is presented in [4] which is based on clustering of neighboring channels. The cross-channel correlation method computes the correlation of normalized correlogram and merges T-F units if their correlation exceeds a certain threshold (cf. Eq. 7). The correlogram is a running autocorrelation of the filter response and the response envelope (see [14]). In addition, neighboring time frames are merged. Figure 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 by the proposed system and those by the cross-channel correlation method. The cross-channel correlation method yields much lower P_C+P_U scores. This is primarily because the correlation method fails to merge resolved harmonics of target speech efficiently; specifically, neighboring harmonics often yield different filter responses. Since cross-channel correlation was proposed mainly 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 determination algorithm for clean speech [2]. 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. The correlation method 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 the cross-channel correlation method for auditory segmentation.

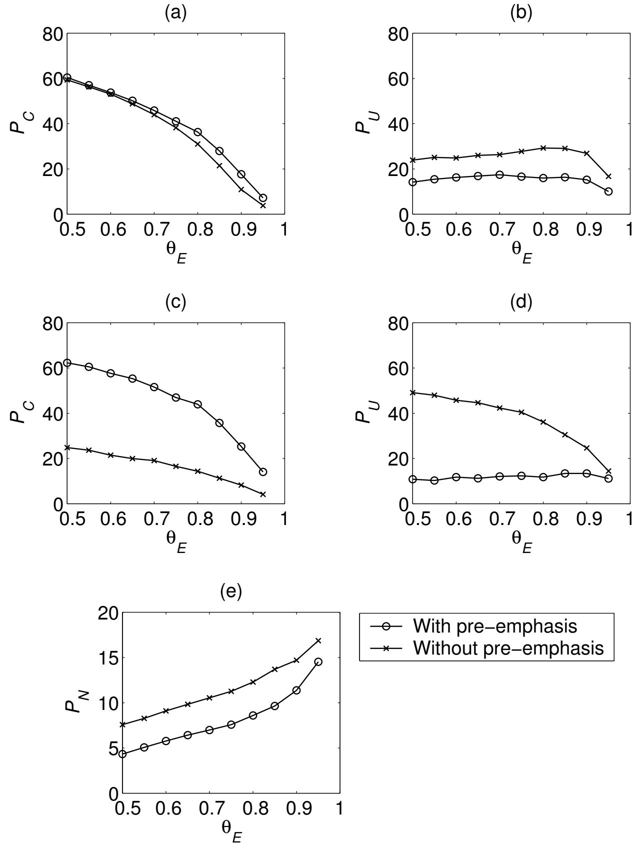

Fig. 9. The results of auditory segmentation with and without pre-emphasis. Target and interference are mixed at 0 dB SNR. (a) The average correct percentage for all the phonemes. (b) The average percentage of under-segmentation for all the phonemes. (c) The average correct percentage for stops, fricatives,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

VI. Discussion

To determine ideal segments of target speech, we need to decide what constitutes acoustic events of a speech utterance (see Sect. II). Here we treat 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 burst 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, and there are reasons to treat strongly coarticulated phonemes as a single event. As a result, coarticulation may lead to unnatural boundaries in ideal segments, and in this case under-segmentation can be more desirable. This problem is partly

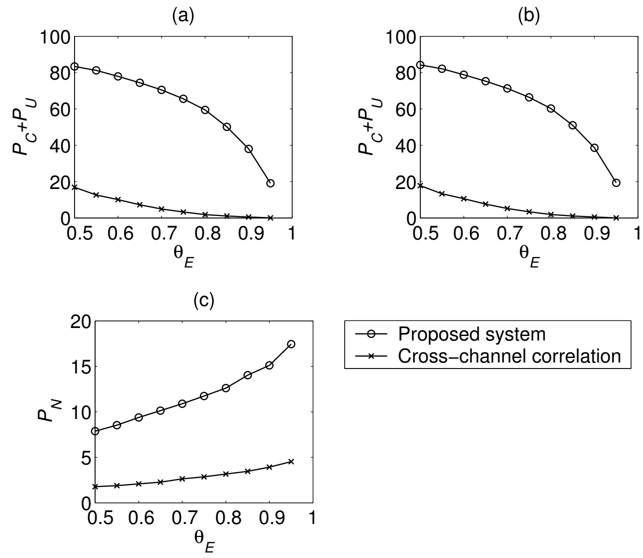

Fig. 10. The results of auditory segmentation for the proposed system and the segmentation result from the cross-channel correlation method. Target and interference are mixed at 0 dB SNR. (a) The average correct percentage plus the average percentage of under-segmentation for all the phonemes. (b) The average correct percentage plus the average percentage of under-segmentation for voiced portions of utterance. (c) The average percentage of mismatch.

taken care of in our evaluation which does not consider under-segmentation as error. Alternatively, one may define a syllable, a word, or even a whole utterance from the same speaker as an acoustic event. In such a definition coarticulation is no longer an issue. However, many valid acoustic boundaries between phonemes are not taken into account, and over-segmentation becomes an issue. In other words, it is not clear whether an instance of over-segmentation is caused by a true boundary between two phonemes or a genuine error.

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 segmentation. Although we have tested only on speech segmentation, the system should be easily extended to other signal types, such as music, because the model is not based on specific properties of speech. 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 performance improvements [4]. In a previous study, we demonstrated the utility of the onset cue for segregating stop consonants [12]. This study on auditory segmentation further shows that event onsets and offsets may play a fundamental role in sound organization. Third, we have extended scale-space theory to the auditory domain. To our knowledge, it is the first time this theory has been used for auditory analysis. 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.

Acknowledgment

This research was supported in part by an AFOSR grant (FA9550-04-01-0117), an AFRL grant (FA8750-04-1-0093), and an NSF grant (IIS-0081058). A preliminary version of this work was presented in the 2004 ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing.

References

[1] J.P. Barker, M.P. Cooke, and D.P.W. Ellis, "Decoding speech in the presence of other sources," *Speech Comm.*, vol. 45, pp. 5-25, 2005.

[2] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, Version 4.2.31, http://www.fon.hum.uva.nl/praat/, 2004.

[3] A.S. Bregman, *Auditory scene analysis*, Cambridge MA: MIT Press, 1990.

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

[5] 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, Ed., New York: Springer, in press, 2005.

[6] K. Chen, D.L. Wang, and X. Liu, "Weight adaptation and oscillatory correlation for image segmentation," *IEEE Trans. Neural Net.*, vol. 11, pp. 1106-1123, 2000.

[7] M.P. Cooke, *Modelling auditory processing and organisation*, Cambridge: Cambridge University Press, 1993.

[8] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-time processing of speech signals*, New York: Macmillan, 1993.

[9] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, pp. 1053-1064, 1994.

[10] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. Dissertation, MIT, Dept. Elec. Engg. and Comput. Sci., 1996.

[11] A. Hoover, G. Jean-Baptiste*, et al.*, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 18, pp. 673 - 689, 1996.

[12] G. Hu and D.L. Wang, "Separation of stop consonants," in *Proc. ICASSP*, Vol. II, pp. 749-752, 2003.

[13] G. Hu and D.L. Wang, "Auditory segmentation based on event detection," in *ISCA Tutorial and Research Workshop on Stat. and Percept. Audio Process.*, 2004.

[14] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, vol. 15, pp. 1135-1150, 2004.

[15] 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 PTR, 2001.

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

[17] M.C. Killion, "Revised estimate of minimal audible pressure: Where is the 'missing 6 dB'?," *J. Acoust. Soc. Am.*, vol. 63, pp. 1501-1510, 1978.

[18] R.F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. ICASSP*, Vol. II, pp. 1282 - 1285, 1982.

[19] B.C.J. Moore, *An introduction to the psychology of hearing*, 5th ed., San Diego, CA: Academic Press, 2003.

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

[21] J.O. Pickles, *An introduction to the physiology of hearing*, 2nd ed., London: Academic Press, 1988.

[22] D. Pierre, Ed., *Speech separation by humans and machines*, Norwell MA: Kluwer Academic, 2004.

[23] 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 and Audio Process.*, vol. 6, pp. 445-455, 1998.

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

[25] J. Weickert, "A review of nonlinear diffusion filtering," in *Scale-space theory in computer vision,* B.H. Romeny, L. Florack, J. Koenderink, and M. Viergever, Ed., Berlin: Springer, pp. 3-28, 1997.

[26] M. Weintraub, *A theory and computational model of auditory monaural sound separation*, Ph.D. Dissertation, Stanford, Dept. Elec. Engg., 1985.