ITRW on Statistical and
Perceptual Audio Processing
ICC Jeju, Korea
October 3, 2004

Auditory Segmentation Based on Event Detection

Guoning Hu

Biophysics Program
The Ohio State University
395 Dreese Lab., 2015 Neil Ave.
Columbus, OH 43210
hu.117@osu.edu

DeLiang Wang

Department of Computer Science and Engineering &
Center for Cognitive Science
The Ohio State University
395 Dreese Lab., 2015 Neil Ave.
Columbus, OH 43210
dwang@cis.ohio-state.edu

ABSTRACT

Acoustic signals from different sources in a natural environment form an auditory scene. Auditory scene analysis (ASA) is the process in which the auditory system segregates an auditory scene into streams corresponding to different sources. Segmentation is an important stage of ASA where an auditory scene is decomposed into segments, each of which contains signal mainly from one source. We propose a system for auditory segmentation based on analyzing onsets and offsets of auditory events. Our system first detects onsets and offsets, and then generates segments by matching corresponding onsets and offsets. This is achieved through a multiscale approach based on scale-space theory. Systematic evaluation shows that much target speech, including unvoiced speech, is correctly segmented, and target speech and interference are well separated into different segments.

1. INTRODUCTION

In a natural environment, multiple sounds from different sources form an auditory scene. Many applications, including automatic speech recognition and hearing aids design, require an effective system that segregates speech in a complex scene. Currently, speech segregation with one microphone (or monaurally) remains a major challenge [7] [9].

On the other hand, the auditory system shows a remarkable capacity in monaural segregation of different sources. This perceptual process is referred to as *auditory scene analysis* (ASA) [1]. In general, ASA takes place in two stages, segmentation and grouping. In segmentation, the auditory system decomposes the complex scene into a collection of segments (or sensory elements), each of which mainly arises from one source. In grouping, segments that are likely to arise from the same source are grouped together. Considerable research has been carried out to develop *computational auditory scene analysis* (CASA) systems for sound separation [2] [4] [5] [7] [13].

In our view, a successful CASA system needs to perform effective segmentation, which provides a foundation for grouping. However, no existing CASA system performs segmentation consistently well. In fact, no system has addressed segmentation for unvoiced speech at all.

The goal of segmentation is to decompose an auditory scene into contiguous time-frequency (T-F) regions, each of which should contain signal mainly from one source. From a computational standpoint, auditory segmentation is similar to image segmentation, which is 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 image properties, such as color and texture. In auditory segmentation, the corresponding task is to find onsets and offsets of individual auditory events. The onsets and offsets generally 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 [1] and there is strong evidence for onset detection by auditory neurons [12]. Our analysis is based on scale-space theory, which is a multiscale analysis widely used in image segmentation [14]. The advantage of using a multiscale analysis is to provide different levels of details for an auditory scene so that one can detect and localize auditory objects at appropriate scales. Our system performs segmentation in three stages. First, an auditory scene is smoothed through a diffusion process. The smoothed scenes at different scales, or different diffusion times, compose a scale space. Second, the system detects onsets and offsets at certain scales, and obtains segments by matching individual onset and offset fronts. Third, the system generates a final set of segments by combining segments obtained at different scales. Note that to determine from which source a segment arises is the task of grouping, which is not addressed in this paper.

This paper is organized as follows. In Sect. 2, 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. 3. In Sect. 4, we propose a quantitative measure to evaluate segmentation, and report our results for speech segmentation. A brief discussion is given in Sect. 5.

2. WHAT IS AN AUDITORY EVENT?

Consider the signal from one source that contains a series of acoustic events. 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, 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 must 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 [8]. Second, when there are multiple sounds in the same environment, a weaker sound tends to be masked by a stronger one [8]. 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 different listeners [8]. For simplicity, we take a constant value, 15 dB sound pressure level (SPL), as the absolute threshold.

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 [1]. Thus the computational goal of segmentation is to generate segments for contiguous T-F regions from the same auditory event. To make this goal concrete requires a specific T-F decomposition of an auditory scene; here we decompose an input in the frequency domain using a filterbank with 128 gammatone filters centered from 50 Hz to 8 kHz [10]. In the time domain, we decompose the signal into consecutive 20-ms windows with 10-ms window shifts. Fig. 1(a) shows such decomposition for a mixture of a target female utterance with an acoustic crowd and music in the T-F domain. The overall signal-to-noise ratio (SNR) is 0 dB. Fig. 1(b) shows the ideal segments for the mixture. Here each phoneme is considered as an acoustic event, and the corresponding 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.

3. SYSTEM DESCRIPTION

The proposed system contains three stages: smoothing, onset/offset detection and matching, and multiscale integration. An acoustic mixture is first normalized at 60 dB SPL. Then it is passed through a bank of gammatone filters (Sect. 2). The input to the system is the average intensity of each filter output at every 1.25-ms frame. (Note the difference between a frame and a window described in Sect. 2.)

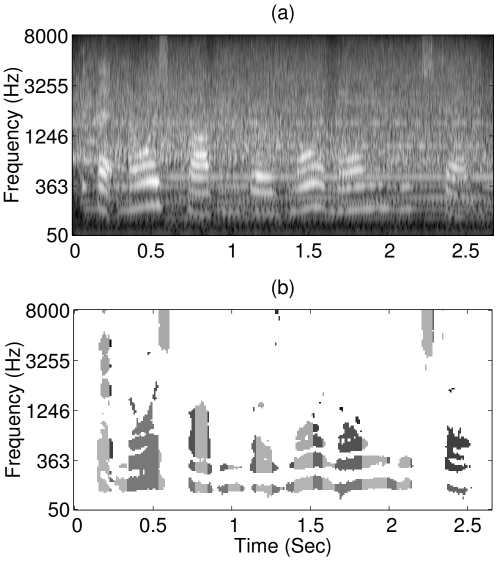

Figure 1. (a) T-F decomposition of a mixture of a female utterance, “That noise problem grows more annoying each day,” and an acoustic crowd with music. (b) The ideal segments for the speech. The total number of ideal segments is 87.

3.1 Smoothing

Onsets and offsets generally correspond to sudden intensity increases and decreases. To find these sudden intensity changes, one may take the derivative of intensity with respect to time and then find 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 actual onsets and offsets. Therefore, the intensity is smoothed over time to reduce the intensity fluctuation. Since an event usually has synchronized onsets and offsets in the frequency domain, the intensity is further smoothed over frequency to enhance common onsets and offsets in adjacent channels. The system performs the smoothing through a diffusion process [14]. A one-dimensional diffusion of a quantity v across a physical dimension x can be described as:

$$\partial_t v = \partial_x (D(v) \cdot \partial_x v) \,, \tag{1}$$

where ∂_t represents the partial derivative with respect to time t, and ∂_x that to x. D is a function controlling the diffusion process. Eq. (1) describes a process that the change of v is determined by the gradient of v across x. When D satisfies certain conditions, v will change so that the gradient of v across x in desired regions gradually approaches a constant, i.e., v is gradually smoothed over x in these regions [14]. The longer t is, the smoother v is. The diffusion time t is referred to as the scale parameter. The smoothed v at different t composes a scale space.

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

$$\partial_t v = \partial_x^2 v \ . \tag{2}$$

According to Eq (2), the change of v forces $\partial_x^2 v$ gradually approach 0. In other words, as t increases, v becomes smoother over x. In fact, Eq. (2) is equal to Gaussian smoothing [14]:

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

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

Let the input intensity be the initial value of v, and let v diffuse across time frames and filter channels. That is:

$$v(c,m,0,0) = I(c,m) \,, \tag{4}$$

$$\partial_{t_m} v = \partial_m (D_m(v) \cdot \partial_m v) \,, \tag{5}$$

$$\partial_{t_c} v = \partial_c (D_c(v) \cdot \partial_c v) \,, \tag{6}$$

where $I(c, m)$ is the logarithmic average intensity of the mixture in channel c at frame m. t_c is the scale for the diffusion across filter channels, and t_m for the diffusion across time frames. Note that the diffusion time t_c and t_m is different from the time of input sound, which corresponds to time frames represented by m. With appropriate D_c and D_m, the output of the diffusion process at each scale, $v = v(c, m, t_c, t_m)$, will be a smoothed version of the input intensity. Since time and frequency are different physical dimensions, the system undergoes the diffusion process across time frames and across filter channels separately. More specifically, to obtain $v(c, m, t_c, t_m)$, the intensity first diffuses across time frames for time t_m, which yields $v(c, m, 0, t_m)$. Then this smoothed intensity diffuses across filter channels for time t_c.

Two forms of $D_m(v)$ are employed here. The first one is a constant $D_m(v)$, i. e., Gaussian smoothing. The second one is the commonly used Perona-Malik model [11]:

$$D_m(v) = 1/[1 + |\partial_m v|^2 / \lambda^2], \quad (7)$$

where λ is a parameter. Compared with Gaussian smoothing, the Perona-Malik model is able to keep better locations of onset and offset. As an example, Fig. 2 shows the smoothed intensity from these two forms of diffusion at different t_m. The input is the mixture of speech, crowd sound and music filtered by a gammatone filter centered around 350 Hz. For the Perona-Malik model, λ=1. A constant $D_c(v)$ is applied for the diffusion across frequency.

3.2 Onset/offset detection and matching

At a certain scale (t_c, t_m), onset and offset candidates are detected by marking peaks and valleys of the difference of v between consecutive time frames. An onset candidate is removed if the corresponding difference is very small, which suggests that the candidate is likely to relate to an intensity fluctuation rather than an event onset.

In order to merge T-F regions in adjacent channels from the same event, the system first combines common onsets and offsets into onset and offset fronts since an event usually has synchronized onsets and offsets. More specifically, an onset candidate is connected with the closest onset candidate in an adjacent channel if their distance in time is smaller than 20 ms, and so is it for an offset candidate. If an onset front occupies less than three channels, we do not further process it because it is 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. For an onset front, the system first determines a corresponding offset for it in each corresponding channel. Then it uses these offsets to determine the corresponding offset front. Let $m_{ON}[c, i]$ and $m_{OFF}[c, j]$ represent the frame for the ith onset candidate and jth offset candidate in channel c, respectively. For each onset candidate, the system identifies the corresponding offset among the offset candidates located between $m_{ON}[c, i]$ and $m_{ON}[c, i+1]$. The decision is simple if there is only one offset candidate in this range. When there are multiple offset candidates, the system chooses the one with the largest intensity decrease, i.e., with the smallest difference of v. We have also considered choosing either the first or the last offset candidate, and they perform slightly worse. Let $(m_{ON}[c_1, i_1], m_{ON}[c_1+1, i_2], \quad, m_{ON}[c_1+n-1, i_n])$ be an onset front occupying n channels, and $(m_{OFF}[c_1, j_1], m_{OFF}[c_1+1, j_2], \quad, m_{OFF}[c_1+n-1, j_n])$ the corresponding offsets determined above. The system compares $(m_{OFF}[c_1, j_1], m_{OFF}[c_1+1, j_2], \quad, m_{OFF}[c_1+n-1, j_n])$ with each offset front, and the offset front with the largest overlap is chosen as the matching offset front. The T-F region between them yields a segment.

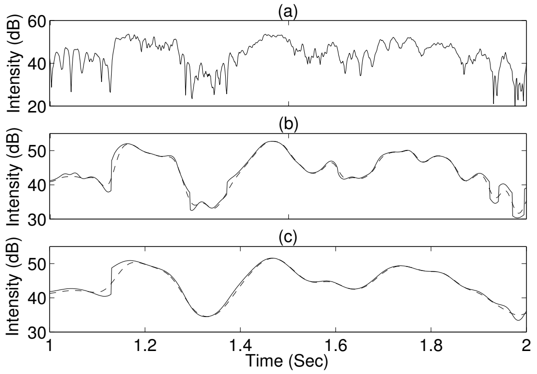

Figure 2. Smoothed intensity at different scales in a channel centered around 350 Hz. (a) Intensity at scale (0, 0), i.e., the initial intensity. (b) Smoothed intensity at scale (0, 50) for the Perona-Malik model (solid line) and that for Gaussian smoothing (dash line). (c) Smoothed intensity at scale (0, 200) for the Perona-Malik model (solid line) and that for Gaussian smoothing (dash line). The input is the mixture of speech, crowd sound and music.

3.3 Multiscale integration

As a result of smoothing, event onsets and offsets occupying small T-F regions may be blurred at a larger (coarser) scale. Consequently, the system tends to miss small events or to generate segments combining different events, which is a case of under-segmentation. On the other hand, at a smaller (finer) scale, the system may be sensitive to intensity fluctuations within individual events. Consequently, the system tends to separate an event into several segments, which is a case of over-segmentation. Therefore, it is difficult to obtain a satisfactory result of segmentation with a single scale. Our system handles this problem by integrating segments generated across different scales in an iterative manner. First, it forms segments by matching onset and offset fronts at a larger scale. Then, at a smaller scale, it locates more accurate onset and offset positions for these segments. In addition, new segments are formed according to the onset and offset fronts detected at the current scale. Then the system goes to an even smaller scale if necessary. Here the integration starts from a large scale and then moves to smaller scales. One could also start from a small scale and then move to larger scales. However, in the latter case, the chances of over-segmenting an input mixture is much higher, which is undesirable since in subsequent grouping larger segments are preferred.

Fig. 3 shows the bounding contours of obtained segments for the mixture of speech, crowd sound and music at different scales. Comparing it with Fig. 1(b), we can see that at the largest scale, the system captures most speech events, but misses some small segments. As the scale decreases, more

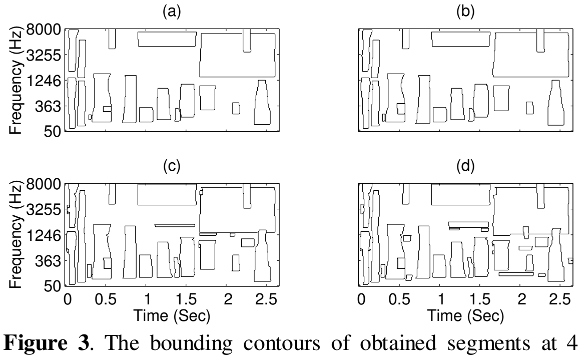

Figure 3. The bounding contours of obtained segments at 4 different scales: (a) (32, 200), (b) (18, 200), (c) (32, 100), and (d) (18, 100), for the mixture of speech, crowd sound and music.

speech segments are generated; at the same time, some segments for interference are also generated.

4. EVALUATION

We have applied our system to 20 speech utterances mixed with 10 intrusions. The utterances are randomly selected from the TIMIT database. The intrusions are: white noise, electrical fan, rooster crow and clock alarm, traffic sound, crowd sound in a playground, crowd sound and music, crowd clapping, bird chirping and waterflow, wind, and rain.

Only a few previous models have explicitly addressed the problem of auditory segmentation [2] [13], but none have 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 true segments and that of computed 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 a region-based definition by Hoover *et al.* [6], which has been widely used for evaluating image segmentation systems. The general idea is to examine the overlapping 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 mismatching. Fig. 4 illustrates these cases, where we let ovals represent ideal segments (numbered with Arabic numerals) and rectangles estimated segments (numbered with Roman numerals). Segment I well covers segment 1, and the overlapping region is labeled 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 combined with ideal segment 1 into an estimated segment. This black region is labeled as mismatching. So is the black region in segment III. Segment II is well covered by the ideal background. Here we do not consider this type of regions in the evaluation because our evaluation focuses on target speech, not interference. We expect that segment-II type regions will be eliminated in subsequent processing. Much of segment VI is covered by the ideal background and therefore we treat the white region of segment VI the same as segment II. (Note the difference between segment I and VI.)

Quantitatively, let $\{r_{GT}[k]\}$, $k=0,1,\ldots,K$, be the set of ideal segments, where $r_{GT}[0]$ indicates the ideal background and others the segments of target speech. These ideal segments are obtained according to the definition of an auditory event (Sect. 2), with the knowledge of target speech and interference before mixing. Note that we consider here each phoneme as an acoustic event, and all the interference as the background, which contains little speech. Let $\{r_S[l]\}$, $l=0,1,\ldots,L$, be the estimated segments, where $r_S[l]$, $l>0$, corresponds to an obtained segment and $r_S[0]$ the obtained background. Let $r[k, l]$ be the overlapping region between an ideal segment, $r_{GT}[k]$, and an estimated segment, $r_S[l]$. Furthermore, let $E[k, l]$, $E_{GT}[k]$, and $E_S[l]$ denote the corresponding energy in these regions. Given the threshold $\theta \in [0.5, 1)$, we say that an ideal segment $r_{GT}[k]$ is well-covered by an estimated segment $r_S[l]$ if the overlapping region, $r[k, l]$, includes most of the energy of $r_{GT}[k]$. That is, $E[k,l] > \theta \cdot E_{GT}[k]$. Similarly, we say $r_S[l]$ is well-covered by $r_{GT}[k]$ if $E[k,l] > \theta \cdot E_S[l]$. Then we label an overlapping region as follows.

- A region $r[k, l]$, $k>0$ and $l>0$, is labeled as correct if $r_{GT}[k]$ and $r_S[l]$ are mutually well-covered.
- Let $\{r_{GT}[k']\}$, $k'=k_1, k_2, \ldots, k_{K'}$, and $K'>1$, be all the ideal segments of target speech that are well-covered by an 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:

$$\sum\nolimits_{k'} E[k',l] > \theta \cdot E_S[l],\ k' = k_1, k_2, \ldots, k_{K'} \tag{8}$$

- Let $\{r_S[l']\}$, $l'=l_1, l_2, \ldots, l_{L'}$, and $L'>1$ be all the obtained segments that are well-covered by an ideal segment of target speech, $r_{GT}[k]$, $k>0$. The corresponding overlapping regions, $\{r[k, l']\}$, $l'=l_1, l_2, \ldots, l_{L'}$, are labeled as over-segmented if these regions include most of the energy of $r_{GT}[k]$, that is:

$$\sum\nolimits_{l'} E[k,l'] > \theta \cdot E_{GT}[k],\ l' = l_1, l_2, \cdots, l_{L'} \tag{9}$$

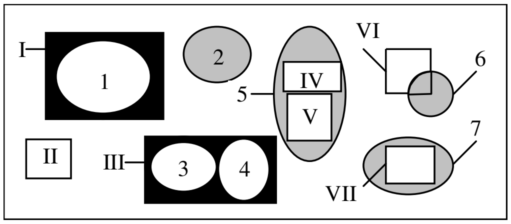

Figure 4. Illustration for correct segmentation, under-segmentation, over-segmentation, missing, and mismatch. Here an oval indicates an ideal segment and a rectangle an estimated one.

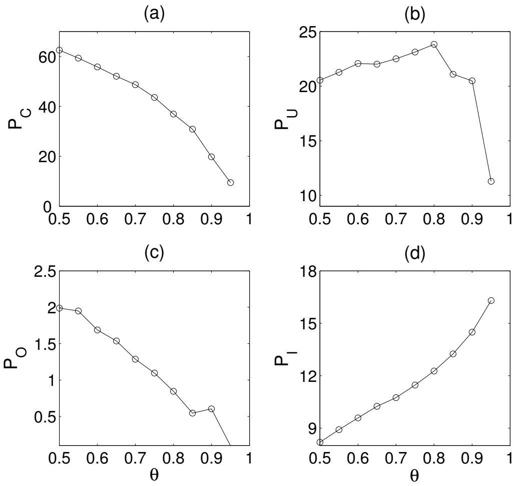

Figure 5. The result of auditory segmentation for the proposed system using the Perona-Malik model. The speech and interference are mixed at 0 dB SNR. (a) The average correct percentage. (b) The average percentage of under-segmentation. (c) The average percentage of over-segmentation. (d) The average percentage of mismatch.

- 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-segment, or over-segment, it is labeled as missing.
- For a region $r[0, l]$, the overlap between the ideal background $r_{GT}[0]$ and an estimated segments $r_S[l]$, it is labeled as mismatching if the estimated segment $r_S[l]$ is not well-covered by the ideal background.

To avoid labeling a region more than once, we stipulate that a region can only take one label with the following order of precedence: correct, under-segmented, over-segmented, missing, and mismatching.

Let E_C be the summated energy in all the regions labeled as correct, and E_U, E_O, E_M, and E_I in the regions labeled as under-segmented, over-segmented, missing, and mismatching respectively. Further let E_{GT} be the total energy of all ideal segments, except for the ideal background, and E_S that of all estimated segments, except for the estimated background. We use the following measurements for evaluation:

- The correct percentage is the percentage of correctly segmented speech to the total energy of ideal speech segments, i.e., $P_C=E_C/E_{GT}\times100\%$.
- The percentage of under-segmentation is the percentage of under-segmented speech to the total energy of ideal speech segments, i.e., $P_U=E_U/E_{GT}\times100\%$.
- The percentage of over-segmentation is the percentage of over-segmented speech to the total energy of ideal speech segments, i.e., $P_O=E_O/E_{GT}\times100\%$.
- The percentage of mismatch is the percentage of interference in the generated segments for target speech, i.e., $P_I=E_I/E_S\times100\%$.

There is no need to provide a separate measure for the missing category since $E_C + E_U + E_O + E_M = E_{GT}$. The advantage of evaluation according to each category is that it clearly shows each type of error. In image segmentation, the region corresponding to each segment is used for evaluation literally. Here, we use the energy of each segment instead. This is because for acoustic signal, T-F regions with strong energy are much more important than those with weak energy.

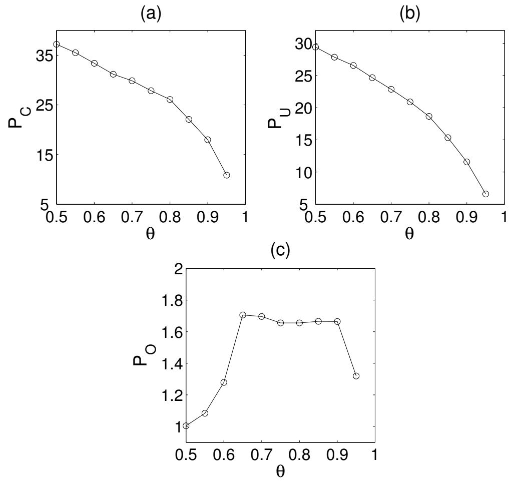

Figure 6. The result of auditory segmentation for stops, fricatives, and affricates. The speech and interference are mixed at 0 dB SNR. (a) The average correct percentage. (b) The average percentage of under-segmentation. (c) The average percentage of over-segmentation.

Fig. 5 shows the average P_C, P_U, P_O, and P_I for different θ. Note that the evaluation is more stringent for higher θ. Speech and interference are mixed at 0 dB SNR. Here the Perona-Malik model is used with λ=3, and the segments are generated in the order of the four scales: (32, 200), (18, 200), (32, 100) and (18, 100) (see Sect. 3). We have also considered segmentation using more scales, but results are not significantly better. As shown in Fig. 5, the correct percentage is about 63% when θ is 0.5, and it decreases to 10% as θ increases to 0.95. A significant amount of speech is under-segmented, which is due mainly to strong coarticulation between phonemes. Luckily, under-segmentation is not really an error since it basically gives larger segments for target speech, good for subsequent grouping. By combining P_C and P_U together, we have 83% of speech correctly segmented when θ is 0.5. Still more than 50% of speech is correctly segmented when θ increases to 0.85. In addition, we can see from Fig. 5 that over-segmentation is not a serious problem. The major error comes from missing, which indicates that portions of target speech are buried in the background. Compared with the SNR of the mixture, which is 0 dB, the percentage of mismatch is not significant. This shows that the interference and the target speech are well separated in the generated segments.

Fig. 6 shows the average P_C, P_U, P_O for stops, fricatives, and affricates, which constitute the major sources of unvoiced speech. The overall performance on these phonemes is worse than that for other phonemes. The average P_C+P_U for them is about 65% when θ is 0.5, and it drops below 50% when θ is larger than 0.75.

Fig. 7 compares the above result with that from a system using Gaussian smoothing over time (Sect. 3) and that from the segmentation stage preformed by the Wang-Brown model [13], which is chosen for comparison because it is a representative CASA model that includes segmentation as an explicit stage. Note that with Gaussian smoothing, the system amounts in fact to a Canny edge detector for onset and offset detection [3]. As

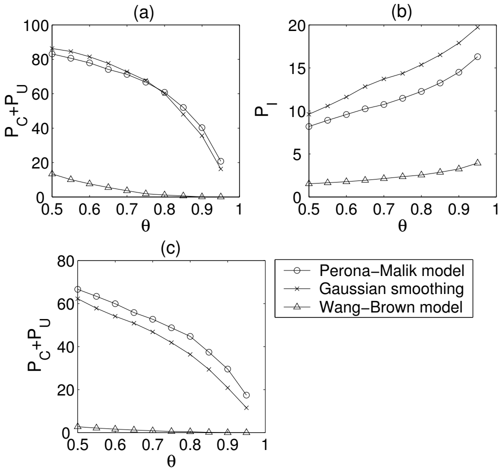

Figure 7. The result of auditory segmentation for the proposed system using the Perona-Malik model and Gaussian smoothing, and the segmentation result from the Wang-Brown model. Speech and interference are mixed at 0 dB SNR. (a) The average correct percentage plus the average percentage of under-segmentation for all the phonemes. (b) The average percentage of mismatch. (c) The average correct percentage plus the average percentage of under-segmentation for stops, fricatives, and affricates.

shown in the figure, Gaussian smoothing gives similar P_C+P_U performance for all the phonemes compared to the Perona-Malik model. However, it performs slightly worse than the Perona-Malik model on average P_I and on average P_C+P_U for stops, fricatives, and affricates. Our model outperforms the Wang-Brown model by large margins, except for P_I where their model performs better. Fig. 8 shows the performance of the system at different SNR levels using the Perona-Malik model. As SNR increases, the system performs better as expected. The improvement is most pronounced from 0 dB to 5 dB.

5. DISCUSSION

We have proposed a system for auditory segmentation and tested it on speech segmentation. This system correctly segments a majority of speech, including unvoiced speech. In addition, speech and interference are well separated to different segments.

Since there is no common definition for an acoustic event, we treat a phoneme, which is accepted as the basic unit of speech, as an acoustic event for target speech. In addition, a closure of a stop or an affricate is treated as a phoneme on its own. By our definition, the acoustic signal within each phoneme is generally stable and over-segmentation is undesirable. However, neighboring phonemes can be coarticulated, and it may also be appropriate to treat coarticulated phonemes as a single event. Coarticulation may cause false boundaries in ideal segments, and as a result under-segmentation can sometimes be more desirable. One may also define the whole utterance from the same speaker as one event. With this definition, we would have lower correct and under-segmentation percentages and much higher over-segmentation percentages. On the other hand, there are well-delineated boundaries between phonemes, and segmenting such boundaries should not be treated as an error.

Compared with previous CASA systems [2] [4] [5] [7] [13], our model makes four novel contributions. First, it provides a general framework for segmentation. Although we have only tested it 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 CASA cues, their utility has not been clearly demonstrated previously. Third, we have employed scale-space theory in the auditory domain. To our knowledge, it is the first time this theory is used in CASA. Finally, our system generates segments for both unvoiced and voiced speech. Little previous research has been conducted on organization of unvoiced speech, and yet speech segregation must address unvoiced speech.

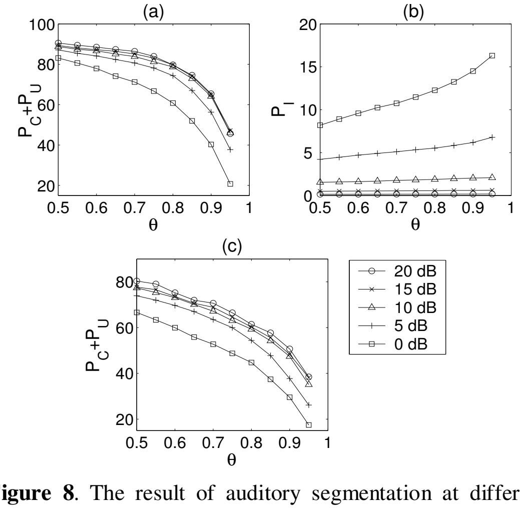

Figure 8. The result of auditory segmentation at different levels of SNR. (a) The average correct percentage plus the average percentage of under-segmentation for all the phonemes. (b) The average percentage of mismatch. (c) The average correct percentage plus the average percentage of under-segmentation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

6. ACKNOWLEDGEMENT

This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (FA9550-04-01-0117).

7. REFERENCE

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

[2] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, 8: 297-336, 1994.

[3] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, 8: 679-698, 1986.

[4] M. Cooke, *Modelling auditory processing and organization*, Cambridge, U.K.: Cambridge University Press, 1993.

[5] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. Dissertation, MIT Department of Electrical Engineering and Computer Science, 1996.

[6] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, 18: 673-689, 1996.

[7] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, to appear, 2004.

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

[9] D. O'Shaughnessy, *Speech communications: Human and machine*, 2nd ed., Piscataway, NJ: IEEE Press, 2000.

[10] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS final report, part B: Implementing a gammatone filterbank," Rep. 2341, *MRC Applied Psychology Unit*, 1988.

[11] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Machine Intell.*, 12: 629-639, 1990.

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

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

[14] 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., Springer press, 1997.