In: *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP'05), 2005

SEPARATION OF FRICATIVES AND AFFRICATES

Guoning Hu

Biophysics Program
The Ohio State University
395 Dreese Lab., 2015 Neil Ave.
Columbus, OH43210, USA
hu.117@osu.edu

DeLiang Wang

Department of Computer Science & Engineering and
Center of Cognitive Science
The Ohio State University
395 Dreese Lab., 2015 Neil Ave.
Columbus, OH43210, USA
dwang@cis.ohio-state.edu

ABSTRACT

Separating speech from acoustic interference is a very challenging task. In particular, no system successfully addresses the separation of unvoiced speech. Fricatives and affricates are two main categories of consonants that contain a significant amount of unvoiced signal. We propose a novel system that separates fricatives and affricates from non-speech interference. The system first decomposes the input mixture into segments, each of which contains signal mainly from one source. Then it detects segments dominated by unvoiced portions of fricatives and affricates with a feature-based Bayesian classifier, and groups these segments with voiced speech separated by a previous system. The proposed system is evaluated with various types of interference and produces promising results.

1. INTRODUCTION

Various sounds in the daily environment interfere with target speech. Separating speech from interference is required in many applications, such as robust speech recognition and hearing aids design, which has proved to be challenging. This task is more difficult in the monaural (one microphone) situation. However, a monaural solution is often necessary or desirable in practice.

Natural speech contains both voiced and unvoiced portions. Compared with voiced speech, unvoiced speech is severely weaker and more vulnerable to interfering sounds. In addition, unvoiced speech lacks harmonic structure, which is an effective cue for voiced speech separation. As a result, separating unvoiced speech is significantly more challenging. Currently no system is effective for unvoiced speech separation in the monaural situation. Speech enhancement techniques, such as spectral subtraction [9] and the subspace method [5], can deal with unvoiced speech only when prior information for interference is available, or interference satisfies specific statistical properties. Hence their applications are limited.

On the other hand, human listeners show impressive abilities in separating target speech in various environments, through a process referred to as auditory scene analysis (ASA) [3]. ASA generally takes place in two stages: segmentation and grouping. In segmentation, the acoustic mixture is decomposed into a collection of segments. Each segment occupies a contiguous time-frequency (T-F) region, containing signal mainly from one source. In grouping, the segments originated from the same source are grouped together. Psychophysical research on ASA has motivated computational systems of speech separation based on ASA principles, with success in separating voiced speech [4] [8]. However, these systems generally utilize harmonicity as the major ASA cue, and cannot deal with unvoiced speech.

To separate unvoiced speech, ASA cues other than harmonicity need to be employed. Therefore, we propose to separate unvoiced speech based on event onset analysis and acoustic-phonetic property of speech, which play important roles in speech perception [3]. Unvoiced speech mainly comes from three categories of phonemes: fricatives, affricates, and stops. Separating stop consonant has been addressed previously [6]. Considering the similarity between fricatives and affricates, we propose a system to separate fricatives and affricates together. In this paper, we will focus on situations where target speech is corrupted by non-speech intrusions.

Our system follows the two stages of ASA: segmentation and grouping. In segmentation, a previous system for auditory segmentation is applied [7], which forms segments for both voiced and unvoiced speech based on onset and offset analysis of auditory events. The next step is to detect segments dominated by unvoiced portions of fricatives and affricates, and group them with corresponding voiced portions. For non-speech intrusions, we may treat grouping as a classification task, i.e., to classify segments as dominated by fricatives, affricates, or other signal. Since each segment shall mainly originate from one source, segments dominated by fricatives and affricates are likely to have similar acoustic-phonetic characters as those from clean speech, while segments dominated by interference are likely to have different characters. Therefore, the system groups segments according to their acoustic-phonetic features. More specifically, it distinguishes segments dominated by fricatives, affricates, or interference through a Bayesian classifier based on segment spectrum and segment duration.

This paper is organized as follows. Sect. 2 discusses the computational goal for the proposed system. Sect. 3 and 4 describe details of segmentation and grouping. Sect. 5 presents evaluation results. A brief discussion is given in Sect. 6.

2. COMPUTATIONAL GOAL

An input mixture is first normalized at 60 dB SPL. It then passes through a 128-channel gammatone filterbank [11], with frequencies centered from 50 Hz to 8 *k*Hz. The input is further divided into 20-ms frames with 10-ms overlapping between consecutive frames. The T-F area in a certain channel within a certain frame is referred to as a T-F unit.

With the above signal decomposition, the computational goal of our system is to retain T-F units where target speech is more intense than interference and cancel other units. In other words, the goal is to identify a binary mask, referred to as the ideal binary mask, where 1 indicates that target is stronger than interference in the corresponding T-F unit and 0 otherwise. Target speech can then be resynthesized with the mask by retaining the acoustic energy from T-F regions corresponding to 1's and rejecting other energy (see [4] for more details). This computational goal is supported by the masking phenomenon of the auditory system [10] and researches on automatic speech recognition with the missing data technique [2]. For more detailed justification, see [8].

As an example, Fig. 1(c) and 1(d) show a mixture of a female utterance and crowd noise with music. Fig 1(e) shows the ideal binary mask. The speech resynthesized from the ideal binary mask is shown in Fig. 1(f), which is very close to the clean utterance shown in Fig. 1(b).

Some fricatives and affricates contain both voiced and unvoiced signal. The T-F regions dominated by their voiced portions are estimated with the Hu-Wang model [8] with pitch information obtained from clean speech. The output of the Hu-Wang model comprises two streams: target stream and interfering stream, corresponding to voiced speech and periodic components of interference respectively. Generally, unvoiced speech and non-periodic interference are not included in either stream. The proposed system is focused on determining T-F regions dominated by the unvoiced portions of fricatives and affricates.

3. SEGMENTATION

In this stage, we apply a previous system for segmentation based on analysis of event onset and offset [7]. The onsets and offsets generally correspond to sudden intensity increases and decreases. The input to the system is the average intensity of each gammatone filter output at every 1.25-ms window. It is smoothed to reduce the intensity fluctuations that do not correspond to actual onsets and offsets through a diffusion process [12]. The output of the diffusion process at a particular diffusion time, referred to as the scale, yields intensity smoothed to a particular degree. The larger the scale is, the smoother the output intensity is.

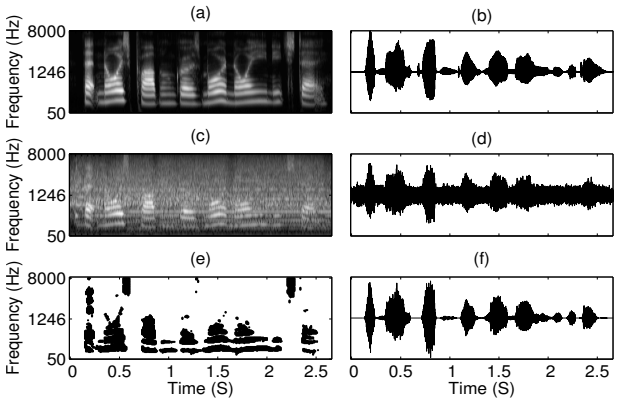

Figure 1. (a) Energy distribution across T-F units and (b) waveform of a female utterance, "That noise problem grows more annoying each day." (c) Energy distribution across T-F units and (d) waveform of the utterance mixed with crowd noise with music at 0 dB. (e) The ideal binary mask of the mixture. (f) The speech resynthesized from the ideal binary mask.

At a certain scale, onsets and offsets are detected as the peaks and valleys of the derivative of the smoothed intensity. Then the system combines common onsets and offsets into onset and offset fronts, and matches individual onset and offset fronts. The T-F region between an onset front and the matching offset front yields a segment.

Finally, the system undertakes a multi-scale integration of segmentation. It first forms segments at a larger scale. Then, at a smaller scale, it locates more accurate onset and offset positions for these segments, and adding new segments formed at the current scale. Then the system goes to an even smaller scale.

As an example, Fig. 2 shows the bounding contours of obtained segments for the mixture of speech and crowd noise with music. Compared with Fig. 1(e), the formed segments cover most speech dominant regions, including those dominated by fricatives and affricates. Some segments for the intrusion are also formed. For more details of this stage, see [7] .

4. SEGMENT GROUPING

The task for this stage is to detect segments dominated by the unvoiced portions of fricatives and affricates, and group them with target stream obtained with the Hu-Wang model [8] (see Sect. 2). It is executed in two steps: segment reduction and segment categorization. In segment reduction, the system removes all the segments with significant energy within time frames that could not contain fricatives and affricates. More specifically, the system first uses target stream to find time intervals containing voiced phonemes other than fricatives and affricates and then removes segments with significant energy within these intervals. As a result, the majority of segments dominated by signal other than fricatives and affricates are removed. This helps to increase the robustness of the system and greatly reduces the computation burden for segmentation categorization. In segment categorization, the system classifies the remaining segments as dominated by fricatives, affricates, or interference.

Each step of grouping involves a classification task. In segmentation reduction, the task is to label each frame within target stream as containing a fricative, an affricate, or any other phoneme. In segment categorization, the task is to distinguish

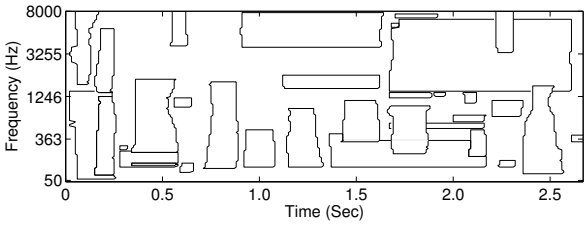

Figure 2. The bounding contours of obtained segments for the mixture of speech and crowd noise with music.

segments dominated by fricatives and affricates from those dominated by interference. For both classification tasks, the key is to choose distinctive features that characterize fricatives and affricates. Previous research suggests the following features to characterize the unvoiced portions of fricatives and affricates: spectrum, which includes the spectrum shape and spectrum intensity, duration, and transition (see[1] for example). The formant transition from a fricative or an affricate to the neighboring voiced phoneme is very difficult to obtain, and it is closely related to the spectrum. Therefore, we will use the first two features for classification.

Let H_0 be the hypothesis that a T-F region is dominated by interference, $H_{1,k}$ a T-F region dominated by a fricative or an affricate, indicated by k, and $H_{2,l}$ a T-F region dominated by other phoneme, indicated by l. Let $X(m)$ be the power spectrum for the input at frame m, which is obtained by an average of 64-point FFT over frame m, and $X_T(m)$ the corresponding power spectrum within the target stream. Frame m is labeled as non-fricative and non-affricate if

$$\max_k P(H_{1,k} \mid X_T(m)) < \max_l P(H_{2,l} \mid X_T(m)) \quad (1)$$

By applying the Bayesian rule, we have

$$\max_k [p(X_T(m) \mid H_{1,k}) P(H_{1,k})] < \max_l [p(X_T(m) \mid H_{2,l}) P(H_{2,l})] \quad (2)$$

Note that since the system labels individual frames, duration information is not used for this classification.

For a segment s, let E_S be the total energy included in the time frames labeled as non-fricative and non-affricate. If E_S is larger than 50% of the total energy of segment s, or it is larger than the average energy of fricatives and affricates in the training data, segment s is removed.

For a retained segment s, which lasts from frame m_1 to m_2, let $X_S(m)$ be the power spectrum within s at frame m, and $\boldsymbol{X}_S = (X_S(m_1), X_S(m_1+1), \ldots, X_S(m_2))$. Similar to (2), s is classified as dominated by a fricative or an affricate if:

$$\max_k [p(\boldsymbol{X}_S \mid H_{1,k}) P(H_{1,k})] > p(\boldsymbol{X}_S \mid H_0) P(H_0) \quad (3)$$

Because segments have varied sizes, the complexity for computing $p(\boldsymbol{X}_S|H_{1,k})$ and $p(\boldsymbol{X}_S|H_0)$ directly is extremely high. Fortunately, by considering only the dependence between two consecutive frames, we already have a good estimation of $p(\boldsymbol{X}_S|H_0)$. That is,

$$\begin{aligned} p(\boldsymbol{X}_S \mid H_0) &= p(X_S(m_1), X_S(m_1+1), \cdots, X_S(m_2) \mid H_0) \\ &\approx p(X_S(m_1) \mid H_0) \bullet p(X_S(m_1+1) \mid X_S(m_1), H_0) \\ &\quad \ldots \bullet p(X_S(m_2) \mid X_S(m_2-1), H_0) \\ &= p(X_S(m_1) \mid H_0) \prod_{m=m_1}^{m_2-1} p(X_S(m+1) \mid X_S(m), H_0), \end{aligned} \quad (4)$$

and the same for $p(\boldsymbol{X}_S|H_{1,k})$. Then (4) becomes

$$\max_k [p(X_S(m_1) \mid H_{1,k}) \; P(H_{1,k}) \prod_{m=m_1}^{m_2-1} p(X_S(m+1) \mid X_S(m), H_{1,k})]$$

$$> p(X_S(m_1) \mid H_0) \; P(H_0) \prod_{m=m_1}^{m_2-1} p(X_S(m+1) \mid X_S(m), H_0) \quad (5)$$

In addition, a Gaussian mixture model (GMM) is used for $p(\boldsymbol{X}(m)|H_0)$, $p(\boldsymbol{X}(m)|H_{1,k})$, and $p(\boldsymbol{X}(m)|H_{2,l})$ so that $p(\boldsymbol{X}_S(m)|H_0)$, $p(\boldsymbol{X}_S(m)|H_{1,k})$, and $p(\boldsymbol{X}_T(m)|H_{2,l})$ can be calculated directly from the corresponding marginal distributions.

In (5), segment duration information is implicitly utilized. To emphasis the contribution of duration in classification, we add duration as an auxiliary feature into (5) as follows:

$$\max_k [p(X_S(m_1), d_S \mid H_{1,k}) \; P(H_{1,k}) \prod_{m=m_1}^{m_2-1} p(X_S(m+1), d_S \mid X_S(m), H_{1,k})]$$

$$> p(X_S(m_1), d_S \mid H_0) \; P(H_0) \prod_{m=m_1}^{m_2-1} p(X_S(m+1), d_S \mid X_S(m), H_0), \quad (6)$$

so that the contribution from spectrum and that from duration are well balanced. Here d_S is the duration of segment s,

The prior distributions and probabilities required for calculating (2) and (6) are obtained from training. The speech samples are from the training part of the TIMIT database. For interference, we collected 100 environmental intrusions, including crowd noise, traffic noise, and wind, etc. 90 of them are used for training, and the remaining 10 are used for evaluation. A GMM with 8 mixtures and a full covariance matrix for each mixture is used for $p(\boldsymbol{X}(m), d|H_0)$, $p(\boldsymbol{X}(m), d|H_{1,k})$, and $p(\boldsymbol{X}(m)|H_{2,l})$.

One problem of the above approach for segment categorization is the potential mismatching between some real interference and intrusions used for training. Since only limited intrusions can be included in training, in a realistic environment some intrusions may not fit the trained interference model. As a result, these intrusions may fit the fricative or affricate model better than the interference model, though it does not fit either model well. Therefore, we introduce a confidence measure here. More specifically, a segment s is classified as dominated by a fricative or an affricate when both (6) is satisfied and the corresponding likelihood $p(\boldsymbol{X}_S, d_S|H_{1,k})$ is larger than a threshold, which guarantees that segment s has a good fit with the corresponding fricative or affricate model. The threshold is chosen to be exactly above the corresponding likelihood of 2% of training samples with the same segment size.

All the segments identified as dominated by fricatives or affricates are included into target stream. Then a binary mask is constructed by assigning 1 to a T-F unit within the target stream and 0 otherwise. The target speech is then resynthesized with the mask, which retains the acoustic energy from the mixture corresponding to 1's and rejects that corresponding to 0's.

As an illustration, Fig. 3 shows the target stream and the

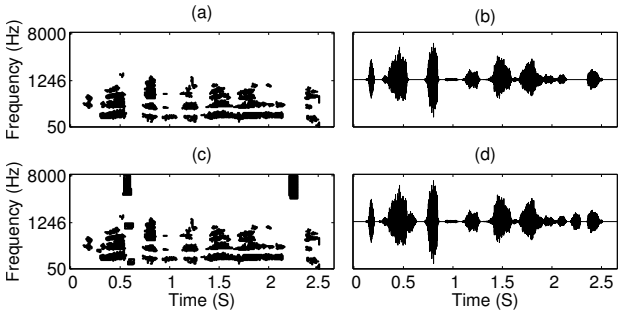

Figure 3. (a) The mask obtained from the Hu-Wang model for the mixture of speech and crowd noise with music, and (b) the corresponding resynthesized speech. (c) The mask obtained from the proposed system for the same mixture, and (d) the corresponding resynthesized speech.

resynthesized speech for the mixture of speech and crowd noise with music. Compared with Fig. 1(b), we can see that the majority of fricative and affricate signal is recovered, which is missing from the output of Hu-Wang model shown in 3(b). At the same time, a little interference is included into the resynthesized speech.

5. EVALUATION

The system is tested with 20 utterances from the testing part of the TIMIT database, mixing with 10 intrusions at different SNR levels. The intrusions are white noise, electrical fan, rooster crowing and clock alarm, traffic noise, crowd noise in playground, crowd noise with music, crowd noise with clapping, bird chirping and water flow, wind, and rain.

We evaluate the system with the following two measures: the percentage of energy missed by the system among the total energy of fricatives and affricates, referred to as P_{EL}, and the percentage of interference energy among the total energy retained by the system as fricatives or affricates, referred to as P_{NR}. An ideal binary mask is used as the ground truth for target speech, which is the computational goal for the system, as described previously (see Sect. 2). Table 1 shows the average P_{EL} and P_{NR} for the proposed system, and the average P_{NR} of the original mixture. Note that P_{EL} of the original mixture is 0%. As shown in the table, the proposed system is able to extract about 70% of the fricative and affricate energy from the mixture under different SNR situations. It also retains a certain amount of interference, which is not significant compared to the interference included in the original mixture.

Table 1. P_{EL} and P_{NR}

Overall SNR (dB)	Proposed system		Mixture
	P_{EL} (%)	P_{NR} (%)	P_{NR} (%)
0	33.48	35.11	82.17
5	32.39	21.19	61.38
10	29.39	8.47	36.05
15	29.60	5.34	16.39
20	29.88	3.30	6.21

Table 2. Segmental SNR for affricates, fricatives, stops, and silence

Overall SNR (dB)	Proposed system (dB)	SS (dB)	Mixture (dB)
0	-0.59	-10.40	-17.98
5	0.58	-7.59	-12.98
10	1.71	-4.72	-7.98
15	2.42	-1.35	-2.98
20	2.93	2.54	2.02

Table 2 shows the segmental SNR of the resynthesized speech averaged over time frames containing affricates, fricatives, stops, and silence, using clean speech as the signal. Regions for stops and silence are included since they may contain interference accepted by the proposed system. Other regions are not considered since they contain little interference accepted by the proposed system, and voiced speech is dominant in these regions. For comparison, Table 2 also shows the segmental SNR of the original mixture and that of the speech enhanced using spectral subtraction (SS), a standard method for speech enhancement [9]. As shown in the table, both the proposed system and spectral subtraction obtain average SNR improvement in every situation. The proposed system performs significantly better than spectral subtraction, especially under low SNR situations, which mainly due to the fact that spectral subtraction cannot deal with non-stationary interference.

6. DISCUSSION

Based on analysis of event onset and acoustic-phonetic properties of speech, the proposed system is able to separate most fricative and affricate consonants without including much interference into the separated speech. To our knowledge, it is the first system that aims explicitly at separating fricatives and affricates. Together with our previous research [6], we have shown that unvoiced speech can be separated through onset-based segmentation, feature-based classification, and subsequent grouping. Currently, the system deals with only non-speech interference. If the interference is an utterance from another speaker, a further process of assigning speech segments to corresponding speakers is required. This problem will be addressed in future research.

7. ACKNOWLEDGEMENT

This research was supported in part by an NSF grant (IIS-0081058), an AFOSR grant (FA9550-04-01-0117), an STTR grant from AFOSR, and an AFRL grant (FA8750-04-1-0093).

8. REFERENCES

[1] A.M.A. Ali and J. Van der Spiegel, "Acoustic-phonetic features for the automatic classification of fricatives," *JASA*, vol. 109, pp. 2217-2235, 2001.

[2] J. Barker, M.P. Cooke, and D.P.W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, in press, 2004.

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

[4] M. Cooke and G.J. Brown, "Computational auditory scene analysis: Exploiting principles of perceived continuity," *Speech Communication*, vol. 13, pp. 391-399, 1993.

[5] Y. Ephraim and H.L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Proc.*, vol. 3, pp. 251-266, 1995.

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

[7] G. Hu and D.L. Wang, "Auditory segmentation based on event detection," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

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

[9] 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.

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

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

[12] 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, pp. 3-28, 1997.