UNVOICED SPEECH SEGREGATION

DeLiang Wang

Department of Computer Science & Engineering
and Center of Cognitive Science
The Ohio State University
Columbus, OH43210, USA
dwang@cse.ohio-state.edu

Guoning Hu

Biophysics Program
The Ohio State University
Columbus, OH43210, USA
hu.117@osu.edu

ABSTRACT

Speech segregation, or the cocktail party problem, has proven to be extremely challenging. While efforts in computational auditory scene analysis have led to considerable progress in voiced speech segregation, little attention has been given to unvoiced speech which lacks harmonic structure and has weaker energy, hence more susceptible to interference. We describe a novel approach to address this problem. The segregation process occurs in two stages: segmentation and grouping. In segmentation, our model decomposes the input mixture into contiguous time-frequency segments by analyzing sound onsets and offsets. Grouping of unvoiced segments is based on Bayesian classification of acoustic-phonetic features. The proposed model yields very promising results.

1. INTRODUCTION

Segregation of target speech from its acoustic background, or the cocktail party problem, is one of the most challenging problems in signal processing. This problem is especially difficult in the monaural (one microphone) situation, where one cannot utilize spatial filtering to separate sounds from different locations. Various methods, such as spectral subtraction [2] and subspace analysis [6], have been proposed for monaural speech enhancement. However, these methods usually make strong assumptions on acoustic interference and therefore cannot address the variability of interference in real environments.

On the other hand, human listeners with normal hearing are capable of dealing with sound intrusions, even in the monaural condition. According to Bregman, the human auditory system segregates a target sound from interference through a process called auditory scene analysis (ASA) [3]. ASA generally takes place in two stages: segmentation and grouping. In the first stage, the auditory system decomposes the acoustic mixture into segments, each of which corresponds to a contiguous time-frequency (T-F) region and contains sound energy mainly from one source. In grouping, the auditory system integrates the segments from the target source to form a target stream. Bregman's ASA account has motivated the emerging area of computational ASA (CASA) study, which aims to achieve the human sound segregation performance by incorporating ASA principles [4].

There have been many efforts on developing CASA systems for speech segregation. Most of the previous studies use harmonicity as the major organizational cue and have achieved considerable success in dealing with voiced speech [4]. However, few studies have dealt with unvoiced speech. Unvoiced speech segregation is a more difficult problem because of two reasons. First, unvoiced speech lacks the harmonic cue and is often noise-like acoustically. Second, sound energy of unvoiced speech is usually much weaker than that of voiced speech; as a result, unvoiced speech is more susceptible to interference.

In this paper we describe a monaural CASA system that segregates unvoiced speech from non-speech interference. This system extends our previous effort in this direction [11] [13]. Our system follows the two established stages of ASA: segmentation and grouping. In segmentation, we generate segments for both voiced and unvoiced speech using a multiscale analysis of onsets and offsets of auditory events. In grouping, we detect segments dominated by the target and group them into the target stream. A key part of the second stage is Bayesian classification of acoustic-phonetic features in order to distinguish segments dominated by unvoiced speech from those dominated by non-speech interference. The features for classification include segment spectrum and segment duration.

The remainder of the paper is organized as follows. Sect. 2 addresses the question of how much speech is unvoiced. Sect. 3 describes the details of the proposed system. Sect. 4 presents evaluation results. Sect. 5 concludes the paper.

2. HOW MUCH SPEECH IS UNVOICED?

For English, unvoiced speech sounds come from the following consonant categories [14]:

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

Eight among them, i.e. /t/, /p/, /k/, /s/, /f/, /ʃ/, /θ/, and /tʃ/, are considered unvoiced. In addition, /h/ may be pronounced either in the voiced or the unvoiced manner.[1] Note that an affricate is a stop followed by a fricative so stops and fricatives are the two main categories comprising unvoiced speech.

Dewey conducted an extensive analysis of the relative frequencies of individual phonemes in written English [5] and concluded that unvoiced sounds account for 21.0% of all

[1] The TIMIT corpus [8] separates the voiced and unvoiced versions of /h/, from which we have estimated that 56.0% of /h/ pronunciations are unvoiced. This ratio is used in our subsequent estimates.

phonemes. For spoken English, a similar analysis by French, Carter, and Koenig on 500 telephone conversations containing a total of about 80,000 words [7] concluded that unvoiced phonemes account for about 24.0%. Another extensive, phonetically labeled corpus is the TIMIT, which contains 6,300 sentences read by 630 different speakers from various dialect regions in America [8]. Many of the same sentences in the TIMIT are read by multiple speakers and there are a total of 2,342 different sentences. We have performed an analysis of relative phoneme frequencies for distinct sentences in the TIMIT corpus, and found that unvoiced phonemes account for 23.1**%.** Table 1 shows the occurrence percentages of six phoneme categories from these studies. It is remarkable that these percentages are quite comparable despite that the fact that written, read, and conversational speech are different in many ways. In particular, the total percentages of the six consonant categories are nearly the same for the three different kinds of speech.

Table 1. Occurrence percentages of six consonant categories

Phoneme types	**Conversational**	**Written**	**TIMIT**
Voiced Stop	6.7	6.9	7.9
Unvoiced Stop	15.1	11.9	12.8
Voiced Fricative	7.5	9.5	7.7
Unvoiced Fricative	8.6	8.6	9.8
Voiced Affricate	0.3	0.4	0.6
Unvoiced Affricate	0.3	0.5	0.5
Total	38.5	37.8	39.3

Table 2. Duration percentage of six consonant categories

Phoneme types	**Conversational**	**TIMIT**
Voiced Stop	5.6	5.2
Unvoiced Stop	16.2	12.9
Voiced Fricative	5.3	5.8
Unvoiced Fricative	9.6	12.0
Voiced Affricate	0.3	0.6
Unvoiced Affricate	0.4	0.7
Total	37.4	37.2

A related question is the relative duration of unvoiced speech in spoken English. Unfortunately, the reported data on the telephone conversations in [7] do not contain durational information. To get an estimate, we use the durations obtained from a phonetically transcribed subset of the Switchboard corpus [9], which also consists of phone conservations. The amount of labeled data in [9], i.e. seventy-two minutes of conversation, is much smaller than that in [7]. Hence we do not use the labeled Switchboard corpus for phoneme frequency analysis; instead we insert the median durations from the transcription to the occurrence frequency data in [7] to deduce the relative durations of unvoiced sounds. Table 2 shows the resulting duration percentages of six phoneme categories, along with those from the TIMIT corpus. Once again, the percentages from the conversational speech are comparable with those from the read speech. In terms of overall time duration, unvoiced speech accounts for 26.2% in phone conversations and 25.6% in the TIMIT corpus.

The above two tables show that unvoiced sounds account for more than 20% of spoken English in terms of both occurrence frequency and time duration. In addition, voiced stops, fricatives, and affricates are often not totally voiced Therefore, unvoiced speech may occur more than suggested by the data shown above. Unvoiced consonants provide crucial information for speech recognition.

3. MODEL DESCRIPTION

Our model for unvoiced speech segregation first decomposes the input signal into T-F units, each corresponding to a bandpass filter response within a time frame. It then segregates unvoiced speech in two stages: segmentation and grouping. Specifically, the input signal is decomposed in the frequency domain with a 128-channel gammatone filterbank [15] with center frequency ranging from 50 Hz to 8 kHz. The filtered signal is further divided into 20-ms frames with 50% overlapping between neighboring frames.

The computational goal of our system is to identify the ideal binary mask [16], which equals 1 for all the T-F units that contain more target energy than interference energy and 0 for all the other units. With a binary T-F mask, one can resynthesize target speech by retaining the signals within T-F units labeled 1 and rejecting others. For more discussion of this computational goal, see [16]. As an example, Fig. 1(a) shows a male utterance, "He then offered his own estimate of the weather, which was unenthusiastic," and Fig. 1(b) shows a mixture of this utterance and crowd noise from a playground. Figs. 1(c) and 1(d) show, respectively, the ideal binary mask for this mixture and the corresponding resynthesized target, which is very similar to the clean utterance in Fig. 1(a).

3.1 Segmentation

We segment the input signal via a multiscale analysis of event onsets and offsets. Onset and offset are important ASA cues [3], corresponding to sudden intensity increases and decreases. For this we apply the system described in [11]. First, the intensity of each gammatone filter output is smoothed to different degrees. The smoothing process reduces the intensity fluctuations that do not correspond to actual onsets and offsets. The degree of smoothing is referred to as the scale and a larger scale yields a smoother output. Second, at each scale, the system marks the peaks and valleys of the first-order derivative of the smoothed intensity as onsets and offsets. Close onsets and offsets at adjacent frequency channels are connected into onset and offset fronts. The system then matches individual onset and offset fronts to form segments. Finally, the system performs multiscale integration from the largest scale to the smallest scale in an iterative manner. More specifically, at each scale, the system first locates more accurate boundaries for the segments obtained at a larger scale. Then, it forms new segments outside the existing segments.

Fig. 1(e) shows the bounding contours of the obtained segments for the aforementioned mixture of speech and crowd noise. Compared with Fig. 1(c), the computed segments cover most speech-dominant regions, including those dominated by unvoiced speech. On the other hand, some segments corresponding to the interference are also formed.

3.2 Grouping

A segment obtained in the previous stage may be dominated by voiced target, unvoiced target, or interference. Since our goal is to segregate unvoiced speech, we need to group segments dominated by unvoiced target. In addition, we also need to group segments dominated by voiced target since these segments may also include unvoiced target. Note that an unvoiced sound is often coarticulated with a neighboring voiced sound, and therefore both sounds may be put into the same segment during segmentation [11].

We first segregate voiced target using our previous voiced segregation system [12] with target pitch obtained from a clean utterance using *Praat* [1]. The resulting target stream is denoted by S_T^1. Then we identify the segments dominated by voiced target according to this stream. We consider a segment to be dominated by voiced target if:

- More than half of its total energy is included in the voiced time frames, and
- More than half of its energy in the voiced frames is included in the stream, S_T^1.

All the segments dominated by voiced target are grouped into the segregated voiced stream, yielding a new stream, denoted by S_T^2.

For remaining segments, we group the segments dominated by unvoiced speech by using an algorithm recently proposed for fricative and affricate segregation [13].

A segment dominated by unvoiced target is likely located at unvoiced time frames, though it may contain some T-F units at voiced time frames since stops, fricatives, and affricates often contain both voiced and unvoiced signal (see Sect. 2). This property is, however, not shared by many interference-dominated segments that may have significant energy in voiced frames. Such segments are removed as follows.

Let H_0 be the hypothesis that a segment is dominated by interference, H_1 that the segment dominated by a stop, a fricative, or an affricate, and H_2 that the segment dominated by any other phoneme. Let u_{cm} denote a T-F unit at frequency channel c and frame m, and $X(c, m)$ the intensity of signal in this unit. We label voiced frames that unlikely contain fricatives, affricates, or stops, according to the segregated voiced target S_T^1. A voiced frame m is so labeled if

$$P(H_1 \mid X_T(m)) < P(H_2 \mid X_T(m)) \tag{1}$$

where $X_T(m) = [X_T(1, m), X_T(2, m), \ldots, X_T(N_c, m)]$, $N_c = 128$ is the total number of channels, and

$$X_T(c,m) = \begin{cases} X(c,m) & if\ u_{cm} \in S_T^1 \\ 0 & else \end{cases} \tag{2}$$

In other words, $X_T(m)$ corresponds to the spectrum of S_T^1 at frame m. A segment is removed if its energy in these labeled frames is greater than 50% of its total energy or if it occupies more than 5 labeled frames; the latter condition amounts to that the T-F region of the segment in the labeled frames is longer than 50% of the average duration of unvoiced phonemes. As a result of this step, most of the segments dominated by interference are removed. We find that this step increases the robustness of the system and greatly reduces the computational burden for the following segment classification.

We classify the remaining segments as dominated by either unvoiced speech or interference, on the basis of segment spectrum and segment duration. Let s be a remaining segment lasting from frame m_1 to m_2. Let $X_s(m) = [X_s(1, m), X_s(2, m), \ldots, X_s(N_c, m)]$ and $\boldsymbol{X}_s = [X_s(m_1), X_s(m_1+1), \ldots, X_s(m_2)]$, where

$$X_s(c,m) = \begin{cases} X(c,m) & if\ u_{cm} \in s \\ 0 & else \end{cases} \tag{3}$$

s is classified as dominated by unvoiced speech if:

$$P(H_1 \mid \boldsymbol{X}_s) > P(H_0 \mid \boldsymbol{X}_s) \tag{4}$$

As in [13], to simplify the computation of $P(H_0|\boldsymbol{X}_s)$ and $P(H_1|\boldsymbol{X}_s)$, we consider the dependence only between consecutive frames and add the duration as an additional feature. As a result, (4) becomes:

$$P(H_1 \mid X_s(m_1), d_s) \prod_{m=m_1}^{m_2-1} \frac{P(H_1 \mid X_s(m+1), X_s(m), d_s)}{P(H_1 \mid X_s(m), d_s)}$$

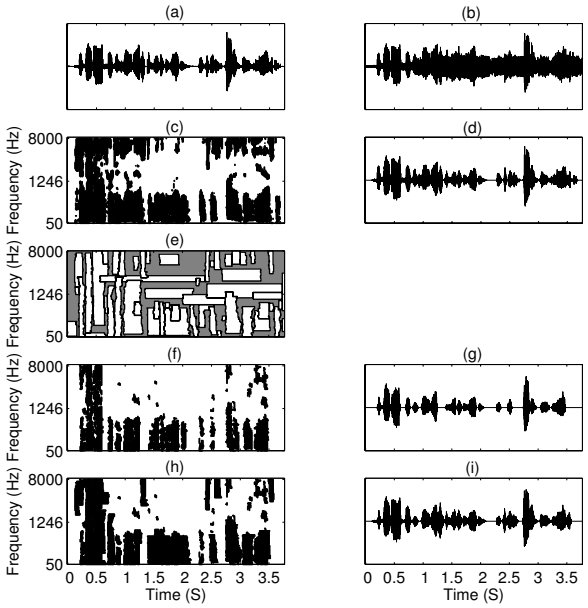

Fig. 1. Speech segregation. (a) Waveform of a male utterance, "He then offered his own estimate of the weather, which was unenthusiastic." (b) Waveform of the utterance mixed with crowd noise from a playground at 0 dB SNR. (c) Ideal binary mask of the mixture. (d) Speech resynthesized from the ideal binary mask. (e) Bounding contours of obtained segments for the mixture. The background is indicated by gray. (f) Segregated voiced target, S_T^1, and (g) the corresponding resynthesized speech. (h) Final segregated target, S_T^3, and (i) the corresponding resynthesized speech.

$$> P(H_0 \mid X_s(m_1), d_s) \prod_{m=m_1}^{m_2-1} \frac{P(H_0 \mid X_s(m+1), X_s(m), d_s)}{P(H_0 \mid X_s(m), d_s)} \tag{5}$$

where d_s is the segment duration.

The probabilities required for calculating (1) and (5) are obtained from training. A multilayer perception (MLP) with 1 hidden layer and 5 hidden units is trained to distinguish fricatives, affricates, and stops from other phonemes. Another MLP with the same configuration is trained to distinguish unvoiced stops, fricatives, and affricates from interference. The output from the two MLPs provides the required posterior probabilities.

In [13], Gaussian mixture models (GMMs) are trained to model different phonemes and interference. The classification is based on the likelihoods from the GMMs. To deal with the potential mismatching between the actual interference and the interference used for training, a confidence measure had to be used. Here, instead of modeling target and interference separately, we train MLPs to distinguish between target and interference. Therefore, no confidence measure is needed.

All the segments classified as unvoiced speech are grouped into the target stream, yielding the final target stream, denoted by S_T^3. To illustrate the system performance, Figs. 1(f), 1(g), 1(h), and 1(i) show S_T^1, S_T^3, and their corresponding waveform signals, respectively, for the mixture of speech and crowd noise in Fig. 1(b). As seen from the figure, S_T^3 includes a majority of the unvoiced target, which is missing from S_T^1 – the segregated voiced target. At the same time, S_T^3 also includes a little more interference.

4. EVALUATION

We use the training part of the TIMIT corpus for our training purposes. Twenty utterances from the testing part of the TIMIT corpus are used for testing. We have collected 100 environmental intrusions, 90 for training and 10 for testing. These intrusions have a large variety, including traffic and wind noises (see [11] [13]).

We evaluate the system performance by comparing the segregated target with the ideal binary mask – the stated computational goal of our system. Two error measures are used here: energy loss, which is the relative target energy missed by the system, and noise residue, which is the relative interference energy retained by the system [12]. Table 3 shows the percent energy loss of stops with respect to their total energy, and that of fricatives and affricatives. Fricatives and affricatives are evaluated together because they are quite similar and there are not many affricates in the testing data. Each value represents the average of 200 mixtures. Our system recoveres about 60% ~ 70% of stops and near 80% of fricatives and affricates for a range of SNR levels. Most of the stop energy is recovered by grouping segments dominated by voiced target, which suggests that stops tend to be coarticulated with neighboring voiced sounds. Most of the energy from fricatives and affricates is recovered from the segments dominated by unvoiced target.

Table 3. Percent energy loss for stops, fricatives, and affricates. "F&A" refers to fricatives and affricates

SNR (dB)	S_T^1		S_T^2		S_T^3	
	Stop	F&A	Stop	F&A	Stop	F&A
0	76.1	86.5	47.9	77.4	39.5	23.4
5	75.9	86.3	44.9	77.6	36.5	22.0
10	75.3	86.2	40.3	77.6	31.2	19.1
15	74.7	86.1	38.1	77.5	31.8	21.9
20	74.5	86.0	35.1	76.9	30.2	22.8

Table 4. Percent total energy loss and total noise residue

SNR (dB)	S_T^1		S_T^2		S_T^3	
	P_{EL}	P_{NR}	P_{EL}	P_{NR}	P_{EL}	P_{NR}
0	21.4	4.5	13.3	11.5	9.2	13.2
5	18.4	1.6	8.6	5.2	6.4	5.8
10	16.6	0.5	7.2	1.8	4.9	2.0
15	15.5	0.2	6.5	0.6	4.3	0.7
20	15.2	0.1	5.8	0.2	3.8	0.3

Table 4 shows the percent energy loss (denoted by P_{EL}) and the percentage of noise residue (denoted by P_{NR}) for all the speech sounds in the testing corpus. Again, each value is the average of 200 mixtures. As shown in the table, our overall system groups much more target energy in comparison with voiced segregation only. Although the current system also includes a certain amount of interference into the target stream when segregating unvoiced speech, the amount is not significant compared to the interference in original mixtures. By taking advantage of coarticulation between neighboring phonemes our system recovers significantly more stop energy than a previous system that segregates stop consonants using onset detection and feature-based classification, especially at low SNR levels [10]. In addition, our model recovers nearly 10% more of fricative/affricate energy than the system described in [13].

5. CONCLUSION

After an analysis of unvoiced speech, we have described a CASA study on unvoiced speech segregation. The proposed model generates segments with a multiscale analysis of event onsets and offsets. Unvoiced speech with strong coarticulation with neighboring voiced speech is segregated by grouping segments dominated by voiced target. Other unvoiced sounds are further segregated by classifying acoustic-phonetic features of the speech signal. The model successfully groups most of unvoiced speech without including much interference.

Our work represents the first systematic effort on unvoiced speech segregation. Our results, plus earlier CASA successes on voiced speech segregation, suggest that CASA is a very promising approach to addressing the cocktail party challenge.

6. ACKNOWLEDGEMENT

This research was supported in part by an AFOSR grant (FA9550-04-01-0117) and an AFRL grant (FA8750-04-1-0093).

7. REFERENCES

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

[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech & Signal Process.*, vol. 27, pp. 113- 120, 1979.

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

[4] 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, NY: Springer, pp. 371-402, 2005.

[5] G. Dewey, *Relative frequency of English speech sounds,* Cambridge, MA: Harvard University Press, 1923.

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

[7] H. Fletcher, *Speech and hearing in communication,* New York, NY: Van Nostrand, 1953.

[8] J. Garofolo*, et al.*, "Darpa TIMIT acoustic-phonetic continuous speech corpus," *NISTIR 4930*, 1993.

[9] S. Greenberg, J. Hollenback, and D.P.W. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *ICSLP'96*, pp. 24-27, 1996.

[10] G. Hu and D.L. Wang, "Separation of stop consonants," in *Proc. ICASSP*, Vol. 2, pp. 749-752, 2003.

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

[12] 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.

[13] G. Hu and D.L. Wang, "Separation of fricatives and affricates," in *Proc. ICASSP*, Vol. 1, pp. 1101-1104, 2005.

[14] P. Ladefoged, *Vowels and consonants,* Oxford, UK: Blackwell, 2001.

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

[16] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines,* P. Divenyi, Ed., pp. 181-197, 2005.