INCORPORATING SPECTRAL SUBTRACTION AND NOISE TYPE FOR UNVOICED SPEECH SEGREGATION

Ke Hu and DeLiang Wang

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{huk, dwang}@cse.ohio-state.edu

ABSTRACT

Unvoiced speech poses a big challenge to current monaural speech segregation systems. It lacks harmonic structure and is highly susceptible to interference due to its relatively weak energy. This paper describes a new approach to segregate unvoiced speech from nonspeech interference. The system first estimates a voiced binary mask, and then performs unvoiced speech segregation in two stages: segmentation and grouping. In segmentation, time-frequency units labeled as 0 in the voiced binary mask are first used to estimate the noise energy and spectral subtraction is then performed to generate time-frequency segments in unvoiced intervals. Based on the type of noise, unvoiced segments are grouped either by selecting segments consistent with those generated by onset/offset analysis or by Bayesian classification of acoustic-phonetic features. Systematic evaluation and comparison show that the proposed approach improves the performance of unvoiced speech segregation considerably.

Index Terms— Unvoiced speech segregation, nonspeech interference, spectral subtraction, onset/offset analysis, Bayesian classification

1. INTRODUCTION

In real-world listening environments, speech reaching our ears is often corrupted by various types of acoustic interference. Segregating speech monaurally from interference is very useful for many applications. While previous research has led to considerable advances in voiced speech segregation, unvoiced speech segregation remains a major challenge. In this paper, we study monaural segregation of unvoiced speech from nonspeech interference.

Motivated by the auditory scene analysis theory of Bregman [1], computational auditory scene analysis (CASA) aims to achieve sound organization based on perceptual principles [2]. A reasonable goal of CASA is the ideal binary mask (IBM) [3], which assigns values of 0 and 1 in the time-frequency (T-F) domain by comparing the local signal-to-noise ratio (SNR) within each T-F unit against a threshold using premixed source signals. Subject tests have shown that speech segregated by IBM leads to dramatic intelligibility improvements for both normal-hearing and hearing-impaired listeners [4, 5, 6].

Segmentation and grouping are two main stages of CASA. In segmentation, the input is decomposed to segments, each of which is a contiguous T-F region primarily originating from a single sound source. The second stage combines segments that likely arise from the same source into a stream. Fundamental frequency ($F0$) has been used as a primary cue for speech segregation; however, systems that employ $F0$ only cannot deal with unvoiced speech segregation.

Hu and Wang recently studied the unvoiced speech segregation problem in the CASA framework and successfully extracted a majority of unvoiced speech from nonspeech interference [7]. However, auditory segmentation based on onsets and offsets may miss weak portions of unvoiced speech. From another perspective, monaural speech enhancement methods enhance noisy speech based on certain assumptions or models of speech and interference [8]. Speech enhancement methods improve speech quality. However, they have a limited ability to improve speech intelligibility [9], probably because generated masks show large deviations from the IBM [6].

In this paper, we describe a speech segregation algorithm that directly estimates the unvoiced IBM. In the first step, our system estimates a voiced binary mask using a supervised learning approach [10]. Then unvoiced speech segregation takes place in two stages: segmentation and grouping. In segmentation, noise energy is first estimated using the voiced binary mask and spectral subtraction is then performed to generate unvoiced T-F segments. Considering that noise characteristics vary in different environments, we propose to use different grouping methods based on noise type. Noise is categorized into three classes based on the variance of noise energy: stationary, nonstationary and highly nonstationary. For each type of interference, an appropriate grouping method is applied to grouping target segments.

The rest of the paper is organized as follows. The next section describes the proposed system in detail. Section 3 shows the systematic evaluation results and conclusion is given in Section 4.

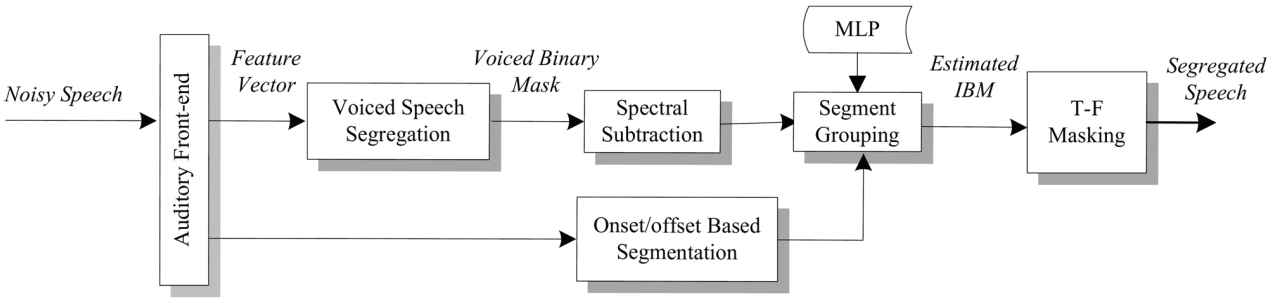

Fig. 1. Schematic diagram of the proposed CASA system. Voiced segregation lays the foundation for unvoiced speech segregation. The unvoiced segregation consists of two stages: segmentation and grouping. In segmentation, the system performs spectral subtraction on noise estimate from voiced binary mask. Grouping is done based on the noise type either by using segments from onset/offset analysis or Bayesian classification of acoustic-phonetic features.

2. SYSTEM DESCRIPTION

The complete system is illustrated in Fig. 1. The input of the noisy speech is first processed by an auditory front-end, which models cochlear filtering and auditory nerve transduction. Specifically, cochlear filtering is performed using a gammatone filterbank with 128 frequency channels whose center frequencies range from 50 Hz to 8 kHz. Each channel output is then processed by the Meddis hair cell model to simulate auditory nerve transduction. The output from the Meddis model is divided into 20-ms-long time frames with a 10-ms frame shift. The resulting representation is called a cochleagram. Details of cochleagram analysis and synthesis can be found in [2]. The envelope of each cochleagram channel is further extracted using a bandpass filter with passband from 50 Hz to 550 Hz. In the following subsections, we describe the voiced speech segregation algorithm and detail the unvoiced speech segregation algorithm.

2.1. Voiced Speech Segregation

Voiced speech segregation is performed first for the purpose of aiding the unvoiced speech segregation. Following [10], a 6-dimensional feature vector is extracted to represent the harmonic structure within the T-F unit of frequency channel c and time frame m in voiced frames:

$$X_{c,m}=\begin{pmatrix} A(c,m,\tau_m) \\ \text{int}(\bar{f}(c,m)\cdot\tau_m) \\ \left|\bar{f}(c,m)\cdot\tau_m-\text{int}(\bar{f}(c,m)\cdot\tau_m)\right| \\ A_E(c,m,\tau_m) \\ \text{int}(\bar{f}_E(c,m)\cdot\tau_m) \\ \left|\bar{f}_E(c,m)\cdot\tau_m-\text{int}(\bar{f}_E(c,m)\cdot\tau_m)\right| \end{pmatrix}, \quad (1)$$

where $A(c,m,\tau_m)$ is the autocorrelation function of the front-end response with time lag of the estimated pitch period τ_m, and the function int(x) returns the nearest integer. $\bar{f}(c,m)$ denotes the estimated average instantaneous frequency of the response within the T-F unit. We estimate the harmonic number as $\bar{f}(c,m)\cdot\tau_m$, which is close to an integer (greater than or equal to 1) when the unit responds to a harmonic. The third feature measures the deviation from the nearest harmonic. Essentially, the first three features are extracted from auditory front-end responses, and the last three are extracted from response envelopes (indicated by the subscript E). Given the feature vector, we train a Multilayer Perceptron (MLP) for each channel to directly maximize the SNR of segregated speech [11]. During MLP training, IBM provides the desired output. Since the input SNR of all mixtures is set to 0 dB, we choose the local SNR threshold (LC) [4] to be 0 dB. Trained MLP's are then used to label T-F units in voiced frames. A T-F unit is labeled as target speech if the posterior probability that the unit contains stronger target energy is greater than the posterior probability that the unit contains stronger interference energy. For convenience, we call target dominant units as active units (with label 1), and those dominated by interference as inactive units (with label 0).

2.2. Unvoiced Speech Segmentation Based on Spectral Subtraction

Obviously, the feature vector in [10], which encodes harmonic structure, cannot be used to segregate unvoiced speech. Nevertheless, voiced speech segregation result can assist in segregation of unvoiced speech. Our unvoiced speech segregation algorithm follows the CASA framework of segmentation and grouping.

In segmentation, we first estimate the noise energy for each channel in unvoiced intervals (sets of consecutive unvoiced frames) by averaging mixture energy in inactive T-F units in neighboring voiced intervals:

$$NE_c=\frac{1}{(l_1+l_2)}\left(\sum_{\substack{m=m_1-l_1 \\ m\notin[m_1,m_2]}}^{m_2+l_2} E_c(m)\cdot(1-y_c(m))\right), \quad (2)$$

where m_1 and m_2 denote the indices of first and last frames of the current unvoiced interval respectively, and l_1 and l_2 are the frame lengths of preceding and succeeding voiced intervals, respectively. $E_c(m)$ is the mixture energy at frame m and channel c, and $y_c(m)$ the actual output from the MLP. Given estimated noise energy, we calculate the local SNR in each T-F unit in unvoiced intervals. A T-F unit is labeled as target if and only if its local SNR exceeds the LC. Given labeled T-F units, segments are formed by simply merging neighboring T-F units in both temporal and spectral dimensions.

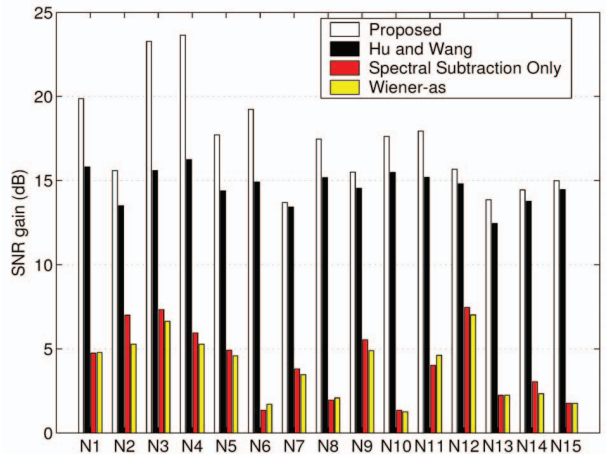

Fig. 2. Unvoiced speech SNR gain of four algorithms. On average, the proposed system achieves an SNR gain of 17.3 dB across all interferences. It has a 2.6 dB SNR improvement over Hu and Wang system. Also, we observe that there is a substantial gap between speech enhancement methods and CASA algorithms.

2.3. Noise Type Based Grouping

Spectral subtraction based segmentation captures most of the unvoiced speech segments, but still retains many T-F units dominated by interference. To further remove residual noise after subtraction is the task of grouping. We find that in order to obtain good grouping performance, different grouping methods are needed for different noise types.

We classify interference into three classes – stationary, nonstationary and highly nonstationary. The classification is based on noise energy variance as the amount of noise energy fluctuation directly defines the type of noise. We estimate the nonstationarity of interference by calculating the variance of noisy speech energy in inactive units in the voiced mask for each channel:

$$\sigma_c^2 = \frac{1}{N_c} \sum_{m \in VF_c} \left(E_c(m) - \bar{E}_c \right)^2, \tag{3}$$

where VF_c denotes the set of frame indices of inactive units in voiced intervals at channel c and $\bar{E}_c$ represents the mean energy for all inactive units at channel c. We use the mean variance $\bar{\sigma}$ and the maximum variance $\sigma_{\max}$ across all channels to classify current interference into three classes

- Stationary: $\bar{\sigma} < C_1$
- Nonstationary: $\bar{\sigma} > C_1$ and $\sigma_{\max} < C_2$
- Highly nonstationary: $\sigma_{\max} > C_2$

where C_1 is chosen to be 3dB to represent the threshold that the energy fluctuation is as large as the average and C_2 is empirically chosen as 10dB. After the noise type classification, a grouping method is used for each noise type to combine segments as follows.

For stationary noise, we use segments generated by onset/offset analysis [12] to select target segments. Specifically, if a subtraction-based segment overlaps with an onset/offset based segment so that at least 90% of the latter energy is contained in the overlapping region, the segment is kept; otherwise, it is discarded. Since segments marked by onsets and offsets only correspond to speech in stationary noise, the selection removes segments that are likely dominated by background noise. Furthermore, noise estimation by averaging is relatively accurate in stationary noise and spectral subtraction based segments also capture weak portions of unvoiced speech, which can be missed in onset/offset analysis.

For nonstationary and highly nonstationary noises, segments produced by onset-offset analysis also contains those corresponding to interference, hence providing little help to identify the T-F boundaries of unvoiced speech segments. In this case, we train two different Bayesian classifiers based on acoustic-phonetic features to classify segments for the two types of noise. Following [7], we use the masked spectral vector

$$Y_m = \{Y(c,m), \forall c\} \tag{4}$$

as the input to the MLP's, where $Y(c,m)=E_c(m)\cdot y_c(m)$. Specifically, $y_c(m)$ is the binary label from spectral subtraction. By making the independence assumption of frames in one segment, a segment S lasting from frame m_1 to m_2 is classified as unvoiced speech if the posterior probability that it belongs to unvoiced speech is higher than that it belongs to interference

$$\prod_{m=m_1}^{m_2} P(H_0(m) \mid Y_m) > \prod_{m=m_1}^{m_2} P(H_1(m) \mid Y_m), \tag{5}$$

where H_0 is the hypothesis that the segment is dominated by unvoiced speech and H_1 is the hypothesis that it is dominated by interference. Each of the two classifiers is trained for its corresponding type of noise (nonstationary or highly nonstationary). MLP's have the same architecture as those in voiced segregation and IBM provides the desired output in training. The trained MLP's are then used to classify segments into speech dominant or interference dominant, hence grouping target segments.

3. EVALUATION

We evaluate our system on mixtures of IEEE sentences [13] and 15 nonspeech interferences. The noises are N1–electric fan, N2–white noise, N3–crowd noise at a playground, N4–crowd noise with clap, N5–rain, N6–babble noise, N7–clock alarm, N8–cocktail party noise, N9–rock music, N10–siren, N11–traffic noise, N12–wind, N13–machine noise, N14–bird chirp with water flowing, and N15–telephone ring. These are chosen to cover a wide variety of real-world interferences. The interferences can be found at [14] and [15]. The IEEE sentence corpus contains 720 phonetically-balanced sentences with relatively low word-context predictability. All sentences were recorded by a single female speaker at a 20 kHz sampling frequency. We downsample the signals to 16 kHz. Each target utterance is mixed with a noise sample randomly cut out from an individual interference at the input SNR of 0 dB. For supervised learning in voiced speech segregation, the training set contains 100 mixtures for MLP training and the remaining 620 are used to evaluate system performance. Feature extraction requires the knowledge of $F0$ at each frame, and we use a pitch tracking algorithm [10] to obtain the pitch contours.

Given that the computational objective of our proposed system is to estimate the unvoiced IBM, we use the same SNR measure in [7] and use the resynthesized speech from the unvoiced IBM as the ground truth

$$SNR = 10\log_{10}\left(\sum_n S_I^2[n] / \sum_n (S_I[n] - S_E[n])^2\right), \quad (6)$$

where $S_I[n]$ and $S_E[n]$ are signals resynthesized from the unvoiced IBM and the mask estimated by our system, respectively. The unvoiced IBM is taken as the part of IBM in unvoiced intervals, which are determined by the pitch contours extracted from clean speech using *Praat* [16]. To obtain only unvoiced speech, we consider segments that are below 1 kHz and connected with voiced mask as voiced speech segments.

Segregation performance of the proposed system in terms of SNR gain for unvoiced speech is summarized in Fig. 2. Each white bar in the figure shows the average SNR gain of all test mixtures for one interference. Across all interferences, the proposed system achieves an average SNR gain of 17.3 dB.

To put the performance of our system in perspective, we compare the performance of the Hu and Wang [7]. In their system, segmentation is performed by onset-offset analysis and grouping is based on Bayesian classification of acoustic-phonetic features. As clear in the figure, the proposed algorithm performs uniformly better for all interferences. On average, the proposed algorithm has a 2.6 dB SNR improvement over the Hu and Wang system. The improvement is more pronounced for stationary and nonstationary noises. We attribute the improvement to the successful capture of weak portions of unvoiced speech and noise type based grouping. Accurate noise energy estimation in stationary noises enables the extraction of weak unvoiced speech. In nonstationary and highly nonstationary noises, Bayesian classifiers specially designed for different noise types increase the discriminant power.

To further isolate the effects of noise type based grouping, we also show the segregation performance for unvoiced speech with spectral subtraction only in Fig. 2. Finally, we compare with a Wiener algorithm based on *a priori* SNR estimation (Wiener-as), which is reported as the best performing speech enhancement algorithm in sentence and consonant recognition tasks [9]. For all the algorithms, we use the same pitch tracking algorithm [10] to produce pitch contours. As can be observed, there is a substantial gap between the speech enhancement methods and the CASA algorithms. This indicates that the CASA framework of segmentation and grouping is effective for speech segregation.

4. CONCLUSION

Separation of unvoiced speech is a very challenging problem. This paper proposes a CASA-based approach for unvoiced speech segregation from nonspeech interference. We take advantage of voiced speech segregation to estimate noise and apply spectral subtraction to produce unvoiced segments. Segments produced this way are subsequently grouped into a speech stream that takes noise types into account. Systematic evaluation and comparison show our algorithm performs considerably better than previous approaches.

Acknowledgements. This research was supported in part by an NSF grant (IIS-0534707), an AFOSR grant (FA9550-08-1-0155), and the VA Biomedical Laboratory Research and Development Program. We would like to thank Z. Jin and G. Hu for providing their programs in this work.

6. REFERENCES

[1] S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT press, 1990.

[2] D. L. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms and applications,* IEEE Press/Wiley-Interscience, 2006.

[3] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, P. Divenyi Ed., Norwell, MA: Kluwer Academic 2005, pp. 181-197.

[4] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[5] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing*, vol. 27(5), pp. 480–492, 2006.

[6] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of Acoustic Society of American*, vol. 123, pp. 1673–1682, 2008.

[7] G. Hu and D.L. Wang, "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, vol. 124, pp. 1306-1319, 2008.

[8] P. Loizou, *Speech Enhancement: Theory and Practice*, Taylor and Francis, Boca Raton FL, 2007.

[9] Y. Hu and P. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of Acoustical Society of America,* 122(3), pp. 1777–1786, 2007.

[10] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, Ph.D. dissertation, Biophysics Program, The Ohio State University, 2006.

[11] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *Technical Report* OSU-CISRC-5/08-TR27, Department of Computer Science and Engineering, The Ohio State University, Columbus Ohio, USA, 2008.

[12] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 396–405, 2007.

[13] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[14] http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

[15] http://www.dcs.shef.ac.uk/~martin/corpora/cookephd.tar.gz

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2005. Available at http://www.fon.hum.uva.nl/praat