Unvoiced Speech Segregation Based on CASA and Spectral Subtraction

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 separation is an important and challenging problem that has not received much attention. We propose a CASA based approach to segregate unvoiced speech from nonspeech interference. As unvoiced speech does not contain periodic signals, we first remove the periodic portions of a mixture including voiced speech. With periodic components removed, the remaining interference becomes more stationary. We estimate the noise energy in unvoiced intervals on the basis of segregated voiced speech. Spectral subtraction is employed to extract time-frequency segments in unvoiced intervals, and we group the segments dominated by unvoiced speech by simple thresholding or Bayesian classification. Systematic evaluation and comparison show that the proposed method considerably improves the unvoiced speech segregation performance under various SNR conditions.

Index Terms: unvoiced speech segregation, CASA, spectral subtraction

1. Introduction

Background noise interferes with target speech and poses a serious problem for many applications. Unvoiced speech is highly susceptible to interference due to relatively weak energy and lack of harmonic structure. While recent research in computational auditory scene analysis (CASA) has made considerable advances in voiced speech segregation, unvoiced speech segregation has been little studied (see [1] for an exception). In this paper, we propose a new approach to segregating unvoiced speech monaurally from nonspeech interference.

Speech enhancement methods have been proposed to enhance noisy speech based on a single recording [2]. Representative algorithms include spectral subtraction, Wiener filtering, minimum mean square error based estimator, and subspace analysis. Such methods work with the whole noisy utterance and therefore have the potential to deal with unvoiced speech. However, they often assume that interference satisfies certain statistical properties and lack the ability to deal with general interference. Another class of methods models source speakers and searches for the best combination to match the mixture and estimate source speech. For example, Radfar et al. [3] use a Gaussian mixture model to represent each source speaker and derive a minimum mean square error estimator to separate individual speech signals. Model-based techniques apply to unvoiced speech, but the assumption that the mixture consists of only speech utterances of trained speakers limits the scope of their applications. It is also unclear that how the system performs when two speakers utter unvoiced speech simultaneously.

On the other hand, CASA aims to achieve sound organization based on human auditory perceptual principles [4]. Segmentation and grouping are the two main stages of CASA. In segmentation, the input is decomposed to segments, each of which is a contiguous time-frequency (T-F) region originating mainly from a single sound source. The grouping stage combines segments that likely arise from the same source into a stream. Ideal binary mask (IBM) has been suggested as a main goal of CASA [5]. The IBM takes the value of 0 or 1 in each T-F unit, where 1 indicates target dominance and 0 interference dominance. Subject tests have shown that IBM-segregated mixtures lead to dramatic intelligibility improvements for both normal-hearing and hearing-impaired listeners [6], [7], [8].

Hu and Wang recently studied unvoiced speech segregation [1]. They utilize onset and offset cues to extract unvoiced speech segments. Acoustic-phonetic features are then used to segregate unvoiced speech in a classification stage. In [9], we incorporated spectral subtraction and noise type in unvoiced speech segregation. Although results are promising, the noise-type dependent grouping involves a large amount of training and the system is evaluated only at one SNR level.

In this work, we propose a simpler method for unvoiced speech grouping. Different from [9], our method first segregates voiced speech and removes it along with periodic portions of interference before segregating unvoiced speech. With periodic signal removal, the remaining interference becomes more stationary and we estimate interference energy in unvoiced intervals by averaging the mixture energy of masked units (those labeled as 0) in neighboring voiced intervals. Estimated noise energy is then used by spectral subtraction to generate T-F segments, and unvoiced speech is grouped based on thresholding or Bayesian classification.

The rest of the paper is organized as follows. We first describe voiced speech segregation in the next section. Unvoiced speech segregation is described in Section 3. Systematic evaluation and comparison are given in Section 4 and we conclude the paper in Section 5.

2. Voiced speech segregation

Our system is shown in Fig. 1. Noisy speech is first decomposed in the T-F domain using a 64-channel gammatone filterbank. Each filter response is then transduced by the Meddis hair cell model and divided to 20-ms time frames with 10-ms overlapping (see [5] for details of the peripheral analysis). Voiced speech segregation is carried out by a tandem algorithm [10] which segregates voiced speech

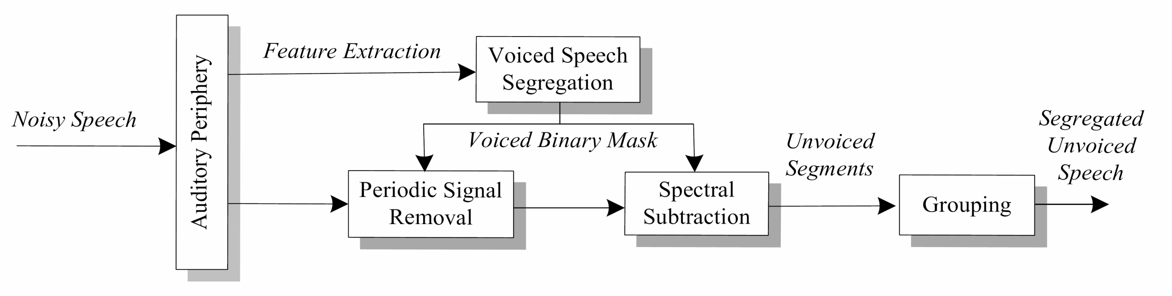

Figure 1: *Schematic diagram of the proposed unvoiced speech segregation system. The system first estimates voiced speech and removes it together with periodic portions of interference. Spectral subtraction is employed to generate T-F segments in unvoiced intervals and unvoiced speech segments are subsequently grouped.*

and detects pitch contours jointly and iteratively. The tandem algorithm requires a multilayer perceptron (MLP) for voiced speech segregation. Therefore, we extract 6-dimensional pitch-based feature vectors [10] to train an MLP for each frequency channel. The training corpus is created by mixing 100 utterances from the training part of the TIMIT database [11] with 100 nonspeech interferences [12] at 0 dB. In training, feature extraction requires pitch information so that we extract pitch contours from clean utterances using Praat [13]; IBM provides the desired output. All 64 MLPs have the same architecture of 6 input nodes, one hidden layer of 5 nodes and 1 output node.

3. Unvoiced speech segregation

3.1. Periodic signal removal

T-F units dominated by periodic signals cannot originate from unvoiced speech and should be removed. Let $u_{c,m}$ denote a T-F unit at channel c and frame m. We consider $u_{c,m}$ to be dominated by a periodic signal if the unit is included in the segregated voiced speech, or it has a high cross-channel correlation [5]. The cross-channel correlation is calculated on the basis of autocorrelation responses (correlogram), and deemed high if it is above a certain threshold

$$C(c,m) > \theta_R \text{ or } C_E(c,m) > \theta_E \qquad (1)$$

where $C(c,m)$ is the cross-channel correlation between filter responses in $u_{c,m}$ and $u_{c+1,m}$, and $C_E(c,m)$ denote that of the response envelopes. θ_R and θ_E represent their corresponding thresholds.

The above thresholds determine the balance between periodic signal removal and unvoiced speech preservation. To find appropriate thresholds, we vary θ_R and θ_E from 0.86 to 1, respectively, and calculate the loss of unvoiced speech during periodic signal removal. Specifically, unvoiced speech is taken as the unmasked portions of the IBM across non-pitched frames, and pitch contours are detected from clean utterances using Praat in this analysis. To exclude inharmonic voiced speech, segments extending below 1 kHz are excluded. We mix 100 sentences from the IEEE sentence database [14] recorded by a single female speaker with 15 nonspeech interferences (see Section 4 for specific interference types) to provide the analysis data. Given this systematic analysis, we set θ_R to 0.9 and θ_E to 0.96 to achieve a good compromise between periodic signal removal and unvoiced speech preservation. In this case, less than 2% of the unvoiced speech is lost on average.

3.2. Unvoiced speech segmentation based on spectral subtraction

After periodic signal removal, we deal with the mixtures of only unvoiced speech and aperiodic interference. Letting $X(c,m)$ be noisy speech energy and $\hat{N}(c,m)$ the estimated noise energy in $u_{c,m}$, we estimate the local SNR (in dB) in this unit as

$$\xi(c,m) = 10\log_{10}\left(\left[X(c,m) - \hat{N}(c,m)\right]^+ \big/ \hat{N}(c,m)\right) \qquad (2)$$

where the function $[z]^+ = z$ if $z \geq 0$ and $[z]^+ = 0$ otherwise. A T-F unit is then labeled as 1 if $\xi(c,m)$ is greater than 0 dB, or 0 otherwise. Interference energy $\hat{N}(c,m)$ in an unvoiced T-F unit is estimated by averaging the mixture energy (in the dB scale) of masked T-F units in two neighboring voiced intervals. For the unvoiced interval at the start or end of an utterance, estimation is only based on the succeeding or preceding voiced interval, respectively. If neighboring voiced intervals contain no masked unit, we continue to search the two further neighboring voiced intervals until at least one contains masked units. In case no masked unit is found in a channel, the mixture energy of the first 5 frames is averaged to obtain the noise estimate. We have also tried linear interpolation and smoothing spline interpolation in noise estimation, but got no better performance. On the other hand, we have investigated the over-subtraction technique to attenuate music noise [2]. We find an over-subtraction factor of 2 to be a good tradeoff and double the noise estimate in (2) during labeling. Unvoiced speech segments are subsequently formed by merging neighboring unmasked T-F units within unvoiced intervals.

The accuracy of the average-based noise estimation improves with periodic signal removal. To show this, we calculate the root mean square (RMS) error of estimated noise over unvoiced intervals with or without periodic signal removal. We mix 100 utterances of the IEEE corpus, which are different from those used in the previous subsection, with the bird chirp noise [12] at 0 dB for evaluation. Fig. 2 shows the mean RMS errors. The RMS error with periodic signal removal is uniformly smaller than that without the removal, especially at high frequencies where the energy of the bird chirp noise is concentrated. Two reasons are suggested. First, inharmonic voiced speech is eliminated during periodic signal removal, leading to more accurate noise estimation. Second, voiced harmonics in high frequencies are relatively weak and

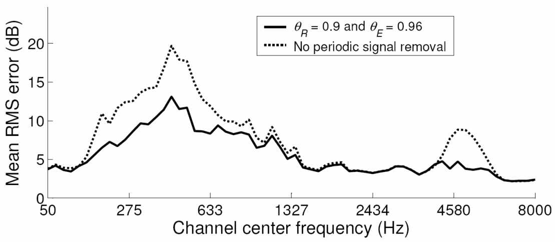

Figure 2: *Mean RMS errors of noise energy estimation with respect to frequency for bird chirp noise. The estimation performance with the chosen thresholds (solid line) is better than that without periodic signal removal (dotted line).*

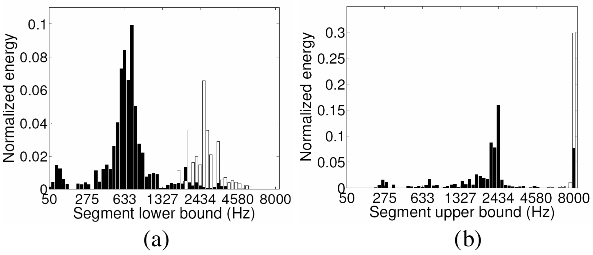

Figure 3: *Normalized energy distributions of unvoiced speech segments (white) and interference segments (black) over (a) segment lower bound and (b) segment upper bound. At each frequency, a lower bar is displayed in front of a higher bar.*

thus we have sufficiently many masked T-F units for noise estimation.

3.3. Unvoiced segment grouping

Spectral subtraction based segmentation captures most of unvoiced speech, but some segments correspond to residual noise. In this subsection, we propose a method to group only unvoiced speech segments. Unvoiced speech consists of unvoiced fricatives, stops and affricates. In speech production, the acoustic cavity of an unvoiced fricative is often small so that resonance concentrates at high frequencies [15]. For example, the alveolar fricative (/s/) often has a spectral peak around 4.5 kHz. An unvoiced stop is generated by forming a closure in the vocal tract and releasing it abruptly [15]. At the stop release multiple acoustic events happen, including a transient, a burst of frication noise, and aspiration noise. As a result, the energy of an unvoiced stop is often located in middle (1.5 kHz–3 kHz) and high frequency bands (3 kHz–8 kHz). The unvoiced affricate, /tʃ/, can be treated as a composite of a stop and a fricative. Therefore, the energy of unvoiced speech is often distributed at middle and high frequencies. This property, however, is not shared by noise residues, which often do not appear in high frequencies due to the relatively accurate noise estimation in that range.

To differentiate unvoiced speech and noise residues, we analyze their energy distributions with respect to frequency. Lower and upper frequency bounds of a segment are used to characterize its frequency span. We perform an ideal classification using 0-dB mixtures of 100 speech utterances and 15 interferences described in Section 3.1. A segment is classified as unvoiced speech if more than half of its energy is retained by the unvoiced IBM, which corresponds to the portions of the IBM with non-pitched frames. Fig. 3(a) shows the normalized energy distribution of segments with respect to their segment lower bounds and Fig. 3(b) the upper bounds. As we expected, unvoiced speech segments mainly reside at high frequencies while interference dominates at low frequencies. Based on this observation and acoustic-phonetic characteristics of unvoiced speech [15], we can simply perform grouping by thresholding: selecting segments with a lower bound higher than 2 kHz or an upper bound higher than 6 kHz as unvoiced speech and removing others as noise.

We can also formulate grouping as a hypothesis test and perform Bayesian classification. Let *S* denote the segment to be classified and two hypotheses be H_0: *S* is dominated by unvoiced speech, and H_1: *S* is dominated by interference. For classification, we construct 3 features for *S*

$$\mathbf{X}_S = (f_L^S, f_U^S, \|S\|) \qquad (3)$$

where f_L^S and f_U^S denote the frequency lower and upper bounds of *S*, respectively, and the third feature represents the size of segment S. We retain *S* as unvoiced speech if

$$P(H_0 \mid \mathbf{X}_S) > P(H_1 \mid \mathbf{X}_S)\,. \qquad (4)$$

We use the MLP in training and adopt the SNR-based objective function in [16] to maximize the output SNR. The same 0-dB mixtures described above are used for training. The MLP classification yields similar performance to simple thresholding. It is probably because the distributions of two types of segments are so well separated that the two thresholds we choose are already effective. We have also tried to incorporate the prior probability ratio or use the acoustic-phonetic features in [1] for classification but also obtained similar results.

4. Evaluation and comparisons

We evaluate the proposed algorithm using a noisy speech corpus composed of 100 utterances and 15 nonspeech interferences. The interference set comprises electric fan, white noise, crowd noise at a playground, crowd noise with clapping, crowd noise with music, rain, babble noise, rock music, wind, cocktail party noise, clock alarm, traffic noise, siren, bird chirp with water flowing, and telephone ring [12]. They are chosen to cover a wide variety of real-world noise types. The 100 test sentences are randomly selected from those of the IEEE sentences which are not used for analysis or training before. All utterances are downsampled from 20 kHz to 16 kHz and each is mixed with an individual interference at the SNR levels of –5, 0, 5, 10, and 15 dB. In training, the first half of an interference is mixed with speech, while in testing the second half is used.

The computational goal of our system is to estimate the unvoiced IBM. Therefore, we adopt the SNR measure in [1] and consider the resynthesized speech from the unvoiced IBM as the ground truth

$$\text{SNR} = 10\log_{10}\left(\sum_n S_I^2[n] \Big/ \sum_n (S_I[n] - S_E[n])^2\right), \qquad (5)$$

where $S_I[n]$ and $S_E[n]$ are the signals resynthesized using the unvoiced IBM and estimated unvoiced IBM, respectively. In estimation, pitch contours are estimated using the tandem algorithm.

We compare our method with the unvoiced speech

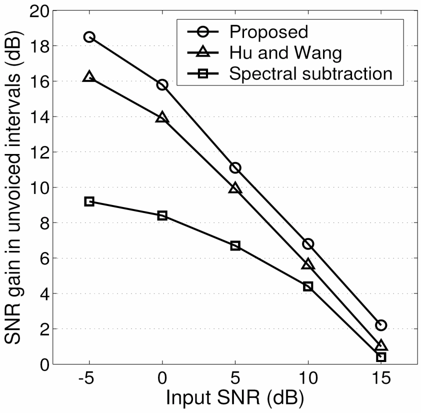

Figure 4: *SNR performance of three unvoiced speech segregation algorithms. The ordinate denotes the SNR gain, which is computed from the output SNR of segregated speech subtracted by the initial SNR of the mixture over unvoiced intervals.*

segregation system proposed by Hu and Wang [1], the only previous system directly dealing with unvoiced speech segregation to our knowledge. In particular, we retrain their acoustic-phonetic feature based MLP classifier using the 1500 mixtures described in Section 3.1. Fig. 4 shows the comparative results. Our algorithm performs better than their system with an average of 1.6 dB SNR improvement over all input SNR levels. In terms of computational complexity, our spectral subtraction based segmentation is more efficient than their method employing multiscale onset-offset analysis which needs to analyze the signal in different scales. In grouping, the proposed thresholding method is computationally much simpler, requiring no MLP training in segment removal and classification.

Since spectral subtraction plays a major role in the segmentation stage of our system, it is informative to evaluate its performance alone. For this evaluation, noise is estimated as described in Section 3.2 but without periodic signal removal. As in our method, an over-subtraction factor of 2 is used and portions of the estimated unvoiced mask below 1 kHz are removed to evaluate unvoiced speech segregation. The performance of the spectral subtraction algorithm is shown in Fig. 4. As can be seen in the figure, the proposed algorithm performs much better than spectral subtraction. The largest gap is about 9.3 dB when the input SNR is -5 dB and the gap is about 1.8 dB as the input SNR increases to 15 dB. We note that the performance without over-subtraction is even worse. We have also evaluated the performance of spectral subtraction directly, i.e. without binary masking, and obtained similar results. It is worth mentioning that large gains at low input SNR levels are particularly useful for people with hearing loss [17]. Here the need to improve SNR in these conditions is more acute than at high input SNRs.

The proposed method also has advantages over our previous system in [9]. First, the proposed method is computationally simpler since it does not use either noise type detection or noise-type dependent grouping. Second, the proposed system has been demonstrated to be effective under various SNR conditions, especially at low SNRs.

5. Conclusions

Unvoiced speech separation is a challenging task. Our proposed CASA system first removes periodic signals and subsequently estimates interference energy on the basis of segregated voiced speech. Spectral subtraction is employed to extract T-F segments, and unvoiced speech is grouped by thresholding or classification. Systematic comparisons show that the proposed system outperforms a recent system over a range of input SNR levels and performs substantially better than spectral subtraction.

6. Acknowledgements

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

7. References

[1] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[3] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft masking filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, 2007.

[4] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT press, 1990.

[5] D. L. Wang and G. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*. Hoboken NJ: Wiley & IEEE Press, 2006.

[6] 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," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.

[7] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, pp. 1673–1682, 2008.

[8] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.

[9] K. Hu and D. L. Wang, "Incorporating spectral subtraction and noise type for unvoiced speech segregation," in *Proc. IEEE ICASSP*, 2009, pp. 4425–4428.

[10] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, in press, 2010.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. (1993) DARPA TIMIT acoustic phonetic continuous speech corpus. [Online]. Available: http://-www.ldc.upenn.edu/Catalog/LDC93S1.html

[12] G. Hu. (2006) 100 nonspeech sounds. [Online]. Available: http://-www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

[13] P. Boersma and D. Weenink. (2007) Praat: doing phonetics by computer (version 5.0.02). [Online]. Available: Online: http://-www.fon.hum.uva.nl/praat

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

[15] K. N. Stevens, *Acoustic Phonetics*. Cambridge MA: MIT Press, 1998.

[16] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 625–638, 2009.

[17] H. Dillon, *Hearing Aids*. New York, NY: Thieme Medical Publishers, 2001.