AN APPROACH TO SEQUENTIAL GROUPING IN COCHANNEL SPEECH

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

Model-based methods for sequential organization in cochannel speech require pretrained speaker models and often prior knowledge of participating speakers. We propose an unsupervised approach to sequential organization of cochannel speech. Based on cepstral features, we first cluster voiced speech into two speaker groups by maximizing the ratio of between- and within-group distances penalized by within-group concurrent pitches. To group unvoiced speech, we employ an onset/offset based analysis to generate time-frequency segments. Unvoiced segments are then labeled by the complementary portions of segregated voiced speech. Our method does not require any pretrained model and is computationally simple. Evaluations and comparisons show that the proposed method outperforms a model-based method in terms of speech segregation.

Index Terms— sequential grouping, clustering, unvoiced speech, cochannel speech separation

1. INTRODUCTION

In everyday listening, speech reaching our ears is often corrupted by various sounds, such as machine noise, music or another voice. Cochannel speech refers to the mixture of two speech signals transmitted simultaneously in a single channel. Under cochannel conditions, two talkers are usually not aware of each other and the resultant speech mixture often has a large amount of overlap. Such a condition poses a big challenge to speech separation and recognition.

Previous studies on cochannel speech separation employ model-based methods. For example, Shao and Wang extend the framework of single speaker identification to the two-talker case and group speech components by maximizing the joint speaker recognition score [1]. Similarly, hidden Markov models are employed to model speakers in [2] and speech is separated by coupling segregation and recognition. Related model-based methods directly estimate individual speech signals [3, 4]. Model-based methods can achieve satisfactory performance when trained models match those of participating speakers. However, this condition is often not met in practice.

We aim to separate cochannel speech in an unsupervised way that requires no prior speaker knowledge. Based on computational auditory scene analysis [5], we first decompose the speech mixture into time-frequency (T-F) segments, which are further grouped across frequency to form simultaneous streams. Each simultaneous stream is mainly dominated by a single speaker and continuous in time. Different simultaneous streams are generally separated in time and how to group them into individual speakers is the task of sequential grouping [5].

In speaker diarization, unsupervised speaker clustering aims to organize homogeneous speech sections into different speaker groups [6]. Sequential grouping resembles speaker clustering but has two unique challenges. First, simultaneous streams consist of spectrally separated components while speech sections in speaker clustering contain spectrally complete frames. Second, a simultaneous stream is much shorter than a speech section in speaker clustering. As pointed out in an analysis based on intra- and inter-speaker distances in [7], a minimum of 5 phones is needed for speaker separability. Therefore, short simultaneous streams generally do not contain enough acoustic information for direct speaker clustering. To verify this, we have directly applied speaker clustering methods for sequential grouping but found unsatisfactory results.

In [8], we proposed a clustering-based method for unsupervised sequential organization of voiced speech. However, feature reconstruction in this method requires a speech prior and unvoiced speech is not handled. In this work, we introduce a cepstral feature generated directly from the mixture based on T-F masking. On the other hand, unvoiced speech poses a big challenge for sequential grouping due to its weak energy and lack of harmonicity. We propose to group unvoiced speech based on segregated voiced speech. First, we employ an onset/offset analysis to extract T-F segments from the whole speech mixture. Then, the portions of segments overlapping with segregated voiced speech are removed, and the remaining portions are re-segmented to produce unvoiced speech segments. For each unvoiced segment, we calculate its overlap with the complementary portions of the segregated voiced speech of each speaker. We then label each unvoiced segment by comparing overlap patterns. Without using any prior models, our method is completely unsupervised and applies to both voiced speech and unvoiced speech.

We briefly describe early processing and feature extraction in Section 2. The sequential organizations of voiced speech and unvoiced speech are detailed in Section 3 and 4, respectively. Evaluation and comparison are given in Section 5, and we conclude the paper in Section 6.

2. EARLY PROCESSING AND FEATURE EXTRACTION

Cochannel speech is first decomposed in the frequency domain using a bank of 128 gammatone filters [5], whose center frequencies are equally spaced in equivalent rectangular bandwidth from 50 Hz to 8000 Hz. Then, the output of each channel is downsampled to 100 Hz in time and compressed by a cubic root operation to obtain the gammatone features (GF) [9].

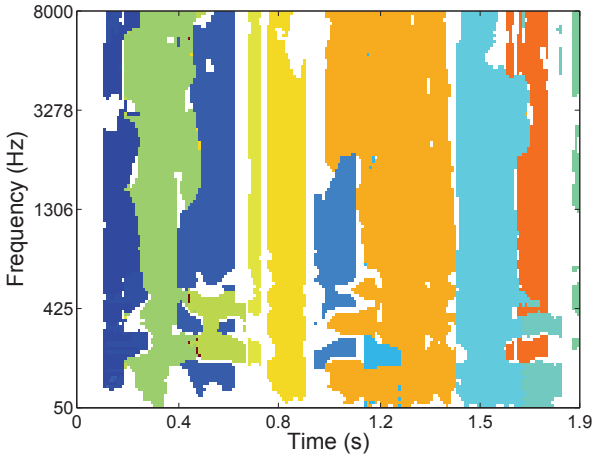

Fig. 1. An example of simultaneous streams generated using the tandem algorithm.

We perform simultaneous grouping using a recently developed tandem algorithm [10]. The algorithm starts by estimating two initial masks based on harmonicity and temporal continuity. Given the initial estimates, pitch contours and simultaneous streams are re-estimated by expanding current pitch contour estimates. The resulting simultaneous streams are represented by binary masks, which are estimates of the ideal binary mask (IBM) [5]. In the IBM, 1 indicates an unmasked T-F unit and 0 a masked one. An example of estimated simultaneous streams is shown in Fig. 1 for a cochannel speech mixture. Each simultaneous stream is indicated by a distinct color.

To group simultaneous streams, we extract masked gammatone frequency cepstral coefficients (GFCC). First, the binary mask associated with each simultaneous stream is used to filter the noisy GFs. Unmasked T-F units are retained while masked T-F units are zeroed out. Then, for each frame, the filtered GF is converted to the GFCC using the discrete cosine transform. Compared to the enhanced GFCC based on a reconstruction method in [8], the masked GFCC does not need any speech prior and can be calculated more efficiently.

3. SEQUENTIAL ORGANIZATION OF VOICED SPEECH

We formulate sequential organization of voiced speech as a problem of unsupervised clustering: voiced simultaneous streams are clustered into two speaker groups. Our clustering objective function is based on the ratio of between- and within-group distances

$$O(\mathbf{g}) = \mathrm{tr}(\mathbf{S}_W^{-1}(\mathbf{g})\mathbf{S}_B(\mathbf{g})) \tag{1}$$

where $\mathbf{g}$ is a hypothesized binary label vector for all voiced simultaneous streams, and $\mathbf{S}_W(\mathbf{g})$ and $\mathbf{S}_B(\mathbf{g})$ are within- and between-group scatter matrices, respectively. Specifically, according to $\mathbf{g}$, simultaneous streams are divided into two groups and we pool the masked GFCC features in individual groups to calculate $\mathbf{S}_W(\mathbf{g})$ and $\mathbf{S}_B(\mathbf{g})$. Definitions of both matrices can be found in [11]. The trace of $\mathbf{S}_W^{-1}(\mathbf{g})\mathbf{S}_B(\mathbf{g})$ is used to measure the group distance, which can be interpreted as the ratio of the between- and within-group scatter matrices along the eigenvector dimensions [11].

When maximizing (1), two simultaneous streams with overlapping pitch contours should not be assigned to the same speaker. We thus penalize any $\mathbf{g}$ with m within-group overlapping pitch frames by

$$P(\mathbf{g}) = 1/(1 + e^{a(m_{\mathbf{g}} - b)}), \quad a < 0 \text{ and } b \geq 0 \tag{2}$$

where $m_{\mathbf{g}}$ denotes the total length of within-group overlapping pitch frames with respect to $\mathbf{g}$, and a and b are constants controlling the steepness of the penalty and tolerance to overlapping errors, respectively. Notice that a is negative; $P(\mathbf{g})$ will saturate to 1 as $m_{\mathbf{g}}$ increases and to zero when $m_{\mathbf{g}}$ is significantly smaller than b.

Combining (1) and (2), the objective function becomes

$$J(\mathbf{g}) = \lambda O(\mathbf{g}) - (1 - \lambda)cP(\mathbf{g}), \quad 0 \leq \lambda \leq 1 \tag{3}$$

where c is a constant which scales $P(\mathbf{g})$ to the range of $O(\mathbf{g})$, and λ controls the tradeoff between these two terms. We set c to be $\max_{\mathbf{g}} O(\mathbf{g})$, and empirically λ needs to be greater or equal than 0.5 to achieve good results.

Given the objective function, the clustering can be formulated as an optimization problem, and the optimal grouping can be found by maximizing (3) based on an exhaustive search. In [8], we show that a genetic algorithm (GA) approximates the optimal solution satisfactorily and is computationally more efficient. Thus, we use the GA search to maximize $J(\mathbf{g})$ in this work. In GA, each chromosome corresponds to a binary label vector $\mathbf{g}$. The fitness function for evaluating partitions is based on the objective function in (3). The chromosome with the highest fitness score in the final population is taken as the GA solution. In particular, the initial population size, number of generations, and the crossover probability are set to be 500, 50, and 0.8, respectively. Details about the GA-based search can be found in [8].

Besides search-based grouping, we have also considered clustering simultaneous streams iteratively using a Gaussian mixture model based likelihood function [12] but obtained worse results. It is probably because a single simultaneous stream does not contain sufficient speaker information. A sum-of-squared-error objective function is not chosen due to its sensitivity to outliers [11].

4. SEQUENTIAL ORGANIZATION OF UNVOICED SPEECH

Unvoiced speech constitutes more than 20% of spoken English in terms of both phoneme occurrences and time durations [13]. Without any speaker models, the unsupervised sequential grouping of unvoiced speech is extremely challenging due to the relatively weak energy and lack of harmonicity of unvoiced speech. In this work, we propose to use segregated voiced streams to provide labeling information for unvoiced speech grouping.

Before grouping, we extract unvoiced speech segments using a multiscale onset/offset analysis [13]. Onsets and offsets correspond to sudden energy changes and thus reflect boundaries of auditory events. The method in [13] first detects onset/offset points in each frequency channel and then connects them across frequency to form onset/offset fronts. Segments are then generated by matching the onset/offset fronts and integrating the results in multiple scales. Since segmentation is based on energy changes, output segments correspond to both voiced and unvoiced speech. To retain only unvoiced segments, we remove the portions of output segments overlapping with segregated voiced speech. Specifically, any T-F unit within an onset/offset based segment and also included in segregated voiced speech is removed. The remaining parts are re-segmented to produce unvoiced segments. Fig. 2 shows unvoiced segments obtained from the input signal used in Fig. 1.

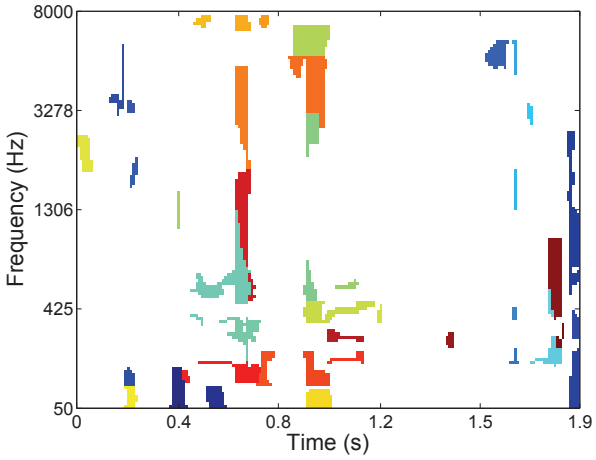

Fig. 2. Segments produced by onset/offset based segmentation followed by voiced signal removal.

We utilize the complementary mask of segregated voiced speech to label unvoiced segments. For two speakers a and b, we denote their segregated voiced streams as V_a and V_b, respectively. Then, we generate a complementary mask C_a for speaker a by flipping all binary values in V_a, and a complementary mask C_b for speaker b similarly. Frames in cochannel speech are of three kinds: two-pitch frames, one-pitch frames and pitchless frames. Among them, pitchless frames contain no unmasked units in either V_a or V_b and thus their corresponding frames in C_a and C_b contain no labeling information. These frames are thus removed in both C_a and C_b. In two-pitch frames, both speakers utter voiced speech and no unvoiced speech exists. Therefore, such frames are also removed from C_a and C_b. Each of the remaining complementary masks contains only single-pitch frames, where unmasked T-F units may include the unvoiced speech of the other speaker. For each unvoiced segment, we calculate its overlapping energy with C_a and C_b to yield E_a and E_b, respectively. When $E_a > 0$ or $E_b > 0$, the segment is assigned to speaker a if $E_b > E_a$ and to speaker b otherwise.

We note that the above method deals with only unvoiced-voiced (or voiced-unvoiced) portions of the mixture but not unvoiced-unvoiced portions (i.e., when $E_a = E_b = 0$). According to [13], unvoiced speech accounts for about 25% of spoken English in time duration so, in principle, unvoiced-unvoiced portions should account for a very small percentage (about $6-7\%$) of total mixture frames. We have analyzed the percentage of unvoiced-unvoiced portions based on 100 cochannel speech mixtures generated using the speech separation challenge (SSC) corpus [14]. Based on estimated unvoiced segments, unvoiced-unvoiced portions constitute about 10% of total unvoiced portions in energy. To separate unvoiced-unvoiced portions is a future research topic.

5. EVALUATION AND COMPARISON

As in [15], we evaluate our algorithm by measuring the target speaker segregation performance. We take the resynthesized speech from the overall IBM as the ground truth and measure the SNR of segregated target speech as

$$\mathrm{SNR} = 10\log_{10}(\sum_n S_I^2[n] / \sum_n (S_I[n] - S_E[n])^2), \qquad (4)$$

where $S_I[n]$ and $S_E[n]$ are the target signals resynthesized from the IBM and an estimated IBM, respectively. We also evaluate the system using ideal simultaneous streams derived from ground-truth pitch contours and IBMs. Here, ground-truth pitch contours are detected from premixed utterances using Praat [16] and the corresponding portions of the IBM are taken as ideal simultaneous streams. Since our algorithm is unsupervised, we treat the segregated signal matching $S_I[n]$ better as the estimated target.

We create cochannel speech mixtures using the SSC corpus [14], which contains 34 speakers of both males and females. All utterances are first downsampled from 25 kHz to 20 kHz. For each utterance deemed a target, another utterance is randomly selected from a different speaker and mixed with the target. The interfering utterance is either cut or concatenated with itself to match the length of the target signal. In total, we have created 100 mixtures at 0 dB using the test part of the corpus for evaluation, among which 49 are different gender (DG) mixtures, and 51 are same gender (SG) mixtures. For the penalty term in (2), a and b are set to -10 and 0.5, respectively, for ideal simultaneous streams. However, since the tandem algorithm may overdetect pitches for a single speaker, we set a and b to -0.3 and 15, respectively, to tolerate such errors.

We compare our method to the background model (BM) based method of [15] since both algorithms operate on simultaneous streams for sequential organization of voiced speech. In separation, the BM method needs to know the target speaker model while our method is completely speaker independent. For sequential organization of unvoiced speech, we compare with a model-based method in [17], which uses a detected speaker pair from voiced speech sequential grouping to group unvoiced speech.

The segregation results are shown in Table 1, where the "BM" row shows the performance using the background model and the "Proposed" row describes that of our method. First, sequential grouping performance based on estimated and ideal voiced simultaneous streams are presented in "ESS" and "ISS" columns, respectively. Compared with the BM method, our algorithm improves the average segregation performance by 0.4 dB in the "ESS" case and 3 dB in the "ISS" case. The larger improvement in the "ISS" case indicates that our method benefits more from improved simultaneous streams. Compared to the improvements in [8], the results also indicate that masked GFCC features performs comparably with reconstructed features. Then, we present the results including unvoiced speech segregation in columns "ESS+Unvoiced" and "ISS+Unvoiced". By comparing the "ESS" and "ESS+Unvoiced" columns (and also "ISS" and "ISS+Unvoiced" columns), we obtain the improvements due to unvoiced speech grouping: 0.2 dB and 0.5 dB in the estimated case for the BM method and our method, respectively, and 0.4 dB and 3.7 dB in the ideal case, respectively. Compared to the BM method, the proposed method gains larger improvements due to grouping of unvoiced segments. The significant improvement in the "ISS+Unvoiced" case is probably because the complementary-mask based method benefits more from improved segregating of voiced speech. Altogether, in the "ESS+Unvoiced" case, our approach outperforms the BM method by 0.7 dB, and the improvement becomes substantial in the "ISS+Unvoiced" case: 6.3 dB. Note that our method performs comparably or better for both SG and DG mixtures. To establish a performance upper bound, we also perform ideal sequential grouping (ISG). In ISG, a simultaneous stream (or segment) is grouped as target if more than half of its energy is retained by the IBM. The results are shown in the "ISG" row in Table 1.

To isolate the performance of unvoiced speech grouping, we evaluate the SNR gains in unvoiced frames for both methods, where

Table 1. Comparisons of output SNRs (in dB) between the proposed method and the model-based method

Speech type	ESS			ISS			ESS+UNVOICED			ISS+UNVOICED		
	SG	DG	Both	SG	DG	Both	SG	DG	Both	SG	DG	Both
BM	3.7	6.0	4.8	8.9	11.8	10.3	3.9	6.2	5.0	9.2	12.3	10.7
Proposed	3.7	6.8	5.2	11.4	15.2	13.3	3.9	7.5	5.7	15.1	19.0	17.0
ISG	5.7	8.0	6.9	14.4	15.7	15.0	7.9	10.6	9.2	22.9	23.1	23.0

Table 2. Comparisons of SNR gains (in dB) in unvoiced intervals between the proposed method and the model-based method

Speech type	ESS+UNVOICED			ISS+UNVOICED		
	SG	DG	Both	SG	DG	Both
BM	6.0	10.0	7.9	7.7	10.5	9.1
Proposed	6.0	9.7	7.8	12.1	14.0	13.0
ISG	11.6	15.4	13.5	18.0	18.1	18.0

unvoiced intervals are determined by ground-truth pitch contours. The results are presented in Table 2. As shown in the table, on average, our performance is 0.1 dB lower than the BM method in the estimated case. Although no improvement is obtained in this case, we emphasize that our method is completely speaker independent while the method in [17] still needs target speaker information. When using ideal voiced simultaneous streams, our method outperforms the BM method by 3.9 dB. In addition, although our method does not improve unvoiced SNR in the estimated case, it does improve the overall SNR as shown in Table 1. This is because onset/offset segments also contribute positively to voiced speech segregation.

6. CONCLUSION

We have proposed a novel unsupervised method for sequential organization of both voiced and unvoiced speech in cochannel conditions. Our method groups voiced speech by clustering using masked GFCC features. Unvoiced speech is segmented by an onset/offset based analysis and then grouped using the complementary portions of segregated voiced speech. Our method does not need any pretrained models and is computationally efficient. Systematic evaluations and comparisons show that our method outperforms a model-based method with both ideal and estimated simultaneous streams.

7. ACKNOWLEDGEMENTS

This research was supported by an AFOSR grant (FA9550-08-1-0155) and the VA Biomedical Laboratory Research and Development Program.

8. REFERENCES

[1] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 289–298, 2006.

[2] J. Barker, A. Coy, N. Ma, and M. Cooke, "Recent advances in speech fragment decoding techniques," in *Proc. Interspeech*, 2006, pp. 85–88.

[3] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, 2007.

[4] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 16–29, 2010.

[5] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: Wiley-IEEE Press, 2006.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] U. O. Ofoegbu, A. N. Iyer, R. E. Yantorno, and S. Wenndt, "Unsupervised indexing of conversations with short speaker utterances," in *Proc. IEEE Aerospace Conf.*, 2006, pp. 1–11.

[8] K. Hu and D. L. Wang, "Unsupervised sequential organization for cochannel speech separation," in *Proc. Interspeech-10*, 2010, pp. 2790–2793.

[9] Y. Shao, *Sequential organization in computational auditory scene analysis*, Ph.D. thesis, Dept. of Comput. Sci. & Eng., The Ohio State Univ., 2007.

[10] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 2067–2079, 2010.

[11] R. Xu and D. C. Wunsch, *Clustering*, Hoboken NJ: Wiley & IEEE Press, 2009.

[12] B. Narayanaswamy, R. Gangadharaiah, and R. M. Stern, "Voting for two speaker segmentation," in *Proc. Interspeech*, 2006, pp. 2086–2089.

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

[14] M. Cooke and T. Lee, "Speech separation challenge website," Online: http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm, 2006.

[15] Y. Shao and D. L. Wang, "Sequential organization of speech in computational auditory scene analysis," *Speech Comm.*, vol. 51, pp. 657–667, 2009.

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

[17] Y. Shao, S. Srinivasan, Z. Jin, and D. L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Comput. Speech Lang.*, vol. 24, pp. 77–93, 2010.