Cochannel Speech Separation Using Multi-pitch Estimation and Model Based Voiced Sequential Grouping

Ming Li, Chuan Cao, Di Wang, Ping Lu, Qiang Fu, and Yonghong Yan

ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
{ming.li, ccao, dwang, plu, qfu, yyan}@hccl.ioa.ac.cn

Abstract

In this paper, a new cochannel speech separation algorithm using multi-pitch extraction and speaker model based sequential grouping is proposed. After auditory segmentation based on onset and offset analysis, robust multi-pitch estimation algorithm is performed on each segment and the corresponding voiced portions are segregated. Then speaker pair model based on support vector machine (SVM) is employed to determine the optimal sequential grouping alignments and group the speaker homogeneous segments into pure speaker streams. Systematic evaluation on the speech separation challenge database shows significant improvement over the baseline performance.

Index Terms: Auditory scene analysis, cochannel speech, multi-pitch estimation, sequential grouping

1. Introduction

Cochannel speech is termed as a combination of speech utterances from two talkers, usually produced when two speech signals are transmitted over a single communication channel. Speech separation algorithms can be used to mitigate the effect of interference on automatic speech recognition and voice communication. Bregman [1] contends that auditory scene analysis (ASA) performed by listeners can be conceptualized as a two-stage process: segmentation and grouping. Grouping is composed of simultaneous and sequential organization. The former involves grouping of segments at a particular time and the latter integrates components across time into the same perceptual stream. Most of the existing computational auditory scene analysis (CASA) systems involve both simultaneous organization and sequential organization [2, 3].

Recently, promising results have been reported on cochannel speaker identification (SID) algorithms using usable speech [2], binary time-frequency masks, and auditory feature uncertainties [4]. In Shao's model based sequential grouping method [2], usable speech, defined as consecutive frames of speech minimally corrupted by interfering speech (single-speaker speech) [2], is adopted for SID and then aligned into two different speaker streams. However, corrupted speech with large portions overlapped by interfering speech is not processed. In the proposed method, not only usable speech but also corrupted speech can be handled in the sequential grouping stage. By utilizing segment based multi-pitch extraction and voiced simultaneous grouping in CASA system, corrupted speech is decomposed into usable speaker homogeneous speech segments which are grouped into different streams by speaker models.

In the paper, we focus on how to use prior speaker pair identity information generated by SID methods to perform sequential organization in voiced cochannel speech. Firstly the input speech mixture is segmented by a multi-scale onset and offset analysis. Segmented multiple pitch sequences are obtained by a frame-level multi-pitch estimation method, followed by a intra-segment pitch tracking method using not only pitch value similarity but also energy dominance assumption. While in sequential grouping stage, binary generalized linear discriminative sequence (GLDS) kernel based support vector machine is adopted to discriminate auditory segments and search for an optimal hypothesis. Specifically, missing data training algorithm is employed to improve SID performance and reduce the amount of training data. Finally, the output estimate of ideal binary mask is used to reconstruct target speaker's speech signal.

The GLDS kernel is based on an explicit expansion in feature space which gives a very concise way of storing speaker pair model and makes scoring function just an inner product which can dramatically reduce the computational complexity of hypothesis scoring in sequential organization.

2. System Framework

In the proposed framework demonstrated in Fig.1, after the auditory segmentation based on onset and offset analysis, a robust multipitch tracking algorithm is performed and used to segregate the corresponding voiced portions in each segment. Besides, SVM based speaker pair models are used to group the speaker homogeneous segments across time to produce binary time-frequency (T-F) masks.

3. Speech Segmentation and Multi-pitch Estimation

Most previous works on cochannel speech separation [5, 2] are mainly based on the target pitch information, so pitch accuracy is the key factor for the separation systems' performance. Since the existence of overlapped speech parts, multi-pitch estimation

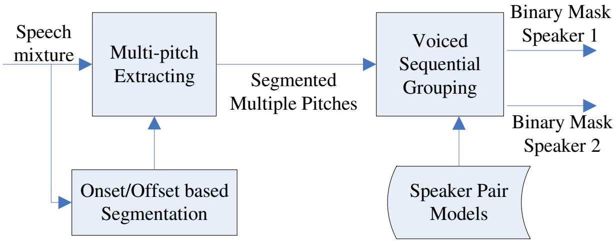


Figure 1: System Framework.

This work is supported by The National High Technology Research and Development Program of China (863program 2006AA010102 2006AA01Z195), MOST (973program 2004CB318106), National Natural Science Foundation of China (10574140 60535030).

methods are necessary to obtain target pitch sequences. Usually, multiple pitch candidates are estimated frame by frame and then tracked into several pitch sequences. Previous multi-pitch tracking methods for speech utterances such as HMM decoding method [6] are simply based on pitch value similarity constrains, not taking other information into account. However, those kind of methods have vital defects that intersected pitch sequences can not be handled and the inevitable false active pitch candidates are treated the same as other true candidates. So a new pitch tracking method is proposed in this paper, which utilizes not only pitch value information but also energy information as well. If we know priorly that one talker's voice is louder than the other, the energy dominance information could be used in the pitch tracking procedure, along with other tracking constrains. However, energy dominance of cochannel speech utterances is not known priorly and inconstant likely. But dominance relationship during a small segment could be considered to stay invariable. Therefore, a segmentation approach which could divide cochannel speech into small "stable" pieces is necessary.

3.1. Speech Segmentation

Energy dominance of a cochannel speech segment could be considered stable unless a new sound source get in or an old sound source fade off. And the auditory onset and offset events are oriented to represent the start and end of a sound source respectively. Therefore, an onset-offset based segmentation method can be utilized to divide the cochannel speech utterance into small "stable" segments we need. The segmentation algorithm in this paper is generally based on Hu's auditory multi-scale analysis [7] and our implementation just simply segments the utterance on the temporal axis.

3.2. Multi-pitch Estimation

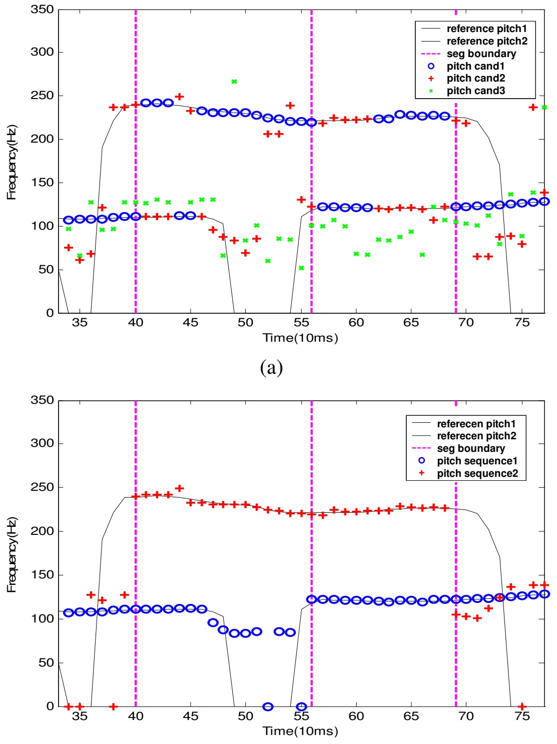
Method described in this section is for intra-segment analysis. Firstly, multiple pitch candidates are obtained by a frame-level multi-pitch estimation method. Then dynamic programming (DP) technique is used to group these candidates into pitch sequences within the segment.

3.2.1. Pitch Candidates Estimation

The frame-level multi-pitch estimation method in this paper is generally based on our previous work in [8], which is grounded on the subharmonic summation (SHS) method and a spectral cancelation framework. Three fundamental frequencies (F_0) candidates are extracted iteratively for every frame and added into the F_0 candidates pool with their saliency, which can be simply considered as the energy of the harmonic family corresponding to the pitch hypothesis.

3.2.2. Intra-segment Pitch Grouping

Based on the F_0 candidates pool, we group the F_0 hypotheses into two independent pitch sequences. Though energy dominance with-in a segment is relatively stable, we still cannot guarantee the speech energy of one talker is larger than the other on every frame, so a frame-level DP algorithm is used for grouping, aiming to track out a pitch sequence with high saliency and slow pitch change rate. Every F_0 in the candidates pool is seemed as a DP node with its pitch value and saliency value and path scoring is defined relevant to the saliency values of the connected nodes and the pitch value change rate between them (F_0 candidates within a frame cannot be connected). Firstly, we track out the pitch sequence with the path of highest score (noted as P^1), and then weaken the saliency value of all the nodes on this path. Secondly we calculate again all the path scores and track out another different pitch sequence (noted as P^2). An DP pitch tracking example can be seen in Fig.2.



(a)

(b)

Figure 2: An example of multi-pitch estimation and intra-segment pitch grouping for file 't14_bwir7p_m4_pgwy6s' in the challenge two talker database [9]. Figure 2(a) shows pitch candidates for every frame (3 candidates per frame in different line styles). Figure 2(b) shows the pitch sequences after the intra-segment grouping process. The reference pitches are extracted by SHS method under clean condition.

4. Voiced Sequential Grouping

Speaker homogeneous voiced segments are generated by CASA simultaneous grouping [5] using segmented multiple pitches, and unvoiced speech is not processed, therefore the voiced simultaneous segments should be composed of T-F units from the same speaker [3]. Thus sequential grouping is used to organize these voiced segments into corresponding speaker streams.

4.1. Sequential Grouping in Cochannel Speech

The goal of voiced sequential grouping is to find two complimentary pitch sequence alignments by which the two reconstructed speaker streams maximize the posterior probability on the specific speaker pair model. For a cochannel mixture, our segment based multi-pitch extraction method extracts 2 pitch sequence, each of which consists of N pitch segments: $\mathbf{x}^1 = \{P_1^1, \ldots, P_i^1, \ldots, P_N^1\}$ and $\mathbf{x}^2 = \{P_1^2, \ldots, P_i^2, \ldots, P_N^2\}$. Our motivation is to mutually exchange two pitch sequence in each corresponding segment to form new pitch sequence pool, $(\mathbf{y}^1, \mathbf{y}^2 \in \mathbf{Y})$ and then search for the optimal pitch segment alignment that maximize the posterior probability given the speaker pair model λ.

$$\hat{\mathbf{y}}^1, \hat{\mathbf{y}}^2 = \underset{\mathbf{y}^1, \mathbf{y}^2 \in \mathbf{Y}}{\operatorname{argmax}} P(\mathbf{y}^1, \mathbf{y}^2 \mid \lambda) \tag{1}$$

Assuming independence of each segment and the segments with different labels come from different speakers, then the probability of the entire two sequence can be denoted as follows:

$$P(\mathbf{y}^1, \mathbf{y}^2 \mid \lambda) = \prod_{i=1}^{N} P(y_i^1 \mid \lambda = 1) \cdot \prod_{i=1}^{N} P(y_i^2 \mid \lambda = 2) \tag{2}$$

Without loss of generality, for each pitch segment alignment $\{y_i\}$, the probability can be modified as:

$$\prod_{i=1}^{N} P(y_i \mid \lambda) = \prod_{i=1}^{N} \frac{P(\lambda \mid y_i) P(y_i)}{P(\lambda)} \tag{3}$$

Without prior knowledge of segments alignment and for the purpose of discriminative classification, we can discard $P(y_i^1)$. Furthermore, we take the logarithm of both sides and use two terms of the Taylor series $log(x) \approx x - 1$ to obtain the discriminative object function:

$$D(\{y_i\} \mid \lambda) = \frac{1}{N} \sum_{i=1}^{N} \frac{P(\lambda \mid y_i)}{P(\lambda)} \tag{4}$$

Note that the term -1 in Taylor series is discarded and the discriminative object function is normalized by the number of pitch segments because these modifications will not affect the discriminative decision. Then, assuming prior probabilities of two speakers in cochannel speech are the same[2]. Thus, to consider two speakers in cochannel speech, the discriminative object function corresponding to (2) is denoted as follows:

$$D(\mathbf{y}^1, \mathbf{y}^2 \mid \lambda) = \frac{1}{N} \sum_{i=1}^{N} P(\lambda = 1 \mid y_i^1) + \frac{1}{N} \sum_{i=1}^{N} P(\lambda = 2 \mid y_i^2) \tag{5}$$

4.2. Support Vector Machine with GLDS Kernel

An SVM is a two-class classifier constructed from sums of a kernel function $K(\cdot, \cdot)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d \tag{6}$$

The original form of the GLDS kernel [10] involves a polynomial expansion $b(x)$, with monomials (between each combination of vector components) up to a given degree p. The GLDS kernel between two sequences of vectors $X = \{x_t\}_{t=1 \cdots N_x}$ and $Y = \{y_t\}_{t=1 \cdots N_y}$ is denoted as a rescaled dot product between average expansions:

$$K(X, Y) = \frac{1}{N_x} \sum_{i=1}^{N_x} b(x_i)^t \cdot \overline{R}^{-1} \cdot \frac{1}{N_y} \sum_{j=1}^{N_y} b(y_j) = \overline{b_x}^t \cdot \overline{R}^{-1} \cdot \overline{b_y} \tag{7}$$

where $\overline{R}$ is the second moment matrix of polynomial expansions and its diagonal approximation is usually used for more efficiency. In addition, the scoring function of GLDS kernel can be simplified with the following compact technique [10].

$$f(\{x_i\}) = (\sum_{i=1}^{N} \alpha_i t_i \overline{R}^{-1} \overline{b_i} + \overline{d})^t \cdot \overline{b_x} = w^t \cdot \overline{b_x} \tag{8}$$

Therefore, the score of target model on a sequence of observations can be calculated just using the averaged observation. Furthermore, by collapsing all the support vectors down into a single model vector w , each target score can be calculated by just a simple inner product which makes this framework suited well for hypothesis searching framework with critical request of computational complexity and large hypothesis space.

GLDS kernel SVM is used as the discriminative classifier, and two classes (1,$-$1) represent two speakers (λ=1,2) respectively. Therefore, the object function (5) can be denoted as:

$$D(\mathbf{y}^1, \mathbf{y}^2 \mid \lambda) = f(\{y_i^1\}) - f(\{y_i^2\}) \tag{9}$$

4.3. Missing Data Training

The binary mask achieves substantial performance gains under cochannel and additive noise conditions in both robust speech recognitionand speaker identification tasks[4]. To account for the deviations of corrupted features from clean ones, we reconstruct the auditory features from estimated binary masks.

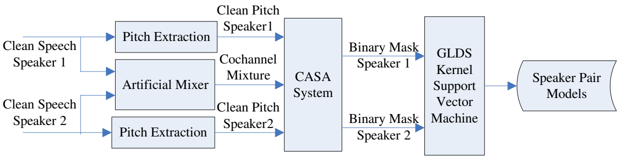


Figure 3: Missing Data Training Framework.

As seen in Fig.3, firstly prior pitch sequences are extracted from clean training speech. Then mix the clean training utterances from two speakers in the speaker pair into multiple cochannel mixtures. Using the prior clean pitch sequences, estimated ideal binary mask and separated speech can be constructed from cochannel mixtures by CASA system. After speech activity detection, nonspeech frames are eliminated and 13 dimension mel-frequency cepstral coefficient (MFCC) feature appended with delta [10] are extracted. Finally, speaker pair model is trained by a GLDS kernel binary SVM classifier.

5. Experiments

5.1. Multi-Pitch Accuracy Evaluation

The performance of the proposed multi-pitch estimation method is evaluated with the two-talker part of the challenge database [9],under five SNR conditions from -6dB to $+6$dB (600 utterances for each condition). Along with the mixing two-talker data, clean speech signals of the forward talker are also provided, which are utilized to obtain the reference pitches using SHS method. In this evaluation, the forward talker is supposed to be the target talker that our system aims to separate from the mixture and the multi-pitch estimation's performance is tested by comparing estimated pitch sequences to reference pitch sequences at the voiced parts under the criterion of 20% tolerance.

Three kinds of pitch accuracy is tested. One is the multi-pitch candidates accuracy, which is measured by whether any of the multiple candidates is in the 20% tolerance of the target pitch. Another is the ideal-grouping accuracy, which measures the pitch accuracy under the optimal sequential grouping alignments. And the last is the baseline grouping accuracy that reflects the performance of our baseline sequential grouping method. In addition, our multi-pitch system is compared with Wu's multi-pitch method (released by Wu) [6] as well.

As seen in Table.1, the average pitch accuracy of multi-pitch candidates and ideal grouping is 90.64% and 81.15% respectively, compared with our baseline sequential grouping of

Table 1: Multi-pitch Accuracy Evaluation Results.

SNR	pitch_cand	ideal_group	base_group	Wu's method
6dB	93.39%	84.59%	75.89%	72.15%
3dB	92.49%	83.54%	74.53%	69.95%
0dB	91.04%	81.65%	73.15%	67.33%
m3dB	89.20%	79.31%	70.97%	64.36%
m6dB	87.06%	76.64%	68.36%	61.71%
Mean	90.64%	81.15%	72.58%	67.10%

72.58% accuracy. The gap between pitch candidates accuracy and ideal grouping accuracy is due to the true pitch loss in the intra-segment grouping process, while the gap between ideal grouping accuracy and baseline grouping accuracy could be attributed to the alignment errors of the baseline system and speaker model based sequential grouping method is aimed to make up this big gap (≈8.5%). It is worth noting that even our baseline system (average 72.58%) is much better than Wu's system (average 67.10%).

5.2. Speaker Pair Model Evaluation

There are 34 speakers in Challenge database [9], to evaluate SVM speaker pair models, we randomly choose 5 male and 5 female speakers to generate 45 speaker pairs. For each speaker in a certain speaker pair, 400 utterances (normally 1∼2 seconds) are used for training the binary classifier, while the other 100 utterances are for testing. The average binary SID performance is listed in Table.2, categorized by different binary mask conditions [5] in speech resynthesis process.

Table.2 shows that the binary mask condition in both training and testing process should be identical. Since the test condition of speaker pair model in voiced sequential grouping is binary mask estimated from hypothesis segments alignment, speech should be resynthesized using binary mask estimated by clean pitch before training.

5.3. Voiced Sequential Grouping Evaluation

An objective and straightforward criterion to quantitatively evaluate the performance of a CASA system is to measure signal-to-noise ratio (SNR) before and after segregation, using clean target speech resynthesized with an all-one mask as signal [5]. For this evaluation, only cochannel mixtures from different speakers with SNR equal to 0dB are considered to simulate real cochannel situations. In challenge two talker test database [9], there are 363 cochannel test utterances within 257 speaker pairs. For speaker pair model training, the artificial mixer generates up to 1000 usable speech from only 100 training utterance of each speaker. All the test and training data are from standard challenge database. Results from both speakers in test utterances are used for analysis and shown in Table.3. In the experiment, hypotheses with very low probabilities are pruned in search space which reduces the computation time dramatically.

Table 2: Speaker Pair Model Evaluation.

Train Condition	Test Condition	Accuracy
all-one-mask	all-one-mask	99.9%
all-one-mask	binary mask by clean pitch	88.0%
binary mask by clean pitch	all-one-mask	92%
binary mask by clean pitch	binary mask by clean pitch	99.5%

Table 3: Voiced Sequential Grouping Evaluation.

Grouping strategy	Multi-pitch accuracy	SNR gain
ideal_grouping	79.62%	7.41
base_grouping	70.60%	3.11
speaker_grouping	76.23%	5.61

In Table.3, ideal sequential grouping achieves 79.62% pitch accuracy and 7.41 dB SNR improvement which is the upper limit on the performance of sequential grouping. The second row presents the results of baseline sequential grouping in Section5.1. Speaker model based sequential grouping achieves 76.23% pitch accuracy and 5.61 dB SNR gain which shows significant improvement over the baseline performance and reflects the effectiveness of employing speaker pair prior information to reduce the cross speaker pitch error.

6. Conclusion

In the proposed framework, after auditory segmentation based on onset and offset analysis, robust multipitch tracking algorithm is performed and used to segregate the corresponding voiced portions of each segment. Furthermore, support vector machine based speaker pair models are used for sequential grouping in voiced cochannel speech. Subsequently, binary mask based missing data speaker model training and GLDS sequential kernel are used to improve SID performance and reduce computation complexity, respectively.

7. References

[1] A. Bregman, *Auditory scene analysis: the perception and organization of sound.* MIT Press, 1990.

[2] Y. Shao and D. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, and Lang. Proc*, vol. 14, pp. 289–298, 2006.

[3] S. Srinivasan, Y. Shao, Z. Jin, and D. Wang, "A computational auditory scene analysis system for robust speech recognition," *ICSLP*, pp. 73–76, 2006.

[4] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," *ICASSP*, vol. 1, pp. 645–648, 2006.

[5] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[6] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

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

[8] C. Cao, M. Li, J. Liu, and Y. Yan, "Multiple F0 estimation in polyphonic music (MIREX 2007)," *Third Music Information Retrieval Evaluation eXchange (MIREX 2007)*, 2007.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, p. 2421, 2006.

[10] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.