ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

MODEL-BASED SEQUENTIAL ORGANIZATION FOR COCHANNEL SPEAKER IDENTIFICATION

Yang Shao and DeLiang Wang

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

10.21437/Interspeech.2004-544

Abstract

It is difficult to directly apply traditional speaker identification (SID) systems to cochannel speech, mixtures from two speakers. Previous work demonstrates that extraction of usable speech segments significantly improves SID performance if speaker assignment, or sequential organization of the segments, is known. We derive a joint computational objective for speaker assignment and cochannel SID, leading to a problem of search for the optimum hypothesis. We propose a hypothesis pruning method based on speaker models to make the search computationally feasible. Evaluation results show that the proposed algorithm approaches the ceiling SID performance obtained with prior pitch information, and yields significant improvement over alternative approaches on speaker assignment.

1. Introduction

Cochannel speech is composed of speech utterances from two talkers. Unlike conversations, talkers from different channels are not aware of each other in cochannel speech. Consequently, speech from both channels has large overlap, which presents a considerable challenge to automatic speaker recognition. On the other hand, for a cochannel recording that has comparable energies from both talkers (e.g. target-to-interferer ratio, or TIR, is zero), human listeners can readily select and follow one speaker's voice. Even in worse scenarios such as a cocktail-party, listeners can select and follow the voice of a particular talker [3]. Bregman [3] describes this process of auditory perception as auditory scene analysis, which consists of simultaneous organization and sequential organization. The former integrates concurrent sound components and the latter integrates components across time into the same perceptual stream. Most of the existing computational auditory scene analysis systems, e.g. [6], address only simultaneous organization. In this paper, we study how to sequentially organize spectral components of the same speaker into a single stream in cochannel speech based on speaker models, for the purpose of improving speaker identification (SID) performance.

Research has been carried out for decades to extract speakers from cochannel speech by either enhancing target speech or suppressing interfering speech [9], [8]. However, as pointed out in [7], for speaker recognition the intelligibility and quality of extracted speech are not as important as in traditional cochannel speech enhancement systems. In a closed-set SID task, what the system needs are portions of the speech that contain speaker characteristics, which are unique, classifiable and long enough for the system to make identification or verification decisions. These portions, or segments, are defined as consecutive frames of speech that are minimally corrupted by interfering speech, and thus called usable speech [7].

Previously, we proposed a usable speech extraction method [12] based on the pitch contours obtained from a robust multipitch tracking algorithm [13]. Our method removes the segments with overlapping pitch tracks as well as those classified as silence or unvoiced; the segments that have a single pitch track are regarded as usable and thus retained. Afterwards, each usable speech segment is assigned to one of the two speaker streams. Finally the streams are fed to a standard speaker recognizer. The evaluation results show a significant improvement for SID across various TIRs from -20 dB to 20 dB. A key limitation in the study is the assumption that the pitch contours of each speaker are *a priori* known and speaker assignment is done with such prior information in order to test whether the extracted segments are useful for SID.

Studies on speaker detection, tracking and clustering have been conducted in multi-speaker environments such as conversational speech and broadcast news. Various methods, supervised or unsupervised, have been explored. A typical method [5] is to use log-likelihood ratio scores from speaker models and a universal background model, to partition a recording into homogeneous segments and then cluster the segments. However, such methods cannot be applied to cochannel speech because, as mentioned earlier, cochannel talkers strongly overlap, resulting in very short speaker-homogenous segments. In the case of 0 dB TIR, such segments typically last 30 *ms* to 300 *ms*, far shorter than the typical minimum length of 1 *sec* for speaker clustering [5]. As pointed out in [7], a speaker recognizer's ability to identify talkers based on pooled frame-level scores is sharply reduced when the overall length is less than 500 *ms*. To verify this, we have explored segment clustering ourselves, and found that the result is barely above the chance level.

In this paper, we propose a model-based speaker assignment method for cochannel SID. We develop a computational objective for joint speaker assignment, or sequential grouping, and SID. Our formulation leads to a search problem to find an optimal hypothesis in the joint speaker and assignment space. Exhaustive search finds the optimal hypothesis though it is computationally infeasible. We propose a hypothesis pruning method, which iteratively removes hypotheses with low probabilities and thus reduces the search space and computation time greatly. We show that the pruning method achieves a performance level close to that of exhaustive search and ceiling performance with prior pitch information.

In Section 2, we describe the proposed speaker assignment method. Section 3 gives evaluation results and the comparisons with alternative approaches. Section 4 concludes the paper.

2. Model-based speaker assignment

Maximum-likelihood classification is well established for speaker identification [10]. However, in order to recognize talkers in cochannel speech, the traditional probability framework needs to be extended to multiple speakers.

2.1. Speaker identification

Given a set of reference speaker models $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_K\}$, the goal of SID is to find the speaker model that maximizes the posterior probability for an observation sequence, $O=\{o_1,o_2,\ldots,o_M\}$. Cepstral features are widely used as observations for speech signals. The SID decision rule is

$$\hat{\lambda}=\arg\max_{\lambda\in\Lambda} P(\lambda\mid O). \tag{1}$$

Applying the Bayesian rule, we have

$$\hat{\lambda}=\arg\max_{\lambda\in\Lambda}\frac{P(O\mid\lambda)P(\lambda)}{P(O)}. \tag{2}$$

Typically, prior probabilities of speakers are assumed equal, and the probability of observing O is the same for all speakers in Λ. Using pre-trained speaker models and assuming independence between observations, (2) can be rewritten as

$$\hat{\lambda}=\arg\max_{\lambda\in\Lambda}\sum_{m=1}^{M}\log p(o_m\mid\lambda) \tag{3}$$

after taking the log operation. Here m is the index of a total of M observations. $p(o\mid\lambda)$ is the Gaussian mixture model (GMM) estimated from training speech of specific talkers using the EM algorithm [10].

2.2. Extension to cochannel speech

Cochannel speaker identification aims to find two speaker models that maximize the posterior probability for the observations. For a cochannel mixture, our usable speech extraction method [12] extracts N usable segments, $X=\{S_1,\ldots,S_i,\ldots,S_N\}$, each of which contains consecutive speech frames, $S_i=\{x\}$, with a single pitch contour. Given X, (1) can be modified as follows

$$\hat{\lambda}_1,\hat{\lambda}_{II}=\arg\max_{\lambda_1,\lambda_{II}\in\Lambda} P(\lambda_1,\lambda_{II}\mid X), \tag{4}$$

which is to find a pair of speaker models, $\hat{\lambda}_1$ and $\hat{\lambda}_{II}$ from the speaker set Λ, that maximize the posterior probability given usable speech segments. As mentioned earlier, these segments must be assigned to two speaker streams because they may be produced by different speakers in cochannel speech. For example, a possible assignment (grouping) may look like $\{S_1^0,S_2^1,\ldots,S_i^1,\ldots,S_N^0\}$, where superscripts, 0 and 1, do not represent the speaker identities but only denote that the segments marked with the same label are from the same speaker. Therefore, the joint objective of sequential grouping and SID may be stated as finding a pair of speaker models, $\hat{\lambda}_1$ and $\hat{\lambda}_{II}$, together with a segment assignment, $\hat{y}$, that jointly maximize the posterior probability:

$$\hat{\lambda}_1,\hat{\lambda}_{II},\hat{y}=\arg\max_{\lambda_1,\lambda_{II}\in\Lambda,\,y\in Y} P(\lambda_1,\lambda_{II},y\mid X), \tag{5}$$

where Y is the assignment space, which includes all possible assignments (labelings) of the segments.

2.3. Derivation

The posterior probability in (5) can be rewritten as

$$P(\lambda_1,\lambda_{II},y\mid X)=\frac{P(\lambda_1,\lambda_{II},y,X)}{P(X)}=\frac{P(\lambda_1,\lambda_{II},X^y)}{P(X)}, \tag{6}$$

in which X^y represents the labeled segments according to a specific segment assignment y. (6) can be further written as

$$P(\lambda_1,\lambda_{II},y\mid X)=P(X^y\mid\lambda_1,\lambda_{II})\frac{P(\lambda_1,\lambda_{II})}{P(X)}. \tag{7}$$

Assuming the independence of speaker models and using the same assumption from traditional SID that prior probabilities of speaker models are the same, we insert equation (7) into (5) and remove the constant terms. The objective then becomes finding two speakers and an assignment that have the maximum probability of assigned usable speech segments given the corresponding speaker models as follows.

$$\hat{\lambda}_1,\hat{\lambda}_{II},\hat{y}=\arg\max_{\lambda_1,\lambda_{II}\in\Lambda,\,y\in Y} P(X^y\mid\lambda_1,\lambda_{II}). \tag{8}$$

Note, the conditional probability is essentially the joint SID score of assigned segments. Assuming that any two segments, e.g. S_i and S_j, are independent of each other given the speaker models, the conditional probability in (8) can be written as

$$P(X^y\mid\lambda_1,\lambda_{II})=\prod_{S_i\in X^0}P(S_i\mid\lambda_1)\prod_{S_j\in X^1}P(S_j\mid\lambda_{II}), \tag{9}$$

in which X^0 is the subset of segments labeled with 0; and X^1 the subset labeled with 1. Assuming the observations are independent of each other. The probability of having a segment, S, from a pre-trained speaker model λ is the product of likelihoods of that speaker model generating each individual observation x of the segment. In other words,

$$P(S\mid\lambda)=\prod_{x\in S}p(x\mid\lambda). \tag{10}$$

In the experiments of this paper, speakers are modeled as 16-mixture GMMs, which are tested to be sufficient for the data, and the observations or features used are cepstral coefficients and their first-order dynamic coefficients.

Instead of the formulation in (6), the posterior probability can also be expanded as follows

$$P(\lambda_1,\lambda_{II},y\mid X)=P(\lambda_1,\lambda_{II}\mid y,X)P(y\mid X). \tag{11}$$

The first term on the right side is the probability of two speaker models given an assignment, which is essentially the joint SID score of the assigned segments. The second term is the conditional probability of a specific assignment, y, given the usable segments, and this probability may be called the assignment model. The decomposition of the posterior in (11) is analogous to HMM-based speech fragment grouping in [2].

2.4. Computational methods

The computational objective in (8) is to find two speakers and one assignment that yield the maximal probability using (9)-(10). Given the usable segments and speaker models, the maximization amounts to a search for the globally optimal hypothesis in the joint speaker and assignment space, Λ and Y.

The brute-force way to find the maximum is exhaustive search. For a cochannel mixture, this involves calculating the probability of the assigned segments given a pair of speaker

models, $P(X^y \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$, for every possible pair out of K speakers in Λ, and for every assignment in Y of N segments. Each segment can take either label, 0 or 1. Therefore, let the calculation of $P(X^y \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ take a unit time, the total computation time is on the order of $O(K^2 \cdot 2^N)$. Clearly, exhaustive search is computationally prohibitive with a large number of usable speech segments though it produces theoretically optimal results.

However, in the search space, some hypotheses have very low probabilities. Subsequently, if these hypotheses could be pruned from consideration, the computation time could be greatly reduced. The results of exhaustive search indicate peaky distributions with each peak occupied by several assignment hypotheses in the search space. Thus, although only retaining the best hypothesis is not optimal, keeping a small number of hypotheses appears sufficient. We propose an iterative hypothesis pruning method to keep the two best hypotheses that integrates all the assigned segments till current iteration. Specifically, the algorithm is as follows.

Hypothesis Pruning Algorithm

Step 0. Order the segments in $X = \{S_1, S_2, \ldots, S_N\}$ in time.

Step 1. Label S_1 in X with 0 (assign it to X^0). This initial assignment is arbitrary.

Step 2. For S_2 in X, form two hypotheses: H_0, H_1, and create a label path for each of them. H_0 assumes that the current segment belongs to set X^0, and H_1 that the current segment belongs to X^1. The label paths are

$$Path[2][H_0] = (0,0), \; Path[2][H_1] = (0,1).$$

$Path[n][\bullet]$ records labels for the assignments of the past n-1 segments and the hypothesized assignment of the current segment.

Step 3. For an unprocessed segment S_n, n>2, form H_0 and H_1. Then expand the label path for H_0 and H_1 as follows,

$$Path[n][H_0] = (\, Path[n\text{-}1][\underset{H \in \{H_0, H_1\}}{\operatorname{argmax}} L(Path[n\text{-}1][H],0)\,],0),$$

$$Path[n][H_1] = (\, Path[n\text{-}1][\underset{H \in \{H_0, H_1\}}{\operatorname{argmax}} L(Path[n\text{-}1][H],1)\,],1),$$

where the L function, as defined below, estimates the joint SID score by considering the best partial segment assignment from 1 to n.

$$L\big(Path[n-1][H],l\big) = \max_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda} P(X^{(Path[n-1][H],l)} \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}), \quad (12)$$

l = 0 or 1, refers to the hypothesized labeling of the current segment.

Step 4. Repeat Step 3 until the last segment S_N is processed. For S_N, compare the likelihood values returned by L for H_0 and H_1. The final winning hypothesis is the one with the higher likelihood. Obtain the corresponding two speaker identities that maximize (12) and the segment assignment for this hypothesis.

The L function in (12) is the same as (8) except that L only considers the partial assignment of S_1 to S_n. Since at each iteration, hypotheses are pruned according to the partial scores, this is a greedy algorithm and it is a form of beam search [11]. For each unlabeled segment, it retains two hypotheses, each of which calculates $P(X^y \mid \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ twice in the worst case, resulting in polynomial time complexity on the order of $O(K^2 \cdot N)$.

3. Evaluation and comparison results

As in previous studies [7], [12], we employ the evaluation data from the TIMIT corpus. The speaker set consists of 38 speakers: 14 females and 24 males. Each speaker has 10 utterances, ranging from 1.5 *sec* to 6.2 *sec* in length. For each speaker, 5 files are used for training and the remaining 5 are used to create cochannel mixtures for testing. For each speaker deemed as the target speaker, 1 out of 5 test file is randomly selected and mixed with randomly selected files from every other speaker, which are regarded as interferers. The overall TIR of the speech mixture is calculated as the ratio of target speech energy over the interfering speech energy. Here we only consider mixtures with TIR equal to 0 dB and a total of 1406 cochannel mixtures are created.

3.1. Evaluation

As there are two types of output jointly produced from the algorithm, we show the results in two tables. Table 1 shows the correct rate of speaker assignment by counting correctly assigned frames, which count those from the same speaker and marked with the same label. Table 2 shows the SID performance with three different criteria because there are up to two speakers in a cochannel mixture. Criterion I records the percentage of mixtures where both speakers are correctly identified; this is the most stringent criteria. Sometimes the speaker from a specified channel is of interest. Thus, criterion II displays target identification correct rate. Criterion III counts the files where either of the two speakers in the mixture is identified correctly.

In Table 1, under the test condition without any usable speech processing, each frame can take two possible labels; so the baseline rate of assignment is 50%. The second row shows that ideal assignment by prior pitch achieves 94.1% correct rate. Note that ideal assignment is applied at the segment level and a segment takes the label of a majority of the frames in it, each decided by comparing the detected pitch with the prior pitch before mixing. The imperfect result reflects that a single-pitch segment does not always contain frames from the same speaker, which is not surprising considering the nature of cochannel speech.

Exhaustive search achieves 77.4% correct rate. It reflects the effectiveness of using speaker characteristics for sequential organization. From the derivation it is evident that exhaustive search places an upper limit on the performance of model-based sequential grouping. Our proposed hypothesis pruning method achieves 76.2% correct rate, approaching the upper limit.

Similar observations can be made from SID results in Table 2. The first row gives the baseline performance with unprocessed mixtures. As there is only the mixture, criterion I does not apply. Ideal assignment produces the ceiling performance though it is not 100% correct because of imperfect assignment and limited segment lengths. Exhaustive search approaches the ceiling performance, and the hypothesis pruning method performs almost as well as exhaustive search, while drastically cutting the computation time — from an average of 7 *min* per file to 0.7 *sec*. Since the search is based on SID scores, the performance gap between the model-based method and ideal assignment is smaller than that of sequential grouping performance.

We have also explored variations of the hypothesis pruning algorithm. First, we have evaluated retaining just one hypothesis instead of two. This essentially degrades the

TABLE 1 CORRECT ASSIGNMENT RATE FOR DIFFERENT SEQUENTIAL GROUPING METHODS.

	Correct rate (%)
W/o usable speech processing	50.0
Ideal assignment by prior pitch	94.1
Exhaustive search	77.4
Hypothesis pruning	76.2
Pitch dynamics	68.2
Spectral divergence	66.2

TABLE 2 COCHANNEL SPEAKER IDENTIFICATION CORRECT RATE.

	SID correct rate (%)		
SID Criteria	I	II	III
W/o usable speech processing	N/A	50.0	82.5
Ideal assignment by prior pitch	43.3	72.0	93.7
Exhaustive search	40.2	70.4	93.9
Hypothesis pruning	37.5	68.8	93.0
Pitch dynamics	22.3	52.5	90.4

algorithm to local decision-making, and it performs significantly worse than keeping two hypotheses. Instead of keeping the two best hypotheses ending with different labels, we have tried retaining two or three best hypotheses out of a total of four considering the previous assignments. The results are similar to those in the tables. The peaky hypothesis distribution in the search space is a main reason why our pruning method approaches the ceiling performance.

3.2. Comparison

In this section, we compare with alternative sequential grouping methods, namely one that employs pitch dynamics and one based on spectral divergence.

One reasonable alternative is to utilize pitch information, particularly since pitch contours have already been obtained. Previous studies have demonstrated the importance of pitch contours for speaker recognition, e.g. [1]. We collect pitch differences between the end-point of a segment and the start-point of the following segment from the training data. Considering that the longer is the gap between two segments the less likely they belong to the same speaker, we multiply the difference by the time lag between them. The resulting product describes the pitch change dynamics between neighboring segments. The product distribution is modeled as a mixture of Gaussian and uniform distribution [13]. Segments are grouped by comparing the likelihoods of the dynamics product given the distribution. The speaker assignment and SID results are also presented in Tables 1 and 2. This method clearly performs worse than the pruning algorithm.

We have also compared with a spectrum-based method, specifically, the speaker assignment algorithm of Morgan et al. [8] that also addresses sequential organization. Their system aims to enhance cochannel speech by separating two talkers and subsequently assigning separated speech components to two speaker streams. The assignment is based on a frame-level spectral comparison using the spectral divergence measure of Carlson and Clement [4]. Since our system considers a usable segment to belong to one speaker, we employ their algorithm to perform only speaker assignment; that is segments are organized using their spectrum-based method. The assignment result is shown in Table 1. With 66.2% correct rate, the spectral method is comparable to the pitch dynamics method, and it lags behind our proposed method.

4. Conclusion

We have proposed a model-based approach for sequential organization, and applied it to improve cochannel SID performance. We have shown that the proposed hypothesis pruning algorithm achieves SID performance close to the ceiling performances with prior pitch information or exhaustive search, performing significantly better than alternative approaches. It is worth noting that our sequential grouping algorithm can handle the situation where only one speaker is present in a cochannel mixture. Since segments may all take the same label after assignment, our algorithm can produce only one speaker identity. Also, the probabilistic framework proposed in here can be extended to situations with more than two speakers in a mixture.

Acknowledgments. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an NSF grant (IIS-0081058).

5. References

[1] B.S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Am.*, vol. 52, pp. 1687-1697, 1972.

[2] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Comm.*, in press.

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

[4] B.A. Carlson and M.A. Clements, "A computationally compact divergence measure for speech processing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 1-6, 1991.

[5] R.B. Dunn, D.A. Reynolds, and T.F. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *Digit. Sig. Process.*, vol. 10, pp. 93-112, 2000.

[6] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, to appear, 2004.

[7] J.M. Lovekin, R.E. Yantorno, K.R. Krishnamachari, D.S. Benincasa, and S.J. Wenndt, "Developing usable speech criteria for speaker identification," in *Proc. ICASSP*, pp. 421-424, 2001.

[8] D.P. Morgan, E.B. George, L.T. Lee, and S.M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 407-424, 1997.

[9] T.F. Quatieri and R.G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 56-69, 1990.

[10] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Comm.*, vol. 17, pp. 91-108, 1995.

[11] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*. 2nd ed., Upper Saddle River, NJ: Prentice Hall, 2003.

[12] Y. Shao and D.L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. ICASSP*, pp. 205-208, 2003.

[13] M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11(3), pp. 299-241, 2003.