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

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A COMBINED ADAPTIVE AND DECISION TREE BASED SPEECH SEPARATION TECHNIQUE FOR TELEMEDICINE APPLICATIONS

*Yunxin Zhao, Xiao Zhang, Xiaodong He and Laura Schopp**
Dept. of Computer Engineering & Computer Science
*Dept. of Physical Medicine & Rehabilitation
University of Missouri, Columbia, MO 65211, USA

ABSTRACT

We present a novel technique for separation of doctor and patient's speech in conversations over a telemedicine network. The mixed speech signals acquired at doctor's site is first broken into single talkers' speech segments and background by using thresholds of energy and duration. The speech segments are then identified as spoken by doctor or patient in two steps. In the first step, Gaussian mixture models (GMM) of doctor and patient are used, where the doctor's model is obtained from his/her training speech, and the patient's model is initialized by a general speaker model and then adapted by the patient's speech. In the second step, a decision tree that uses contextual and confidence features is applied to refine the identification results. Preliminary experiments were performed on three data sets collected in telemedicine. Without adaptation and decision tree, error rates at the segment-level and frame-level were 25.44% and 16.53%, respectively. With adaptation, segment and frame error rates were reduced to 13.11% and 7.85%, and with decision tree, the error rates were further reduced to 10.48% and 6.73%, respectively.

1. INTRODUCTION

In talker identification and tracking, Gaussian mixture modeling of talkers has been shown as a successful approach [1,2]. Recently, new efforts appeared in speaker separation and turn detection. For example, a combined speaker identification and adaptation strategy was taken to detect talker change in manually segmented utterances of broadcast-news [3], and a joint optimal decoding of speech and identification of talkers was accomplished by using talker-dependent language models [4].

We are currently investigating applications of spoken language processing in telemedicine to provide assistance to hearing impaired users. In the telemedicine scenario, the doctor's speech stream is mixed by the patient's speech stream since the patient's speech is played through loud speakers at the doctor's site and hence propagates into doctor's microphone. The separation of doctor-patient's conversation speech into two speech streams is therefore needed. In the current work, initial speech segmentation is made by applying energy and duration thresholds to the input speech. The speaker ID associated with each segment is then determined by a GMM classifier and a decision tree. In the telemedicine scenario, doctor's GMM model can be estimated from his or her training speech, but GMMs of each patient cannot be trained prior to conversation. Therefore a general speaker model is initially estimated from several talkers' speech excluding those of the intended doctors and patients, and each patient's model is subsequently adapted using Maximum a Posterior (MAP) estimation. The decision tree uses confidence and contextual features to further improve accuracy of speaker ID since many speech segments are very short and the IDs cannot be reliably identified by the GMM method. The speech separation performance is evaluated as the error rates at the segment and frame levels.

This paper is organized as five sections. In section 2, an overview of the speech separation system is described. In section 3, details of the system components are discussed. Experiments and results are presented in section 4 and a conclusion is made in section 5.

2. SYSTEM OVERVIEW

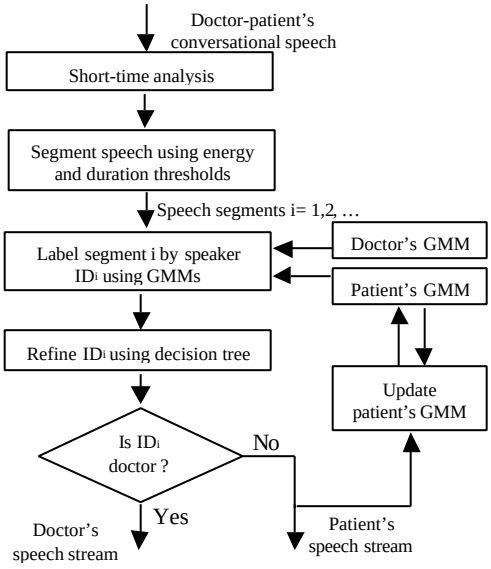

Figure 1 System overview

The on-line speech separation system is illustrated in Figure 1. Doctor-patient's conversation is acquired by a microphone at the doctor's site. Short-time analysis is performed to produce a sequence of speech feature

vectors, and the sequence is then broken into segments using thresholds of energy and duration. For each segment, a GMM-based speaker classifier is applied to determine the speaker ID as doctor or patient, and a decision tree that utilizes contextual and confidence features is next employed to refine the speaker ID. If the speaker is identified as patient, the speech segment is used to update the patient's GMM. The system outputs two separate speech streams, one for doctor, the other for patient.

3. SPEECH SEPARATION

3.1. Energy and Duration Based Segmentation

We assume that during a conversation, doctor and patient do not speak simultaneously and there is a brief stop interval between each speech turn. This assumption is proven as valid in the types of telemedicine applications we investigated, and the speech segmentation method is designed based on this characteristic of doctor-patient's conversation speech. An energy threshold is estimated from the background signal recorded prior to each conversation. The input speech signal is blocked into short-time frames to compute features of energy and cepstral coefficients. When the energy values of a number of successive frames are below the threshold, a breaking point of the current speech stream is identified. In this way, speech input is broken into short segments. The segments are merged into longer segments that consist of a silence followed by speech spoken by a single talker. Furthermore, a short segment located in the middle of a long silence is merged with the silence and a short segment without enough separation from its neighboring segment is merged into its neighboring segments. The merging reduced the total segments by 20% without compromising subsequent speaker ID accuracy.

3.2. Adaptive GMM-based Speaker Identification

Denote the GMM of doctor by Θ_1 and that of the patient by Θ_2. The probability density function of GMM Θ evaluated for a feature vector x_t is

$$p(x_t|\Theta) = \sum_{i=1}^{M} a_i \mathrm{N}(x_t; q_i) \qquad q_i = (m_i, c_i)$$

For a speech segment that consists of feature vectors $\{x_{t_1}, x_{t_1+1}, ..., x_{t_2}\}$, a score function is defined as

$$S(t_1, t_2) = \log \frac{\prod_{t=t_1}^{t_2} p(x_t|\Theta_1)}{\prod_{t=t_1}^{t_2} p(x_t|\Theta_2)}$$

The segment speaker ID is determined as

$$ID(t_1, t_2) = \begin{cases} 1 & S(t_1, t_2) \geq 0 \\ 2 & S(t_1, t_2) < 0 \end{cases}$$

with 1 and 2 representing the doctor and the patient, respectively. The GMMs are estimated using the Expectation-Maximization (EM) algorithm [5], where the patient's model is initially taken from a general speaker model, and MAP estimation is used to adapt the model to the patient's speech according to the following procedure [6][7]:

1. Initialize the patient's model by a general speaker model $\Theta_2^{(0)}$. Set n=0.
2. Collect statistics from patient's speech for all component Gaussians:

$$g_{t,i}^{(n)} = a_i p(x_t|q_i^{(n)}) / \sum_{j=1}^{M} a_j p(x_t|q_j^{(n)})$$

$$l_i^{(n)} = \sum_{t=1}^{T} g_{t,i}^{(n)} / (\sum_{t=1}^{T} g_{t,i}^{(n)} + b n_i)$$

$$m_{i,x}^{(n)} = \sum_{t=1}^{T} g_{t,i}^{(n)} x_t / \sum_{t=1}^{T} g_{t,i}^{(n)}$$

$$c_{i,x}^{(n)} = \sum_{t=1}^{T} g_{t,i}^{(n)} (x_t - m_{i,x}^{(n)})(x_t - m_{i,x}^{(n)})' / \sum_{t=1}^{T} g_{t,i}^{(n)}$$

   where n_i is the sample size for q_i and b is a parameter that controls the adaptation rate.

3. Update the Gaussian mixture model parameters:

$$a_i^{(n+1)} = (b n_i + \sum_{t=1}^{T} g_{t,i}^{(n)}) / (b \sum_{j=1}^{M} n_j + \sum_{j=1}^{M} \sum_{t=1}^{T} g_{t,j}^{(n)})$$

$$m_i^{(n+1)} = (1 - l_i^{(n)}) m_i^{(0)} + l_i^{(n)} m_{i,x}^{(n)}$$

$$c_i^{(n+1)} = (1 - l_i^{(n)}) c_i^{(0)} + l_i^{(n)} c_{i,x}^{(n)} + l_i^{(n)} (1 - l_i^{(n)}) (m_{i,x}^{(n)} - m_i^{(0)})(m_{i,x}^{(n)} - m_i^{(0)})'$$

4. If convergence criterion is not met, increment n and go back to step 2.

3.3. Decision Tree Refinement on Speaker IDs

To improve the accuracy of segment speaker IDs obtained using the GMMs, a decision tree is constructed to relabel the segment speaker IDs. The complete set of features used in the decision tree are defined below for the i-th segment, assuming that in this segment the silence portion covers t_0 to t_1 and the speech portion covers t_1 to t_2.

Feature 1: the speaker identifier ID_i obtained from the GMM classifier.

Feature 2: the relative log-likelihood score L_r, i.e., $L_r = |S(t_1,t_2)|/L_v$, with $L_v=t_2-t_1$. This feature relates to confidence of speaker ID as measured by the likelihood score.

Feature 3: the length of speech L_v. This feature also carries confidence information, since the accuracy of speaker ID as obtained from the GMM classifier depends on the speech length.

Features 4, 5, 6, 7: the speaker IDs of the neighboring segments ID_{i-2}, ID_{i-1}, ID_{i+1}, ID_{i+2} obtained from the GMM classifier. These four features, combined with the first feature ID_i, carry contextual information.

Feature 8: the length of silence, i.e. $L_s=t_1-t_0$. This feature carries contextual information since it describes the time separation of speech between this segment and the previous segment.

The "thresholding on one feature" binary decision tree is adopted. In building the tree, an exhaustive search is performed to find the best feature-threshold pair at each node, and the Maximum Impurity Reduction Criterion [8] is used as the optimization criterion. Let X represent the training subset at the current node, and X_i represent those training samples in X which belong to class C_i, with $X=X_1 \cup X_2$. Define $p(C_i \mid X) = \frac{\#(X_i)}{\#(X)}$ with "#" denoting the size of a set. The impurity of a node is measured as $I = p(C_1 \mid X)p(C_2 \mid X)$.

At each non-terminal node of the tree, the optimal feature-threshold pair is determined to maximize the impurity reduction when splitting the data into the left and right children nodes, i.e. X^L and X^R. The impurity reduction at this node is computed as $\Delta I = I - I'$, where $I' = p(C_1 \mid X^L)p(C_2 \mid X^L) + p(C_1 \mid X^R)p(C_2 \mid X^R)$. This process continues until a termination condition is satisfied, A node becomes a terminal node if its impurity is zero or if the number of data samples is below a threshold.

After the tree is built, each terminal node will be marked by a class that the node belongs to or by "invalid". A terminal node is marked as "invalid" if it meets one of the following criterions:

(1) the number of training samples is less than MIN_SAMPLE_NUM,
(2) the impurity is larger than MAX_IMPURITY,

where MIN_SAMPLE_NUM and MAX_IMPURITY can be obtained by experimentation.

When determining speaker ID for a segment i, a sequence of questions are asked, starting from the root node and ending at one of the terminal nodes. If the terminal node is "invalid", no correction is made on ID_i, otherwise, ID_i will be assigned by the class marked on that node.

4. EXPERIMENTS

4.1. Data

Three sets of conversation speech data were collected from one female doctor's conversations with three different female patients over a telemedicine network. These data were used as test data to evaluate the proposed speech separation system.

Test set I: a 765-second conversation between the doctor and a patient, where 54% was doctor's speech and 46% was patient's speech. This conversation was characterized by doctor and patient each speaking 1~3 sentences in a turn.

Test set II: a 766-second conversation between the doctor and a patient, where 57% was doctor's speech and 43% was patient's speech and pause. The patient answered the doctor's questions in a brief manner, but she took quite a bit of thinking time.

Test set III: a 146-second conversation between the doctor and a patient, where 73% was doctor's speech and 27% was patient's speech. The patient was a non-native English speaker. The doctor advised the patient with 3~5 sentences each time the patient asked a short question.

The training set consisted of dictation speech from the doctor and two females that were not any of the three patients, with each dictation being 10 minutes long. The training set also consisted of a 539-second conversation between a male doctor and a male patient.

4.2. System Implementation

Speech signal was sampled at 6.8 KHz with 16 bits per sample. The short-time analysis frame length was 20 msec, and the shift of frames was 10 msec.

Speech segmentation

The speech segmentation energy threshold was 160000 for the test sets I and II, and 22000 for the test set III. The speech separation window size was fixed as 2 for all three test sets.

GMMs

The speech features consisted of 16 Mel-frequency cepstral coefficients and their 1st-order time-derivatives. The doctor model was trained from the 10-minute dictation speech of the female doctor, and the general speaker model (for patient) was trained from the speech of the rest training talkers. The doctor's GMM was chosen to have 16 Gaussian densities, and the patient's GMM had 4 Gaussian densities.

Decision tree

Due to limited experimental data, decision trees were trained and used for speaker identification in the following manner:

1. The decision tree was built using the data of test set I and applied to test sets II and III. This tree had 15 layers, 219 nodes and 220 terminal nodes (including 182 invalid terminal nodes).
2. The decision tree was built using the data of test set II and applied to test sets I and III. This tree had 15 layers, 387 nodes and 388 terminal nodes (including 336 invalid terminal nodes).

In the experiment, the MIN_SAMPLE_NUM was 4, and the MAX_IMPURITY was 0.

4.3. Evaluation Methods

In measuring the performance of the proposed speech separation system, we observed two types of errors: mixed false classification (MFC) which occurred when a segment consisted of both doctor and patient's speech, resulting in uncertainty in speaker ID, and single false classification (SFC) which occurred when a segment had a single speaker's speech but its ID was wrong. We define the Segment-level Error Rate (SER) as

$$SER = \frac{\text{number of MFC + number of SFC}}{\text{total number of segments}}$$

To take into consideration of segment length variations, we also define the Frame-level Error Rate (FER) as

$$FER = \frac{\text{number of MFC frames + number of SFC frames}}{\text{total number of frames}}$$

4.4. Results

We first evaluated the case of using the general speaker model as the patient's models in GMM-based speaker identification. The result is summarized in Table I, where FNM denotes the total number of segments. After energy and duration based segmentation, 0.63% segments contained mixed doctor-patient's speech, which contributed to the MFCs. Most of the segment lengths were between 0.1s ~ 3s, which related the SERs with the FERs.

TABLE I
Separation error rates by using a general speaker model in GMM classification

	MFC	SFC	FNM	$SER_{\%}$	$FER_{\%}$
Test set I	8	206	987	21.68	13.77
Test set II	2	299	922	32.65	18.71
Test set III	3	6	151	5.96	19.60

We next improved the patients' models through MAP adaptation. We found that using a small amount of adaptation data from each patient's speech could significantly reduce the error rate, and therefore in the experiment we used only the beginning 10 seconds of each patient's speech for MAP adaptation of each patient's GMM. The results are summarized in Table II. On average, the adaptation helped reducing the SER and FER (averaged over three test sets) by 48.47% and 52.51%, respectively.

TABLE II
Separation error rates by using MAP adaptation on patient's GMM

	MFC	SFC	FNM	$SER_{\%}$	$FER_{\%}$
Test set I	8	97	987	10.64	6.24
Test set II	2	157	922	17.25	8.13
Test set III	3	3	151	3.98	14.76

Finally we applied the decision tree to refine the ID output from the GMM classifier using MAP adaptation, and the results are summarized in Table III. The results indicate that the decision tree reduced the average SFC and the average FER by 20.23% and 14.24%, respectively.

TABLE III
Separation error rates after decision tree refinement

	MFC	SFC	FNM	$SER_{\%}$	$FER_{\%}$
Test set I	8	73	987	8.21	5.16
Test set II	2	130	922	14.32	6.79
*Test set III	3	2	151	3.31	14.55
+Test set III	3	2	151	3.31	14.56

* result of the tree built from data of test set I
+ result of the tree built from data of test set II

5. CONCLUSION

In this paper, a novel technique is proposed to separate doctor-patient's conversational speech into separate speech streams for telemedicine applications. GMM with patient model adaptation is employed to identify the speaker of each speech segment. Context and confidence features are used by a decision tree to refine the speaker IDs. Our preliminary experimental results showed that this technique is promising for telemedicine applications.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under the grant NSF EIA 9911095.

REFERENCES

[1] D. A. Reynolds. "Speaker identification and verification using Gaussian mixture speaker models." Speech Communication, vol.17, (1-2): pp. 91-108, August 1995.

[2] I. Magrin-Chagnolleau et. al., "Detection of target speakers in audio databases," Proceedings of ICASSP, pp. 821-824, Phoenix, AZ, April 1999.

[3] Z.-H. Zhang, S. Furui and K. Ohtsuki, "On-line incremental speaker adaptation with automatic speaker change detection," Proceedings of ICASSP, pp. 961-964, Istanbul, Turkey, June 2000.

[4] N. Murai and T. Kobayashi, "Dictation of multiparty conversation using statistical turn taking model and speaker model," Proceedings of ICASSP, pp. 1575-1578, Istanbul, Turkey, June 2000.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood estimation from incomplete data via the EM algorithm," J. Royal Statist. Soc. Ser. B, vol. 39, pp. 1-38, 1977.

[6] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," IEEE Trans. On Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, April 1994.

[7] Y. Zhao, "Self-learning speaker/channel adaptation based on spectral variation source decomposition," Speech Communication, vol. 18 (1), pp. 65-78, January 1996.

[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks, 1986.