Adaptive Decision Tree-based Phone Cluster Models for Speaker Clustering

Chia-Hsin Hsieh, Chung-Hsien Wu and Han-Ping Shen

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, R.O.C.
{nsgnall, chwu, happy}@csie.ncku.edu.tw

Abstract
This study presents an approach to speaker clustering using adaptive decision tree-based phone cluster models (DT-PCMs). First, a large broadcast news database is used to train a set of phone models for universal speakers. The multi-space probability distributed-hidden Markov model (MSD-HMM) is adopted for phone modeling. Confusing phone models are merged into phone clusters. Next, for each state in the phone MSD-HMMs, a decision tree is constructed to store the contextual, phonetic, and speaker characteristics for data sharing over all speakers. For speaker clustering, each input speech segment is used to retrieve the Gaussian models from the DT-PCMs to construct the initial speaker-dependent phone cluster models. Finally, all the corresponding adapted speaker-dependent phone cluster models are used for speaker clustering via a cross-likelihood ratio measure. The experimental results show the DT-PCMs outperform the conventional GMM-based approach.

Index Terms: Speaker clustering, adaptive decision tree-based phone cluster models, confusion phone cluster models, speaker clustering, adaptive decision tree-based phone cluster models, confusion phone cluster models, speaker clustering.

1. Introduction
While the spoken documents, such as broadcast news or meeting records, have been drastically increased in recent years, retrieval and management of the spoken documents become more and more important. Accordingly, speaker segmentation and clustering [2] has been developed to annotate an input audio stream with type information, such as speaker identity, foreground or background audio types. Speaker clustering can be applied to improve the performance of speech recognition and speaker identification or further speaker clustering. In contrast, speaker clustering can achieve a better performance if sufficient training data are available to train the speaker-dependent phone models for recognition. However, the length of each speech segment in the broadcast news is not long enough to train a speaker-dependent phone models for speaker clustering. It is beneficial to construct a set of initial universal phone models and then adapt the universal models to a speaker-related phone models based on the input speech segment. Nevertheless, the universal phone models are too general to be effectively adapted to a specific speaker using only a short speech segment. In order to reduce the model number and eliminate the problem of data sparseness, confusing phone models are merged into phone clusters. Besides, speaker-related, contextual information should be considered into the universal phone cluster models for effective speaker adaptation. Based on this idea, this approach first constructs a set of universal phone cluster models based on a decision tree, in which the leaf nodes contain the Gaussian models with speaker-related information. The decision tree is constructed based on the contextual, phonetic, and speaker characteristic information. In speaker clustering, the speaker-related Gaussian models can be retrieved from the adaptive decision tree-based universal models based on the input speech segment and can be further adapted to the phone cluster models related to the speaker of the input speech segment. The adapted speaker-related phone cluster models obtained from the input broadcast news are finally used for speaker clustering.

2. The proposed framework for speaker clustering
Figure 1 shows the proposed framework for speaker clustering. In the training phase, a large speech database with several speakers collected from the broadcast news with transcriptions is used to train the universal phone models for speech recognition. Then the confusing phone models are further merged into phone cluster models to reduce the number of phone models to avoid speech recognition errors. Finally, an adaptive decision tree-based phone cluster models (DT-PCMs) is constructed to store the contextual, phonetic, and speaker characteristics for data sharing over universal speakers.

In the speaker clustering phase, the audio segmentation module [9] detects all the boundaries with speaker changes or foreground/background audio changes first. Then a speech activity detection module is applied to identify all speech segments which will be used for clustering. After speech recognition on the input speech segments, the corresponding top-N phone sequences are transcribed for each speech segment. For each speech segment, the most similar Gaussian model in the DT-PCMs for each state is retrieved to construct the initial phone cluster models based on the contextual, phonetic and speaker characteristics of the input speech segment. Then the initial phone cluster model is adapted to its corresponding speaker-related phone cluster models using maximum likelihood linear regression (MLLR) method. Finally, a bottom-up agglomerative clustering algorithm is applied via the cross-likelihood ratio on all speaker-related phone cluster models; each represents one input speech segment, for speaker clustering.
3. Adaptive decision tree-based phone cluster model

This study first constructs an adaptive decision tree-based phone cluster models (DT-PCMs), in which the leaf nodes contain the Gaussian models to be retrieved for initial model construction. The DT-PCMs store the phonetic, contextual and speaker characteristic information from a broadcast news audio database. The construction of the DT-PCMs consists of two steps. The first is phone clustering based on the phoneme confusion information and the second is the decision tree construction for each phone cluster model using the phonetic, contextual and speaker characteristic information.

3.1. Context-dependent phone clustering

In this step, a confusion phone set is obtained on a large speech database using the universal phone models first. There are 150 right context dependent sub-syllables in Mandarin speech and accordingly a 150×150 confusion matrix is built for each speaker via a sub-syllable similarity measure as follows:

\[
[\Omega_{\text{subsyl}i}^{\text{subsyl}j}] = \Phi_{\text{subsyl}i}(O, \lambda_j)
\]

\[
\Phi_{\text{subsyl}i}(O, \lambda_j) = \frac{1}{N_i} \sum_{o \in O} \left[-\log p(O_{o}, \lambda_j)\right]
\]

\[
\Phi_{\text{subsyl}i}(O, \lambda_j) = \frac{1}{N_i} \sum_{o \in O} \left[-\log \sum_{s_{o}, i_{o}} p(O_{o}, s_{o}, i_{o}) \lambda_j\right]
\]

\[
\Phi_{\text{subsyl}i}(O, \lambda_j) = \frac{1}{N_i} \sum_{o \in O} \left[-\log p(O_{o}, \hat{s}_{o}, \hat{i}_{o} | \lambda_j)\right]
\]

where \([\Omega_{\text{subsyl}i}^{\text{subsyl}j}]\) denotes the confusion matrix and \(\Phi_{\text{ subsyl}i}(\cdot)\) denotes the sub-syllable similarity measure function. \(O_{o}\) is the \(n^{\text{th}}\) sample in sub-syllable \(i\). \(\lambda_j\) denotes the MSD-HMM model parameters of sub-syllable \(j\). \(N_i\) is the number of occurrences of sub-syllable \(i\) in the training database. \(\hat{S}_{o}\) and \(\hat{I}_{o}\) is the optimal state sequence and mixture sequence for the \(n^{\text{th}}\) sample in sub-syllable \(i\), respectively.

Then the phone models are merged using the conventional agglomerative clustering algorithm based on the similarity measures. The phone models are divided into three classes: vowels, voiced consonants and unvoiced consonants in order to construct the MSD-HMM, since only the unvoiced consonants do not have the pitch information. Figure 2 shows an example of the merged phone cluster models for three speakers. In the phone cluster models, the voiced and unvoiced consonants have high degree of phone confusion and obtain a smaller number of phone clusters.

3.2. Decision Tree-based Phone Cluster Model

In general, for a small-sized training database, smaller number of phone models can eliminate the data sparseness problem. Nevertheless, the original contextual, phonetic and speaker characteristics cannot be preserved for speaker clustering if all the confusion phones and speakers are merged into a phone model. Therefore the contextual, phonetic and speaker characteristics can be modeled by the Gaussian models for different speakers and can be stored in the leaf nodes of a decision tree [3] for data sharing.

In this study, a 3-state left-to-right MSD-HMM with single Gaussian distribution is used to model each phone cluster. In the training phase DT-PCMs, the minimum likelihood gain is adopted as the stopping criterion for each splitting. Figure 3 illustrates an adaptive decision tree-based phone cluster model for MFCC and F0 features. The Gaussian models for each state of the phone cluster MSD-HMMs which contain the contextual and speaker information are characterized by a leaf node in a decision tree. Therefore, a speaker-related initial phone cluster model can be constructed by retrieving the Gaussian models from the DT-PCMs via the features from the recognition results and the input speech segment.

The definition of the question set for node splitting of decision tree construction is important. This study defines
four types of question features for DT-PCMs construction and retrieval.

1) The contextual feature in phone level: i.e. whether the preceding (left) phone, current phone or succeeding (right) phone in the speech segment belongs to a phone cluster or not.

2) The phone confusion feature. This study employs the top-N recognition candidates for a speech segment to extract the phone confusion features which can capture the phone confusion information, such as “if a phone model appears in top-N candidates.

3) The prosodic feature. This study utilizes the ratio of durations between two contiguous states in a phone as the prosodic feature. The value of the ratio is divided into 10 sub-intervals as the threshold of a question.

4) The cepstral and F0 features for Gaussian models in the leaf nodes. Since the recognition results and the original speech segment are used as the decision tree input, the mean vector of MFCCs and F0 features for each state is utilized to obtain the most similar Gaussian models which belong to the same leaf node but different training speakers.

4. Speaker clustering using DT-PCMs

The speaker clustering phase consists of four steps. The first step is audio segmentation. This step detects all the boundaries with speaker changes or foreground/background changes in the broadcast news audio files. The minimum description length (MDL)-based Gaussian model with multiple change-points window strategy [9] is employed to detect the audio changes. This approach can achieve a very low miss probability and a satisfactory false alarm rate so as to keep more speech data and less background audio in a speech segment. After detecting all audio changes, in the second step, the conventional speech activity detection is utilized to identify the speech segments among all audio segments and then obtain the phone sequence using a speech recognizer.

The third step is to adapt the initial phone cluster models, constructed based on the DT-PCMs, into a set of new speaker phone cluster models for the input speech segment using MLLR. Figure 4 illustrates the speaker phone cluster model adaptation. In Fig. 4, each speech segment is transcribed into a top-N phone lattice to retrieve the initial phone cluster models from DT-PCMs.

Speaker clustering is performed using the conventional agglomerative clustering algorithm based on the cross-likelihood ratio (CLR) [4]. The CLR measure is defined as follows. Given two clusters C_i and C_j, C_i comprises N_i speech segments: \{S_{seg}^i | 1 \leq i \leq N_i, i \in \mathbb{N}\} and C_j comprises N_j speech segments: \{S_{seg}^j | 1 \leq j \leq N_j, j \in \mathbb{N}\}. The similarity between clusters C_i and C_j with their own speaker-related phone cluster models can be obtained as follows

\[
\text{CLR}_{\text{rounded}}(C_i, C_j) = \frac{1}{N_i} \log \frac{L(S_{seg}^i | \theta)}{L(S_{seg}^i | \theta_{\text{new}})} - \frac{1}{N_j} \log \frac{L(S_{seg}^j | \theta)}{L(S_{seg}^j | \theta_{\text{new}})}
\]

where

\[
L(S_{seg}^i | \theta) = \prod_{k=1}^{N_i} p(S_{seg}^i_k | \theta), \quad p(S_{seg}^i_k | \theta) = \prod_{q=1}^{K} p(S_{subSyl}^i_k | \theta^{q=})
\]

\[
L(S_{seg}^j | \theta) = \prod_{l=1}^{N_j} p(S_{seg}^j_l | \theta), \quad p(S_{seg}^j_l | \theta) = \prod_{q=1}^{K} p(S_{subSyl}^j_l | \theta^{q=})
\]

\(\theta_{\text{new}}\) and \(\theta^{q=}\) denote the corresponding phone cluster model parameters of SubSyl_i and SubSyl_j, respectively. Furthermore, a threshold-based stopping criterion is utilized for speaker clustering, i.e. if the ratio between the current iteration and its preceding iteration is smaller than a threshold, the clustering process will be terminated. Therefore a tuning set for threshold tuning is required to determine the threshold for testing.

5. Experimental results

5.1. Training, tuning, test data and evaluation metric

For performance evaluation, MATBN (Mandarin Across Taiwan-Broadcast News) corpus collected by Academia Sinica [8] was adopted as the benchmark. MATBN corpus comprises 198 hours broadcast news audio files collected during 2001 to 2003 with transcribed textual contents and speaker identities. All the speech segments from the anchors and field reporters of the broadcast news audio files collected during 2001 and 2002 were extracted for training. The speech segments from twenty-five speakers were utilized to construct the universal phone models. Furthermore, one hour show: 20030306 was utilized for parameter tuning and another one hour show: 20030129 was for testing. The tuning data comprises 14 speakers and 77 speech segments, and the test data comprises 11 speakers and 74 speech segments. The average length of each speech segment is about 20s.

The evaluation metric for speaker clustering is the
average cluster purity \((acp)\) and average speaker purity \((asp)\) [1]. The equations of \(acp\) and \(asp\) are defined as follows:

\[
acp = \frac{1}{N} \sum_{j=1}^{R} p_j \times n_j = \frac{1}{N} \sum_{j=1}^{R} \left( \frac{\sum_{m=1}^{M} n_{mj}}{n_m} \right) \times n_j, \quad p_j = \frac{\sum_{m=1}^{M} n_{mj}^2}{\sum_{m=1}^{M} n_m}
\]

\[
asp = \frac{1}{N} \sum_{j=1}^{R} p_j \times n_j = \frac{1}{N} \sum_{j=1}^{R} \left( \frac{\sum_{m=1}^{M} n_{mj}}{n_j} \right) \times n_j, \quad p_j = \frac{\sum_{m=1}^{M} n_{mj}^2}{\sum_{m=1}^{M} n_j}
\]

where \(M\) denotes the number of clusters, \(R\) denotes the number of speakers and \(N\) denotes the number of speech segments. \(n_m\) represents the number of speech segments in cluster \(m\), and \(n_{mj}\) is the number of speech segments belonging to speaker \(j\) in cluster \(m\). Also, the cluster number \((C#)\) and a joint measure is estimated to give a single evaluation value by applying a geometric mean on \(acp\) and \(asp\) as follows:

\[
K = \sqrt{acp \times asp}
\]

5.2. Performance of speech recognizer and audio segmentation

The MATBN training database was adopted for speech recognizer construction. The bi-gram language model was trained with a 30k vocabulary dictionary on a newswire text corpus consisting of 20 million Chinese characters from a News website at the same time period. A syllable accuracy of 82% and word accuracy of 80% were achieved for ASR.

The speaker segmentation was tuned to obtain a low miss probability. A miss probability of 0.05 and a false alarm probability of 0.25 were achieved. In the segmentation results, most missed boundaries occurred if two contiguous speech segments uttered by a speaker but belonged to different stories. In contrast, most false alarm boundaries occurred during background change with the same speaker segment.

5.3. Comparison between GMM and DT-PCMs

For performance comparison, Figures 5 shows the tuning results of speaker clustering for GMM-based and DT-PCMs-based approaches. The best ratio of CLR between two contiguous iterations was used as the stopping parameter for testing.

Table 1 presents the evaluation results based on the best stopping parameter from the tuning set. By comparing the cluster number of the best clustered results and the true cluster number, the GMM-based approach has the under-clustering problem. In contrast, the DT-PCMs-based approach has slighter under-clustering problem and keeps higher \(acp\) than the GMM-based approach.

6. Conclusions

This study presents a model-based speaker clustering approach on broadcast news using DT-PCMs. The DT-PCMs is constructed based on the contextual, phone confusion and speaker characteristic information. Each phone is modeled by an MSD-HMM to model the MFCC and pitch features simultaneously. The DT-PCMs is used for initial phone cluster model construction. The MLLR method is then employed to adapt the initial models into the speaker-related phone cluster models for the input speech segment. Finally, for the input broadcast news, speaker clustering is performed by clustering all the corresponding adapted speaker-dependent phone cluster models based on the cross-likelihood ratio. The experimental result shows that the proposed DT-PCMs approach outperforms the conventional GMM-based approaches in speaker clustering.

7. References


