SPEAKER IDENTIFICATION USING MULTILINGUAL PHONE STRINGS

Qin Jin, Tanja Schultz, Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University
E-mail: {qjin, tanja, ahw}@cs.cmu.edu

ABSTRACT

Far-field speaker identification is very challenging since varying recording conditions often result in unmatching training and test situations. Although the widely used Gaussian Mixture Models (GMM) approach achieves reasonable good results when training and testing conditions match, its performance degrades dramatically under non-matching conditions. In this paper we propose a new approach for far-field speaker identification: the usage of multilingual phone strings derived from recognizers in eight different languages. The experiments are carried out on a database of 30 speakers recorded with eight different microphone distances. The results show that the multilingual phone string approach is robust against non-matching conditions and significantly outperforms the GMMs. On 10-second test chunks, the average closed-set identification performance achieves 96.7% on variable distance data.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing a speaker by machines using the speaker's voice. It can operate in two modes: identifying a parti-cular speaker or verifying a speaker's claimed identity [1]. Furthermore, speaker recognition can be subdivided into closed-set and open-set problems [2], depending on whether the set of speakers is known or not. It can also be text-dependent or text-independent. In this paper closed-set text-independent speaker identification is considered.

The techniques developed for text-independent speaker identification include Nearest Neighbor, Vector Quantization, discriminative Neural Networks and Gaussian Mixture Models [3]. Nowadays, the latter is the most widely and successfully used method for speaker identification. However, for the use of speaker identification in real world applications, some challenging problems needed to be solved. Among them is the robust identification of speakers in far field. Although GMM has been applied successfully to closed-speaking microphone scenarios under matching training and testing conditions, its performance degrades dramatically under unmatching conditions. In this paper, we propose a new approach, which is based on the idea of using multilingual phone strings as input feature for speaker identification. By using phone strings, we expect to model the pronunciation idiosyncrasy of a speaker. The phone strings are decoded applying phone recognizers from eight different languages. By using multiple languages for decoding, we expect to obtain more robust and language independent speaker identification. Two variations of this approach are compared to the traditional acoustic feature GMM. Results are given for matching and unmatching conditions using data recorded on variable distances. The remaining paper is organized as follows: the next section describes the database used for carrying out all experiments. After a brief repetition of GMMs in section 3, the multilingual phone string approach is introduced in section 4. Section 5 gives an overview of the experiments and results before section 6 summarizes and concludes the paper.

2. DATABASE DESCRIPTION

Real-world applications are expected to work under unmatching circumstances, i.e. the testing conditions e.g. in terms of microphone distances might be quite different from what had been seen during training. Therefore, methods for robust speaker identification under various distances needed to be explored. For this purpose a database containing speech recorded from various microphone distances had been collected at the Interactive Systems Laboratories. The database contains 30 speakers in total. From each speaker five sessions had been recorded where the speaker sits at a table in an office environment, reading an article, which is different for each session. Each session is recorded using eight microphones in parallel: one closed-speaking microphone (Sennheizer headset), one Lapel microphone worn by the speaker, and six other Lapel microphones. The latter six are attached to microphone stands sitting on the table, at distances of 1 foot, 2 feet, 4 feet, 5 feet, 6 feet and 8 feet to the speaker, respectively. Tables and graphs shown in this paper use "Dis 0" to represent closed-speaking microphone distance data, and "Dis n" (n>0) to refer to the n-feet distance data.

The data of the first four sessions, together 7 minutes of spoken speech (about 5000 phones) are used for training the multilingual phone string approach, whereas only one minute of the first session was used as training data for the GMM approach. Testing was carried out on the remaining fifth session adding up to one minute of spoken speech (about 1000 phones). The GMM approach was tested only on 10-second chunks, whereas the phone string approach was also tested on longer and shorter chunks.

3. GAUSSIAN MIXTURE MODELS APPROACH

The GMM approach has been widely studied and used in speaker recognition tasks [3]. A multi-variate GMM density, $P(\vec{x}|\lambda)$, is a weighted sum of uni-modal multi-variate Gaussian density $P(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i p(\vec{x}|\lambda_i)$, where λ_i is the parameter set of one Gaussian $\{\mu_i, \Sigma_i\}$ and M is the number of mixture of components. 13-dimension LPC cepstra are used as speaker's feature vectors and are clustered into 32 centers using K-means. These centers are used to initialize the Gaussian mixture centers. We use EM algorithm to produce the most likely estimates of mean vectors, covariance matrices and mixture weights. In the recognition stage, the unknown speaker is identified as speaker J if: $J = \arg\max_{j} \sum_{t=i}^{T} \log P\left(\vec{x}_t \middle| \lambda^j\right)$. T refers to the number of feature vectors in the training speech and λ^j is the GMM of speaker j.

Test \ Train	Dis 0	Dis 1	Dis 2	Dis 6
Dis 0	**100**	43.3	30	26.7
Dis 1	56.7	**90**	76.7	40
Dis 2	56.7	63.3	**93.3**	53.3
Dis 6	40	30	60	**83.3**

Table 1: SID rate (% correct)

Table 1 shows the GMM Speaker IDentification Rate in percentage correct for matching and non-matching distance conditions in training and testing. Under matching conditions (numbers are given in bold) the GMM approach achieves reasonable good results, however under non-matching conditions the performance degrades dramatically. We conclude from these results that the GMM approach lacks robustness in the case where the models are tested on distances, which are not covered from the training data.

4. MULTILINGUAL PHONE STRING APPROACH

Phone recognition and n-gram modeling has been successfully used for language identification [4,5] in the past, whereas its application to speaker identification is introduced very recently [6]. In this paper we extend the approach proposed in [6] to tackle the non-matching distance and channel conditions. Furthermore, we introduce two different methods based on the multilingual phone string approach and compare these to the GMM approach.

The basic idea of the multilingual phone string approach is to take phone strings decoded by phone recognizers of several different languages as features instead of using the conventional acoustic feature vectors. Throughout this experiments we applied phone recognizers of eight different languages. By using information derived from phone strings, we expect to cover speaker dependent idiosyncrasy of pronunciation. We expect features derived from the pronunciation idiosyncrasy to be more robust against non-matching conditions than acoustic features. Furthermore we aim to increase the robustness by providing supplementary information from eight different languages.

4.1. Phone Recognizer in eight Languages

The experiments are based on phone recognition engines built in the eight languages: Mandarin Chinese (CH), Croatian (KR), German (DE), French (FR), Japanese (JA), Portuguese (PO), Spanish (SP), and Turkish (TU). For each language, the acoustic model consists of a 3-state HMM per phone with a mixture of 128 Gaussian components per state. The Gaussians are on 13 Mel-scale cepstral coefficients with first and second order derivatives, power, and zero crossing rate. After cepstral mean subtraction a linear discriminant analysis reduces the input vector to 32 dimensions. All engines are trained and evaluated in the framework of the *GlobalPhone* project, which provides 15 to 20 hours word-level transcribed training data per language [7]. Table 2 shows the number of phones per language and the resulting Phone Error Rates on each language. See [7] for further details.

Language	Phones	PER	Language	Phones	PER
CH	137	48.8	KR	41	41.1
DE	43	46.1	PO	46	45.0
FR	38	46.7	SP	40	33.0
JA	31	32.6	TU	29	42.8

Table 2: Phone error rate (PER %) for eight languages

4.2 Phone Language Model Training

For the following experiments we trained Phone Language Models (PLM) for each training speaker as showed in figure 1 for speaker J. The label L1 PR in figure 1 refers to the phone recognizer of language No.1, and L8 PR refers to the phone recognizer of language No.8. The training data of speaker J is decoded by the phone recognizers of each language to produces sequences of phone strings. The n-gram phone language model PLM L1 for speaker J is created from the phone sequence of all training utterances spoken by speaker J decoded by the phone recognizer of language L1.

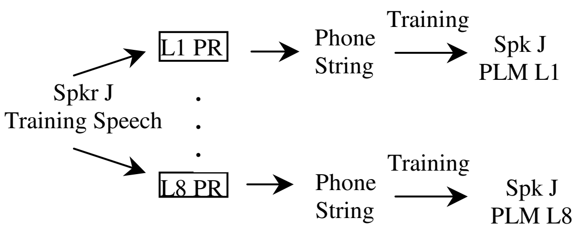

Figure 1: Diagram of training the Phone Language Model

We present two multilingual phone string approaches named SID-MPLM and SID-SPMPLM, respectively. Both will be explained in detail in the following sub-sections. These approaches have the above described phone language model training step in common. The difference between SID-MPLM and SID-SPMPLM is how the PLM of each speaker is applied.

4.3. SID-MPLM

The PLM of each speaker, which was trained as explained in figure 1, is now used to determine the identity of a speaker. Figure 2 shows for the SID-MPLM (Speaker IDentification using Multilingual Phone Language Model) approach how the incoming test speech of an unknown speaker is processed by the PLM of speaker J.

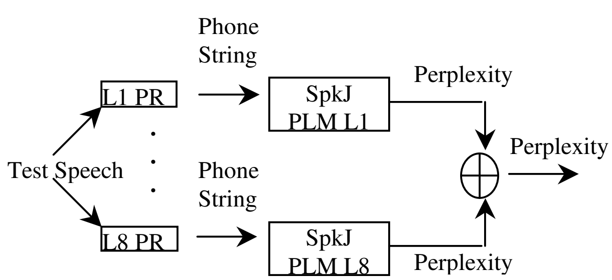

Figure 2: Block Diagram of SID-MPLM

Firstly, the phone recognizers of eight languages decode the test speech and produce eight phone strings, one per language. Secondly, these phone strings are fed into the speakers' PLM of the corresponding language to calculate the perplexities. This process results in eight perplexities (one per language) for each speaker. In the third step these eight perplexities are interpolated to build a final perplexity for each speaker. The training speaker, which produces the lowest perplexity, is identified as the test speaker. In our experiments we used trigram PLMs and equal weight linear interpolation.

4.4. SID-SPMPLM

In the SID-MPLM approach, both training and test data are decoded using equal distribution phone language model. The speakers PLM is then used to compute the perplexity of test data. The idea for the SID-SPMPLM approach is to use the speaker-dependent PLM directly to decode the test speech. The underlying assumption is, that a speaker achieves a lower decoding distance score on a matching PLM than for a non-matching PLM. In other words, the training step in the SID-SPMPLM approach is identical to the one in the SID-MPLM approach, but the testing step differs: for the SID-SPMPLM approach the test data is decoded multiple times using one speaker-dependent PLM each time. Thus in our experiments, test data will be decoded 30 times for each language, each time with one speaker's PLM. We use an equal weight linear interpolation scheme to combine the decoding scores from all languages. The speaker to which the PLM belongs, which produces the lowest interpolated decoding distance score, is hypothesized.

5. EXPERIMENTS AND RESULTS

5.1. SID-MPLM performance

Language	60s	40s	10s	5s	3s
CH	100	100	56.7	40	26.7
DE	80	76.7	50	33.3	26.7
FR	70	56.7	46.7	16.7	13.3
JA	30	30	36.7	26.7	16.7
KR	40	33.3	30	26.7	36.7
PO	76.7	66.7	33.3	20	10
SP	70	56.7	30	20	16.7
TU	53.3	50	30	16.7	20
Int. of all LM	**96.7**	**96.7**	**96.7**	**93.3**	**80**

Table 3: SID Rate with different test length at Dis 0

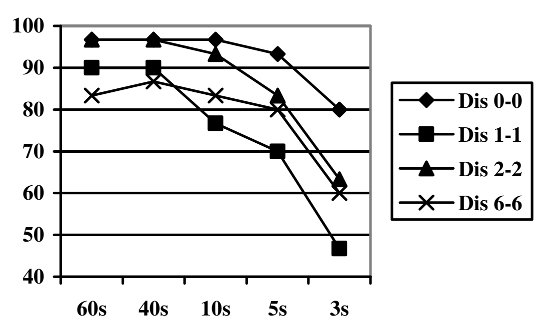

Figure 2: SID Rate with different test length (seconds)

Table 3 shows the identification accuracy of SID-MPLM approach with decreasing test utterance length when both testing and training distance is Dis0. As the test utterance becomes shorter, the combination of all languages becomes more important. Figure 2 shows the identification accuracy of combining all languages at different distances. These are the results under matched conditions (Dis n-n manes both training data and test data are from distance n feet). On 10 seconds test, the performance is comparable to GMM.

5.2. SID-SPMPLM results

Language	(% correct)	Language	(% correct)
CH	53.3	KR	26.7
DE	40	PO	30
FR	23.3	SP	26.7
JA	26.7	TU	36.7
Int. of all LM		**60**	

Table 4: SID rate (% correct) of SID-SPMPLM

SID_SPMPLM is more expensive than SID-MPLM. But the performance is not as good. One reason is that speaker's PLM is not fully trained. We train the speaker's PLM using around 8-minute speech. This amount of data is not enough for training a good PLM. So we will also try this idea on Switch Board data on a small set of speakers.

5.3. Matched Condition

Language	Dis0-0	Dis1-1	Dis2-2	Dis6-6
Int. of all LM	96.7	90	96.7	83.3

Table 5: SID rate (% correct) of SID-MPLM under matched conditions

Test-train distance	Dis1-1	Dis1-2	Dis1-0
Int. of all LM	90	80	50

Table 6: SID rate (% correct) of SID-MPLM under unmatched conditions

Test distance	Dis1	Dis2	Dis6
Int. of all LM	96.7	96.7	83.3

Table 7: SID rate of SID-MPLM under unmatched conditions with combination of PLM at all distances

From table 5, we can see the performance is good under matched conditions. However under unmatched conditions, if we only use the phone model of unmatched distance, the performance degrades as it is showed in table 6. But if we combine all phone models at difference distances, we can make up the loss as shown in table 7.

6. CONCLUSIONS

In this paper we described two speaker identification approaches using phone strings decoded by multiple language phone recognizers and evaluated them on variable distance data. The experiments results show that the SID-MPLM approach is robust under unmatched conditions and outperforms GMM. The reason behind the multilingual phone string idea is that we expect phone strings can capture the pronunciation idiosyncrasy of speakers. And this feature will be robust over different conditions. The experiments results indicate that phone string is an appropriate feature for speakers. And by using multilingual phone strings we obtain robustness and language independence. Furthermore we think speaker's pronunciation idiosyncrasy will be more dominant in spontaneous speech. Thus multilingual phone strings will capture this feature more efficiently in spontaneous speech. So next step we will try these two approaches on spontaneous data.

7. ACKNOLEGDMENTS

8. REFERENCES

[1] Campbell, Joseph P., Jr., "Speaker Recognition: A Tutorial", Proceeding of the IEEE, IEEE, vol. 85, no. 9, pp 1437-62, Sept. 1997.

[2] Herbert Gish and Michael Schmidt, "Text-Independent Speaker Identification", IEEE Signal Processing Magazine, IEEE, pp 1437-62, Oct. 1994.

[3] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Volume 3, No. 1, January 1995.

[4] Marc A. Zissman and Elliot Stinger, "Automatic Language Identification of Telephone Speech Messages Using Phone Recognition and N-gram Modeling", Proceedings of IEEE ICASSP, Volume 1, pp 305-308, Minneapolis, USA, 1994.

[5] Marc A. Zissman, "Language Identification Using Phone Recognition and Phonotactic Language Modeling", Proceedings of IEEE ICASSP, Volume 5, pp 3503-3506, Detroit, MI, May 1995.

[6] Mary A. Kohler, Walter D. Andrews, Joseph P. Compbell, Jaime Hernandez-Cordero, "Phonetic Refraction for Speaker Recognition", Proceedings of Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark, September 2001.

[7] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", Speech Communication, Volume 35, Issue 1-2, pp 31-51, August 2001.