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 un-matching training and testing 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 un-matching conditions. In this paper we propose a new approach for far-field speaker identification: the usage of multilingual phone strings derived from phone 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 multi-lingual phone string approach is robust against un-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 particular 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 need 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 un-matching 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 un-matching 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 need to be explored. For this purpose a database containing speech recorded from microphones at various 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 components. 13-dimension LPC cepstra are used as feature vectors and 32 centers clustered using K-means 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: $J=\arg\max_j \sum_{t=i}^{T} \log P(\vec{x}_t|\lambda^j)$. 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) of GMM

Table 1 shows the GMM Speaker IDentification Rate in percentage correct for matching and un-matching distance conditions in training and testing. Under matching conditions (numbers are given in bold) the GMM approach achieves reasonable good results, however under un-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 un-matching distance and channel conditions. Furthermore, we introduce two different methods based on multilingual phone strings 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 the 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 un-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 produce 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.

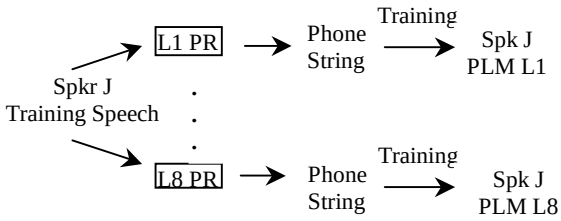

Figure 1: Diagram of training the Phone Language Model

We present two multilingual phone string approaches named MPLM-pp and MPLM-dec, 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 MPLM-pp and MPLM-dec is how the PLM of each speaker is applied.

4.3. MPLM-pp

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 how the incoming testing speech of an unknown speaker is processed by the PLM of speaker J in MPLM-pp approach (Multilingual Phone Language Model used for perplexity calculation).

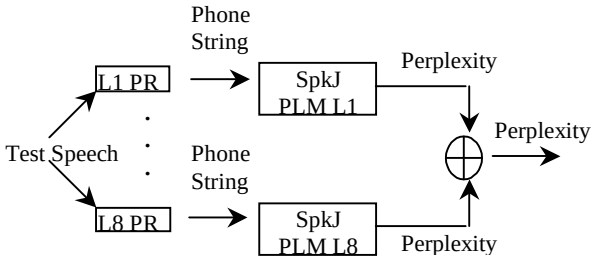

Figure 2: Block Diagram of MPLM-pp

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. MPLM-dec

In the MPLM-pp approach, both training and test data are decoded using equal distribution phone language model. The speaker's PLM is then used to compute the perplexity of testing data. The idea for the MPLM-dec 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 un-matching PLM. In other words, the training step in the MPLM-dec approach is identical to the one in the MPLM-pp approach, but the testing step differs: for the MPLM-dec approach the testing data is decoded multiple times using one speaker-dependent PLM each time. Thus in our experiments, the 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 training speaker who has the PLM, which produces the lowest interpolated decoding distance score, is hypothesized as the identified speaker.

5. EXPERIMENTS AND RESULTS

5.1. MPLM-pp results

Table 3 shows the identification accuracy of MPLM-pp approach at different test utterance length for the matching condition, where both testing and training are recorded at distance Dis0.

Language	500s	50s	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 tested at different test length on Dis 0

With decreasing test utterance length, the performance based on a single language gets very low, however this can be overcome by using the multilingual information derived from all eight languages. After a linear combination of all languages the SID performance clearly outperforms the one on single language. Figure 3 shows the identification accuracy of combining all languages on data recorded at different distances. These are the results under matching conditions (Dis n-m refers to training on distance n and testing on distance m data). On 10 seconds test chunks, the performance is comparable to GMMs.

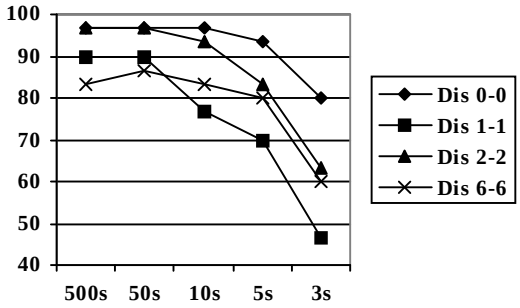

Figure 3: SID Rate over test length for various distances

5.2. MPLM-dec results

MPLM-dec is far more expensive than MPLM-pp, however the performance is worse. One reason might be that the speaker's PLM is undertrained, since we only used about 7 minutes of speech to train it. This amount of data is not enough for an accurate estimation of the speaker's PLM. Therefore, we are planning for future experiments to focus on larger training data, such as SwitchBoard.

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 MPLM-dec

5.3. Matching vs. un-matching condition results

Table 5 shows high performances for MPLM-pp on matching conditions. However table 6 shows the degraded performance for the un-matching case, if only phone language models of un-matching distance are applied. If the phone language models are combined at all distances, the degradation can be compensated as shown in table 7.

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

Table 5: SID rate of MPLM-pp on matching conditions

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

Table 6: SID rate of MPLM-pp on un-matching conditions

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

Table 7: SID rate of MPLM-pp on un-matching conditions with combination of PLM at all distances

6. CONCLUSIONS

In this paper we described two speaker identification approaches using multilingual phone strings and compared them to GMMs. Phone strings capture the pronunciation idiosyncrasy of speakers and are expected to be appropriate features, which are more robust under different conditions. The evaluation on variable distance data proved the robustness of the phone string approach, achieving 96.7% speaker identification accuracy on 30 speakers under un-matching conditions, which clearly outperformed GMMs. The proposed multilingual phone string approach has the additional benefit of being language independent. Furthermore we expect speaker's pronunciation idiosyncrasy to be even more dominant in spontaneous speech. Our future research will therefore investigate multilingual phone strings for speaker identification in spontaneous speech.

7. REFERENCES

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

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

[3] D. A. Reynolds and R. 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] M. A. Zissman and E. Singer, "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] M. A. Zissman, "Language Identification Using Phone Recognition and Phonotactic Language Modeling", Proceedings of ICASSP, Volume 5, pp 3503-3506, Detroit, MI, May 1995.

[6] M. A. Kohler, W. D. Andrews, J. P. Compbell, J. 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.