PHONETIC SPEAKER IDENTIFICATION

Qin Jin, Tanja Schultz, Alex Waibel

Interactive Systems Laboratory
Carnegie Mellon University
{qjin, tanja, ahw}@cs.cmu.edu

ABSTRACT

This paper describes the exploration of text-independent speaker identification using novel approaches based on speakers' phonetic features instead of traditional acoustic features. Different phonetic speaker identification approaches are discussed in this paper and evaluated using two speaker identification systems: one multilingual system and one single language multiple-engine system. Furthermore, text-independent speaker identification experiments are carried out on a distant-microphone database as well as gender identification experiments are investigated on the NIST 1999 Speaker Recognition Evaluation dataset. The results show that phonetic features are powerful for speaker identification and gender identification.

1. INTRODUCTION

Speaker identification is the process of automatically recognizing a speaker by machine using the speaker's voice [1]. It has developed into an increasing more important speech processing technique by providing useful information for speech analysis. The state of the art in speaker identification is based almost exclusively on traditional short-term acoustic features. But there are other levels of rich information capturing speaker characteristics that have not been sufficiently explored, such as pronunciation idiosyncrasy, idiolectal word usage and speaking style, etc. Recently, Kohler, Andrews and Doddington have tried novel approaches for speaker recognition using phonetic and word idiolect features [2][3][4]. In this paper, we discuss our exploration of speaker identification based on phonetic features.

The basic idea of phonetic speaker identification is to identify a speaker using phonetic sequences derived from that speaker's utterance. Although the phonetic sequences are produced using acoustic features, the identification decision is made based solely on the phonetic sequences. The assumption behind the phonetic approach is that phonetic sequences can cover a speaker's idiosyncratic pronunciation. Two novel phonetic speaker identification approaches are described in this paper.

2. PHONETIC SPEAKER IDENTIFICATION

We first developed a speaker identification system using phonetic sequences from phone recognizers trained on multiple languages. We call this our multilingual system. This system uses phonetic sequences produced by context-independent phone recognizers from multiple languages instead of traditional short-term acoustic vectors [5][6]. Since this information comes from complementary phone recognizers, we anticipate greater robustness. Furthermore, this approach is somewhat language independent since the recognizers are trained on data from different languages. We also developed a speaker identification system using phonetic sequences produced by single language multiple phone recognizers, which we call our multi-engine system. This system uses phonetics sequences produced by three different context-independent English phone recognizers.

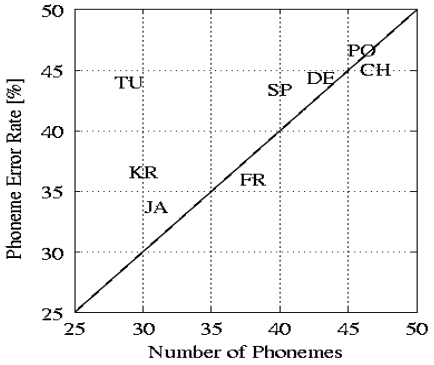

Figure 1: Error rate vs number of phones

2.1. Phone Recognition

In this paper, the multilingual system uses phone recognizers built in eight languages: Mandarin Chinese (CH), German (DE), French (FR), Japanese (JA), Croatian (KR), Portuguese (PO), Spanish (SP) and Turkish (TU). All the phone recognizers are trained and evaluated in the framework of the *GlobalPhone* project. Figure 1 shows phone error rates per language in relation to the number of modeled phones. See [7] for further details.

2.2. Phonetic Language Model (PLM) Training

In identifying an unknown speaker as one of the N target speakers, we need a speaker-dependent Phonetic Language Model (PLM) for each target speaker in each language. In this paper, we use $PLM_{i,j}$ to represent the phonetic language model for speaker *j* in language *i*. Figure 2 shows the procedure of training PLMs for speaker *j*. M phone recognizers $\{PR_1, \ldots, PR_M\}$ decode the training data of speaker *j* to produce M phonetic sequences. From these M phonetic sequences, M PLMs are created for speaker *j*, one for each language. During the decoding of training data, each PR_i uses an equiprobable phone language model. Thus the produced phonetic sequence is

based solely on a phone recognizer's acoustic model. This procedure does not require transcription at any level.

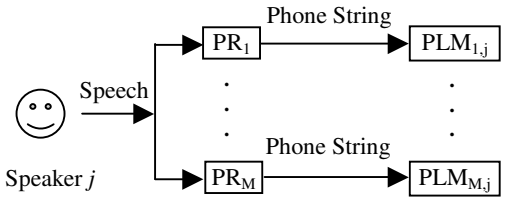

Figure 2: Training Phonetic Language Model for One Speaker

We present two phonetic speaker identification approaches which we call PLM-pp and PLM-score. These two approaches have the above described phonetic language model training step in common. The difference between PLM-pp and PLM-score is how the PLM of each speaker is applied during the identification.

2.3. PLM-pp Speaker Identification

The PLM of each speaker, which was trained as explained in Figure 2, is now used to determine the identity of an unknown speaker. In PLM-pp, each of the M phone recognizers PR_i , as used for PLM training, decodes the test speech and produces a phonetic sequence which is scored against each of N speaker-dependent phonetic language models $PLM_{i,j}$. This results in a perplexity matrix **PP**, whose $PP_{i,j}$ element is the perplexity produced by phonetic language model $PLM_{i,j}$ on the phonetic sequence output from phone recognizer PR_i. Our decision rule is to identify an unknown speaker as speaker *j** given by

$$j^* = \underset{j}{Min}\left(\sum_{i=1}^{8} PP_{i,j}\right).$$

2.4. PLM-score Speaker Identification

In the PLM-pp approach, both training and test data are decoded by M phone recognizers PR_i using equiprobable phone language models. When decoding a test utterance in PLM-score, however, we replace the equiprobable phone language model used during PLMs training with each of the N speaker-dependent $PLM_{i,j}$ in turn. The test utterance is therefore decoded by each of the M phone recognizers N times, resulting in a decoding score matrix **SC**, whose $SC_{i,j}$ element is the decoding score produced jointly by phone recognizer PR_i and phonetic language model $PLM_{i,j}$ during decoding. The same decision rule is applied as in PLM-pp. The key idea for the PLM-score approach is to use the speaker-dependent $PLM_{i,j}$ directly to decode the test speech. The underlying assumption is that a speaker achieves a lower decoding distance score on a matched PLM than for a mismatched PLM. The disadvantage of the PLM-score approach is that the test utterance will be decoded M*N times as opposed to M times for PLM-pp. Furthermore, the success of this approach relies on the ability to produce reliable phonetic models from the training data.

3. MULTILINGUAL VS. MULTI-ENGINE

3.1. Distant –Microphone Database

In order to explore methods for robust speaker identification under various distances, a distant-microphone database containing speech recorded from various microphone distances has been collected at the Interactive Systems Laboratory. The database contains 30 native English speakers in total. For each speaker, five sessions have been recorded with the speaker sitting 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 close-speaking microphone (Sennheizer headset, Dis 0), one lapel microphone (Dis L) worn by the speaker, and six other lapel microphones which are attached to microphone stands sitting on the table at distances of 1, 2, 4, 5, 6 and 8 feet from the speaker. The data of the first four sessions, together 7 minutes of spoken speech (about 5000 phones), are used for training the PLMs and the remaining fifth session adding up to one minute of spoken speech (about 1000 phones) is used for testing. The PLM-pp approach was also tested on longer and shorter chunks.

3.2. Multilingual PLM-pp Speaker Identification

Test Length	60s	40s	10s	5s
Dis 0	96.7	96.7	96.7	93.3
Dis L	96.7	96.7	86.7	70
Dis 1	90	90	76.7	70
Dis 2	96.7	96.7	93.3	83.3
Dis 4	96.7	93.3	80	76.7
Dis 5	93.3	93.3	90	76.7
Dis 6	83.3	86.7	83.3	80
Dis 8	93.3	93.3	86.7	66.7

Table 1: PLM-pp SID rate on varying test lengths at matched training and testing distances

Test Length	60s	40s	10s	5s
Dis 0	96.7	96.7	96.7	90
Dis L	96.7	100	90	66.7
Dis 1	93.3	93.3	80	70
Dis 2	96.7	96.7	86.7	80
Dis 4	96.7	96.7	93.3	80
Dis 5	93.3	93.3	86.7	70
Dis 6	93.3	86.7	83.3	60
Dis 8	93.3	93.3	86.7	70

Table 2: PLM-pp SID rate on varying test lengths at mismatched training and testing distances

Table 1 and 2 compare the multilingual PLM-pp identification results for all distances on different test utterance lengths under matched and mismatched conditions, respectively. Under matched conditions, training and testing data are from the same distance. Under mismatched conditions, we do not know the test speech distance; we make use of all D sets of phonetic language models ($PLM_{i,j}$), where D is the number of distances (D = 8 in this paper), and modify our decision rule to make decision by selecting $j^* = \underset{j}{Min}\left(\underset{k}{Min}\sum_{i=1}^{8} PLM_{i,j,k}\right)$, where *i* is the index over phone recognizers, *j* is the index over speaker, and $k \in \{1,2,\cdots,D\}$. These two tables indicate that the

performance of PLM-pp is comparable under matched and mismatched conditions.

3.3. Multi-Engine PLM-pp Speaker Identification

In order to investigate whether the reason for the success of the multilingual PLM-pp approach is related to the fact that different languages contribute useful information or that it simply lies in the fact that different recognizers provide complementary information, we conducted the following set of experiments. We replaced the eight multilingual phone recognizers with three English phone recognizers which were trained on very different conditions, namely: Switchboard (telephone, highly conversational), Broadcast News (various channel conditions, planed speech), and Verbmobil English (high quality, spontaneous). For a fair comparison between the three English engines and the eight multilingual engines, we generated all possible language triples out of the set of eight languages (56 triples) and calculated the average, minimum and maximum performance for each. Table 3 compares the results of the multilingual system to the multi-engine system. The results show that the best performance of the multilingual triples always outperforms the performance of the multi-engine triple. From these results we draw the conclusion that multiple English phone recognizers provide less useful information for the classification task than do multiple language phone recognizers. This is at least true for our given choice of multiple English engines in the context of speaker identification. The multiple languages have the additional benefit of being language independent. This results from the fact that the actual spoken language is not covered by the used multiple language phone recognizers.

Approach	Multilingual	Multi-Engine
Dis 0	87.92 (66.7 – 100)	93.3
Dis L	88.21 (63.3 – 96.7)	86.7
Dis 1	83.57 (66.7 – 93.3)	86.7
Dis 2	93.63 (86.7 – 96.7)	76.7
Dis 4	81.43 (56.7 – 96.7)	86.7
Dis 5	86.07 (66.7 – 96.7)	83.3
Dis 6	81.96 (66.7 – 93.3)	63.3
Dis 8	87.14 (63.3 – 93.3)	63.3

Table 3: Comparison of SID rate using PLM-pp by Multilingual triples and Multi-Engine

3.4. Combination of Multilingual and Multi-Engine System

In order to investigate whether combining the multilingual system and the multi-engine system can provide more improvement for the speaker identification task, we conducted a second set of experiments. Table 4 compares the speaker identification performance of using the multilingual system alone with that of combining the multilingual system with all the three multiple English phone recognizers as well as with each of the three English phone recognizers. The combination is made as adding more languages to the multiple languages and the same lowest perplexity decision rule is applied. In Table 4, we use ML to represent the multilingual system and ME to represent the multi-engine system. SWB, BN and VE are used to represent single English phone recognizer trained on Switchboard, Broadcast News and Verbmobil English respectively. The results indicate that the interpolation of multilingual and multi-engine doesn't give any further improvement. But we cannot conclude from these results that adding English language won't provide more complimentary information for speaker identification, since the three English phone recognizers are trained differently from those 8 language phone recognizers. To clarify this question, further investigation needs to done.

	ML	ML+ME	ML+SWB	ML+BN	ML+VE
Dis 0	96.7	93.3	93.3	93.3	93.3
Dis L	96.7	96.7	96.7	93.3	96.7
Dis 1	93.3	90.0	90.0	90.0	90.0
Dis 2	96.7	96.7	96.7	96.7	96.7
Dis 4	96.7	93.3	93.3	93.3	93.3
Dis 5	93.3	93.3	93.3	93.3	93.3
Dis 6	93.3	80.0	80.0	83.3	83.3
Dis 8	93.3	90.0	90.0	93.3	93.3

Table 4: Comparison of SID rate of Multilingual system and combination of Multilingual with Multi-Engine

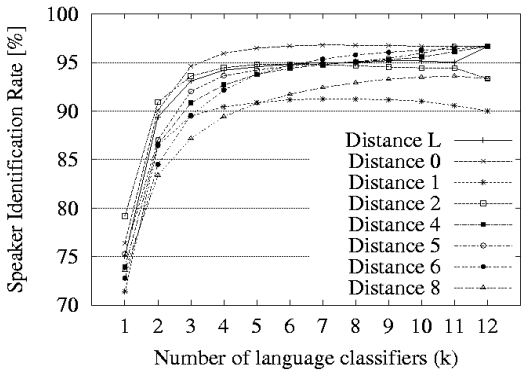

Figure 3: Speaker Identification rate vs number of phone recognizers

3.5. Number of Languages vs. Identification Performance

In a third set of experiments, we investigated the influence of the number of phone recognizers on speaker identification performance. These experiments were performed on an improved version of our phone recognizers in 12 languages trained on the above described GlobalPhone data (Arabic(AR), Korean(KO), Russian(RU) and Swedish(SW) are available in this version in addition to the 8 languages named in section 2.1). Figure 3 plots the speaker identification rate over the number k of languages used in the identification process on matched 60-second data for all distances. The performance is given in average and range over the k out of 12 language k-tuples. Figure 3 indicates that the average speaker identification rate increases with the number of involved phone recognizers. It also shows that the maximum performance of 96.7% can already be achieved using only two languages; in fact two (2 out of 12 = 66) language pairs gave optimal results: CH-KO, and CH-SP. However, the lack of a strategy for finding the best

suitable language pair does not make this very helpful. On the other hand, the increasing average indicates that the probability of finding a suitable language-tuple that optimizes performance increases with the number of available languages. While only 4.5% of all 2-tuples achieved best performance, as many as 35% of all 4-tuples, 60% of all 6-tuples, 76% of all 8-tuples and 88% of all 10 tuples were likewise found to perform optimally in this sense.

4. PLM-$_{PP}$ VS. PLM-$_{SCORE}$

Approach	PLM-pp (%)	PLM-score(%)
CH	100	53.3
DE	80	40
FR	70	23.3
JA	30	26.7
KR	40	26.7
PO	76.7	30
SP	70	26.7
TU	53.3	26.7
Int. of all Language	**96.7**	**60**

Table 5: Comparison of SID rate using PLM-pp and PLM-score on distance data

Table 5 compares the performance of the PLM-score approach at Dis0 under matched conditions with that of PLM-pp on 60-second test data. Even though PLM-score is far more expensive than PLM-pp, its performance (60%) is much worse than PLM-pp (96.7%). The poor performance of PLM-score seems to support the assumption made earlier that the phonetic language models we produced, which perform well within the PLM-pp framework, are not sufficiently reliable to be used during decoding as required by PLM-score.

Approach	PLM-pp (%)	PLM-score(%)
CH	88.5	89.5
DE	89.5	88.5
FR	89	91
JA	86.5	89
KR	87.5	88
PO	89	91.5
SP	92	92
TU	90	89
Int. of all Language	**94**	**94**

Table 6: Comparison of SID rate using PLM-pp and PLM-score for gender identification on SWB data

In order to test the performance of the PLM-score approach when enough data for training a reliable PLM is given, we conducted the following experiments. We used NIST 1999 speaker recognition evaluation dataset. There are a total of 309 female and 230 male target speakers. For each target speaker there are two minutes of training speech with each minute from one telephone channel type and one-minute test speech of unknown channel type. Although two minutes of speech are far from enough to train a reliable PLM for each target speaker, we have enough data to train the PLM for each gender. We conducted gender identification using both the PLM-pp and PLM-score approaches. We randomly choose 200 test trials containing 100 females and 100 males. The results in Table 6 indicate that given enough training data from which we can get a reliable phonetic language model, the PLM-pp and PLM-score produce comparable results. The conditions for which PLM-score is likely to perform well are under investigation.

5. CONCLUSIONS

We have investigated speaker identification based on phonetic information extracted from the speakers' utterances. We described two different phonetic speaker identification approaches. Our results are very encouraging and indicate that phonetic features as captured by a phonetic language model are very powerful for discriminating between speakers. Our evaluation on the distant microphone database proved the robustness of this novel approach.

Both of the described approaches achieved a good identification performance under the condition that enough training data are available. However, it needs to be clarified in further experiments what amount of data is necessary for training reliable speaker-dependent phonetic models.

Furthermore, the question about what information other than acoustic features are appropriate to extract from speech data to solve the task of speaker identification are not sufficiently studied yet. Features such as speaking style, pronunciation idiosyncrasy, word idiolect etc. are natural features that are used by human to discriminate speakers but have not been efficiently explored in speaker recognition community so far. Phonetic speaker identification is one initial investigation in this direction. More investigation about this higher-level rich speaker information is under exploration.

6. REFERENCES

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

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

[3] Walter D. Andrews, Mary A. Kohler, Joseph P. Campbell and John J. Godfrey, "Phonetic, Idiolectal, and Acoustic Speaker Recognition", Proceedings of Odyssey Workshop 2001.

[4] George Doddington, "Some Experiments on Idiolectal Differences Among Speakers, http://www.nist.gov/speech/tests/spk/2001/doc/, January 2001.

[5] Qin Jin, Tanja Schultz, Alex Waibel, "Speaker Identification Using Multilingual Phone Strings", Proceedings of IEEE ICASSP, Orlando, Florida, 2002.

[6] Tanja Schultz, Qin Jin, Kornel Laskowski, Alicia Tribble, Alex Waibel, "Speaker, Accent and Language Identification Using Multilingual Phone Strings", Proceedings of Human Language Technologies Conference, San Diego, California, 2002.

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