SPEAKER, ACCENT, AND LANGUAGE IDENTIFICATION USING MULTILINGUAL PHONE STRINGS

Tanja Schultz, Qin Jin, Kornel Laskowski, Alicia Tribble, Alex Waibel

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

1. INTRODUCTION

The identification of an utterance's non-verbal cues, such as speaker, accent and language, can provide useful information for speech analysis. In this paper we investigate far-field speaker identification, as well as accent and language identification, using multilingual phone strings produced by phone recognizers trained on data from different languages.

Currently, approaches based on Gaussian Mixture Models (GMMs) [4] are the most widely and successfully used methods for speaker identification. Although GMMs have been applied successfully to close-speaking microphone scenarios under matched training and testing conditions, their performance degrades dramatically under mismatched conditions. The term "mismatched condition" describes a situation in which the testing conditions, e.g. microphone distance, are quite different from what had been seen during training. For language and accent identification, phone recognition together with phone N-gram modeling has been the most successful approach in the past [6]. More recently, Kohler introduced an approach for speaker recognition where a phonotactic N-gram model is used.

In this paper, we extend this idea to far-field speaker identification, as well as to accent and language identification. We introduce two different methods based on multilingual phone strings to tackle mismatched distance and channel conditions and compare them to the GMM approach.

2. THE MULTILINGUAL PHONE STRING APPROACH

The basic idea of the multilingual phone string approach is to use phone strings produced by different context-independent phone recognizers instead of traditional short-term acoustic vectors [1]. For the classification of an audio segment into one of n classes of a specific non-verbal cue, m such phone recognizers together with $m \times n$ phonotactic N-gram models produce an $m \times n$ matrix of features. A best class estimate is made based solely on this feature matrix. The process relies on the availability of m phone recognizers, and the training of $m \times n$ N-gram models on their output.

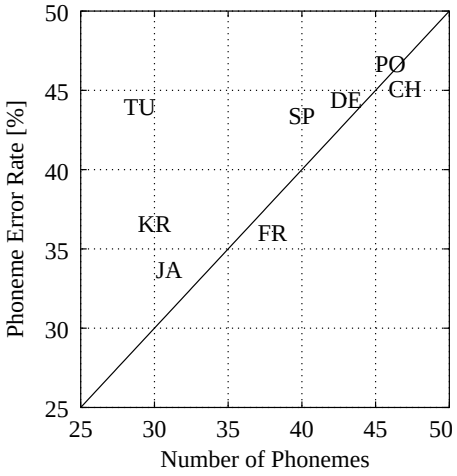

Fig. 1. *Error rate vs number of phones in 8 languages*

By using information derived from phonotactics rather than directly from acoustics, we expect to cover speaker idiosyncrasy and accent-specific pronunciations. Since this information is provided from complementary phone recognizers, we anticipate greater robustness under mismatched conditions. Furthermore, the approach is somewhat language independent since the recognizers are trained on data from different languages.

2.1. Phone Recognition

For the experiments presented here, the m phone recognizers were borrowed without modification from among the eight available within the GlobalPhone project: Mandarin

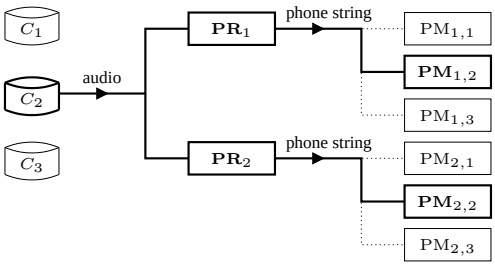

Fig. 2. *Training of feature-specific phonotactic models*

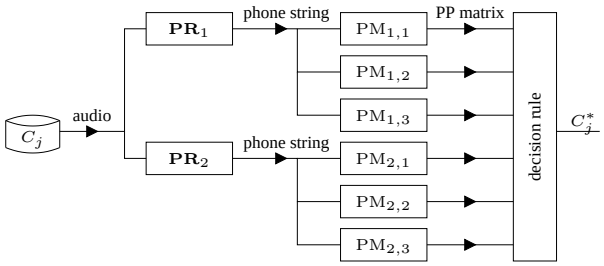

Fig. 3. *Block diagram of MPM-pp*

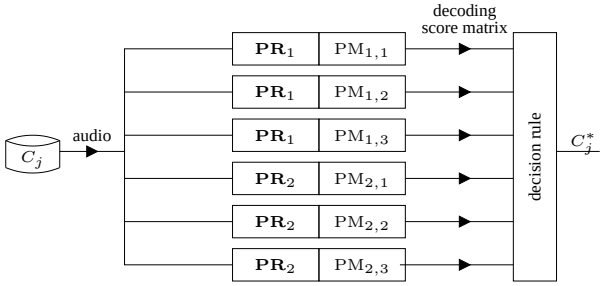

Fig. 4. *Block diagram of MPM-dec*

Chinese (CH), German (DE), French (FR), Japanese (JA), Croatian (KR), Portuguese (PO), Spanish (SP) and Turkish (TU). Figure 1 shows phone error rates per language in relation to the number of modeled phones. See [5] for further details.

2.2. Phonotactic Model Training

In classifying a non-verbal cue C into one of n classes, C_j, our feature extraction scheme requires $m \times n$ distinct phonotactic models $\mathrm{PM}_{i,j}$, $1 \leq i \leq m$ and $1 \leq j \leq n$, one for each combination of phone recognizer PR_i with output class C_j. $\mathrm{PM}_{i,j}$ is trained on phone strings produced by phone recognizer PR_i on C_j training audio as shown in Figure 2. During the decoding of the training set, each PR_i is constrained by an equiprobable phonotactic language model. This procedure does not require transcription at any level.

2.3. Classification

We present two multilingual phonotactic model (MPM) approaches to feature extraction, MPM-pp and MPM-dec.

In MPM-pp, each of m phone recognizers $\{\mathrm{PR}_i\}$, as used for phonotactic model training, decodes the test audio segment. Each of the resulting m phone strings is scored against each of n phonotactic models $\{\mathrm{PM}_{i,j}\}$. This results in a perplexity matrix $\mathbf{PP}$, whose $(PP)_{i,j}$ element is the perplexity produced by phonotactic model $\mathrm{PM}_{i,j}$ on the phone string output of phone recognizer PR_i. Although we have explored some alternatives, our generic decision algorithm is to propose a class estimate C_j^* by selecting the lowest $\sum_i (PP)_{i,j}$. Figure 3 depicts the MPM-pp procedure.

In MPM-dec, we also use all m phone recognizers $\{\mathrm{PR}_i\}$, but this time when decoding a test utterance we replace the equiprobable phonotactic language model used during phonotactic training with each of the n phonotactic models $\mathrm{PM}_{i,j}$ in turn. The test audio segment is therefore decoded by each of the m phone recognizers n times, resulting in a decoding score matrix $\mathbf{SCORE}$, whose $(SCORE)_{i,j}$ element is the decoding score produced jointly by phone recognizer PR_i and phonotactic model $\mathrm{PM}_{i,j}$ during decoding. As in MPM-pp, the class C_j^* whose $\sum_i (SCORE)_{i,j}$ is lowest is hypothesised. The key behind this method is that a phonotactic model $\mathrm{PM}_{i,j}$ is used directly in the decoding; however, this means that a test utterance must be decoded $m \times n$ times as opposed to only m times for MPM-pp. Furthermore, this procedure relies on the ability to produce reliable phonotactic models $\{\mathrm{PM}_{i,j}\}$ from the training data which are suitable for decoding.

3. EXPERIMENTS

3.1. Speaker Identification (SID)

Real world speaker identification is expected to work under mismatched conditions, regardless of the microphone distances during training and testing. To investigate robust speaker ID, a database has been collected in our lab containing 30 speakers reading different articles. Each of the five sessions per speaker are recorded using eight microphones in parallel: one close-speaking microphone

(Dis 0), one lapel (Dis L) microphone worn by the speaker, and six other lapel microphones at distances of 1, 2, 4, 5, 6, and 8 feet from the speaker. About 7 minutes of spoken speech (approximately 5000 phones) is used for training the PMs, while for training the GMMs one minute was used. The different amount of training data for the two approaches seems to make the comparison quite unfair; however, the training data is used for very different purposes. In the GMM approach, the data is used to train the Gaussian mixtures. In the MPM approach, the data is solely used for creating phonotactic models; no data is used to train the Gaussian mixtures of the phone recognizers.

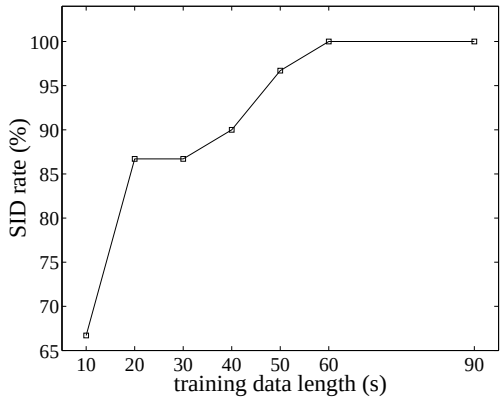

Fig. 5. *GMM performance with increasing training data*

Figure 5 shows the performance of the GMM approach with increasing amounts of training data, from 10 seconds to 90 seconds, on 10 seconds of test data. The graph indicates that for a fixed configuration of GMM structure, adding more training data is not necessary.

Testing	Training			
	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. *GMM performance under matched and mismatched conditions*

The GMM approach was tested on 10-second chunks, whereas the phone string approach was additionally tested on shorter and longer (up to one minute) chunks. We report results for closed-set text-independent speaker identification. Table 1 shows the GMM results with one minute training data on 10 seconds of test data. It illustrates that the performance under mismatched conditions degrades considerably when compared to performance under matched conditions.

Language	60s	40s	10s	5s	3s
CH	100	100	56.7	40.0	26.7
DE	80.0	76.7	50.0	33.3	26.7
FR	70.0	56.7	46.7	16.7	13.3
JA	30.0	30.0	36.7	26.7	16.7
KR	40.0	33.3	30.0	26.7	36.7
PO	76.7	66.7	33.3	20.0	10.0
SP	70.0	56.7	30.0	20.0	16.7
TU	53.3	50.0	30.0	16.7	20.0
Int. of all LM	**96.7**	**96.7**	**96.7**	**93.3**	**80**

Table 2. *MPM-pp SID rate on varying test lengths at Dis 0*

Table 2 shows the identification results of each phone recognizer and the combination results for eight language phone recognizers for Dis 0 under matched conditions. This shows that multiple languages compensate for poor performance on single engines, an effect which becomes even more important on shorter test utterances.

Test Length	60s	40s	10s	5s
Dis 0	96.7	96.7	96.7	93.3
Dis L	96.7	96.7	86.7	70.0
Dis 1	90.0	90.0	76.6	70.0
Dis 2	96.7	96.7	93.3	83.3
Dis 4	96.7	93.3	80.0	76.7
Dis 5	93.3	93.3	90.0	76.7
Dis 6	83.3	86.7	83.3	80.0
Dis 8	93.3	93.3	86.7	66.7

Table 3. *MPM-pp SID rate on varying test lengths at matched training and testing distances*

Table 3 and Table 4 compare the 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 segment distance; we make use of all $p = 8$ sets of $\mathrm{PM}_{i,j}$ phonotactic models, where p is the number of distances, and modify our decision rule to estimate $C_j^* = \min_j (\min_k \sum_i \mathrm{PM}_{i,j,k})$, where i is the index over phone recognizers, j is the index over speaker phonotactic models, and $1 \leq k \leq p$. These two tables indicate that the performance of MPM-pp, unlike that of GMM, is

Test length	60s	40s	10s	5s
Dis 0	96.7	96.7	96.7	90.0
Dis L	96.7	100	90.0	66.7
Dis 1	93.3	93.3	80.0	70.0
Dis 2	96.7	96.7	86.7	80.0
Dis 4	96.7	96.7	93.3	80.0
Dis 5	93.3	93.3	86.7	70.0
Dis 6	93.3	86.7	83.3	60.0
Dis 8	93.3	93.3	86.7	70.0

Table 4. *MPM-pp SID rate on varying test lengths at mismatched training and testing distance*

comparable for matched and mismatched conditions.

Language	MPM-pp (%)	MPM-dec (%)
CH	100	53.3
DE	80	40.0
FR	70	23.3
JA	30	26.7
KR	40	26.7
PO	76.7	30.0
SP	70	26.7
TU	53.3	26.7
Int. of all PM	**96.7**	**60**

Table 5. *Comparison of SID rate using MPM-pp and MPM-dec*

Table 5 compares the performance of MPM-dec at Dis 0 under matched conditions with that of MPM-pp on test utterances of 60 seconds in length. Even though MPM-dec is far more expensive than MPM-pp, its performance is only 60% under matched conditions for close-speaking data while MPM-pp yields 96.7%. The considerably poorer performance of MPM-dec seems to support the assumption made earlier that the phonotactic models we produced, which perform well within the MPM-pp framework, are not sufficiently reliable to be used during decoding as required by MPM-dec. These findings led us to focus on the use of the MPM-pp approach for accent and language identification.

3.2. Accent Identification (AID)

In this section we apply our non-verbal cue identification framework to accent identification. In a first experiment, we use the MPM-pp approach to differentiate between native and non-native speakers of English. Native speakers of Japanese with varying English proficiency levels make up the non-native speaker set [2]. Each speaker was recorded reading several news articles aloud; training and testing sets are disjoint with respect to articles as well as speakers. The data used for this experiment is shown in Table 6.

	use	native	non-native
n_{spk}	training	3	7
	testing	2	5
$\sum n_{\mathrm{utt}}$	training	318	680
	testing	93	210
$\sum \tau_{\mathrm{utt}}$	training	23.1 min	83.9 min
	testing	7.1 min	33.8 min

Table 6. *Number of speakers, total number of utterances and total length of audio for native and non-native classes*

We employ 6 of the GlobalPhone phone recognizers, $\mathrm{PR_i} \in \{\mathrm{DE}, \mathrm{FR}, \mathrm{JA}, \mathrm{KR}, \mathrm{PO}, \mathrm{SP}\}$. In training, native utterances are used to produce 6 phonotactic models $\mathrm{PM_{i,nat}}$; the same is done for non-native speech resulting in 6 $\mathrm{PM_{i,non}}$. During classification, the 6×2 phonotactic models produce a perplexity matrix for the test utterance to which we apply our lowest average perplexity decision rule; the class with the lower perplexity is identified as the class of the test utterance.

On our evaluation set of 303 utterances, this system classifies with an accuracy of 93.7%. The separability of the two classes is demonstrated in the average perplexity of each class of phonotactic model over all test utterances. The average perplexity of non-native models on non-native data is lower than the perplexity of native models on that data. Similarly, native models give lower scores to native data than do non-native models. Table 7 shows these averages.

Phonotactic model	Utterance class	
	non-native	native
non-native	29.1	31.7
native	32.5	28.5

Table 7. *Average phonotactic perplexities for native and non-native classes*

The accented speech experiment is unique among our classification tasks in that it attempts to determine the class of an utterance in a space that varies continuously according to the English proficiency of its speaker. Although classification among native and non-native speakers is discrete, it can be described as identifying speakers

who are clustered at the far ends of this proficiency axis. In a second experiment, we attempt to further classify non-native utterances according to proficiency level.

The original non-native data was labelled with the proficiency of each speaker on the basis of a standardized evaluation procedure conducted by trained proficiency raters [2]. All speakers received a floating point grade between 0 and 3, with a grade of 4 reserved for native speakers. The distribution of non-native training speaker proficiencies shows that they fall into roughly three groups and we create three corresponding classes for our new discrimination task. Class 1 represents the lowest proficiency speakers, class 2 contains intermediate speakers, and class 3 contains the high proficiency speakers.

We apply the MPM-pp approach to classify utterances from non-native speakers according to assigned speaker proficiency class. The phonotactic models are trained as before, with models in 6 languages for each of 3 proficiency classes; our division of data is shown in Table 8.

	use	class 1	class 2	class 3
n_{spk}	training	3	12	4
	testing	1	5	1
$\sum n_{\mathrm{utt}}$	training	146	564	373
	testing	78	477	124
$\sum \tau_{\mathrm{utt}}$	training	23.9 min	82.5 min	40.4 min
	testing	13.8 min	59.0 min	13.5 min
ave. prof	training	1.33	2.00	2.89
	testing	1.33	2.00	2.89

Table 8. *Number of speakers, total number of utterances, total length of audio and average speaker proficiency score per proficiency class*

Phonotactic model	Utterance proficiency		
	Class 1	Class 2	Class 3
Class 1	28.35	23.85	25.46
Class 2	23.85	23.86	24.17
Class 3	25.46	23.94	23.91

Table 9. *Average phonotactic perplexities per proficiency class*

Our results indicate that discriminating among proficiency levels is a more difficult problem than discriminating between native and non-native speakers. Table 9 shows that the class models in this experiment were more confused than the native and non-native models, and classification accuracy suffered as a result. We were able to achieve 84% accuracy in differentiating between class 1 and class 3 utterances, but accuracy on 3-way classification ranged from 34% to 59%.

Overall, the phone string approach worked well for classifying utterances from speaker proficiency classes that were sufficiently separable. Like the other applications of this approach, accent identification requires no hand-transcription and could easily be ported to test languages other than English/Japanese.

3.3. Language Identification (LID)

In this section, we apply the non-verbal cue identification framework to the problem of multiclassification of four languages: Japanese (JA), Russian (RU), Spanish (SP) and Turkish (TU).

We employed a small number of phone recognizers in languages other than the four classification languages in an effort to duplicate the circumstances common to our other non-verbal cue experiments, and to demonstrate a degree of language independence which holds even in the language identification domain. Phone recognizers in Chinese (CH), German (DE) and French (FR), with phone vocabulary sizes of 145, 47 and 42 respectively, were borrowed from the GlobalPhone project as discussed in [5].

The data for this classification experiment, also borrowed from the GlobalPhone project but not used in training the phone recognizers, was divided up as shown in Table 10. Data set 1 was used for training the phonotactic models, while data set 4 was completely held-out during training and used to evaluate the end-to-end performance of the complete classifier. Data sets 2 and 3 were used as development sets while experimenting with different decision strategies.

	Set	JA	RU	SP	TU
n_{spk}	1	20	20	20	20
	2	5	10	9	10
	3	3	5	5	5
	4	3	5	4	5
$\sum n_{\mathrm{utt}}$	all	2294	4923	2724	2924
$\sum \tau_{\mathrm{utt}}$	all	6 hrs	9 hrs	8 hrs	7 hrs

Table 10. *Number of speakers per data set, total number of utterances and total length of audio per language*

For phonotactics, utterances from set 1 in each $L_j \in \{\mathrm{JA}, \mathrm{RU}, \mathrm{SP}, \mathrm{TU}\}$ were decoded using each of the three phone recognizers $\mathrm{PR}_i \in \{\mathrm{CH}, \mathrm{DE}, \mathrm{FR}\}$ and 12

separate trigram models were constructed with Kneser/Ney backoff and no explicit cut-off. The training corpora ranged in size from 140K to 250K tokens, and the resulting models were evaluated on corpora constructed from set 2 utterances, of size 27K to 140K tokens. Trigram coverage for all 12 models fell between 73% to 95%, with unigram coverage below 1%.

In order to explore classification in a timeshift-invariant setting, we elected to extract features from segments of audio selected from anywhere in each utterance. For each of $\mathrm{PR}_i \in \{\mathrm{CH}, \mathrm{DE}, \mathrm{FR}\}$, phone strings for all utterances of each speaker in data set 4 were concatenated following decoding. Overlapping windows representing durations of 5, 10, 20 and 30 seconds, offset by 10% of their width, were identified for classification, each leading to a matrix of 3×4 perplexities. Duration was approximated using each speaker's average phone production rate per second for each recognizer PR_i. The number of testing exemplars is depicted per segment length in Table 11.

Set	5 s	10 s	20 s	30 s
4	23541	11689	5765	3789

Table 11. *Number of test exemplars per segment length*

Classification using lowest average perplexity led to 94.01%, 97.57%, 98.96% and 99.31% accuracy on 5s, 10s, 20s and 30s data respectively, as shown in Figure 6.

For comparison with our lowest average perplexity decision rule, we contructed a separate 4-class multi-classifier, using data set 2, for each of the four durations $\tau_k \in \{5\mathrm{s}, 10\mathrm{s}, 20\mathrm{s}, 30\mathrm{s}\}$; data set 3 was used for cross-validation. With each speaker's utterances concatenated and then windowed as was done for data set 4, this led to audio segments as in Table 12. These were subjected to the same feature extraction as before, yielding a 3×4 perplexity matrix per datum.

A class space of 4 classes induces 7 unique binary partitions. For each of these, we trained an independent multilayer perceptron (MLP) with 12 input units and 1

Set	5 s	10 s	20 s	30 s
2	48504	24092	11883	7815
3	28180	14003	6917	4556

Table 12. *Number of ECOC/MLP training and cross-validation exemplars per segment length*

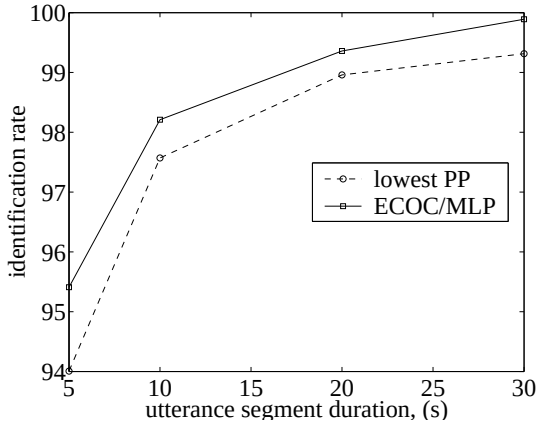

Fig. 6. *Language identification rate vs audio segment duration*

output unit using scaled conjugate gradients on data set 2 and early stopping using the cross-validation data set 3. In preliminary tests, we found that 25 hidden units provide adequate performance and generalization when used with early stopping. The output of all 7 binary classifiers was concatenated together to form a 7-bit code, which in the flavor of error-correcting output coding (ECOC) was compared to our four class codewords to yield a best class estimate. Based on total error using the best training set weights and cross-validation set weights on the cross-validation data, we additionally discarded those binary classifiers which contributed to total error; these classifiers represent difficult partitions of the data. Performance of this ECOC/MLP classification scheme on 5s, 10s, 20s and 30s data from set 4 was 95.41%, 98.33%, 99.36% and 99.89% respectively, shown in Figure 6.

4. LANGUAGE DEPENDENCIES

Implicit in our non-verbal cue classification methodology is the assumption that phone strings originating from phone recognizers trained on different languages yield crucially complementary information. Thus far we have not explored the degree to which the phone recognizers must differ, nor can we state how performance varies with the number of phone recognizers used. In this section we report on two experiments in the speaker identification arena intended to answer these questions.

4.1. Multi-lingual vs Multi-engine

We conducted one set of experiments to investigate whether the reason for the success of the multilingual phone

string approach is related to the fact that the different languages contribute useful classification information or that it simply lies in the fact that different recognizers provide complementary information. If the latter were the case, a multi-engine approach in which phone recognizers trained on the same language but on different channel or speaking style conditions might do a comparably good job. To test this hypothesis, we used a multi-engine approach based on three English phone recognizers which were trained on very different conditions, namely: Switchboard (telephone, highly conversational), Broadcast News (various channel conditions and speaking styles), and English Spontaneous Scheduling Task (high quality, spontaneous). The experiments were carried out on two different distances, Dis 0 and Dis 6, for the speaker identification task. For a fair comparison between the three English engines and the eight language engines, we generated all possible language triples out of the set of eight languages ($\binom{8}{3} = 56$ triples) and calculated the average, minimum and maximum performance for each. The results, given in Table 13, show that for Dis 0 the multi-engine approach lies within the range of the multilingual approach, and even outperforms the average. On Dis 6, however, the multi-engine approach is significantly outperformed by all $\binom{8}{3}$ language triples, and the average performance achieves half of the errors. Even if the poor performance of the multi-engine approach on Dis 6 is alarming and may indicate some robustness problems, it cannot be concluded from these results that multiple English language recognizers provide less useful information for the classification task than do multiple language phone recognizers. Further investigations on other distances, as well as on other non-verbal cues, are necessary to fully answer this question.

Approach	Multi-Lingual	Multi-Engine
Dis 0	87.92 (66.7-100)	93.3
Dis 6	81.96 (66.7-93.3)	63.3

Table 13. *Multi-Lingual vs Multi-Engine SID rates*

4.2. Number of involved languages

In a second suite of experiments, we investigated the influence of the number of phone recognizers on speaker identification rate. These experiments were performed on an improved version of our phone recognizers in 12 languages trained on the above described GlobalPhone data. Figure 7 plots the speaker indentification rate over the number m of languages used in the identification process on matched 60 second data at Dis 6. The performance is given in average and range over the $\binom{12}{m}$ language m-tuples. Figure 7 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 (out of $\binom{12}{m} = 66$) language pairs gave optimal results: CH-KO, and CH-SP. However, the lack of a strategy for finding the best suitable pair does not make this very helpful. On the other hand, the increasing average indicates that the probability of finding a suitable language-tuple which 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.

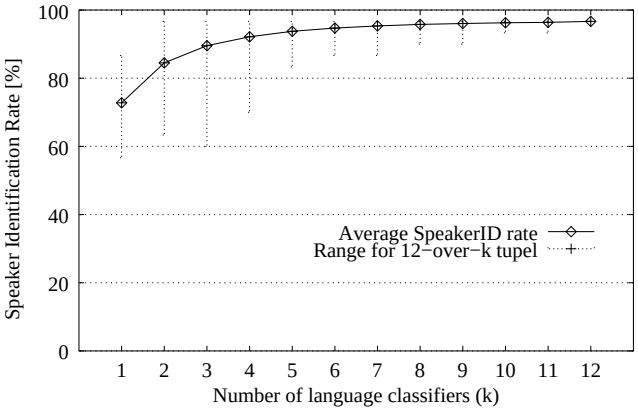

Fig. 7. *SID rate vs number of phone recognizers*

5. CONCLUSIONS

We have investigated the identification of non-verbal cues from spoken speech, namely speaker, accent, and language. For these tasks, a joint framework is developed which uses phone strings, derived from different language phone recognizers, as intermediate features and which performs classification decisions based on their perplexities. Our good identification results validate this concept, indicating that multilingual phone strings could be sucessfully applied to the identification of various non-verbal cues, such as speaker, accent and language. Our evaluation on variable distance data proved the robustness of the approach, achieving a 96.7% speaker identification rate on 10s chunks from 30 speakers under mismatched conditions, clearly outperforming GMMs on large distances. Furthermore, we achieved 93.7% accent discrimination accuracy between native and non-native speakers. The speaker and accent identification experiments were carried out on English data, although none of the applied phone recognizers were trained or adapted to English spoken speech. For language identification, we obtained 95.5% classification accuracy

for utterances 5 seconds in length and up to 99.89% on longer utterances, showing additionally that some reduction of error is possible using decision strategies which rely on more than just lowest average perplexity. Additionally, the language identification experiments were run on languages not presented to the phone recognizers for training. The language independent nature of our experiments suggests that they could be successfully ported to non-verbal cue classification in other languages.

6. REFERENCES

[1] Q. Jin, T. Schultz, and A. Waibel, "Speaker Identification using Multilingual Phone Strings", to be presented in: *Proceedings of ICASSP*, Orlando, Florida, May 2002.

[2] M. A. Kohler, W. D. Andrews, J. P. Compbell, and L. Hernander-Cordero, "Phonetic Refraction for Speaker Recognition", *Proceedings of Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, September 2001.

[3] L. Mayfield-Tomokyo, "Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR", PhD thesis, CMU-LTI-01-168, Language Technologies Institute, Carnegie-Mellon University, 2001.

[4] 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.

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

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