Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 101-108.
Association for Computational Linguistics.

IMPROVEMENTS IN NON-VERBAL CUE 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

Today's state-of-the-art front-ends for multilingual speech-to-speech translation systems apply monolingual speech recognizers trained for a single language and/or accent. The monolingual speech engine is usually adaptable to an unknown speaker over time using unsupervised training methods; however, if the speaker was seen during training, their specialized acoustic model will be applied, since it achieves better performance. In order to make full use of specialized acoustic models in this proposed scenario, it is necessary to automatically identify the speaker with high accuracy. Furthermore, monolingual speech recognizers currently rely on the fact that language and/or accent will be selected beforehand by the user. This requires the user's cooperation and an interface which easily allows for such selection. Both requirements are awkward and error-prone, especially when translation services are provided for many languages using small devices like PDAs or telephones. For these scenarios, front-ends are desired which automatically identify the spoken language or accent. We believe that the automatic identification of an utterance's non-verbal cues, such as language, accent and speaker, are necessary to the successful deployment of speech-to-speech translation systems.

Currently, approaches based on Gaussian Mixture Models (GMMs) [1] 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. For language and accent identification, phone recognition together with phone N-gram modeling has been the most successful approach in the past [2]. More recently, Kohler introduced an approach for speaker recognition where a phonotactic N-gram model is used [3].

In [4], we extended Kohler's approach to accent and language identification as well as to speaker identification 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. In that work, we explored a common framework for the identification of language, accent and speaker using multilingual phone strings produced by phone recognizers trained on data from different languages. In this paper, we propose and evaluate some improvements, comparing classification accuracy as well as realtime performance in our framework. Furthermore, we investigate the benefits that are to be drawn from additional phone recognizers.

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 [6]. 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.

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

The experiments presented here were conducted using two versions of phone recognizers borrowed without modification from the GlobalPhone project [5]. All were trained using our Janus Recognition Toolkit (JRTk).

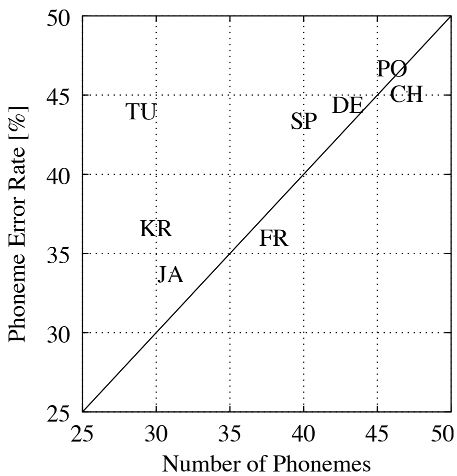

Fig. 1. *Error rate vs number of phones for the baseline GlobalPhone phone recognizer set*

The first set of phone recognizers, which we refer to as our baseline, includes recognizers for: Mandarin Chinese (CH), German (DE), French (FR), Japanese (JA), Croatian (KR), Portuguese (PO), Spanish (SP) and Turkish (TU). For each language, the acoustic model consists of a context-independent 3-state HMM system with 128 Gaussians per state. The Gaussians are on 13 Mel-scale cepstral coefficients with first and second order derivatives and power. Following cepstral mean subtraction, linear discriminant analysis reduces the input vector to 32 dimensions.

The second set consists of extended phone recognizers, available in 12 languages. Arabic (AR), Korean (KO), Russian (RU) and Swedish (SW) are available in this set in addition to the languages named above for the baseline set. The 12 new phone recognizers were derived from an improved generation of context dependent LVCSR systems which also include vocal tract normalization (VTLN) for speaker normalization. For decoding, we used an unsupervised scheme to find the best warp factor for a test speaker and calculate a viterbi alignment based on that speaker's best warp factor. To improve system speed, we reduced the number of Gaussians per state from 128 to 16; in addition, the feature dimension was halved from 32 to 16 using linear discriminant analysis.

Figure 1 shows the phone error rates in relation to the number of modeled phones for eight languages. The error rate correlates with the number of phones used to model this language. Turkish seems to be an exception to this finding. The error analysis showed that this is due to a very high substitution rate between the closed front vowels e, i, y.

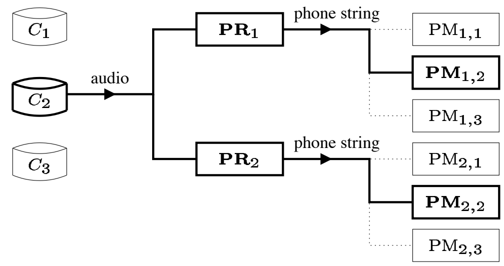

Fig. 2. *Training of feature-specific phonotactic models*

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

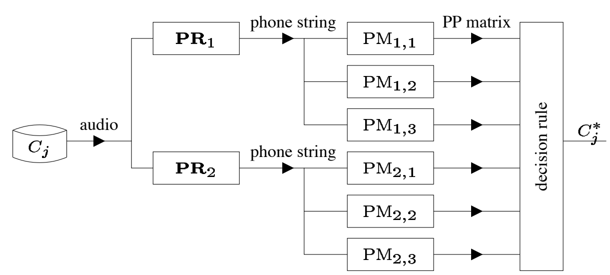

Fig. 3. *MPM-pp classification block diagram*

During classification, 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 this procedure, which we refer to as MPM-pp.

3. EXPERIMENTS

3.1. Speaker Identification (SID)

In order to investigate 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. This database contains 30 native English 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 microphone (Dis L) worn by the speaker, and six other lapel microphones at distances of $1, 2, 4, 5, 6$, and 8 feet from the speaker.

In a first experiment, we compare the performance of the MPM-pp approach using our baseline phone recognizers to the GMM approach. 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. And we have found that with a fixed configuration of GMM structures, adding more training data does not lead to noticeable improvement in performance [4].

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 for 10 second segments*

The GMM approach was tested on 10 seconds of audio, whereas the phone string approach was additionally tested on shorter and longer (up to one minute) segments. 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.

Table 2 shows the identification results using each of the 8 baseline phone recognizers individually and their combination results for Dis 0 under matched conditions. This shows that multiple recognizers collectively compensate for the poor performance of single recognizers, an effect which becomes even more important for shorter test utterances.

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*

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*

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*

Table 3 and Table 4 compare the identification results for all distances for different test utterance lengths under matched and mismatched conditions, respectively. Under matched conditions, training and testing data are drawn from the same microphone. 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 comparable for matched and mismatched conditions.

We conducted additional experiments to determine the impact of the improved GlobalPhone recognizers on the identification rate for this task. To that end, we used all 8 baseline recognizers and only the corresponding 8 of the 12 available improved recognizers. Table 5 compares the speaker identification rate on matched conditions for 60 seconds of audio. The comparison indicates that an improvement in phone error rate leads to slight improvements in speaker identification rate for distances Dis 0 and Dis 5. Performance decreases for Dis L, while for Dis 6 the improved recognizers outperform the baseline significantly. Overall we cannot conclude from these results that better phone recognizers result in a higher identification rate. However, we can summarize that the improved engines show an identification performance of 93.3% or higher on all distances for matched conditions on 60 seconds of audio, in spite of the drastic reduction in acoustic model parameter dimensions.

Distance	phone recognizers	
	baseline	improved
Dis 0	96.7	96.7
Dis L	96.7	93.3
Dis 1	90.0	93.3
Dis 2	96.7	96.7
Dis 4	96.7	96.7
Dis 5	93.3	96.7
Dis 6	83.3	96.7
Dis 8	93.3	93.3

Table 5. *Comparison of MPM-pp classification using baseline and improved phone recognizers on matched conditions for 60 seconds of audio (SID rate in %)*

3.2. Accent Identification (AID)

In previous experiments on accent identification we used the MPM-pp approach to identify native and non-native speakers of English and to identify speakers of varying proficiency levels in English.

	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*

In our current experiments, we have augmented the number of phonotactic models used to classify utterances. We decode training data from each class using the baseline phone recognizer for Chinese and run our original experiments with a new bank of phonotactic models in 7 languages: the original 6 $\mathrm{PR_i} \in \{\mathrm{DE, FR, JA, KR, PO, SP}\}$, plus $\{\mathrm{CH}\}$. During classification, the 7×2 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 for 2-way classification between native and non-native speakers, our classification accuracy improves from 93.7% using models in 6 languages to 97.7% using models in 7 languages. An examination of the average perplexity of each class of phonotactic model over all test utterances reveals the improved separability of the classes. The average perplexity of non-native models on non-native data is lower than the perplexity of native models on that data, and the discrepancy between these numbers grows after adding training data decoded in an additional language. The native models became less separable on average but discriminatory power still improved overall. Table 7 shows these average perplexities for our previous and current experiments.

# of phone recognizers	Phonotactic model	Utterance class	
		non-native	native
6	non-native	29.1	31.7
	native	32.5	28.5
6 + CH	non-native	28.9	34.1
	native	32.8	31.1

Table 7. *Average phonotactic perplexities for native and non-native classes using 6 phone recognizers (top) versus 7 (bottom)*

In the proficiency-level experiments, we apply the MPM-pp approach to classify utterances from non-native speakers according to assigned speaker proficiency class. The original non-native data has been labelled with the proficiency of each speaker on the basis of a standardized evaluation procedure conducted by trained proficiency raters [7], and we attempt to classify non-native speakers from three classes according to their proficiency. Class 1 represents the lowest proficiency speakers, class 2 contains intermediate speakers, and class 3 contains the high proficiency speakers. Profiles of the testing and training data for these experiments are 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*

We have added phonotactic models trained on Chinese recognizer output to this experiment as well, and gained a small improvement over our results using models in 6 languages. Table 9 displays two confusion matrices for this task, one showing original results and one showing results with the added Chinese phone recognizer.

# of Phone recognizers	System hypothesis	Actual proficiency		
		Class 1	Class 2	Class 3
6	Class 1	8	3	19
	Class 2	8	41	61
	Class 3	2	12	99
6 + CH	Class 1	8	5	17
	Class 2	6	53	51
	Class 3	1	20	92

Table 9. *Confusion matrix for 3-way proficiency classification using 6 phone recognizers (top) versus 7 (bottom)*

Classification accuracy in the 3-way proficiency classification task improves somewhat, rising from 59% in the original experiment to 61% using the additional phone recognizer. As the confusion matrix for this experiment shows, the phonotactic models trained in Chinese cause the system to correctly identify more of the class 2 utterances, but at the expense of some class 3 utterances which are also identified as class 2 by the new system.

In both 2-way and 3-way classification, the addition of a seventh phone recognizer improved classification accuracy. 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)

For this task, we applied the non-verbal cue identification framework to the problem of multiclassification of four languages: Japanese (JA), Russian (RU), Spanish (SP) and Turkish (TU). We elected to use a small number of phone recognizers in languages other than the four classification languages in order to duplicate the circumstances common to our other non-verbal cue identification 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.

In this section, we first reiterate our accuracy results using phone recognizers drawn from the baseline set; the details of those experiments are discussed in [4]. We then compare both the identification accuracy and realtime performance with results obtained using the improved GlobalPhone phone recognizers.

The data for this classification experiment, also borrowed from the GlobalPhone project but not used in training the phone recognizers, was divided 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.

We first benchmarked accuracy using our lowest average perplexity decision rule. For comparison, we constructed a separate 4-class multiclassifier, using data set 2, for each of the four durations $\tau_k \in \{5\mathrm{s}, 10\mathrm{s}, 20\mathrm{s}, 30\mathrm{s}\}$. Our multi-classifier combined the output of multiple binary classifiers using an error-correcting output coding (ECOC) technique. A class space of 4 languages induces 7 unique binary partitions. For each of these, we trained an independent multilayer perceptron (MLP) with 12 input units and 1 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 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.

With phone recognizers drawn from the baseline set, classification accuracy using lowest average perplexity led to 94.01%, 97.57%, 98.96% and 99.31% accuracy on 5s, 10s, 20s and 30s data respectively, while with ECOC/MLP classification accuracy improved to 95.41%, 98.33%, 99.36% and 99.89% respectively.

Replacement of the baseline phone recognizers with ones from the extended and improved GlobalPhone set led to classification accuracies, using lowest average perplexity, of 94.83%, 97.89%, 98.98%, and 99.26% on 5s, 10s, 20s and 30s data respectively. All classification rate results are plotted in Figure 4. Comparing the lowest average perplexity results from the old with the new recognizers shows that the improved recognizers lead to higher improvements for the short utterance segments, for the 30s segments the results are slightly worse.

The runtime performance of the phone recognition component was assessed, using set 1 of the data in Table 10, on a dual CPU 933 MHz Pentium III machine with 512 MB of memory and 900 MB of swap with low load. Realtime factors are presented for both the baseline set and the improved set of phone recognizers in Table 11.

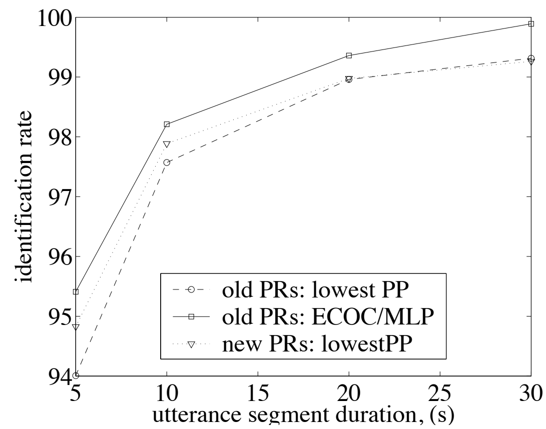

Fig. 4. *Language identification rate vs duration*

PR	vocabulary size	realtime factor	
		baseline	improved
CH	145	9.6	0.94
DE	47	4.0	0.32
FR	42	3.7	0.40

Table 11. *Realtime factors for baseline and improved phone recognizers*

While the difference in classification accuracy between the baseline identification system and that built using the improved phone recognizers is perhaps not statistically significant, the improvements in runtime speed are very dramatic. For vastly different vocabulary sizes, the improvement is almost a whole order of magnitude.

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. In [4] we performed some initial experiments to explore the influence of variation in phone recognizers, and how the identification rate varies with the number of phone recognizers used. In this section we report on two follow-up experiments for the speaker identification task intended to answer these questions.

4.1. Multiple languages vs single language multiple engines

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 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 were 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 three different phone recognizers all trained on English data, but under different channel conditions (telephone, channel-mix, clean) and different speaking styles (highly conversational, spontaneous, planned) [4].

The experiments were carried out on matched conditions on all distances for 60 seconds of audio for the speaker identification task. To compare the three English engines to the multiple language engines, we generated all possible language triples out of the set of languages and calculated the average, minimum and maximum performance over all triples. We evaluated the performance for both recognizer versions, the baseline 8-language phone recognizers and the improved 12-language ones. In the first case, we generated all possible language triples out of the set of eight languages ($\binom{8}{3} = 56$ triples); in the latter, we did the same out of the set of twelve languages ($\binom{12}{3} = 220$ triples). In both cases, we calculated the average, minimum and maximum performance over all triples. The results are given in Table 12.

Dis	Multiple Languages $\binom{8}{3}$	Multiple Languages $\binom{12}{3}$	Multiple EN PRs
Dis 0	87.9 (66.7-100)	94.6 (80.0-100)	93.3
Dis L	88.2 (63.3-96.7)	93.1 (80.0-96.7)	86.7
Dis 1	83.6 (66.7-93.3)	89.5 (76.7-96.7)	86.7
Dis 2	93.6 (86.7-96.7)	93.6 (86.7-96.7)	76.7
Dis 4	81.4 (56.7-96.7)	90.8 (73.3-96.7)	86.7
Dis 5	86.1 (66.7-96.7)	92.0 (73.3-96.7)	83.3
Dis 6	82.0 (82.0-93.3)	89.5 (60.0-96.7)	63.3
Dis 8	87.1 (87.1-93.3)	87.2 (63.3-96.7)	63.3

Table 12. *Multiple languages vs multiple English phone recognizers (SID rates in %)*

The improved versions of the multiple language phone recognizers give significantly better average results for most of the distances. The results also show that the multiple English engine approach in almost all cases lies within the range of the multilingual approach. However, the average performance of the multiple language approach using the improved engines always outperforms the multiple English engine approach. This indicates that most of the language triples achieve better results than the single-language multiple engines.

In summary, table 12 shows that best performance of the multi-language approach always outperforms the multiple English engine approach; moreover, in the case of the 12 improved GlobalPhone engines, even the average always outperforms the multiple English engine approach. From these results, we draw the conclusion that multiple English language recognizers provide less useful information for the classification task than do multiple language phone recognizers. This is at least true for the given choice of multiple engines in the context of speaker identification. The fact that the multiple engines were trained on English, i.e. the same language which is spoken in the speaker identification task, whereas the multiple languages were trained on 12 languages other than English, makes the multiple language approach even more appealing since it indicates a great potential for portability to non-verbal cue identification on other languages.

4.2. Number of involved languages

In this set of experiments, we investigated the influence of the number of phone recognizers on the speaker identification performance. These experiments were performed on the improved version of GlobalPhone phone recognizers in 12 languages. Figure 5 plots the speaker identification rate over the number k of languages used in the identification process on matched conditions on 60 seconds of audio for all distances. The performance given for each distance is an average over the $\binom{12}{k}$ language k-tuples. The results indicate that the average speaker identification rate increases for all distances with the number of involved phone recognizers. For some distances, a saturation effect occurs beyond 6 languages (distance 0 and 1); for other distances, even the 12^{th} language has a positive effect on the average performance (distance 4, 6, L). The increasing average indicates that the probability of finding a suitable language-tuple which optimizes performance increases with the number of available languages. We also analyzed whether the increasing performance is related to the total number of phones used for the classification process rather than the number of different engines, but did not find evidence for a strong correlation.

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 was presented which uses phone strings, derived from different phone recognizers, as intermediate features and which performs classification decisions based on their perplexities. Our good identification results validate this concept, indicating that multilingual

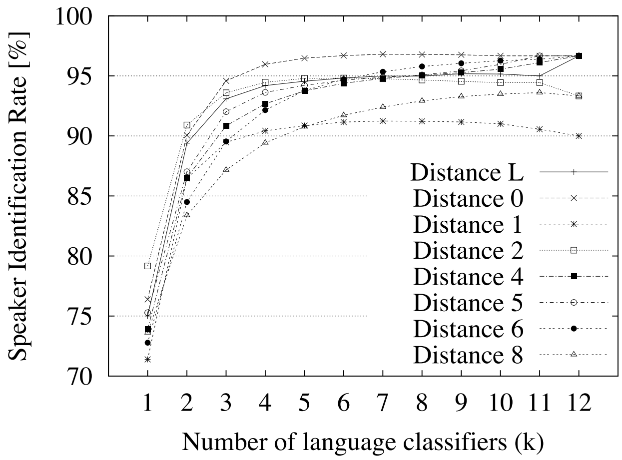

Fig. 5. *SID rate vs number of phone recognizers*

phone strings could be sucessfully applied to the identification of various non-verbal cues, such as speaker, accent and language. The evaluation on our distant microphone database proved the robustness of the approach, achieving a 96.7% speaker identification rate on 10 seconds of audio from 30 speakers under mismatched conditions, clearly outperforming GMMs on large distances. We achieved similar results using phone recognizers with a drastically reduced parameter dimension. Furthermore, in the speaker identification domain we showed that, on average, the use of phone recognizers trained on different languages leads to greater accuracy than do multiple same-language phone recognizers.

Our classification framework performed equally well in the domains of accent and language identification. We achieved 97.7% discrimination accuracy between native and non-native English speakers, showing that the addition of a seventh recognizer to this task, namely Chinese, reduced the error rate by 63%.

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. Furthermore, accuracy was shown to improve, at least for short utterance durations, using phone recognizers which are more accurate but constrained to a much smaller parameter space. While retaining classification accuracy, these phone recognizers run faster than realtime, outperforming the speed of the baseline by almost 90%.

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. Similarly, our 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] 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.

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

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

[4] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel, "Speaker, Accent, and Language Identification Using Multilingual Phone Strings", *HLT 2002*, San Diego, California, March 2002.

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

[7] 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.