MULTILINGUAL ACOUSTIC MODELS FOR THE RECOGNITION OF NON-NATIVE SPEECH

V. Fischer, E. Janke, S. Kunzmann, T. Roß

IBM Voice Systems, European Speech Research,
Vangerowstr. 18, D-69115 Heidelberg, F.R. of Germany
vfischer@de.ibm.com

ABSTRACT

In this paper we report on the use of multilingual Hidden Markov Models for the recognition of non-native speech. Based on the design of a common phoneme set that provides a phone compression rate of almost 80 percent compared to a conglomerate of language dependent phone sets, we create acoustic models that share training data from up to 5 languages. Results obtained on two different data bases of non-native English demonstrate the feasibility of the approach, showing improved recognition accuracy in case of sparse training material, and also for speakers whose native language is not in the training data.

1. INTRODUCTION

With the emergence of new technologies and devices in the field of telecommunications speech is expected to become a predominant input medium for easy and natural access to information from anywhere. Despite large progress in fields like large vocabulary continuous speech recognition or noise robustness, many speech recognizers still reveal accuracy problems if a speaker's pronunication systematically differs from those observed during system training. In applications such as desktop dictation, this problem can successfully be tackled by speaker adaptation methods such as e.g. MAP [8], and these techniques have also been applied for (offline) adaptation of acoustic models to certain dialects, see e.g. [6]. However, there is a growing number of applications that do not allow the collection of a sufficient amount of adaptation data and have to cope with a large variety of individual pronunciations; consider, for example, a tourist information system that is used only once each by many individuals with numerous accents and/or native languages.

Partially motivated by these considerations there is a growing interest in *language independent* and *multilingual* speech recognition. While language independent acoustic modeling [3] utilizes training data from several languages for the fast bootstrapping of monolingual recognizers for an unseen target language, multilingual speech recognition [11, 12, 7] aims on the creation of acoustic models that can decode speech from a variety of languages at one and the same time.

In this study we investigate the use of multilingual acoustic models for the recognition of non-native English. In contrast to e.g. [13], where non-native speech in the target language was included in the training data, we restrict ourselves to the use of native training material from several languages. In doing so, we start with a brief review on our efforts to merge the phone alphabets of seven different languages into a common phone set. Section 3 outlines the training of rank-based multilingual Hidden Markov Models, and in Section 4 we report on some initial recognition experiments on two different data bases, an in-house collection of non native speech data and a part of the Polycost speaker verification database [10]. Finally, Section 5 gives a conclusion and some prospects for further work.

2. COMMON PHONOLOGY

Our initial work towards a seamless multilingual speech recognizer (see [7]) has concentrated on the definition of a common phone alphabet for seven languages: Arabic, (British) English, French, German, Italian, (Brazilian) Portuguese, and Spanish. A common phone set for these languages was created in a two-stage process: using SAMPA notation, we first mapped the language specific phones to their closest IPA [1] equivalent, which required simplifications for some languages, but also resulted in the introduction of new phone models for other languages. For example, we gave up syllabic consonant phones in German, whereas for British English we introduced new diphtong phones. Subsequently, phones that shared the same SAMPA symbol were merged. In order to still achieve high phone compression rates when adding more and more languages, we then defined a reduced common phonology [9]. For that purpose, diphtongs (and most of the long vowels) were replaced by a sequence of two (in case of long vowels identical) short vowels, stressed vowels were dropped, and also some changes to the coding of consonants were made.

Table 1 compares the size of the language specific phone sets of the seven languages to the reduced common phone set, and Table 2 shows the number of vowel and consonant phones that are unique to one of the seven languages. The overall phone compression rate is 78.7 percent, and is slightly larger for consonants (79.9 percent) than for vowels (76.5 percent).

(a)	total	En	Fr	Gr	It	Es	Pt	Ar
vowels	132	18	17	23	22	14	24	14
cons.	224	31	19	37	48	35	24	30
total	356	49	36	60	70	49	48	44

(b)	total	En	Fr	Gr	It	Es	Pt	Ar
vowels	31	13	15	17	7	5	12	11
cons.	45	24	19	23	28	24	22	28
total	76	37	34	40	35	29	34	39

Table 1: Number of vowels and consonants in the phone sets of seven languages (a), and in the common phone set (b). Languages are British English (En), French (Fr), German (Gr), Italian (It), Spanish (Es), Brazilian Portuguese (Pt), Arabic (Ar).

While data driven phone clustering methods (e.g. [4]) usually tend to achieve a less significant phone compression rate, but may produce more accurate recognition results as long as only few languages are considered, we think that the utilization of phonetic expert knowledge is the more promising way if many languages are involved. Moreover, by e.g. the use of language questions [14] for phonetical decision tree growing a data driven differentiation can be introduced in the acoustic model training (cf. Section 3).

	total	En	Fr	Gr	It	Es	Pt	Ar
vowels	14	3	2	2	–	–	3	4
cons.	16	–	1	1	6	–	–	8
total	30	3	3	3	6	–	3	12

Table 2: Number of vowels and consonants in the reduced common phone set that are unique to one of the seven languages.

3. MULTILINGUAL MODELS

Feature extraction, the construction of a set of context dependent allophonic Hidden Markov Models (HMMs), and the estimation of continuous density Gaussian mixture model parameters are the main aspects to consider in the training of a rank-based speech recognizer [2].

The acoustic front end used in this study computes 13 MFCC (including C0) and their first and second order derivative every ten milliseconds. By means of a multilingual bootstrap acoustic model sequences of feature vectors are viterbi-aligned against the transcription of the training data in order to obtain an allophonic label for each vector. Subsequently, a set of binary questions about the phonetic context ("Is the phone in position i in the subset $\mathcal{S}_j$?") is used to identify homogeneous regions of the feature space. Each terminal node (leaf) of the so created polyphone decision network is represented as a context dependent, single-state Gaussian mixture HMM, and a k-means procedure is employed to obtain initial HMM output probabilities from the data at each leaf of the network. Finally, the initial HMM parameters are refined by running a few forward-backward iterations.

In [7] we obtained improved models from several extensions of the sketched procedure which include:

- the transformation of cepstral feature vectors by means of a multilingual linear discriminant analysis (LDA),
- the use of *language questions* ("Is the phone in position i from a subset $\mathcal{L}_j$ of languages") [14] in the creation of the phonetic decision network, and
- the replacement of k-means clustering by a Bayesian Information Criterion (BIC) based cluster procedure [5] that allows a proper determination of acoustic model complexity.

However, since we did not implement LDA and language question support in the viterbi decoder used in Section 4, only clustering via BIC was used for the acoustic model training.

In the experiments described below we utilized a shortcut method which enables the fast bootstrap of acoustic models that make use of training data from an arbitrary subset of languages $\mathcal{L}_j = \{L_i | i = 1 \ldots n\}$. For that purpose we first created a common HMM inventory by growing of a multilingual polyphone decision network from all data. Subsequently, the common HMMs were individually trained with data from each language $L_i, i = 1 \ldots m \geq n$, resulting in m sets of mono-lingual Gaussians that serve as a repository for the creation of true multilingual HMMs. Finally, given the desired set $\mathcal{L}_j$ of languages, the latter were created by the merging of (language dependent) Gaussians that belong to the same leaf of the common decision network.

4. EXPERIMENTS

Investigations on the use of multilingual acoustic models for the recognition of non-native speech were carried out on two different databases, both consisting of (continuous) English digit strings, but dealing with different scenarios. In both experimental setups we used the training procedure outlined above, data from up to five languages (French, German, Italian, Spanish, and UK English) for the creation of multilingual crossword triphone HMMs, and a single-pass, time synchronous viterbi-decoder.

4.1. In-house database

In a first row of experiments we collected test data from 29 non-native and 10 native speakers (16 female, 23 male) that read the same test script (English digits), cf. Table 3.

A modest amount of 11 kHz training data that was chosen from a database of office correspondence, journalism, etc., was used for the creation of acoustic models M1 – M4. All models make use of the same multilingual HMM inventory, but were trained with a different amount of data: M1 is a monolingual model build from UK English training data, M2 and M3 are bilingual models that make additional use of either French or German data, and M4 was trained with data from all languages.

(a)	Es	Fr	Gr	It	En
no.of speakers	250	1105	500	500	699
speech data [h]	16.3	19.5	19.6	19.8	16.2
words [×1000]	13.2	11.1	18.1	33.5	21.6

(b)	Es	Fr	Gr	It	En
test speakers	6	7	10	6	10
digits[×1000]	3	3.5	5	3	5

Table 3: Training (a) and test data (b) overview for the 11 kHz acoustic models.

Recognition results for non-native speakers from Spain, France, Germany, and Italy as well as for native speakers are given in Table 4. For both French and German test speakers we achieved a small improvement from the bilingual models M2 and M3, but also obtained a degradation for the control group of native speakers. In contrast, by using training data from all languages (M4) we obtained an average improvement of 22.4 percent for the non-native speakers, whereas the control group improved by 29.4 percent. Clearly, this has to be attributed to the larger amount of training data and demonstrates the cross-language transfer capabilities of the multilingual models.

WER	Es	Fr	Gr	It	avg.	En
M1	12.13	5.91	9.26	12.03	9.62	6.40
M2	–	5.77	–	–	–	6.60
M3	–	–	8.10	–	–	7.26
M4	10.07	5.17	5.98	10.03	7.47	4.52

Table 4: English digit error rate for non-native speakers from Spain (Es), France (Fr), Germany (Gr), and Italy (It); avg: average over the four groups; En: native control group.

4.2. Polycost database

Since we considered a lack of training data as one reason for the fairly high digit error rates, more recently we also used larger corpora of telephone speech for the training of multilingual acoustic models, cf. Table 5. Those models were evaluated on a subset of the Polycost 250 speech database [10], which was originally designed for speaker verification tasks. For our test we used digit strings from 60 speakers (30 female, 30 male) and calls from 12 different countries: Belgium, Switzerland, Denmark, Spain, France, Ireland, Italy, The Netherlands, Portugal, Sweden, Turkey, and The United Kingdom. Note, however, that the origin of the call is only a vague indication for the speakers' native language, which is in particular true for calls from countries such as Belgium or Switzerland.

	Es	Fr	Gr	It	En
no.of speakers	5541	3004	4544	1940	5077
speech data [h]	40.1	47.2	40.1	43.1	51.7
words [×1000]	163	226	117	128	191

Table 5: Training data overview for the 8 kHz acoustic models.

Table 6 gives results for three different acoustic models: M5 is a monolingual model that was trained with 51.7 hours of speech from 5077 native English speakers, M6 used all data for the creation of a multilingual triphone decision tree, and for M7 data from 5 languages was used for both decision tree growing and the estimation of HMM parameters.

WER [%]	non-native	native
M5	5.51	2.91
M6	3.86	2.91
M7	4.52	3.16

Table 6: Word error rates for non-native (Be, Ch, Dk, Es, Fr, It, Nl, Pt, Se, Tr) and native (Ireland, United Kingdom) English digits for the 8 kHz acoustic models.

Although the multilingual models M6 and M7 performed better for callers from most countries (Switzerland, Denmark, Spain, France, The Netherlands, and Sweden), we observed a 15 percent relative decrease in word error rate for non-native English when comparing M6 and M7. The source of this discrepancy were calls from Turkey, for which the multilingual models showed a 25 percent decrease in accuracy, and — surprisingly — also Italian speakers showed a 10 percent degradation. These results are currently undergoing investigation.

5. CONCLUSION

In this paper we have used multilingual acoustic models for the recognition of non-native speech from two different data bases. The reasonable improvement obtained on the 11kHz in-house data base underlines demonstrates the efficiency of multilingual acoustic models in case of sparse training data. Results for the Polycost data base show that this effect becomes smaller, if a larger amount of training data from the target language is available. However, these results also demonstrate that — aside from Turkish — improved error rates for non-native speech can be obtained without having the speakers' native language in the training data.

We have recently started to explore whether a priori knowledge about the target language, as — for example — given by (sub-)phone unigram or bigram probabilities can be used to improve the recognition accuracy of multilingual models. In case of the in-house data base we found an average improvement of 5 percent relative when combining a speech frame's sub-phone unigram and acoustic score. We wish to report more detailed on this approach, once we have tested it in a true multilingual decoding scenario, that might include language identification as well.

REFERENCES

[1] International Phonetic Association. IPA chart, *revised version of:* The Principles of the International Phonetic Association, 1949. *Journal of the International Phonetic Association*, 1(1), 1993.

[2] L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny. Robust methods for using context-dependent features and models in a continuous speech recognizer. In *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Adelaide, 1994.

[3] P. Beyerlein, W. Byrne, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang. Towards language independent acoustic modeling. In *Proc. of the 1999 Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, 1999.

[4] P. Bonaventura, F. Gallocchio, and G. Micca. Multilingual Speech Recognition for Flexible Vocabularies. In *Proc. of the 5th European Conference on Speech Communication and Technology*, pages 355–358, Rhodes, Greece, 1997.

[5] S. Chen and P. Gopalakrishnan. Clustering via the Bayesian Information Criterion with Applications to Speech Recognition. In *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, pages 645–648, Seattle, 1998.

[6] V. Fischer, Y. Gao, and E. Janke. Speaker independent upfront dialect adaptation in a large vocabulary continuous speech recognizer. In *Proc. of the 5th Int. Conf. on Spoken Language Processing*, Sydney, 1998.

[7] V. Fischer, J. Gonzalez, E. Janke, M. Villani, and C. Waast-Richard. Towards Multilingual Acoustic Modeling for Large Vocabulary Continuous Speech Recognition. In *Proc. of the IEEE Workshop on Multilingual Speech Communications*, Kyoto, Japan, 2000.

[8] J. Gauvain and C. Lee. Maximum a posteriori estimation of multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.

[9] F. Palou Cambra, P. Bravetti, O. Emam, V. Fischer, and E. Janke. Towards a common alphabet for multilingual speech recognition. In *Proc. of the 6th Int. Conf. on Spoken Language Processing*, Beijing, 2000.

[10] D. Petrovska, J. Hennebert, H. Melin, and D. Genoud. Polycost: A Telephone-Speech Database for Speaker Recognition. In *Proc. of the Workshop on Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, 1998.

[11] T. Schultz. *Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen.* Dissertation. Universität Karlsruhe, Institut für Logik, Komplxität und Deduktionssysteme. 2000.

[12] T. Schultz and A. Waibel. Language Portability in Acoustic Modeling. In *Proc. of the IEEE Workshop on Multilingual Speech Communications*, Kyoto, Japan, 2000.

[13] U. Uebler and M. Boros. Recognition of Non-Native German Speech with Multilingual Recognizers. In *Proc. of the 6th Europ. Conf. on Speech Communication and Technology*, volume 2, pages 911–914, Budapest, 1999.

[14] T. Ward, S. Roukos, C. Neti, J. Gros, M. Epstein, and S. Dharanipragada. Towards Speech Understanding Across Multiple Languages. In *Proc. of the 5th Int. Conf. on Spoken Language Processing*, Sydney, 1998.