French–German Bilingual Acoustic Modeling for Embedded Voice Driven Applications

Jozef Ivanecký, Volker Fischer, Siegfried Kunzmann

IBM AIM, European Voice Technology Development
Schönaicher-Str. 220, 71032 Böblingen, Germany
{ivanecky, vfischer, kunzmann}@de.ibm.com

Abstract. Multilingual access to information and services is a key requirement in any pervasive or ubiquitous computing environment. In this paper we describe our efforts towards multilingual speech recognition with a focus on applications that are designed to run on embedded devices, like e.g. a commercially available PDA. We give an overview on speech recognition techniques suited for the special requirements of the expected phonetic and acoustic environments and explore the ability to create multilingual acoustic models and applications that are able to run on embedded devices in real-time.

1 Introduction

Today, multilingual voice access to information is becoming more and more common in real applications. While multilingual ASR systems have reached a certain maturity in telephony ASR systems, where only weak CPU and memory usage constraints exist, the situation is completely different for systems supposed to run on small devices like PDAs, or mobile phones. In this case, both restrictive CPU and memory requirements, and the need to run in real-time still result in high error rates. Also, a very noisy acoustic channel with background noise of many different characteristics is an additional challenge for the creation of small acoustic models that are capable of simultaneously recognizing several languages.

Based on experiences from mono-lingual acoustic modeling and application design for embedded devices, we defined the following requirements for multilingual embedded systems:

- the size of a multilingual system should not be significantly larger than the size of a mono-lingual system,
- the phonology used for multilingual modeling should cover as many languages as possible, or at least it should seamlessly extend to languages not yet under consideration,
- digits recognition — which is of particular interest in many embedded applications — should not suffer too much from multilingual modeling.

As a first step towards these goals, in this paper we describe the construction of a bilingual French–German acoustic model for embedded devices. The remainder of the paper is organized as follows: In Section 2 we give a brief overview over

the used common phone alphabet. In Section 3 we focus on the multi-language specific acoustic modeling parts. Section 4 describes the experiments and a brief summary is given in Section 5.

2 Common Phonetic Alphabets

The definition of a common phonetic alphabet for multilingual speech recognition has to deal with at least two conflicting goals: while on the one hand the phonetic inventory of each language should be covered as precise as possible in order to achieve high recognition accuracy, at the same time as many phones as possible should be shared across languages in order to efficiently utilize the training data and for the creation of reasonably small acoustic models.

Starting from available, disjoint phonetic alphabets for seven languages (Arabic, British English, French, German, Italian, (Brazilian) Portuguese, and Spanish) which are used within our monolingual speech recognition research activities we have designed two common phonetic alphabets of different level of details [6]. In a first step, language specific phone sets were simplified following available SAMPA transcription guidelines (see [8]) which affected each languages phone set to a different degree: While, for example the native French phone set remained unchanged, we gave up syllabic consonants for German, and at the same time introduced new diphthongs for British English. Then, language specific phones mapped to the same SAMPA symbol were merged into a common unit. This resulted in a common phonetic alphabet consisting of 121 phones (65 vowels, 56 consonants) for the seven languages. As can be seen in Table 1, this gave an overall reduction of 60 percent compared to the simplified language specific phonologies.

	Total	En	Fr	Gr	It	Es	Pt	Ar
vowels	65	20	17	23	14	10	20	14
consonants	56	24	19	26	32	30	22	29
Total	121	44	36	49	46	40	42	43

Table 1. Number for vowel and consonant phones for seven languages in the detailed common phone set. Languages are British English (En), French (Fr), German (Gr), Italian (It), Spanish (Es), Brazilian Portuguese (Pt), Arabic (Ar).

To further increase the overlap we also defined a less detailed common phonetic alphabet, cf. Table 2. We achieved this in three steps:

1. we dropped the distinction between stressed and unstressed vowels for Spanish, Italian, and Portuguese
2. we represented all long vowels as a sequence of two (identical) short vowels
3. we splitted diphthongs into their two vowel constituents.

In doing so, the average number of languages that contribute to the training data for each of the 76 phones (the sharing factor) increased from 2.28 to 2.53, and if Arabic is not considered, the sharing factor increased from 2.74 to 3.56.

	Total	En	Fr	Gr	It	Es	Pt	Ar
Vowels	31	13	15	17	7	5	12	11
Consonants	45	24	19	23	28	24	22	28
Total	76	37	34	40	35	29	34	39

Table 2. Number of vowels and consonants for seven languages in the reduced common phone set.

While this more aggressive inventory reduction caused an increase in word error rate by about 7 percents (measured on an in-house database), if compared to the more detailed common phone alphabet, a benefit of the reduced phone inventory stems from the fact that additional languages with can be covered with less new phones as with the detailed inventory. The integration of eight additional languages (Czech, Japanese, Finnish, Greek, Dutch, Danish, Norwegian, and Swedish) required only 2 additional vowels and 12 consonants. The result makes us believe that the slight degradation in accuracy is tolerable and likely to be adjustable by improved acoustic modelling techniques.

Following the merging procedure outlined above, for our bilingual French–German acoustic model we ended up with 57 phones that are used for recognition of utterances from a general domain. However, since digit recognition is of particular importance in many applications for embedded devices, our language specific phonologies for embedded acoustic modeling were already enriched with some digit specific phones. In order to minimize the impact from using a common phone set on digit recognition, we decided to keep additional 59 digit specific phones separate for each language, which finally resulted in a set of 116 phones that are used in the bilingual model.

3 Speech Recognition

Multilingual acoustic modeling facilitates the development of speech recognizers for languages with only little available training data, and also allows reduced complexity of application development by creating acoustic models that can simultaneously recognize speech from several languages [7]. The use and combination of multilingual acoustic models has also proven advantageous for the recognition of accented speech produced by a wide variety of non-native speakers with different commands of the system's operating language [4].

Acoustic modeling for multilingual speech recognition to a large extend makes use of well established methods for (semi-)continuous Hidden-Markov-Model training. Methods that have been found of particular use in a multilingual setting include, but are not limited to, the use of multilingual seed HMMs, the

use of language questions in phonetic decision tree growing, polyphone decision tree specialization for a better coverage of contexts from an unseen target language, and the determination of an appropriate model complexity by means of a Bayesian Information Criterion; see, for example, [7],[3] for an overview and further references.

Having now reached a certain maturity, the benefits of multilingual acoustic models are most evident in applications that require both robustness against foreign speakers and the recognition of foreign words. To see the same benefits also in embedded and mobile domain, the special needs have to be considered.

In the remainder of this section we will review the basic requirements imposed by the scenario under consideration and will describe how these are taken into account in the training of acoustic models as well as in the recognition phase.

The design of an embedded speech recognizer has to deal with only limited computational resources, both in terms of CPU power and memory capacity, that today's embedded devices can offer. While some applications may run entirely on the local device and therefore require a relatively compact acoustic model, others may defer parts of the recognition process to a recognition server, which requires compatibility of at least the client's and server's acoustic front–end.

The latter is ensured by the use of a standard acoustic front–end, that computes 13 Mel Frequency Cepstrum Coefficients (MFCC) every 15 milliseconds. Utterance based cepstral mean subtraction and C0 normalization are applied to compensate for the acoustic channel and the first and second order delta coefficients are computed to capture the temporal dynamics of the speech signal.

Recognizer training comprises the definition of a suitable HMM inventory and the determination of the HMM parameters. For that purpose, the training data is viterbi–aligned against its transcription in order to obtain an allophonic label for each feature vector. To bootstrap the initial multilingual system two monolingual systems had been used in the alignment step.

Context dependent non cross–word triphone HMMs are obtained from the leaves of a decision network [1] that is constructed by asking binary questions about the phonetic context P_i for each feature vector, $i = -1, \ldots, 1$. These questions are of the form:"Is the phone in position i in the subset S_j?", and the subsets are derived from meaningful phone classifications commonly used in speech analysis. Finally, the data at each leaf of the network is used in a k–means procedure to obtain initial output probabilities whose parameters are then refined by running a few iterations of the forward–backward algorithm.

The k–means procedure follows a simple rule of thumb and equally distributes a fixed number Gaussian mixture components across the HMM states. Usually, in a highly dynamic and heterogeneous environment, an increased total number of Gaussians can significantly improve the recognition accuracy. However, this is infeasible for applications that have to deal with a limited amount of memory, and therefore the determination of an appropriate acoustic model size is of particular importance.

The so created acoustic model can run with either IBM's large vocabulary telephony speech recognition engine, which employs a fast pre–selection

of candidate words and an asynchronous stack search algorithm, or with a time–synchronous viterbi–decoder. The latter is the core of IBM's Embedded Speech Engine (ESE), which is designed for the use with a moderate vocabulary size and finite state grammars. The highly portable and scalable ESE can run on any suitable 32 bit general–purpose CPU; see [2] for an overview on design issues and performance.

4 Experiments

We used the procedure outlined in the previous section to generate a bilingual French–German acoustic model and to compare its recognition accuracy on different tasks to the respective monolingual models. For that purpose we used approx. 600.000 training utterances per language that were gathered in several in-car data collection efforts. Being primarily interested in application for embedded devices, we restricted the size of the bilingual acoustic model to that of the monolingual German model which uses roughly 800 triphone HMMs with a total of approx. 16.000 Gaussian mixture densities.

Our in-car speech recognition test scenario comprised both a digit recognition task and a radio command and control task. For the digit recognition task we run four different tests with an close digits grammar:

- German test set with German pronunciation only,
- German test set with mixed German and French pronunciation,
- French test set with French pronunciation only,
- French test set with mixed French and German pronunciation.

While experiments with only native pronunciations allow to measure the influence of multilingual modeling, we consider the experiments with mixed pronunciations as a step towards the development of true multilingual applications where grammars are shared across languages if possible. However, since only very few French–German homographs appear in the command and control grammars, in this case we experimented only with native pronunciations. Our automotive test set comprises 40 different speakers (20 male, 20 female), each of them recorded in 4 different driving conditions at 0 km/h (with engine off), 40 km/h, 80 km/h, and 110 km/h.

Digit recognition results averaged over all driving conditions are presented in Table 3. While a comparison with the mono-lingual baseline systems demonstrate the feasibility of our multilingual modeling approach, we faced a significant increase in word error rate when we allowed pronunciations from the second language. From the table is clear, that this is mainly caused by an increased number of insertions and substitutions. For German we found that most of these errors are due to the insertion of the word "un", which is French for "1". There is also an increased number of substitutions, but — different from our expectations — this problem was not caused by the confusion of the phonetically similar German digit 6 – "sechs" with the French digit 7 "sept".

For French we observed a larger decrease in accuracy when comparing the mono-lingual and the multilingual models. However, this is an expected result, since the French baseline system was specifically tuned for digit recognition and it uses many more Gaussians for digit phones than both the German and the bilingual model.

	Del	Ins	Sub	total
German mono	0.46%	0.31%	2.16%	*2.94%*
German pr. only	0.47%	0.33%	2.30%	*3.10%*
German + French pr.	0.45%	0.63%	3.75%	*4.83%*
French mono	0.40%	1.57%	1.09%	*3.06%*
French pr. only	0.40%	2.18%	1.75%	*4.32%*
French + German pr.	0.39%	2.28%	2.30%	*4.97%*

Table 3. Word error rates for digit recognition with mono- and bilingual acoustic models.

	Del	Ins	Sub	total
German mono	0.10%	0.10%	2.29%	*2.49%*
German + French	0.15%	0.15%	3.46%	*3.75%*
French mono	0.26%	0.53%	3.92%	*4.71%*
French + German	0.39%	2.16%	4.53%	*7.08%*

Table 4. Word error rates for radio command and control phrases.

Results for the command and control test set are given in Table 4. The results are comparable to the digit test, and also show a larger increase in word error rate for the French system. We assumed that the main reason for the loss in accuracy is the system size of the bilingual model, which is equal to the German monolingual model. In order to prove this assumption we therefore increased the number of Gaussians in the bilingual model by approx. 50 percent. Digit recognition results obtained with this model for both open and closed digit grammar are given in Table 5.

The achieved results confirm our assumption that multilingual speech recognition requires larger acoustic models. However, the size of the larger model with 24k Gaussians is already beyond the limit of todays commonly available embedded or mobile devices. Therefore, improved acoustic modeling still remains a need for future research.

	close		open	
System size	16k	24k	16k	24k
German mono	2.94%	–	3.52%	–
German pr. only	3.10%	2.93%	3.78%	3.56%
German + French pr.	4.83%	4.49%	6.31%	5.88%

Table 5. Digits error rates obtained on the German test for different system size and close and open grammar.

5 Summary

In this paper we described various aspects of the development of a bilingual acoustic model for embedded devices. We explored the design of a common phonetic alphabet for up to 15 languages and described techniques for the training of highly noise robust acoustic models necessary for mobile as well as embedded devices. Experiments in bilingual French–German speech recognition demonstrated the feasibility of our approach, but at the same time unveiled the need for further research in acoustic modeling in order to create multilingual system of acceptable footprint without an unwanted decrease in accuracy.

References

1. L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, M. Picheny. Context-dependent Vector Quantization for Continuous Speech Recognition. In *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, 1993.
2. T. Beran, V. Bergl, R. Hampl, P. Krbec, Jan Šedivý, B. Tydlitát, J. Vopička. Embedded ViaVoice. In *Proc. of TSD 2004*, Brno, 2004.
3. V. Fischer, J. Gonzalez, E. Janke, M. Villani, C. Waast-Richard. Towards Multilingual Acoustic Modeling for Large Vocabulary Speech Recognition. In *Proc. of the IEEE Workshop on Multilingual Speech Communications*, Kyoto, 2000.
4. V. Fischer, E. Janke, S. Kunzmann. Likelihood Combination and Recognition Output Voting for the Decoding of Non-Native Speech with Multilingual HMMs. In *Proc. of the 7th Int. Conference on Spoken Language Processing*, Denver, 2002.
5. V. Fischer, S. Kunzmann. Bayesian Information Criterion based Multi-style Training and Likelihood Combination for Robust Hands Free Speech Recognition in the Car. In *Proc. of the IEEE Workshop on Handsfree Speech communication*, Kyoto, 2001.
6. S. Kunzmann, V. Fischer, J. Gonzalez, O. Emam, C. Gnther, E. Janke. Multilingual Acoustic Models for Speech Recognition and Synthesis. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, 2004.
7. T. Schultz, A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition, *Speech Communications*, Vol. 35, 2001.
8. C.J. Wells. Computer Coded Phonemic Notation of Individual Languages of the European Community, *Journal of the International Phonetic Association*, Vol. 19, pp. 32-54, 1989.