MULTILINGUAL ACOUSTIC MODELS FOR SPEECH RECOGNITION IN LOW-RESOURCE DEVICES

Enrique Gil Garcia, Erhan Mengusoglu, Eric Janke

IBM AIM, European Voice Technology Development
IBM UK Laboratories, Hursley Park, Winchester SO21 2JN, England
{egilgarcia, mengusog,eric_janke}@uk.ibm.com

ABSTRACT

Multilingual access to information and services is a key requirement in any pervasive or ubiquitous computing environment. In this paper we review the design of a common alphabet for up to fifteen languages and describe its application to multilingual speech recognition in low-resource devices in real-time. We give an overview of the special requirements for acoustic modeling in such environments and present initial results of a technique that aims on a more efficient discrimination between languages in training while keeping low memory footprint. We also report the usefulness of a multilingual recognizer as a language-independent system to bootstrap a new language.

Index Terms— multilingual speech recognition, common phone alphabet, spoken language identification, language independent recognizer

1. INTRODUCTION

In areas like Europe, with a huge number of consumers speaking different languages, there has arisen a particular interest both in multilingual speech recognition [1] and in the fast bootstrapping of new languages [4]. Moreover, the proliferation of speech applications on low-resource devices like palm-size computers, mobile phones, in-car navigation systems, etc puts additional emphasis on developing ASR systems with small memory footprint and low cpu usage [7, 2]. These constrains make the simple combination of monolingual systems unfeasible. The challenge is then to design a multilingual ASR system with the complexity of a monolingual system which is still capable of recognizing several languages (let's say: UK, GR, FR, ES, IT …) equally well, i.e., with a level of accuracy comparable with their counterpart monolingual systems. Finally (and no less importantly), such a multilingual recognizer can be considered as a language independent or language-neutral system [4] that makes it possible to reduce time-to-market cycle for new languages.

	total	Es	It	En	Gr	Fr	Nl	Se
vow	77	10	10	12	14	17	14	23
con	69	28	28	24	26	19	21	23
total	146	38	38	36	40	36	35	46

Table-1. Number of vowels and consonants for seven languages : Spanish (Es), Italian (It), British English (En) German (Gr), French (Fr), Dutch (Nl) and Swedish (Se).

The remainder of the paper is organized as follows: in Section 2 we give a brief overview of the common phone alphabet used in this work. In Section 3 we focus on multilingual acoustic modeling and present experimental results. In Section 4 we outline a technique aiming to improve the acoustic separation between languages in training. In Section 5 we consider a multilingual system as a language-independent system to bootstrap a new language. Finally, Section 6 gives a conclusion and some prospects of future work.

2. COMMON PHONOLOGY

The definition of a common phonetic alphabet for multilingual speech recognition has to consider two conflicting design issues: on one hand the different sounds of each language should be covered separately in order to achieve high recognition accuracy, while on the other, as many phones as possible should be shared across languages both for efficient utilization of training data and to achieve reasonably small acoustic models.

We designed a common phonetic alphabet starting from the existing phonetic alphabets for seven languages (Arabic, British English, French, German, Italian, (Brazilian) Portuguese, and Spanish) [1,3] with further extensions to cover additional languages (fifteen in total) including American English, European Portuguese, Japanese, Greek, Czech, Finnish, Norwegian. For that purpose, the language specific phone sets were first simplified following available SAMPA transcription guidelines [6]. With this approach, languages were affected to different degrees: while the

	Es	It	En	Gr	Fr	Nl	Se
Es	-	18	19	18	15	17	14
It	10	-	18	20	16	17	14
En	3	3	-	20	17	20	17
Gr	4	5	4	-	17	19	18
Fr	7	7	4	10	-	17	14
Nl	5	5	5	9	9	-	17
Se	2	2	5	5	5	8	-

Table-2. Number of shared consonants (upper right corner) and vowels (lower left corner) of the merged alphabet.

	total	Es	It	En	Gr	Fr	Nl	Se
vow	77	4	6	-	2	1	-	5
con	69	-	-	3	3	4	-	12
total	146	4	6	3	5	5	-	17

Table-3. Number of vowel (vow) and consonant (con) phonetic units that belong to only one of the seven languages.

native French phone set remained unchanged, we gave up syllabic consonants for German, and at the same time introduced new diphthongs for British English. In a second step, language specific phones mapped to the same SAMPA symbol were merged into a common unit. This yielded a common phonetic alphabet of 146 phones (77 vowels, 69 consonants) for the fifteen languages.

Table-1 shows the phones actually used for the seven languages of interest in this work. We represented all long vowels and diphthongs as a sequence of two short vowels (but for Swedish). In doing so, the sharing factor or average number of languages that contribute to the training data for each of the 94 phones (64% of the total 146 phones) turned out to be 2.86. The phone sharing factor becomes a trade-off between enhancing the differentiation between languages and reducing the complexity of the system because of both reliable estimation of its parameters and limited resources.

3. MULTILINGUAL ACOUSTIC MODELING

Acoustic modeling for multilingual speech recognition, to a large extent, makes use of well established methods for (semi-) continuous Hidden-Markov-Model training. Methods that have been found of particular interest in a multilingual setting include, but are not limited to, the use of multilingual seed HMMs, the use of language questions in phonetic 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 (BIC) [10]; cf., for example, [1,5] for an overview and further references.

The training of a rank-based speech recognition system [9] is a bootstrap procedure that comprises feature extraction, the construction of a set of context dependent, allophonic HMMs and the subsequent estimation of the continuous density Gaussian mixture parameters.

The acoustic models reported in this paper were designed for medium-vocabulary, grammar-based recognition in low-resource devices [7]. As a prerequisite, the seven languages considered use a common acoustic front-end that computes 13 MFCC (including C0) and their first and second order derivative every 15 miliseconds. The training data are Viterbi-aligned against their transcription and each acoustic vector is context-tagged. The context for a given phone is restricted to adjacent phones and the system uses within-word context only, i.e., context does not extend over word boundaries. The allophones are identified by growing a decision network (with binary phonetic context questions) using the context-tagged feature vectors. Each terminal node (or leaf) of the network is modeled by a single state HMM with a self loop and a forward transition. The acoustic observations that characterize the training data at each leaf are modeled by a Gaussian Mixture Model (GMM), with diagonal covariance matrices to give an initial acoustic model. The complexity of the model is selected by the use of BIC criterion. Starting with these initial set of GMMs several iterations of the standard Baum-Welch EM algorithm is run to refine the model.

Using the outlined method, the training of the multilingual models was performed as follows :

1. After mapping the language dependent phone sets and phonetic context questions and rewriting of training baseforms initial monolingual models were trained.
2. The language specific training vocabularies were merged (a language tag added to each spelling) and the monolingual models were used to viterbi-align the training data from each language.
3. Based on the (monolingual) alignments a common decision network was constructed to obtain multilingual seed HMMs.
4. Data from all languages was used for the forward-backward refinement of the initial HMM parameters.
5. In the experiments described later on, these sets of multilingual HMMs were also used for the viterbi-alignment step in a further iteration.

For the evaluation we aimed to use roughly the same amount of training data (in-car recordings at different speeds and conditions) for each of the seven languages. However, because of heterogeneous databases, we ended up using 50000 utterances for each language which yielded a minimum of 16.4 hours for Swedish (Speecon database) to a maximum of 44.1 hours for Spanish (~50% phonetically rich

	Es	En	It	Gr	Fr	Nl
Mono	2.2	7.7	11.7	7.7	4.9	7.1
M20	2.3	9.1	-	-	-	-
M21	2.4	-	13.3	-	-	-
M22	-	-	-	8.9	6.4	-
M3	2.5	9.5	13.4	-	-	-
M5	3.0	11.6	14.4	10.3	8.6	-
M7	3.1	13.4	14.9	12.1	9.1	10.1
M7Es	2.4	-	-	-	-	-

Table-4. Word error rate (WER %) for in-car data

text sentences). Other languages were in the range 25h-35h . No attempt was made to standardize the test scenarios.

Our efforts on the creation of acoustic models that allow of the seamless recognition of the seven languages are summarized in Table-4. All systems have ~1K leaves. Monolingual systems comprise of ~14K gaussians, whereas multilingual systems have ~20K gaussians. M20 is a bilingual Spanish-English (Es+En) system, M21 is a bilingual Spanish-Italian (Es+It), M22 is a bilingual German-French (Gr+Fr), M3 is M21+En, M5 is M3+Gr+Fr and M7 is M5+Nl (Dutch)+Se (Swedish).

In the first row of Table-4, we have the monolingual systems decoding their respective monolingual test sets. Those figures give a baseline for the multilingual systems to compare with. The next rows show the degradation when incorporating more languages to the multilingual builds. The degradation observed might be explained to some extend by the use of less dedicated gaussians per language (we aimed to accommodate the languages in ~20K gaussians). Other source of degradation might come from the common/shared leaves (gausssians) between languages. To assess that point, we built a monolingual Es system (M7Es) by re-using the multilingual decision network of M7. WERs of Mono Es (2.2) and M7Es (2.4) are now closer. Both monolingual systems (~14K gaussians) were built with the same data. If we compare again Mono Es (2.2) with M7 (3.1), now we could split the degradation in: 0.2% (~9% relative) due to a multilingual decision network (instead of a true monolingual) and 0.7% (~32% relative) due to the influence of the data of the other languages (in the shared leaves, the gaussians were generated with a mixture of data from several languages instead of monolingual data).

4. LANGUAGE DEPENDENT LABELLING

The comparison of M7Es with M7 in Table-4 suggests that adding more data from other languages to the shared leaves of the multilingual decision network did not help a particular language by introducing confusion (sparse data) for the

	Es	En
1 - Mono Es (14K)	1.9	-
2 - Mono En (12K)	-	7.1
3 - Bilingual Es+En (16K)	2.2	8.6
Models for LDL merging		
4 - Mono Es (8K)	2.1	-
5 - Mono En (8K)	-	8.1
6 - Bilingual Es+En (10K)	2.3	8.9
LDL models		
7 - LDL Es+En (14K)	2.3	10.0
8 - LDL Es+En (15K)	2.3	10.2
9 - LDL Es+En (16K)	2.4	8.9

Table-5. Baseline systems (1,2,3) ; Models for language dependent labeling (4,5,6) ; LDL models (7,8,9)

gaussian generation. It seems to point out that there is a margin for improvement if we manage to enhance the separation/discrimination of gaussians between languages. Then, for a bilingual case, the proposed algorithm, initially presented in [8] for gender clustering, comprises of the following steps:

1. Create three systems: a bilingual system (with its bilingual decision network and its bilingual gaussians) and two monolingual systems re-using the bilingual decision network.
2. For each leaf of the bilingual decision network, compute a measure of statistical distance between the two sets of monolingual gaussians associated at each leaf.
3. Merge the two sets of monolingual gaussians at each leaf. Then, starting from the leaf with the lowest distance, replace the merged monolingual gaussians with their counterpart bilingual ones until a target number of gaussians is reached.

In training, the aim is to enhance the language separation by keeping the bilingual gaussians only for those leaves whose monolingual gaussians are statistically close enough. Otherwise, we merge the monolingual gaussians. This merging mechanism tries to keep the final number of gaussians under a certain limit as well (footprint issues). At runtime, the computational cost is low and the aim is to be able to identify the spoken language so that (only) the gaussians of the correct language are used (highest weight) to label the input audio (Language Dependent Labelling, or LDL).

Table-5 summarizes the acoustic models created to evaluate the performance of the algorithm. The number in parenthesis stands for number of gaussians. Several LDL models were

	Pt
ML5 (Es+It+En+Gr+Fr)	48.5
ML7 (Es+It+En+Gr+Fr+Nl+Se)	40.7
MAP adaptation of ML7	26.5
Mono Pt	10.4

Table-6. Word Error Rate (WER %) of Portuguese decodings

merged considering different target of number of gaussians. The LDL system (9) is just the union of the two monolingual systems (4,5), whereas LDL systems (7) and (8) contain some bilingual gaussians from model (6). The initial results in Table-5 show a LDL system (9) that performs similarly to a non-LDL system of comparable size (3), although we should enhance the language identification process to get real benefits from this technique.

5. LANGUAGE INDEPENDENT SYSTEM

A multilingual recognizer can be considered as a language-independent or language-neutral system [4] for efficient portability to a new target language by reducing time-to-market cycles. Different approaches can be followed depending on the amount of data (and time) available for the new target language:

- Cross-language transfer: Make use of the multilingual system to recognize the new language directly (i.e., without any training data of the new language).
- Language adaptation: The multilingual system is adapted with limited amount of data of the new language. It also assumes that the multilingual system have a good coverage of the phones of the new language.
- Bootstrapping: Rebuild the system with large amount of training data by using the multilingual system as seed model.

Table-6 illustrates the experiments carried out to bootstrap Portuguese (Pt) as the new/unseen language. It is interesting to note that the multilingual system ML7 outperforms ML5. It seems to suggest that we might boost the cross-language capabilities of multilingual systems by incorporating more languages (and then increasing the phone coverage) into the multilingual systems. Not surprisingly, rebuilding the system with large amount of data of the new language yields the best results.

6. CONCLUSION

In this paper we have presented initial efforts towards multilingual speech recognition in low-resource devices. Experiments that aim on the seamless multilingual recognition of up to seven languages have proven the feasibility of the approach. We have also presented a data-driven technique that aims at a more efficient discrimination between languages in training with low computational cost at runtime, which highlights the need for further research in the language identification process in order to boost the performance of the technique. Finally, we have explored different approaches to bootstrap a new language from a multilingual recognizer.

7. REFERENCES

[1] V. Fischer, J. Gonzalez, E. Janke, M. Villani, 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

[2] Jozef Ivanecky, Volker Fischer, Siegfried Kunzmann, *"French-German Bilingual Acoustic Modeling for Embedded Voice Driven Applications"*, in Proc. of TSD 2005, Karlovy Vary, Czech Republic, 2005

[3] F. Palou Cambra *et al*, *"Towards a common alphabet for multilingual speech recognition"*, in Proc. of the 6th Int. Conf. on Spoken Language Processing, Beijing, 2000

[4] T. Schultz and A. Waibel*, "Language independent and language adaptive acoustic modeling for speech recognition",* Speech Communications, vol. 35, 2001

[5] T. Schultz, A. Waibel, *"Polyphone Decision Tree Specialization for Language Adaptation",* in Proc. ICASSP, Istanbul, Turkey, June 2000

[6] 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

[7] Sabine Deligne, Satya Dharanipragada, Ramesh Gopinath, Benoit Maison, Peder Olsen, and Harry Printz*, "A robust high accuracy speech recognition system for mobile applications"*, IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 551-561, November 2002

[8] Peder Olsen, Satya Dharanipragada, *"An efficient integrate gender detection scheme and time mediated averaging of gender dependent acoustic models",* in Eurospeech 2003, pp. 2509-2512, 2003

[9] L. Bahl, P. de Souza, P. Gopalakrishanan, D. Nahamoo, and M . Picheney, *"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

[10] 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, pp. 645-648, Seattle, 1998