IDIAP RESEARCH REPORT

TOWARDS MIXED LANGUAGE SPEECH RECOGNITION SYSTEMS

David Imseng Hervé Bourlard
Mathew Magimai.-Doss

Idiap-RR-15-2010

JULY 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Towards mixed language speech recognition systems

David Imseng, Hervé Bourlard, Mathew Magimai.-Doss

July 6, 2010

Abstract

Multilingual speech recognition obviously involves numerous research challenges, including common phoneme sets, adaptation on limited amount of training data, as well as mixed language recognition (common in many countries, like Switzerland). In this latter case, it is not even possible to assume that one knows in advance the language being spoken. This is the context and motivation of the present work. We indeed investigate how current state-of-the-art speech recognition systems can be exploited in multilingual environments, where the language (from an assumed set of five possible languages, in our case) is not a priori known during recognition. We combine monolingual systems and extensively develop and compare different features and acoustic models. On SpeechDat(II) datasets, and in the context of isolated words, we show that it is actually possible to approach the performances of monolingual systems even if the identity of the spoken language is not a priori known.

Index Terms: speech recognition, multilingual speech recognition, combination of mono-lingual speech recognition systems, mixed language recognition.

1 Introduction

Multilingual speech processing is nowadays witnessing a renewed interest, not only because of real needs, but also thanks to the convergence of automatic speech recognition (ASR) technologies (mainly due to high performance English recognizers) in the form of powerful statistical parametric methodologies such as generative Gaussian Mixture Models (GMMs) [1], discriminative Multilayer Perceptrons (MLP) as employed in hybrid systems [2] or the combination of discriminative and generative approaches as employed in Tandem systems [3].

Different methodologies have been applied to multilingual ASR. For instance GMM-based monolingual recognizers were trained on different languages, with (e.g. [4]) and without (e.g. [5]) sharing data across languages. Hybrid HMM/MLP systems have also been applied to multilingual ASR [6, 7] and multilingual Tandem systems have been presented in [8] for example.

Even if data from multiple languages was used, most studies required to explicitly identify the language in order to process the data with the correct recognizer, properly trained on a particular language. In the presented work, we also consider systems where the language identity is not a priori known. More specifically, we compare different features and acoustic models on a monolingual and a mixed language isolated word recognition task on SpeechDat(II) data. In the monolingual task, we assume that the language identity is known in advance and in the mixed language task, we consider a system that infers the language implicitly, as a by-product of the recognition process, by running multiple recognizers in parallel and performing a score-based output decision.

An advantage of systems that are not aware of the language identity (mixed language task) is that they do not require to explicitly perform language identification. However, usually, the performance of such systems is lower compared to systems that know the language a priori (monolingual task). We compare the difference in terms of performance between the mono- and multi-lingual task using different features, namely, PLP cepstral coefficients, Tandem features, and different acoustic modeling techniques, namely GMM-based and MLP-based. We demonstrate that there are indeed considerable differences between the monolingual scenario and the mixed language scenario. Our study exposes two trends if score-based multilingual output decisions are performed: firstly, MLP-based acoustic modeling seems to be preferable to GMM-based acoustic modeling and secondly, using Tandem features extracted from an MLP trained to classify a set of universal phonemes (created by merging the phoneme sets of the languages considered) yields a better system compared to the case where an MLP is trained for each language individually (to classify the language specific phonemes). We exploit these findings by using Tandem features extracted from an MLP trained to classify universal phonemes and MLP-based acoustic modeling to build a system that yields the best performance on our mixed language isolated word recognition task.

The remainder of this paper is structured as follows: Section 2 presents the databases that are used and defines the monolingual and the mixed language task. Section 3 describes the different features followed by a presentation of the evaluated systems in Section 4. Section 5 discusses the experimental results, and finally Section 6 concludes the paper.

2 Databases and Tasks

In this section we introduce the SpeechDat(II) databases that we used and define the tasks on which we are evaluating and comparing different systems.

2.1 Databases - SpeechDat(II)

We used data from SpeechDat(II) that currently consists of recordings from 14 different European countries. In order to be representative, the SpeechDat(II) databases are gender-balanced, dialect-balanced according to the dialect distribution in a language region and age-balanced. The databases are subdivided into different corpora. We only used *Corpus A*, that contains three isolated read application words per speaker. The term *application words* describes a set of about 30 words such as “help” or “cancel”, which could be used in interactive voice response applications.

To build comparable systems, test sets, that preserve the gender, dialect and age distributions of the original set, were specified for every database and standardized test routines were described in [9]. For this paper, we used the datasets of five languages, namely British English (EN), Swiss French (SF), Swiss German (SZ), Italian (IT), and Spanish (ES). In Swiss German, there are 2000 recorded speakers. As standardized by SpeechDat(II), datasets with a minimum of 2000 speakers have pre-defined test sets that contain the data of 500 speakers. The remaining 1500 speakers are sub-divided into a development set (10%, 150 speakers) and a training set (1350 speakers). To avoid any bias in terms of available amount of data towards a particular language, the same number of speakers was used in all languages, even if other databases provide data from more than 2000 different speakers. For this purpose, a subset of 2000 speakers was chosen from the whole dataset by using the same procedure as for the test set creation and then the subset was split into training, development and test sets. Hence, we did not use the pre-defined test sets, rather used the scripts available at [9] to ensure that the splits can be reproduced.

There are several commonly defined tests on the SpeechDat(II) databases [9]. For our work, we used the *A-test* (test on Corpus A) also referred to as application words test which is a small vocabulary isolated phrase test. Similar to the previous work [5], the utterances with out-of-vocabulary words, mispronunciation, unintelligible speech or truncations were excluded in all procedures and noise markers were ignored. Table 1 summarizes the number of utterances out of the total possible 6000 utterances (three utterances from each of the 2000 speakers) considered for each language and their distribution across the training, test and development set. The total duration of the utterances is also given (in hours).

Table 1: *Number of available utterances (utt.), and total duration in hours (h), for each of the five considered languages. British English (EN), Spanish (ES), Italian (IT), Swiss French (SF) and Swiss German (SZ).*

Lang.	training		dev		test		total	
	utt.	h	utt.	h	utt.	h	utt.	h
EN	3512	1.2	390	0.1	1305	0.4	5207	1.7
ES	3932	1.4	438	0.2	1447	0.5	5817	2.0
IT	3632	1.5	416	0.2	1368	0.6	5416	2.3
SF	3809	1.4	430	0.2	1429	0.5	5668	2.1
SZ	3862	1.3	432	0.1	1426	0.5	5720	1.9
total	18747	6.8	2106	0.8	6975	2.5	27828	10.0

The database provides a lexicon for each language that contains the pronunciations for the words in terms of the SAMPA[1] phoneme set. We use these lexicons for our study. Table 2 displays the number of phonemes that are used for the application words task. Note that some languages do not use all the available phonemes for the application words task.

Table 2: *Number of phonemes used per language for the application words task.*

Language	EN	ES	IT	SF	SZ
# phonemes	33	29	35	36	46

In this work, we build an isolated word/phrase recognizer for each language and compare them on two different tasks, namely, monolingual task and mixed language task.

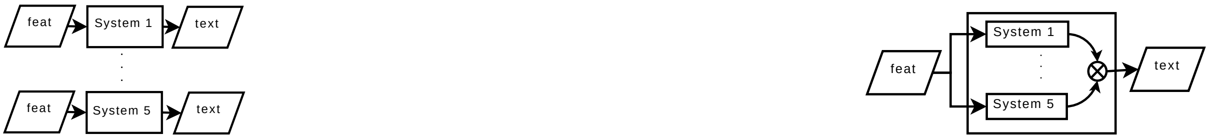
[1] http://www.phon.ucl.ac.uk/home/sampa/index.html

2.2 Monolingual task

In the monolingual task, given the trained ASR systems for all the five languages and the language identity of the test utterance, we select and run the monolingual recognizer corresponding to the language. In other words, the monolingual task is a system that “knows” the language a priori during testing, therefore optimal recognition is performed by decoding each test utterance with the correct monolingual recognizer. The monolingual task serves as the reference task in our studies. Figure 1(a) depicts the monolingual task.

2.3 Mixed Language task

In the mixed language task, we consider a system where the spoken language identity is “not known” a priori and is implicitly inferred by running multiple monolingual recognizers in parallel. In other words, the mixed language task can be seen as a *black box* system as illustrated in Fig. 1(b), where we run all the five monolingual recognizers and select the one with the maximum likelihood as the recognized output.



(a) Monolingual Task

(b) Mixed language Task

Figure 1: *Visualization of the different tasks. Five monolingual recognizers are build. In the monolingual task, the language is known in advance during testing whereas in the mixed language task, no language information is available during testing.*

Both tasks are evaluated using three different feature types and two different acoustic modeling techniques.

3 Features

In this section, we describe the different types of features that are used in our work.

3.1 Perceptual Linear Prediction (PLP)

The first type of features are conventional PLP cepstral features [10]. Twelve cepstral coefficients including the zeroth coefficient are used and additionally, delta and acceleration coefficients are appended. The 39 dimensional PLP features are extracted every 10 ms on a 25 ms window after having performed voice activity detection using Tracter[2].

[2] http://juicer.amiproject.org/tracter/

3.2 Monolingual Tandem Features

Multilayer Perceptrons (MLPs) can be used as feature extractors as in Tandem systems [3]. For each language, an MLP is trained to estimate phoneme posteriors based on the extracted PLP features (Section 3.1). After having taken the logarithm of the posteriors, the Karhunen-Loève transformation (KLT) is applied without performing any dimensionality reduction and then the concatenated feature vectors (PLPs and processed posteriors) are used as input to a monolingual recognizer. The process of extracting Tandem features is done for each language individually, thus we refer to it as monolingual Tandem features. Figure 2 illustrates one of the five systems based on the monolingual Tandem features.

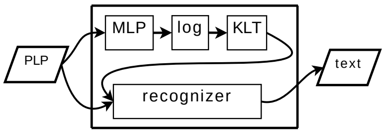


Figure 2: *Monolingual Tandem features. The estimated posteriors are post-processed by taking the logarithm and performing a Karhunen-Loève transformation (KLT) and then used as input to a recognizer together with the conventional PLP features.*

3.3 Multilingual Tandem Features

Instead of extracting Tandem features for every language separately by training a separate MLP, some components of the Tandem feature extraction process can be shared across languages. The dictionaries of the SpeechDat(II) datasets are all in the international SAMPA format. A universal phoneme set was built by merging phonemes across languages that are represented by the same symbol (knowledge-based approach [11]). The universal phoneme set consists of 92 phonemes (more details can be found in [12]). In contrast to the monolingual Tandem features, only a single MLP (instead of five MLPs) is trained to estimate posterior probabilities of the universal phonemes for all languages. KLT is then used to perform a dimensionality reduction for each language individually[3] such that the multi- and the monolingual Tandem features have the same dimensionality. The individually processed posteriors are then concatenated with PLP features and used as input for the monolingual recognizers. Figure 3 illustrates the system based on multilingual Tandem features.

4 System description

We investigate two kinds of acoustic modeling techniques within the framework of HMM-based ASR systems. The first kind of acoustic modeling technique uses Gaussian mixture models (GMM) to model the acoustics/feature observation [1], and the second type of acoustic modeling technique, uses an MLP classifier to model the acoustics/feature observation [2]. Furthermore, we study the two acoustic modeling techniques using three different kinds of features. Thus, we build and compare six different systems (also shown in Table 3):

[3] The transformation matrix of the KLT is estimated for every language separately on the corresponding training data.

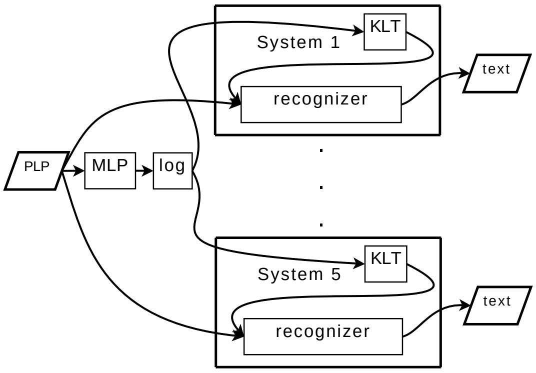


Figure 3: *Multi-Tandem system: only one neural network is trained (instead of five networks) on the data of all involved languages, then the KLT is used to perform dimensionality reduction for each language individually.*

1. HMM/GMM system: HMM/GMM-based ASR using PLP features.
2. HMM/MLP system: Hybrid HMM/MLP-based ASR using PLP features.
3. Mono-Tandem: HMM/GMM-based ASR using monolingual Tandem features.
4. Mono-MLP-Tandem: Hybrid HMM/MLP-based ASR using monolingual Tandem features.
5. Multi-Tandem: HMM/GMM-based ASR using multilingual Tandem features.
6. Multi-MLP-Tandem: Hybrid HMM/MLP-based ASR using multilingual Tandem features.

Note that the systems Mono-MLP-Tandem (4) and Multi-MLP-Tandem (6) are different from conventional Tandem systems in the sense that they use a discriminative classifier in the form of an MLP instead of a generative GMM classifier to model the feature observations.

5 Experimental Results and Discussion

We build context-independent phoneme based isolated word recognition systems, where each context-independent phoneme is modeled by a three state left-to-right HMM. The number of context-independent phonemes for each language can be found in Table 2.

We used the HTK toolkit [13] for the training and recognition of the GMM-based systems, where each state is modeled by 32 mixtures of Gaussians with diagonal covariance matrices.

For the MLP-based systems, a three layer MLP was trained to classify context-independent phonemes with quicknet software[4]. The input to the MLP contained the feature vector at the current time frame plus four frames preceding and following context (i.e., nine frames in total). In case of HMM/MLP systems,

[4] http://www.icsi.berkeley.edu/Speech/qn.html

all the MLPs had 600 hidden nodes. The MLPs of the HMM/MLP systems were used for monolingual Tandem feature extraction. The MLP for multilingual Tandem feature extraction had 524 hidden nodes (this was done in order to ensure that in average sense the number of parameters is comparable to a single monolingual MLP). The MLP classifiers used in the Mono-MLP-Tandem system and in the Multi-MLP-Tandem system contained 600 hidden nodes.

In the case of the mixed language task, before making a decision (i.e. choosing the output word hypothesis that yields maximum likelihood) a recognizer dependent bias was subtracted from the respective log likelihood scores similar to [14]. More specifically, we run all the recognizers on the development set and estimated the average log likelihood, which is used as bias.

The results of the experiments are shown in Table 3. The performance of the systems is expressed as average performance on all five languages (the individual performance of each language can be found in [12]).

Table 3: *Experimental results. The different approaches are described in Section 4. The performance on the monolingual and the mixed language task are shown and also the relative change between the two tasks is given.*

System	Acoustic modeling	Features	Task		Relative change
			monolingual	mixed language	
HMM/GMM	GMM	PLP	98.4	78.2	-21%
HMM/ANN	ANN	PLP	97.5	86.3	-11%
Mono-Tandem	GMM	monolingual Tandem	98.7	77.2	-22%
Mono-MLP-Tandem	ANN	monolingual Tandem	98.5	86.9	-12%
Multi-Tandem	GMM	multilingual Tandem	98.8	82.9	-16%
Multi-MLP-Tandem	ANN	multilingual Tandem	98.5	88.8	-10%

On the monolingual task, the HMM/MLP performance is lower than the performance of the other systems. All the other systems only slightly differ in performance among each other. In literature, it has been typically observed that the use of Tandem features yields performance improvements. However, for the monolingual task we do not observe such improvements. This may be due to the easy nature of the recognition task, i.e., small vocabulary isolated word recognition.

On the mixed language task however, there are considerable differences between the performance of different systems. It can be observed that the multilingual Tandem features yield the best system for both, the GMM-based acoustic model and the MLP-based acoustic model. This may be due to the sharing of information about different languages through the discriminatively trained single MLP which is used for the multilingual Tandem feature extraction.

Further analysis of the performance change between the monolingual task and the mixed language task among different approaches, exposes a general trend, that the MLP-based acoustic modeling technique yields less relative loss than the GMM-based acoustic modeling technique. In case of PLP or monolingual Tandem features, this trend is more pronounced (almost a factor of two when compared to the respective MLP-based systems). Altogether, these results suggest that it may be better to use a discriminative acoustic modeling technique such as MLP for the mixed language task.

6 Conclusion and Future Work

In this paper, we investigated the performance of speech recognition systems with different features and acoustic modeling techniques on a mixed language task (where the language identity of the test utterance is assumed to be unknown), and compared it against the performance on a monolingual speech recognition task (where the language identity of the test utterance is assumed to be known). Our studies on isolated word recognition show that there is a significant performance difference between the monolingual task and the mixed language task. However, this difference may be better bridged by the use of multilingual Tandem features and discriminative acoustic modeling techniques, such as MLP.

In future, we intend to explore other techniques to build a universal phoneme set and propose to extend our study on mixed language recognition to the use of lexicons defined with a universal phoneme set (as opposed to language specific phoneme sets) and to the phonetically rich sentences task of Speech-Dat(II) database.

7 Acknowledgments

This research was supported by the Swiss NSF through the project MultiModal Interaction and Multimedia Data Mining under contract number MULTI-200020-122062 and through the National Center of Competence in Research on ”Interactive Multimodal Information Management” (www.im2.ch). The authors would like to thank Phil Garner for his help with Tracter.

References

[1] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] N. Morgan and H. Bourlard, “Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach,” *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, 1995.

[3] H. Hermansky, D. P. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” in *Proc. of ICASSP*, 2000, pp. 1635–1638.

[4] T. Schultz and A. Waibel, “Language independent and language adaptive acoustic modeling for speech recognition,” *Speech Communication*, vol. 35, pp. 31–51, 2001.

[5] B. Lindberg, F. T. Johansen, N. Warakagoda, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, “A noise robust multilingual reference recogniser based on Speechdat(II),” in *Proc. of ICSLP*, 2000, pp. 370–373.

[6] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, “On the use of a multilingual neural network front-end,” in *Proc. of Interspeech*, 2008, pp. 2711–2714.

[7] S. Dupont, C. Ris, O. Deroo, and S. Poitoux, “Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents,” in *Proc. of ASRU*, 2005, pp. 29–34.

[8] L. Tóth, J. Frankel, G. Gosztolya, and S. King, “Cross-lingual portability of MLP-based tandem features – a case study for English and Hungarian,” in *Proc. of Interspeech*, 2008, pp. 2695–2698.

[9] G. Chollet, F. T. Johansen, B. Lindberg, and F. Senia, “LE2-4001 deliverable identification,” ENST, Telenor, CPK and CSELT, Tech. Rep., 1998.

[10] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1751, 1990.

[11] J. Köhler, “Multilingual phone models for vocabulary-independent speech recognition tasks,” *Speech Communication*, vol. 35, pp. 21–30, 2001.

[12] D. Imseng, M. Magimai.-Doss, and H. Bourlard, “A comparison of acoustic modeling techniques and feature types for multilingual speech recognition,” Idiap, Idiap-RR, April 2010.

[13] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge University Press, 2000.

[14] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 4, no. 1, pp. 31–42, January 1996.