Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 142–151

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Using out-of-language data to improve an under-resourced speech recognizer

David Imseng[a,b,*], Petr Motlicek[a], Hervé Bourlard[a,b], Philip N. Garner[a]

[a] *Idiap Research Institute, Martigny, Switzerland*
[b] *Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland*

Available online 8 February 2013

**Abstract**

Under-resourced speech recognizers may benefit from data in languages other than the target language. In this paper, we report how to boost the performance of an Afrikaans automatic speech recognition system by using already available Dutch data. We successfully exploit available multilingual resources through (1) posterior features, estimated by multilayer perceptrons (MLP) and (2) subspace Gaussian mixture models (SGMMs). Both the MLPs and the SGMMs can be trained on out-of-language data. We use three different acoustic modeling techniques, namely Tandem, Kullback–Leibler divergence based HMMs (KL-HMM) as well as SGMMs and show that the proposed multilingual systems yield 12% relative improvement compared to a conventional monolingual HMM/GMM system only trained on Afrikaans. We also show that KL-HMMs are extremely powerful for under-resourced languages: using only six minutes of Afrikaans data (in combination with out-of-language data), KL-HMM yields about 30% relative improvement compared to conventional maximum likelihood linear regression and maximum a posteriori based acoustic model adaptation.

*Keywords:* Multilingual speech recognition; Posterior features; Subspace Gaussian mixture models; Under-resourced languages; Afrikaans

## 1. Introduction

Developing a state-of-the-art speech recognizer from scratch for a given language is expensive. The main reason for this is the large amount of data that is usually needed to train current recognizers. Data collection involves large amounts of manual work, not only in time for the speakers to be recorded, but also for annotation of the subsequent recordings. Therefore, the need for training data is one of the main barriers in porting current systems to many languages. On the other hand, large databases already exist for many languages.

Previous studies have shown that automatic speech recognition (ASR) may benefit from data in languages other than the target language only under certain conditions such as there being less than one hour of data for the training language (Imseng et al., 2012a; Qian et al., 2011). Usually, a language with large amounts of training data is used to simulate small amounts of target training data (Imseng et al., 2012a; Qian et al., 2011). For instance (Niesler, 2007) studied the sharing of resources on real under-resourced languages, including Afrikaans, inspired by multilingual acoustic modeling techniques proposed by Schultz and Waibel (2001). However, only marginal ASR performance gains were reported.

Standard ASR systems typically make use of phonemes as subword units to model human speech production. A phoneme is defined as the smallest sound unit of a language that discriminates between a minimal word pair (Bloomfield, 1933, p. 78). Although humans are able to produce a large variety of acoustic sounds, we assume that all those sounds across speakers and languages, share a common acoustic space. We found in previous studies (Imseng et al., 2012a; D. Imseng et al., 2011) that the

* Corresponding author at: Idiap Research Institute, Martigny, Switzerland. Tel.: +41 27 721 77 76.
*E-mail address:* dimseng@idiap.ch (D. Imseng).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.01.007

relation between phonemes of different languages can (1) be learned and (2) be exploited for cross-lingual acoustic model training or adaptation. Posterior features, estimated by multilayer perceptrons (MLPs), are particularly well suited for such tasks. Even though previous posterior feature studies that used more than one hour of target language data reported rather small or no improvements (up to 3.5% relative) (Tòth et al., 2008; Grézl et al., 2011), we successfully used posterior features estimated by MLPs that are trained on similar languages such as English, Dutch and Swiss German to boost the performance of an Afrikaans speech recognizer (Imseng et al., 2012b).

In this paper, we show how to significantly boost the performance of an existing Afrikaans speech recognizer that was trained on three hours of within-language data, by using 80 h of Dutch data. We also compare different acoustic modeling techniques and investigate their usefulness if only very limited amounts of within-language data are available.

In our most recent study (Imseng et al., 2012b), we compared two different acoustic modeling techniques for posterior features, namely Tandem (Hermansky et al., 2000) and Kullback-Leibler divergence based hidden Markov models (KL-HMM) (Aradilla et al., 2008). KL-HMM and Tandem both exploit multilingual information in the form of posterior features; we found that they benefit from MLPs that were trained on context-dependent targets, but limited ourselves to MLPs with relatively small numbers of context-dependent targets (about 200). In this study however, we further investigate MLPs trained on context-dependent targets and allow ten times more output units. We also investigate a different (and more suitable) cost function for the KL-HMM framework and compare the aforementioned acoustic modeling techniques to subspace Gaussian mixture models (SGMM), conventional maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptations.

Given three hours of Afrikaans data, KL-HMM, Tandem and SGMM successfully exploit 80 h of Dutch data and yield more than 10% relative improvement compared to the conventional HMM/GMM based monolingual recognizer. Furthermore, we also compare the performance of KL-HMM, Tandem, SGMM, MLLR and MAP if only six minutes of Afrikaans data is available. KL-HMM is demonstrated to be particularly well suited to such low amount of data scenarios and outperforms all other acoustic modeling techniques and also MLLR and MAP adaptations.

We first briefly review Tandem, KL-HMM and SGMM in Section 2. In Section 3, we then present the databases that we used for the training of the MLPs and the shared SGMM parameters as described in Section 4, and give an overview over the investigated systems in Section 5. Experiments and results are then given in Section 6 and discussed in Section 7.

## 2. Acoustic modeling

In this paper, we investigate three different acoustic modeling techniques and also compare them to a conventional HMM/GMM system. The investigated approaches are well suited to exploit out-of-language data. We also compare them to an HMM/GMM system that exploits out-of-language data with the conventional maximum likelihood linear regression (MLLR) approach (Gales, 1998) and with maximum a posteriori (MAP) adaptation (Gauvain and Lee, 1993).

Two of the presented approaches exploit out-of-language data on the feature level, namely Tandem (Hermansky et al., 2000) and Kullback–Leibler divergence based HMM (KL-HMM) (Aradilla et al., 2008). Subspace Gaussian mixture models (SGMM) (Burget et al., 2010) on the other hand exploit out-of-language data on the acoustic model level. The Tandem approach is illustrated in Fig. 1, KL-HMM in Fig. 2 and SGMM in Fig. 3.

The posterior feature based approaches exploit out-of-language information in the form of a Multilayer Perceptron (MLP) which was trained on out-of-language data, whereas the SGMM uses a Universal Background Model (UBM) and shared projection matrices trained on out-of-language data. In the remainder of this section, we will briefly review all three acoustic modeling techniques.

### 2.1. Feature level

Both posterior feature based approaches involve the training/estimation of two different kind of distributions:

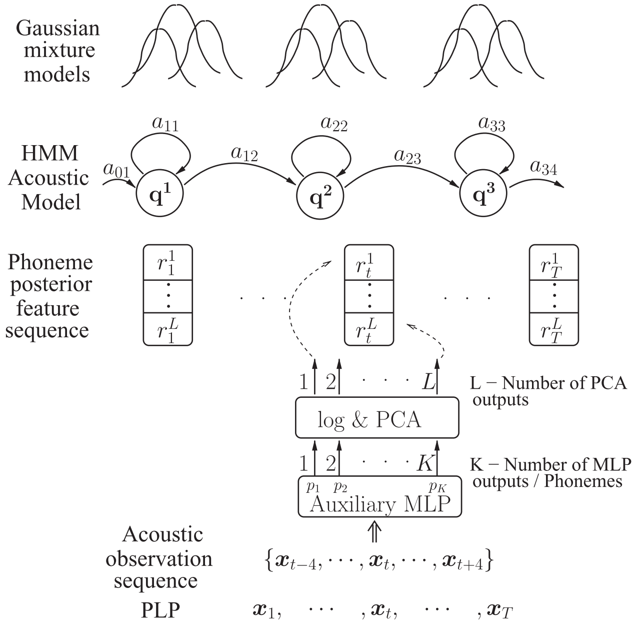

Fig. 1. Tandem – the emission probabilities of the HMM states are modeled with Gaussian mixtures and the MLP output is post-processed. For more details, see Section 5.4.