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*

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, 2011) that the 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 h 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-

*Corresponding author

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 h 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 Figure 1, KL-HMM in Figure 2 and SGMM in Figure 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:

- *Posterior features:* The posterior features are phone class posterior probabilities given the acoustics and estimated with an MLP that can be trained on any auxiliary dataset. Therefore we call it an *auxiliary MLP* and choose an out-of-language dataset with large amounts of available data with which to train. The language of the training data determines the number of output units K (number of phone classes) of the MLP. The phone classes can for example be context-independent monophones or context-dependent triphones. More details about the MLP training are given in Section 4.1.

  Once the MLP is trained, we consider a sequence of T acoustic feature vectors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, namely perceptual linear prediction (PLP) features, extracted from within-language data. As seen in Figure 2, the phone class posterior sequence $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$ is then estimated with the previously trained auxiliary MLP. To estimate $\boldsymbol{z}_t = (z_t^1, \ldots, z_t^K)^\mathsf{T}$, we consider a nine frame temporal context $\{\boldsymbol{x}_{t-4}, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+4}\}$. The described phone class posterior estimation is identical for both posterior feature based acoustic modeling techniques.

- *HMM state distributions:* The HMM states $q^d : d \in \{1, \ldots, D\}$ are associated with the target language. Each phone (mono- or tri-phone) of the target language is modeled with three states, thus the total number of states D is equal to three times the number of phones of the target language.

  The HMM state distributions consist of emission and transition probabilities. Based on anecdotal knowledge, we fix the transition probabilities a_{ij} for both posterior feature based acoustic modeling techniques (see Figures 1 and 2). The emission probabilities however are modeled differently for Tandem and KL-HMM. As we will describe below, Tandem (Section 2.1.1) uses Gaussian mixtures and KL-HMM (Section 2.1.2) uses a categorical distribution. The emission probabilities are trained from within-language data only. Here, we assume that we have access to a limited amount of within-language data.

2.1.1. Tandem

The conventional Tandem approach models the emission probabilities of the HMM states q^d with mixtures of Gaussians. Figure 1 illustrates the HMM associated with a three-state-phone (q^i, q^j, q^k). To model the emission probabilities with Gaussians, the posterior features z_t

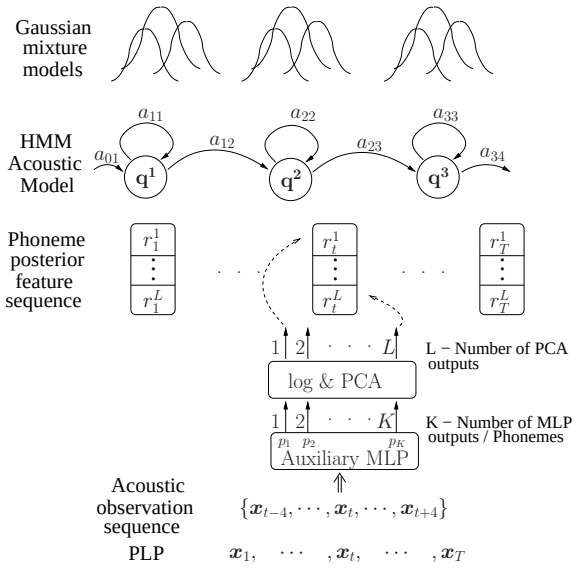

Figure 1: Tandem - the emission probabilities of the HMM states are modeled with Gaussian mixtures and the MLP output is post-processed. For more details, see Section 5.4.

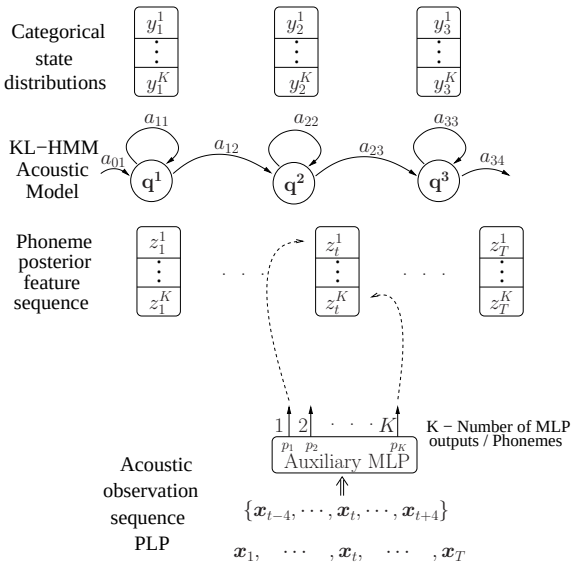

Figure 2: KL-HMM - the emission probabilities are modeled with categorical distributions and the MLP output can directly be used. More details can be found in Section 5.5.

need to be post-processed. More specifically, the log-phone class posteriors are decorrelated with a principal component analysis (PCA). The transformation matrix can be estimated on within-language data. Usually, the resulting feature vector $\boldsymbol{r}_t = (r_t^1, \ldots, r_t^L)^\mathsf{T}$, has a reduced dimensionality L.

2.1.2. Kullback–Leibler divergence based HMM

As illustrated in Figure 2, a KL-HMM is a particular form of HMM in which the emission probability of state q^d is parametrized by a categorical distribution $\boldsymbol{y}_d = (y_d^1, \ldots, y_d^K)^\mathsf{T}$, where K is the dimensionality of the features. A categorical distribution is a multinomial distribution from which only one sample is drawn. In contrast to Tandem that uses Gaussian mixtures and therefore needs the post-processed features $\boldsymbol{r}_t$, the categorical distributions can directly be trained from phone class posterior probabilities $\boldsymbol{z}_t$.

Kullback and Leibler introduced the term *discrimination information* (Kullback and Leibler, 1951; Kullback, 1987) which is nowadays often referred to as the *Kullback–Leibler distance*[1], defined by Cover and Thomas (1991). The divergence of Kullback and Leibler (1951) is today referred to as the symmetric variant of the KL divergence. Aradilla et al. (2008) proposed multiple KL divergence based local scores for KL-HMM training and decoding. In recent studies, we used the symmetric variant of the KL divergence. However, recently we found that the asymmetric KL divergence $KL(x||y)$ is in fact more robust. This is also intuitively reasonable in that the underlying acoustic modeling problem is not symmetric since we observe the posterior features and train the categorical distributions. Therefore, we use the following Kullback–Leibler based distance as local score in this study:

$$d(\boldsymbol{z}_t, \boldsymbol{y}_d) = \sum_{k=1}^{K} z_t^k \log \frac{z_t^k}{y_d^k}. \tag{1}$$

A detailed description of training and decoding algorithms based on the symmetric variant of the KL divergence can be found in (Imseng et al., 2012a). In this paper we use the asymmetric KL divergence as given in (1). For clarity, we briefly review the training and decoding algorithms.

Training

The categorical distributions $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_D\}$ can be learned using an iterative Viterbi segmentation-optimization scheme. The cost function can be defined by integrating the local score, given in (1), over time t and states q^d, resulting in

$$\mathcal{F}(Z, Y) = \sum_{t=1}^{T} \sum_{d=1}^{D} d(\boldsymbol{z}_t, \boldsymbol{y}_d) \delta_t^d, \tag{2}$$

where the Kronecker delta δ_t^d is defined as:

$$\delta_t^d = \begin{cases} 1, & \text{if } \boldsymbol{x}_t \text{ is associated with state } q^d \\ 0, & \text{otherwise.} \end{cases}$$

To associate each $\boldsymbol{x}_t$ with one of the states, the HMM aligns the phone class posterior probabilities Z with the states by minimizing $\mathcal{F}(Z, Y)$, given in (2).

Each $\boldsymbol{z}_t$ is then used to update a particular categorical distribution $\boldsymbol{y}_d$. To minimize $\mathcal{F}(Z, Y)$ subject to $\sum_{k=1}^{K} y_d^k = 1$, we take the partial derivative with respect to each variable y_d^k and set it to zero to find the minimum.

[1] In reality, usually it is referred to as a divergence rather than a distance because it is not a metric.

Then, we introduce the Lagrange multipliers λ to enforce the sum to one constraint:

$$\frac{\partial}{\partial y_d^k}\left[\mathcal{F}(Z,Y)+\lambda\left(\sum_{k=1}^{K} y_d^k - 1\right)\right] = 0. \quad (3)$$

Solving (3) yields:

$$y_d^k = \frac{1}{\lambda}\sum_{\forall t^*} z_t^k, \quad (4)$$

where the sum extends over all t^* such that $\boldsymbol{x}_{t^*}$ is associated with state q_d. Solving (4) for λ yields:

$$\lambda = \sum_{\forall t^*}\sum_{k=1}^{K} y_d^k = \sum_{\forall t^*} 1 = T_d,$$

where T_d stands for the number of frames associated with state q_d. We thus obtain:

$$y_d^k = \frac{1}{T_d}\sum_{\forall t^*} z_t^k. \quad (5)$$

Decoding

During decoding, we minimize:

$$\mathcal{F}_{\mathcal{Q}}(Z,Y) = \min_{\mathcal{Q}} \sum_{t=1}^{T} \left[d(\boldsymbol{z}_t, \boldsymbol{y}_{q_t}) - \log a_{q_{t-1}q_t}\right], \quad (6)$$

where $\mathcal{Q} = \{q_1, \ldots, q_T\}$ stands for all allowed state paths and $\boldsymbol{y}_{q_t}$ is the categorical distribution associated with q_t, the state at time t. The transition probabilities $a_{q_{t-1}q_t}$ are fixed.

2.2. Acoustic model level

In addition to feature level, out-of-language data can also be directly exploited on the acoustic model level to improve ASR performance. In this study we employ SGMMs as an acoustic modeling technique exploiting out-of-language data. Similar to feature level, HMM state distributions associated with the target language are estimated. The transition probabilities are fixed and the emmission probabilities are modeled using probability density function in an SGMM manner.

2.2.1. Subspace Gaussian mixture model (SGMM)

In the SGMM acoustic modeling approach, each speech state associated with an HMM is modeled by a GMM, as is the case for conventional HMM/GMMs. However, the GMMs are not the parameters of the model. Instead, each HMM state q^d (where d represents a state index) is associated with a vector $\boldsymbol{v}_d = (v_d^1, \ldots, v_d^S)^{\mathsf{T}}$, where S is usually similar to the dimension of the acoustic speech features. Mathematically, the SGMM model can be described as follows (Povey et al., 2010):

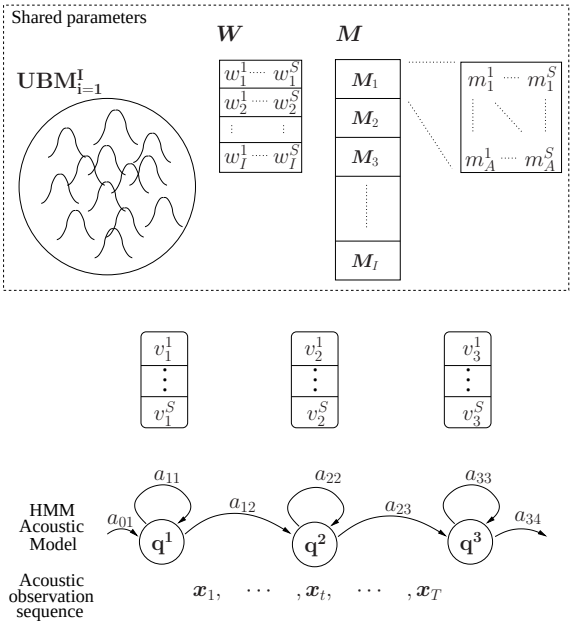

Figure 3: SGMM - the emission probabilities of each context-dependent HMM-state q_d are modeled by GMM. Each HMM-state is parametrized by a vector $\boldsymbol{v}_d$. The parameters $\mathbf{M}$ and $\mathbf{W}$ are globally shared.

$$p(\boldsymbol{x}_t|q^d) = \sum_{i=1}^{I} \omega_i^d \mathcal{N}(x_t; \boldsymbol{\mu}_{i,d}, \boldsymbol{\Sigma}_i), \quad (7)$$

$$\boldsymbol{\mu}_{i,d} = \mathbf{M}_i \boldsymbol{v}_d, \quad (8)$$

$$\omega_i^d = \frac{\exp(\boldsymbol{w}_i \cdot \boldsymbol{v}_d)}{\sum_{l=1}^{I}\exp(\boldsymbol{w}_l \cdot \boldsymbol{v}_d)}, \quad (9)$$

where $\boldsymbol{x}_t \in R^A$ denotes feature vector, q^d is the HMM-state, and $\boldsymbol{v}_d \in R^S$ is the state-specific vector. The model in each HMM state is represented by a simple GMM with I Gaussians, mixture weights $\boldsymbol{\omega}^d = (\omega_1^d, \ldots, \omega_I^d)^{\mathsf{T}}$, means $\boldsymbol{\mu}_{i,d}$, and covariances $\boldsymbol{\Sigma}_i$. $\boldsymbol{\Sigma}_i$ are shared across the states. The state-specific vectors $\boldsymbol{v}_d \in R^S$ together with the globally shared parameters $\boldsymbol{M} = (\boldsymbol{M}_1, \ldots, \boldsymbol{M}_I)^{\mathsf{T}}$ and $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I)^{\mathsf{T}}$ with $\boldsymbol{w}_i = (w_i^1, \ldots, w_i^S)$ are used to derive the means and mixture weights representing the given HMM state. For the initialization of the SGMM, a generic GMM with I Gaussians, denoted as UBM, modeling all the speech is used. SGMM acoustic modeling is depicted in Figure 3 for a single HMM.

Note that the equations above assume (without loss of generality) one state-specific vector $\boldsymbol{v}_d$ to be assigned to each HMM-state. However, as done for the experiments in this study, we can model each state with a mixture of sub-states (Povey et al., 2011), each having its own sub-state specific vector $\boldsymbol{v}_{d_j}$, where $j = 1, \ldots, J_d$ with J_d being the number of sub-states of state d. In that case, we also need to estimate the mixture weights c_j for each sub-state.

ID	Language	Number of phonemes	Amount of training data
AF	Afrikaans	38	3 h
CGN	Dutch	47	81 h

Table 1: Summary of the different languages with number of phonemes and amount of available training data.

ID	Language	Number of output units	Frame accuray on validation set
AF	Afrikaans	187	43.8%
CGN	Dutch	1789	56.5%

Table 2: Summary of the MLPs with number of output units and frame accuracy on the cross-validation set.

3. Databases

We used data from Afrikaans and Dutch as summarized in Table 1. In this section, we describe the two databases.

3.1. LWAZI

The Afrikaans data is available from the LWAZI corpus provided by the Meraka Institute, CSIR, South Africa[2] and described by Barnard et al. (2009). The database consists of 200 speakers, recorded over a telephone channel at 8 kHz. Each speaker produced approximately 30 utterances, where 16 were randomly selected from a phonetically balanced corpus and the remainder consisted of short words and phrases.

The Afrikaans database comes with a dictionary (Davel and Martirosian, 2009) that defines the phoneme set containing 38 phonemes (including silence). The dictionary that we used contained 1585 different words. The HLT group at Meraka provided us with the training and test sets. In total, about three hours of training data and 50 minutes of test data is available (after voice activity detection).

Since we did not have access to an appropriate language model, we trained a bi-gram phoneme model on the training set and only report phoneme accuracies in this study. The bi-gram phoneme model learned the phonotactic constraints of the Afrikaans language and has a phoneme perplexity of 14.5 on the training set and 14.7 on the test set.

3.2. Corpus Gesproken Nederlands

Heeringa and de Wet (2008) reported that standard Dutch seems to be the best language from which to borrow acoustic data for the development of an Afrikaans ASR system. Our studies confirmed that hypothesis (Imseng et al., 2012b). Therefore, we used data of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) (Oostdijk, 2000) that contains standard Dutch pronounced by more than 4000 speakers from the Netherlands and Flanders. The database is divided into several subsets and we only used "Corpus o" because it contains phonetically aligned *read* speech data pronounced by 324 speakers from the Netherlands and 150 speakers from Flanders. "Corpus o" uses 47 phonemes and contains 81 h of data after the deletion of silence segments that are longer than one second. It was recorded at 16 kHz, but since we use the data to perform ASR on Afrikaans, we downsampled it to 8 kHz prior to feature extraction.

[2] `http://www.meraka.org.za/hlt`

4. Multilingual boosting strategies

In this section, we describe how out-of-language data is exploited in case of feature-level and acoustic model-level adaptation.

4.1. Feature level approach

For each language (Afrikaans and Dutch), we trained an MLP from 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features (C0 – C12+Δ+$\Delta\Delta$) in a nine frame temporal-context (four preceding and following frames), extracted with the HTS variant (Zen et al., 2007) of the HTK toolkit. The number of parameters in each MLP was set to 10% of the number of available training frames, to avoid overfitting. We used Quicknet (Johnson, 2005) software to train the MLPs.

We have already shown that systems that use MLPs which are trained on context-dependent targets (triphones) outperform MLPs trained on context-independent monophones (Imseng et al., 2012b). Therefore, we trained both MLPs on triphone targets. To obtain triphone targets, we developed a standard HMM/GMM system with all the training data for both languages and used a standard decision tree approach to tie rare triphones. More specifically, we used the minimum description length criterion to automatically determine the number of tied triphones for each language independently (Shinoda and Watanabe, 1997). As described by Shinoda and Watanabe (1997), the MDL criterion has a hyper-parameter, c, which controls the weight of the term that penalizes models with large amounts of triphones. We tuned c on the Afrikaans database and fixed it to 16 (for both databases). The HMM/GMM systems were then used to align the training data in terms of tied triphones. We used 90% of the training set for training and 10% for cross-validation to stop training. Table 2 shows the number of output units (tied triphones) per MLP and frame accuracy on the cross-validation set.

4.2. Acoustic model level approach

To exploit out-of-language data, the SGMM model parameters can be divided into HMM-state specific and shared parameters, as visualized in Figure 3. As proposed by Burget et al. (2010), $\boldsymbol{M}$ and $\boldsymbol{W}$ projection matrices together with UBM can be perceived as shared (language-independent) and can therefore be trained using large amounts of data from different languages. As already mentioned in Section 2.2.1, we use several sub-states per HMM-state. The sub-state-specific vectors $\boldsymbol{v}_{\boldsymbol{d}_j}$ as well

Afrikaans	Dutch
ɑː	ɑ
ae	ɛ
oe	ʏ
øː	ø
ɦ	h

Table 3: The Afrikaans phonemes without a matching Dutch seed model (same IPA symbol not present in the Dutch phoneme set) are given in the left column. The corresponding manually chosen Dutch seed models are listed in the right column.

as the mixture weights c_j are trained on within-language data.

5. Systems

In this section, we will describe the systems that we investigated to study the exploitation of out-of-language data in the framework of under-resourced ASR. We will compare the performance of the Tandem approach with the performance of KL-HMM and SGMM. Furthermore, we will also compare the proposed systems to an HMM/GMM baseline only trained on within-language data and to an HMM/GMM system trained on Dutch and then adapted to Afrikaans by using MLLR and MAP.

5.1. HMM/GMM

Each context-dependent triphone is modeled with three states (q^i, q^j, q^k). As usually done, we first train context-independent monophone models that serve as seed models for the context-dependent triphone models. We use eight Gaussians per state to model the emission probabilities. To balance the number of parameters with the amount of available training data, we apply conventional state tying with a decision tree that is based on the minimum description length principle (Shinoda and Watanabe, 1997). Training and decoding is performed with HTS.

5.2. Maximum likelihood linear regression (MLLR)

To evaluate whether an under-resourced language could be accommodated by linear transforms, we first train a triphone HMM/GMM system on the Dutch data. Each triphone state is modeled with 16 Gaussians. We investigate the standard MLLR and use a regression tree that allows up to 32 regression classes.

For most Afrikaans phonemes, we use the corresponding Dutch phoneme, represented with the same IPA symbol, as a seed model for MLLR. However, not all the Afrikaans phonemes are also present in the Dutch phoneme set. The Afrikaans phonemes without matching Dutch seed model are given in Table 3 together with the respective manually chosen Dutch seed model. Furthermore, since the diphthongs iə, uə, əu, əi are not present in the Dutch phoneme set, we split them into individual phonemes (monophthongs).

5.3. Maximum a posteriori (MAP) adaptation

Since Köhler (1998) has shown that MAP adaptation is suitable for cross-lingual acoustic model adaptation, we also evaluate MAP adaptation. More specifically, the mean $\mu_{j,m}$ of mixture component m and state j is adapted as follows:

$$\hat{\boldsymbol{\mu}}_{j,m} = \frac{N_{j,m}}{N_{j,m}+\tau}\boldsymbol{\mu}^A_{j,m} + \frac{\tau}{N_{j,m}+\tau}\boldsymbol{\mu}^D_{j,m}, \qquad (10)$$

where $N_{j,m}$ is the occupation likelihood of the Afrikaans data, τ a parameter to tune, $\boldsymbol{\mu}^A$ the mean of the Afrikaans data and $\boldsymbol{\mu}^D$ the mean of the Dutch data.

As seed models, we used the same Dutch context-dependent HMM/GMM models as in Section 5.2. For Afrikaans phonemes without a matching Dutch seed model, we again mapped phonemes as explained in Section 5.2 and Table 3.

5.4. Tandem

Similar to the conventional HMM/GMM system, for the Tandem system, we train context-independent monophone models that serve as seed models for the three-state context-dependent triphone models. We use eight Gaussians per state to model the emission probabilities and use PCA for decorrelation. PCA can also be used to reduce the dimensionality to, for example, 30, as is typically done (Qian et al., 2011; Grézl et al., 2011). In recent studies, we have shown that the dimensionality of the feature vectors greatly affects the performance of the Tandem system (Imseng et al., 2012b). Furthermore, we observed that preserving 99% of the variance yielded similar results to using all the dimensions (Imseng et al., 2012b). Therefore, in this study, we fix the dimensionality such that 99% of the variance is preserved (note that the dimensionality of different systems varies and is given in Tables 4, 5 and 6).

To balance the number of parameters with the amount of available training data, we also use a decision tree that is based on the minimum description length principle (Shinoda and Watanabe, 1997).

5.5. KL-HMM

As for HMM/GMM and Tandem, for the KL-HMM system, we train context-independent monophone models that serve as seed models for the three-state context-dependent triphone models.

For KL-HMM, we applied a decision tree clustering reformulated as dictated by the KL criterion (Imseng et al., 2012c). Since it is not obvious how to apply the minimum description length principle to the modified clustering approach, we tuned the threshold that determines the number of tied states on the development set.

System	Feature dimension	Number of tied states	Phoneme accuracy
HMM/GMM	39	1447	61.2 %
KL-HMM	187	15207	60.6 %
Tandem	48	1846	*64.7* %
SGMM	39	2000	**65.5** %

Table 4: Using 3 h of Afrikaans data to build a monolingual ASR system. Acoustic modeling techniques are described in Section 5. The best result is marked bold; italic numbers point to results that are not significantly worse.

5.6. SGMM

The training of SGMMs is also done from context-independent monophone models that serve as seed models for the three-state context-dependent triphone models.

Decision tree clustering was done automatically, after having specified the number of leaves to be similar to the Tandem system. The UBM has $I = 500$ Gaussians and the dimensionality of the substate phone-specific vectors, S, was fixed to 50.

6. Evaluation

In this section, we analyze the performance of the different systems. We always apply the same bi-gram phoneme model as described in Section 3.1 and report Afrikaans phoneme accuracies on the test set (about 50 min of data). The bi-gram phoneme model scaling factor was determined for each system independently on the cross-validation set (see Section 4.1). In general, we expect that the exploitation of Dutch data will improve the Afrikaans ASR performance. For all the significance tests, we used the bootstrap estimation method (Bisani and Ney, 2004) and a confidence interval of 95%.

6.1. Afrikaans data only

For the first set of experiments, we only used the Afrikaans training set (3 h of data) for the training of the global and local parameters. More specifically, the MLP for the posterior feature extraction as well as the globally shared SGMM parameters were trained on three hours of Afrikaans (see Table 2 for MLP details). In previous studies (Povey et al., 2010), SGMM outperformed HMM/GMM when 15 h of training data was used. We hypothesize that SGMM also outperforms conventional HMM/GMM if only three hours of data is available for training. Furthermore, Tandem outperformed conventional HMM/GMM and KL-HMM systems if three hours of Afrikaans data was available for training (Imseng et al., 2012b).

Table 4 shows the results. Note that the baseline results reported by van Heerden et al. (2009), 63.1% phoneme accuracy, were the first set of results obtained for the data and the official train and test set were compiled after the official database release. Personal communication with the HLT group at Meraka confirmed that the lower performance of our baseline can be attributed to the different data partitioning[3].

The results in Table 4 confirm our hypotheses. On Afrikaans data only, SGMM performs best, followed by Tandem. Bold numbers in tables mark the best result (column-wise) and italic numbers are not significantly different from the best performance. KL-HMM and the HMM/GMM baseline perform significantly worse than SGMM and Tandem.

Table 4 also shows the feature dimensionality of the employed acoustic modeling techniques. HMM/GMM and SGMM are both based on acoustic features (39 dimensions). KL-HMM uses the raw output of the Afrikaans MLP. For the Tandem system however, recall that the posterior features need to be post-processed. Keeping 99% of the variance after PCA results in a 48-dimensional feature vector.

The number of tied states, also shown in Table 4, for HMM/GMM and for Tandem were automatically determined with the MDL criterion. Based on anecdotal knowledge, we fixed the number of tied states for the SGMM system similar to the number of tied states for the Tandem system. The number of tied states for the KL-HMM was tuned on the cross-validation set. Since the categorical distributions of the KL-HMM can be trained with very few data, modeling each triphone state separately performed best. Hence, the decision tree was only used to synthesize unseen triphone contexts during testing.

Due to the extremely high number of states of the KL-HMM system compared to the other systems, the KL-HMM system has the most parameters of the four systems given in Table 4. To increase the number of parameters of the other systems, we increased the number of Gaussians per state for the HMM/GMM as well as for the Tandem system to 16 and doubled the number of sub-states of the SGMM system. However, none of the performances improved.

6.2. Auxiliary Dutch data

For the second set of experiments, we used the Dutch data to train the MLP as well as the globally-shared SGMM parameters. We also trained Dutch seed models for the MLLR and MAP adaptation. The Afrikaans data was used to train the HMM distributions (KL-HMM and Tandem), the sub-state phone-specific vectors $\mathbf{v}_d$ and sub-state mixture weights $\mathbf{c}_j$ (SGMM) and the MLLR adaptation. MAP adaptation was applied as described in (10) and τ was tuned on the development set (see Table 5).

Since three hours seems to be a reasonable amount of training data, we also simulated very low-resourced languages and evaluated three different scenarios: six minutes of data, one hour of data and three hours of data.

[3]The HLT group now also uses the partitioning that we used in this paper and report a lower performance.

System	Feat. dim.	6 min TS	6 min τ	6 min PA [%]	1 h TS	1 h τ	1 h PA [%]	3 h TS	3 h τ	3 h PA [%]
HMM/GMM	39	116	—	38.6	594	—	55.3	1447	—	61.2
MLLR	39	—	—	41.3	—	—	44.4	—	—	44.7
MAP	39	11357	15	39.4	11357	5	46.9	11357	1	50.6
KL-HMM	1789	635	—	**53.1**	13308	—	**61.5**	15207	—	*67.3*
Tandem	286	114	—	41.0	537	—	*61.3*	1846	—	*68.2*
SGMM	39	150	—	40.2	600	—	*60.4*	2000	—	**68.5**

Table 5: Exploiting Dutch data to improve an Afrikaans ASR system. The different acoustic modeling techniques are described in Section 5. TS stands for the number of tied states, PA for phoneme accuracy and τ is the parameter of the MAP adaptation. Best results of each PA column are marked bold; italic numbers point to results that are not significantly worse.

For the sake of comparison, we also evaluated a conventional HMM/GMM system for each scenario. We hypothesize, that KL-HMM performs best for very low amounts of data because we have seen this behavior in previous similar evaluations of KL-HMM (Imseng et al., 2012c). If three hours of data is available, we expect that KL-HMM, Tandem and SGMM are successfully exploiting the out-of-language data and performing similarly well.

Table 5 confirms our hypotheses. The HMM/GMM (only trained on Afrikaans) is clearly outperformed by KL-HMM, Tandem and SGMM, hence all three systems successfully exploit out-of-language information. MLLR and MAP, however, only perform better than HMM/GMM if six minutes of Afrikaans data are available. Note that both approaches are bound to phoneme sets. Köhler (1998) for example had for each phoneme a multilingual seed model that was trained from data associated with a matching IPA symbol. In our case however, we needed to manually map several Afrikaans phoneme models as discussed in Table 3. If there is 1 h or more data available, MAP outperforms MLLR.

For the three hours as well as the one hour scenario, SGMM, KL-HMM and Tandem all perform statistically the same. While SGMM is the most suitable acoustic modeling technique if we train only on within-language data, KL-HMM (which was performing significantly worse in Table 4) benefits most from out-of-language data and seems to be particularly well suited to exploit out-of-language information on this database. Furthermore, KL-HMM using six minutes of data performs almost as well as a conventional monolingual HMM/GMM system using one hour of data. In the case of the SGMM, results are slightly worse than expected. We suppose that the dimensionality of the sub-state-specific vectors is probably too high for only six minutes of data.

6.3. Within- and out-of-language data

We have already shown that properly combining acoustic information from multiple similar languages can boost the performance. Therefore, we hypothesize that the performance can be improved if we concatenate the output of both MLPs or train the shared SGMM parameters on both languages. The concatenated MLP outputs were renormalized to guarantee that the feature vectors can be interpreted as posterior distributions, as assumed by the KL-HMM. For the Tandem systems, we post-process the normalized vectors as already described in Section 5.4. For SGMM, we trained the shared parameters with the data of both languages.

System	Feature dimension	Phoneme accuracy
KL-HMM	1976	**68.8** %
Tandem	308	*68.4* %
SGMM	39	*68.6* %

Table 6: Using the Dutch and Afrikaans MLP (KL-HMM and Tandem) and use Dutch and Afrikaans data to train the shared parameters (SGMM). The best result is marked bold; italic numbers point to results that are not significantly worse.

However, Table 6 shows that the results only marginally improve for Tandem and SGMM. For KL-HMM, they improve by 1.5% absolute. KL-HMM performs best but not statistically differently from the other systems.

7. Discussion

The results in Section 6 have shown that (a) out-of-language data improved an existing Afrikaans speech recognizer and (b) KL-HMM outperforms all other approaches if only 6 min of Afrikaans data are available. In this section, we discuss the two conclusions.

7.1. Out-of-language data

The systems in Table 6 perform significantly better than the HMM/GMM baseline that does not use Dutch data (see Table 4). We hypothesize that Dutch data mostly improve recognition accuracy of phonemes for which the Afrikaans dataset does not contain much training data. Figure 4 shows the relative phoneme accuracy change per phoneme of the systems given in Table 6 with respect to the HMM/GMM baseline that does not use Dutch data. The phonemes on the x-axis are sorted according to their frequency in the Afrikaans training data with the most frequent phonemes on the left. Figure 4 appears to confirm our hypothesis.

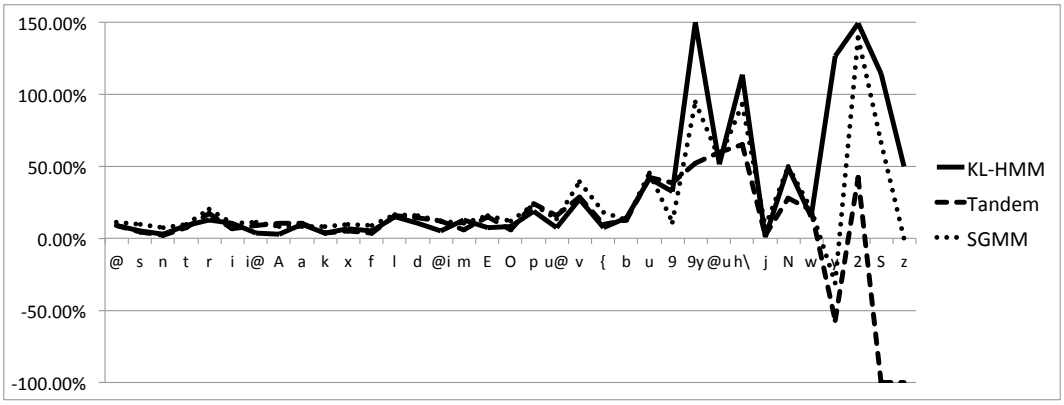

Figure 4: Relative phoneme accuracy change per phoneme of the systems shown in Table 6 with respect to the monolingual HMM/GMM baseline system. The phonemes on the x-axis are sorted according to their frequency in the Afrikaans training data (most frequent phoneme on the left).

7.2. KL-HMM

Even though we performed an extensive error analysis, there was no clear evidence for which phonemes KL-HMM yields most improvement compared to the other modeling techniques. Rather, KL-HMM consistently improves the recognition accuracy across all phonemes. We attribute the improvement to the sophisticated acoustic modeling and the constrained optimization space that are particularly well suited for low amount of data scenarios.

8. Conclusion and future work

We successfully exploited Dutch data and boosted a monolingual speech recognizer that was trained on three h of Afrikaans data. We compared KL-HMM, Tandem, SGMM, MLLR as well as MAP and found that KL-HMM, Tandem and SGMM successfully exploit out-of-language data if at least one hour of within-language data are available. If only six minutes of data are available, KL-HMM outperforms all other acoustic modeling techniques including MLLR and MAP.

Furthermore, we found that if three h of within-language data and 80 h of out-of-language data are available, the proposed systems yield 12% relative improvement compared to a conventional HMM/GMM system only using within-language data. If only six minutes of within-language data and 80 h of out-of-language data are available, KL-HMM performs relatively about 30% better than MLLR and MAP .

We exploited multilingual information on the feature level by applying simple concatenation of MLP outputs. In future, we plan to explore different methods to combine the output of several MLPs. Furthermore, we also exploited multilingual information on the acoustic modeling level. To investigate whether the two approaches are complementary, we plan to implement an SGMM system based on posterior features.

9. Acknowledgement

This research was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1 and the National Centre of Competence in Research (NCCR) in Interactive Multimodal Information Management (IM2) `http://wwww.im2.ch`.

The authors are grateful to the HLT group at Meraka, and especially Dr. Febe de Wet, for providing us with the training and test sets as well as the Afrikaans dictionary.

References

D. Imseng, H. Bourlard, P. N. Garner, Using KL-divergence and multilingual information to improve ASR for under-resourced languages, in: Proc. of ICASSP, 4869–4872, 2012a.

Y. Qian, J. Xu, D. Povey, J. Liu, Strategies for Using MLP based Features with Limited Target-Language Training Data, in: Proc. of ASRU, 354–358, 2011.

T. Niesler, Language-dependent state clustering for multilingual acoustic modelling, Speech Communication 49 (2007) 453–463.

T. Schultz, A. Waibel, Language-independent and language-adaptive acoustic modeling for speech recognition, Speech Communication 35 (2001) 31–51.

L. Bloomfield, Language, New York: Holt, 1933.

D. Imseng, H. Bourlard, J. Dines, P. N. Garner, M. Magimai.-Doss, Improving non-native ASR through stochastic multilingual phoneme space transformations, in: Proc. of Interspeech, 537–540, 2011.

L. Tòth, J. Frankel, G. Gosztolya, S. King, Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian., in: Proc. of Interspeech, 2695–2698, 2008.

F. Grézl, M. Karafiát, M. Janda, Study of Probabilistic and Bottle-Neck Features in Multilingual Environment, in: Proc. of ASRU, 359–364, 2011.

D. Imseng, H. Bourlard, P. N. Garner, Boosting under-resourced speech recognizers by exploiting out of language data - Case study on Afrikaans, in: Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages, 60–67, 2012b.

H. Hermansky, D. Ellis, S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, in: Proc. of ICASSP, III–1635–1638, 2000.

G. Aradilla, H. Bourlard, M. Magimai-Doss, Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task, in: Proc. of Interspeech, 928–931, 2008.

M. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Computer Speech and Language 12 (2) (1998) 75 – 98.

J.-L. Gauvain, C.-H. Lee, Speaker adaptation based on MAP estimation of HMM parameters, in: Proc. of ICASSP, vol. 2, 558–561, 1993.

L. Burget, et al., Multilingual Acoustic Model for Speech Recognition based on Subspace Gaussian Mixture Models, in: Proc. of ICASSP, 4334–4337, 2010.

S. Kullback, R. A. Leibler, On Information and Sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79–86, http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729694.

S. Kullback, The Kullback-Leibler Distance, The American Statistician 41 (4) (1987) 340–341, in Letters to the Editor.

T. Cover, J. Thomas, Elements of information theory, Wiley, New York, 1991.

D. Povey, et al., Subspace Gaussian Mixture Models for Speech Recognition, in: Proc. of ICASSP, 4330–4333, 2010.

D. Povey, M. Karafiát, A. Ghoshal, P. Schwarz, A Symmetrization of the Subspace Gaussian Mixture Model, in: Proc. of ICASSP, 4504–4507, 2011.

E. Barnard, M. Davel, C. van Heerden, ASR Corpus design for resource-scarce languages, in: Proc. of Interspeech, 2847–2850, 2009.

M. Davel, O. Martirosian, Pronunciation dictionary development in resource-scarce environments, in: Proc. of Interspeech, 2851–2854, 2009.

W. Heeringa, F. de Wet, The origin of Afrikaans pronunciation: a comparison to west Germanic languages and Dutch dialects, in: Proc. of the Conf. of the Pattern Recognition Association of South Africa, 159–164, www.let.rug.nl/heeringa/dialectology/papers/prasa08.pdf, 2008.

N. Oostdijk, The Spoken Dutch Corpus. Overview and first evaluation., in: In Proceedings of the Second International Conference on Language Resources and Evaluation, vol. II, 887–894, 2000.

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, K. Tokuda, The HMM-based speech synthesis system version 2.0, http://hts.sp.nitech.ac.jp/, 2007.

D. Johnson, Quicknet, http://www.icsi.berkeley.edu/Speech/qn.html, 2005.

K. Shinoda, T. Watanabe, Acoustic modeling based on the MDL principle for speech recognition, in: Proc. of Eurospeech, vol. I, 99–102, 1997.

J. Köhler, Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks, in: Proc. of ICASSP, vol. 1, 417–420, 1998.

D. Imseng, J. Dines, P. Motlicek, P. N. Garner, H. Bourlard, Comparing different acoustic modeling techniques for multilingual boosting, in: Proc. of Interspeech, 2012c.

M. Bisani, H. Ney, Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Proc. of ICASSP, vol. 1, I–409–412, 2004.

C. van Heerden, E. Barnard, M. Davel, Basic speech recognition for spoken dialogues, in: Proc. of Interspeech, 3003–3006, 2009.