USING KL-DIVERGENCE AND MULTILINGUAL INFORMATION TO IMPROVE ASR FOR UNDER-RESOURCED LANGUAGES

David Imseng[1,2]*, Hervé Bourlard*[1,2]*, Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{dimseng,bourlard,pgarner}@idiap.ch

ABSTRACT

Setting out from the point of view that automatic speech recognition (ASR) ought to benefit from data in languages other than the target language, we propose a novel Kullback-Leibler (KL) divergence based method that is able to exploit multilingual information in the form of universal phoneme posterior probabilities conditioned on the acoustics. We formulate a means to train a recognizer on several different languages, and subsequently recognize speech in a target language for which only a small amount of data is available. Taking the Greek SpeechDat(II) data as an example, we show that the proposed formulation is sound, and show that it is able to outperform a current state-of-the-art HMM/GMM system. We also use a hybrid Tandem-like system to further understand the source of the benefit.

Index Terms— Multilingual speech recognition, neural network features, fast training, Kullback-Leibler divergence

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 needed to train current recognizers [1]. Data collection involves large amounts of manual work, not only in time for the speakers to be recorded, but also for transcription 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 exist already for many languages. However, most current recognition techniques require large amounts of data in the *target* language (the language that the system is supposed to recognize) and are not able to exploit available multilingual information.

The goal of this paper is to show how to use multilingual training data to boost the performance of a speech recognizer for a target language with very little available training data. We propose to exploit multilingual information in the form of universal phoneme posterior probabilities conditioned on the acoustics to improve monolingual system performance. More specifically, we propose to use a hidden Markov model (HMM) that is based on the Kullback-Leibler (KL) divergence (KL-HMM). We compare the novel KL-HMM system to a system based on recently proposed multilingual Tandem features [2], to a conventional HMM/GMM system [3] based on PLP features and some standard adaptation techniques.

To enable evaluation, we use SpeechDat(II) data from five European languages to train a multilayer perceptron (MLP) to estimate universal phoneme posterior probabilities. The Greek SpeechDat(II) database is taken as representative of an unseen language with little available data. Universal phoneme posterior probabilities are then estimated with the previously trained multilingual MLP and used by the Tandem system and the KL-HMM system. Results reveal that multilingual information is successfully exploited by the proposed KL-HMM system. Therefore, if only very little Greek training data is available, the KL-HMM system outperforms these systems.

Work with similar aims already exists: Köhler [4] used maximum a-posteriori (MAP) adaptation with five minutes of data, but explored the behavior on a rather small task on the German Voicemail database with a vocabulary of 62 words and data from 140 speakers. Schultz and Waibel [5] studied rapid adaptation for continuous speech recognition. They used maximum likelihood linear regression (MLLR) in combination with a decision tree specialization technique to adapt language independent acoustic models to Portuguese. They report 71% word accuracy with 25 minutes of training data (compared to 81% word accuracy when a system was trained from scratch with 16.5 hours of data). Le and Besacier [6] used two hours of data to perform fast acoustic modeling of Vietnamese speech. They used phone mapping and a MAP based algorithm to adapt the French seed models. Their Vietnamese speech dialog system yielded word accuracies of 64%. Zhao and O'Shaughnessy [7] studied MLLR based cross-language adaptation from English to Mandarin on broadcast news data and found that it is better to perform training on native speech data instead of performing cross-language adaptation if there are more than eight minutes of data available.

In this paper, we propose a system that is not based on conventional HMM/GMM structures and standard techniques, but on the recently proposed KL-HMM. We will show that it is better to exploit multilingual information with the KL-HMM system if there are less than 75 minutes of data available. Further, we will also show that five minutes of Greek data are sufficient to achieve a performance of 77% word accuracy on continuous read speech with a vocabulary of 10k words, compared to a conventional HMM/GMM system, trained on 13.5 hours of data, that yields 85% word accuracy.

The remainder of the paper is organized as follows: Section 2 presents the KL-HMM framework and Section 3 introduces the systems that are compared. Experimental details and results are given in Section 4. Section 5 then concludes the paper and presents possible future research directions.

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

2. KULLBACK-LEIBLER BASED HMMS

The notion of KL-HMM was introduced by Aradilla [8]. In this section, we briefly summarize the basic training and decoding techniques in the context of our experiments.

A KL-HMM is a particular form of HMM where each state $d : d \in \{1, \ldots, D\}$, where there are D states in the target language, is parametrized by a multinomial distribution $y^d = \left(y_1^d, \ldots, y_K^d\right)^\mathsf{T}$, where K is the dimensionality of the features. The transition probabilities are also parameters of the KL-HMM, but, to minimize their effect on the decoding, we consider them to be fixed. In this paper, each phoneme of the target language is modeled with three states and equal transition probabilities, except silence, which has different, but still fixed, transition probabilities to model longer durations.

The proposed system involves two different phoneme sets:

- A target phoneme set that is used to model the speech during decoding. We assume that there is only a limited amount of data available for the target language.
- A universal phoneme set that is used during the feature extraction. The universal phoneme set was built by merging phonemes that share the same symbol across all training languages. We use universal phoneme posterior probabilities conditioned on the acoustics as features. For the feature extraction, a multilingual MLP was trained on large amounts of data to estimate universal phoneme posterior probabilities.

KL-HMM uses a cost function that is based on the KL divergence. The KL divergence is a measure of difference between probability distributions. Since universal phoneme posterior probabilities conditioned on the acoustics are used as features, and each state of the HMM is parametrized by a probability distribution, the KL divergence is well suited for that setup. We will show later that the proposed system is particularly useful to perform training with small amounts of data.

2.1. Training

For the description of the training and the decoding techniques, we assume to have access to the following:

- A set of T acoustic vector observations $X = \{x_1, \ldots, x_T\}$, where
$$x_t = \left(x_{t,1}, x_{t,2}, \ldots, x_{t,F}\right)^\mathsf{T}$$
with F being the dimensionality of the acoustic vector.
- A set of probability vectors $\mathcal{P} = \{P_1, \ldots, P_T\}$, containing conditional probability distributions
$$P_t = \left(P(u^1|x_t, \theta), P(u^2|x_t, \theta), \ldots, P(u^K|x_t, \theta)\right)^\mathsf{T}$$
with u^k being universal phonemes and K the number of universal phonemes. The probability distributions are estimated with the MLP, whose parameters θ were previously trained on multilingual data.
- Phonetic transcriptions for all the training data (no alignments).

and the following is estimated:

- A set of multinomial distributions $Y = \{y^1, \ldots, y^D\}$, where
$$y^d = \left(y_1^d, y_2^d, \ldots, y_K^d\right)^\mathsf{T}$$
is the multinomial distribution associated with state d.

The multinomial distributions Y can be optimized with the help of a cost function $\mathcal{F}_\mathcal{Q}(\mathcal{P}, Y)$, that minimizes a KL based measure between $\mathcal{P}$ and Y. Like in our previous studies, we use a symmetric variant of the KL divergence [9]:

$$f_{SKL}(P_t, y^d) = \frac{1}{2} f_{KL}(P_t, y^d) + \frac{1}{2} f_{KL}(y^d, P_t) \tag{1}$$

$$\tag{2}$$

where

$$f_{KL}(x, y) = \sum_{k=1}^{K} x(k) \log \frac{x(k)}{y(k)} \tag{3}$$

$$\tag{4}$$

Hence the cost function $\mathcal{F}_\mathcal{Q}(\mathcal{P}, Y)$ can be written as:

$$\mathcal{F}_\mathcal{Q}(\mathcal{P}, Y) = \min_\mathcal{Q} \sum_{t=1}^{T} \left[f_{SKL}(P_t, y^{q_t}) - \log a_{q_{t-1} q_t} \right] \tag{5}$$

where $\mathcal{Q} = \{q_1, \ldots, q_T\}$ stands for all possible state paths allowed by the given phonetic transcriptions and $q_t = d$, i.e., q_t, the state at time t, is one of the D possible states. The term $a_{q_{t-1} q_t}$ stands for the probability of going from state q_{t-1} to q_t.

Y can be optimized with the Viterbi EM algorithm using $f_{SKL}(P_t, y^{q_t})$ as local distances during alignment and $\mathcal{F}_\mathcal{Q}(\mathcal{P}, Y)$ as cost function for the re-estimation of the multinomial distributions. More specifically, each P_t is associated with a particular state d by aligning the feature vectors $\mathcal{P} = \{P_1, \ldots, P_T\}$ with the states by minimizing $\mathcal{F}_\mathcal{Q}(\mathcal{P}, Y)$ (expectation step). The resulting segmentation is then used to update the multinomial distributions Y (maximization step).

2.2. Decoding

Given a test sequence of universal phoneme posterior probabilities of length T', and a set of hypotheses $\mathcal{M}$, the recognized hypothesis $\hat{m}$ is the one with the lowest score:

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \mathcal{F}_{\mathcal{Q}(m)}(\mathcal{P}, Y)$$

and

$$\mathcal{F}_{\mathcal{Q}(m)}(\mathcal{P}, Y) = \min_{\mathcal{Q}(m)} \sum_{t=1}^{T'} \left[f_{SKL}(P_t, y^{q_t}) - \log a_{q_{t-1} q_t} \right]$$

where $\mathcal{Q}(m)$ represents the set of all possible state sequences allowed by hypothesis m.

3. SYSTEM DESCRIPTION

We used two different kinds of features to compare six systems.

3.1. Features

We first introduce the two feature types:

3.1.1. Perceptual Linear Prediction

We extracted 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features ($C_0 - C_{12} + \Delta + \Delta\Delta$), with the HTS variant[1] of the HTK toolkit.

[1] `http://hts.sp.nitech.ac.jp/`

3.1.2. Universal posteriors

We propose to use the large amounts of collected speech data in various languages to train a multilingual MLP to estimate universal phoneme posterior probabilities. A universal phoneme set was built by merging phonemes that share the same symbol across five European languages (British English, Italian, Spanish, Swiss French and Swiss German). The universal phoneme set consists of 116 SAMPA[2] phonemes and silence. Then a multilingual MLP was trained with 63 hours of SpeechDat(II) data, collected in the five aforementioned languages, as explained by Imseng et al. [10]. The universal phoneme posterior probability estimates were obtained by forward passing the Greek MF-PLP features.

3.2. Systems

In total we compare six systems. Three systems based on MF-PLP features and three systems based on universal posteriors.

3.2.1. Monolingual HMM/GMM system

We first built a conventional HMM/GMM system that only used the available Greek data. The system based on context dependent phonemes (triphones) was trained from the MF-PLP features with the HTS toolkit. The tied triphone models were modeled with 2, 4, 8 and 16 Gaussian mixtures with diagonal covariance. Depending on the available amount of training data, the optimal choice for the number of Gaussians may vary. We tuned it on the development set.

3.2.2. Maximum likelihood linear regressions

To evaluate whether the new language could be accommodated by linear transforms, we first trained a triphone HMM/GMM system on the multilingual data. Each triphone was modeled with 16 Gaussians. We investigated the standard maximum likelihood linear regression (MLLR) as well as a constrained version of it (CMLLR). CMLLR has fewer parameters and might therefore be advantageous if we only have access to a limited amount of data. We used a regression tree that allowed up to 32 regression classes. Since not all the Greek phonemes were present in the universal phoneme set, we needed to map the palatal plosives c and ɟ to the velar plosives k and g respectively.

3.2.3. KL-HMM

The KL-HMM system, described in Section 2, was based on triphones and used universal posterior features. Since it is not evident how to cluster KL-HMM states with a decision tree, we limited ourselves to word-internal triphones only (as opposed to cross-word triphones for the HMM/GMM systems). During decoding, we backed off to the context independent model of the center phoneme if a triphone was not seen during training. The absence of a decision tree based tying is certainly a weakness of the proposed approach and will be addressed in future work. Each triphone was modeled with three states.

3.2.4. Multilingual Tandem system

The multilingual Tandem system used conventional HMM/GMM structures to model the universal posterior features. Besides the choice of the features, the training was same as for the monolingual HMM/GMM system (Section 3.2.1). To model universal posteriors with Gaussians, as usually done, we applied logarithm and Karhunen-Loève transformation to de-correlate. To compare the impact of the different modeling techniques (HMM/GMM versus KL-HMM), we did not concatenate PLP features as was done, for example, by Imseng et al. [2], where multilingual Tandem features were used to boost the performance in a mixed language environment.

3.2.5. Linear hidden network based Tandem system

To adapt the universal posteriors to the target language and reduce their dimensionality, we used a technique similar to the linear hidden network (LHN) as proposed by Scanzio et al. [11]. More specifically, we trained a single layer neural network to estimate Greek phoneme posteriors based on the universal posteriors estimated by the multilingual MLP. We then applied the same post-processing and training procedure as for the multilingual Tandem system (Section 3.2.4).

4. EXPERIMENTAL SETUP AND RESULTS

A-priori, we would expect the following:

- The conventional HMM/GMM system should perform best if there is a large amount of training data.
- The KL-HMM system should perform best if there is only very little training data available (because of the efficient modeling of the multilingual MLP features).
- The multilingual information should be exploited better if KL-HMMs are used for the modeling instead of HMM/GMM structures.
- The MLLR systems and the LHN system should perform better than the monolingual HMM/GMM system and the multilingual Tandem system, respectively, for low amounts of data.

The purpose of the experiments, then, is to prove or disprove these hypotheses.

The Greek SpeechDat(II) database contains a relatively large amount of data. As we did for the other SpeechDat(II) databases, we only used *corpus S*, which contains ten read sentences per speaker. In total, we used the data of 2000 speakers, split into training (1500 speakers), development (150 speakers) and testing (350 speakers) sets as described by Imseng et al. [2]. The total amount of training data consists of 808 minutes of speech (13.5 hours). To simulate limited resources, we continuously reduced the amount of available data. To do so, we randomly picked a subset of utterances for both the training and the development set. The amount of training data varies from 13.5 hours to 5 minutes. We did not change the test set and all the systems were evaluated on the same set. The test sentences use 10k different words.

Since we have no access to an appropriate language model, we simply built two different language models: one with all the sentences from the development set and one with all the sentences from the test set. These language models have perplexities of 43 and 44 respectively. The development language model was used during the parameter tuning (language scaling factor and word insertion penalty) on the development set and the test language model was used during the evaluation. In this sense, results should be considered as optimistic. As already explained in Section 3, for the HMM/GMM based systems (monolingual HMM/GMM, Multilingual Tandem, LHN), we also tuned the number of mixtures per Gaussian on the development set.

[2] http://www.phon.ucl.ac.uk/home/sampa/grk-uni.htm

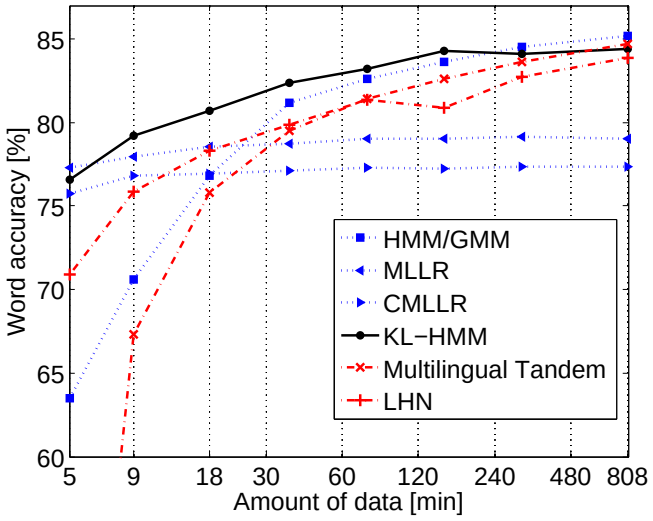

Fig. 1. Word accuracies from the six different systems for different amounts of training data. The x-axis is logarithmic and displays the amount of data and the y-axis is linear and shows word accuracies.

Figure 1 visualizes the results. It can clearly be seen how all the systems except the KL-HMM and both MLLR system collapse with very little data.

As we expected, the conventional HMM/GMM performs best if there are 808 minutes of data available for training. If there are 308 minutes of data available, the performances of the KL-HMM system and the monolingual HMM/GMM system are statistically identical. For the significance test, we used the bootstrap estimation method [12] and a confidence interval of 95%. With decreasing amounts of training data (75 minutes and less), the KL-HMM system performs significantly better than the other systems.

If there are only 5 minutes of training data available, the MLLR system performance is best, but not significantly different from the KL-HMM system performance. However, the MLLR transform saturates quickly with a rather low performance suggesting that the language difference cannot be explained simply by linear transforms. If there are 18 minutes of data available, KL-HMM performs significantly better than MLLR. In contrast to our assumption that CMLLR might perform better than MLLR for low amounts of data, it never performs better than MLLR.

Zhao and O'Shaughnessy [7] found that it is better to perform training on native speech data instead of performing MLLR based cross-language adaptation if there are more than eight minutes of data available. In our study, the monolingual HMM/GMM system performs better than the MLLR system if there are more than 30 minutes of data. However, Zhao and O'Shaughnessy adapted English models to Mandarin, whereas we adapted multilingual models from European languages to Greek. The multilingual Tandem system and the LHN system perform similarly, but for small amounts of data, the LHN system clearly outperforms the multilingual Tandem system.

The KL-HMM system and the multilingual Tandem system use basically the same features, but model them differently. It seems that the Gaussian mixtures of the Tandem system are not able to exploit the previously learned multilingual information in the form of posterior probabilities, because the latter is outperformed by the HMM/GMM system that uses the same modeling technique, but standard PLP features. The KL-HMM system, however, still performs quite well if only five minutes of Greek data are available and degrades by less than ten percent absolute compared to the HMM/GMM state-of-the-art system, trained on 13.5 hours of data. Altogether, the KL-HMM system performs best (or statistically identical to the best system) for all investigated amounts of data except 808 minutes. Hence, we accept all four hypotheses.

5. CONCLUSION

We have shown that a KL-HMM system equals or outperforms current state-of-the-art speech recognition techniques for an unseen language if there is only very little training data available. With only five minutes of data along with word labels, the KL-HMM system yields a performance of 77% word accuracy. Hence we can conclude that the KL-HMM framework is well suited to perform automatic speech recognition for under-resourced languages.

In future, we will investigate KL-divergence based decision tree clustering and expect to improve the KL-HMM system performance.

6. REFERENCES

[1] Tanja Schultz et al., "SPICE: Web-based tools for rapid language adaptation," in *Proc. of Interspeech*, 2007, pp. 2125–2128.

[2] David Imseng, Hervé Bourlard, and Mathew Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proc. of Interspeech*, 2010, pp. 278–281.

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

[4] Joachim Köhler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proc. of ICASSP*, 1998, pp. I–417–420.

[5] Tanja Schultz and Alex Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[6] Viet-Bac Le and L. Besacier, "First steps in fast acoustic modeling for a new target language: Application to Vietnamese," in *Proc. of ICASSP*, 2005, pp. I–821–824.

[7] Xufang Zhao and Douglas O'Shaughnessy, "An evaluation of cross-language adaptation and native speech training for rapid HMM construction based on very limited training data," in *Proc. of Interspeech*, 2007, pp. 1433–1436.

[8] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proc. of Interspeech*, 2008.

[9] David Imseng, Ramya Rasipuram, and Mathew Magimai.-Doss, "Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition," in *Proc. of ASRU*, 2011, pp. 348–353.

[10] David Imseng, Hervé Bourlard, Mathew Magimai.-Doss, and John Dines, "Language dependent universal phoneme posterior estimation for mixed language speech recognition," in *Proc. of ICASSP*, 2011, pp. 5012–5015.

[11] Stefano Scanzio et al., "On the use of a multilingual neural network front-end," in *Proc. of Interspeech*, 2008, pp. 2711–2714.

[12] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, 2004, pp. I–409–412.