Discriminative Optimization of language adapted HMMs for a Language Identification System based on Parallel Phoneme Recognizers

Josef G. Bauer, Bernt Andrassy and Ekaterina Timoshenko

Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81730 Munich, Germany
{Josef.Bauer,Bernt.Andrassy,Ekaterina.Timoshenko.ext}@siemens.com

Abstract

Recently an unsupervised learning scheme for Hidden Markov Models (HMMs) used in acoustical Language Identification (LID) based on Parallel Phoneme Recognizers (PPR) was proposed. This avoids the high costs for orthographically transcribed speech data and phonetic lexica but was found to introduce a considerable increase of classification errors. Also very recently discriminative Minimum Language Identification Error (MLIDE) optimization of HMMs for PPR based LID was introduced that again only requires language tagged speech data and an initial HMM. The described work shows how to combine both approaches to an unsupervised and discriminative learning scheme. Experimental results on large telephone speech databases show that using MLIDE the relative increase in error rate introduced by unsupervised learning can be reduced from 61% to 26%. The absolute difference in LID error rate due to the supervised learning step is reduced from 4.1% to 0.8%.

Index Terms: Language Identification, Minimum Classification Error, unsupervised learning, discriminative training

1. Introduction

Parallel Phoneme Recognizers (PPR, [1]) based on language dependent Hidden Markov Models (HMMs) is a powerful approach to automatic acoustic Language Identification (LID) aiming to identify the language or dialect of spoken utterances. A typical application of such a system would be a call routing system where a user is transferred to an agent or system capable of his preferred language.

In PPR based LID systems a language specific acoustic model (HMM) is build for all target languages. This differs from the PPRLM ([1]) approach where the acoustic model is usually independent from the set of target languages. Following the standard approach to Hidden Markov Model parameter estimation as used for most speech recognition systems large amounts of orthographically transcribed speech data and phonetic lexica in all deployed languages are required. Such speech resources might not be available for all languages or may be time consuming and expensive to acquire.

In [2] a procedure for Unsupervised Adaptation (UnAd) of an existing HMM for LID to a new language was proposed. The algorithm only requires language specific speech data. Neither orthographic transcriptions nor language specific phonetic lexica are needed to build a LID system for a new dialect or language. Yet it was found that a PPR LID system based on unsupervised learning can have an error rate more than 60% higher than a system based on fully supervised training ([2]).

On the other hand there exists a discriminative training procedure for the HMMs of a PPR LID system that also does not require orthographically transcribed speech data nor a language specific phonetic lexicon per se. In [3] this so called Minimum Language Identification Error (MLIDE) training of HMM parameters was only applied to an HMM obtained from fully supervised Maximum Likelihood training.

The described work attempts to combine both unsupervised adaptation of an HMM to a new language and discriminative optimization of such a language adapted HMM. The training procedure should not require orthographically transcribed speech data and language specific phonetic lexica. On the other hand it would be desirable if it led to the same low error rates as supervised training with orthographically transcribed speech data and language specific lexica.

2. General System Architecture

Figure 1 shows an example system for 3 languages. It consists of 3 language dependent phoneme recognizers each with a language dependent HMM and phoneme bigram language model. The feature extraction that transforms the speech signal to a series of feature vectors is common for all phoneme recognizers. The phoneme recognizers are based on one pass Viterbi decoders that deliver the most probable phoneme sequence and a normalized negative log likelihood — also called score — for all languages. In the described PPR LID system the computed phoneme sequences are not further exploited but only the scores are further processed. The scores is related to the a-posteriori probability for the specific language but due to a simple approximation ([4]) for score normalization this relation is rather weak. Therefore an optional ANN is applied in order to better approximate the a-posteriori language probabilities. It delivers the approximated a-posteriori language probabilities as a result of non-linear processing of all languages scores.
a language decision without the optional Artificial Neural Network (ANN) can be taken by deciding on the language with the minimal score. Using the ANN the decision is taken by selecting the language with the maximum a-posteriori probability.

3. Supervised Maximum Likelihood HMM Training

In Maximum Likelihood (ML) training HMM parameters (Gaussian mean vectors) are optimized to lead to maximum production probability of the training data given the speech production model. If the taken model assumptions were true and an unlimited amount of training material was available this should lead to optimal parameters in terms of error rates for Automatic Speech Recognition (ASR) and Language Identification (LID). For a real world HMM system it is known that both requirements are not met and other training criteria may lead to better results in terms of ASR and LID error rates.

For Supervised Maximum Likelihood HMM Training (SuML) orthographically transcribed speech material and a phonetic lexicon covering all words in the training material are necessary. The orthographic transcriptions and the phonetic lexicon are combined to form the phoneme sequence of each utterance. In the described system a forced Viterbi alignment is then performed and HMM parameters are updated iteratively.

4. Unsupervised MLLR+MAP HMM Adaptation

If for some reason orthographic transcriptions of the training speech data or a phonetic lexicon is not available Unsupervised HMM Adaptation (UnAd) can be used to generate a Language Adapted HMM (LA HMM) for a new language. As described in [2] the training speech data can be phonetically transcribed with a phoneme recognizer and an existing HMM.

The HMMs for the phoneme recognizer are usually Maximum Likelihood (ML) trained on some language or a set of languages where the requirements for standard ML training are met. In a practical application one would train the HMM on a set of languages making it a multilingual (MuLi) model covering the phonemes of the training languages. Once the phonetic transcription for the training material is obtained forced Viterbi parameter estimation can take place. As the amount of speech material for HMM adaptation can vary a combination of Maximum Likelihood Linear Regression (MLLR) and Maximum A-Posteriori (MAP) is chosen for parameter optimization because it is less depending on the size of the training material than standard Maximum Likelihood estimation ([5]).

5. Minimum Language Identification Error HMM Training

As described above Maximum Likelihood HMM Training is not guaranteed to lead to minimum classification error rates in a real world setup. A so called discriminative training criterion ([6]) can help to further reduce the language identification error rate. Using the Minimum Classification Error (MCE, [7]) parameter estimation scheme an objective function can be formulated that approximates the language classification error rate on the set of training patterns. The objective function must be differentiable regarding the HMM parameters. With the help of a gradient descend the HMM parameters can then be optimized to lead to a minimum of the classification error rate on the training material. If the training material represents the classification task well it can be expected that Minimum Language Identification Error training (MLIDE) also leads to decreased classification error rates on independent test pattern.

Looking at the procedure required for MLIDE training of HMM parameters it can be seen that the major part of the procedure is identical to the language recognition process. Therefore for MLIDE HMM training only the resources of the recognition system are needed which means that no orthographic transcriptions of the training material and no respective phonetic lexica are necessary. Therefore MLIDE HMM training is per se unsupervised in the sense that only speech material tagged with the spoken language is needed. Phoneme bigrams can be applied but could also be omitted which was shown in [3] to have little influence on the resulting classification performance.

6. Language Model Training

In the described work the parameters of all Language Models (LMs) are obtained by Maximum Likelihood parameter estimation. In order to train language model parameters only a set of phoneme sequences needs to be considered. Further optimization of LM parameters by means of MLIDE is for sure possible and promising but not investigated here.

In the case of Supervised LM Training the phoneme sequences can be compiled from the orthographic transcriptions and the phonetic lexicon. Such a LM could also be denoted as Language Dependent Language Model (LD LM).

For Unsupervised LM Training the phoneme sequences are obtained from a phoneme recognizer applied to the training speech material. The phoneme recognizer can either utilize the initial — usually multilingual (MuLi) — HMM or a language adapted HMM obtained from unsupervised HMM adaptation.

7. ANN Training

All Artificial Neural Network (ANN) parameters are estimated with minimum square error objective. During training the language scores from the parallel phoneme recognizers are presented to the input nodes and a binary pattern is presented to the output nodes. All output nodes are targeting 0 except the one for the spoken language which is set to 1. In this way the ANN should be able to approximate the a-posteriori probabilities of the languages ([8]). ANN training is performed separately for each specific combination of HMMs and LMs.

8. Experimental Setup

8.1. CDHMM System

The language dependent phoneme recognizers are based on Continuous Density Hidden Markov Model (CDHMM) system originally developed for Automatic Speech Recognition (ASR) ([9]). Feature extraction is based on MFCCs and a linear transformation based on Linear Discriminant Analysis (LDA) is applied. The linear transformation optimized for ASR uses a super-vector of 2 MFCC vectors. 3-state context independent phoneme models in Bakis topology with fixed transition penalties are applied. Each phoneme recognizer uses a set of 2048 Gaussian densities with only one variance parameter.
8.2. Databases
HMM training, adaptation, language model training and ANN training are performed on the phonetically rich sentences of the SpeechDat II ([10], [11]) and Polyphone (fixed network part) databases containing only fixed network telephone speech. Recognition tests are either on fixed network speech (data bases mentioned before) or on mobile telephone network speech (SpeechDat II mobile and Polyphone). The set consists also only of phonetically rich sentences with a mean length of about 6 seconds. For the multilingual HMM training 3 non-target languages are chosen: French, Polish and Spanish. The 4 target languages for LID Training / adaptation / recognition are Italian, English, German and Dutch.

Further details of the experimental setup can be found in [2] and [3].

9. Experimental Results
Figure 2 illustrates the three different combinations of HMM and LM training methods. Note that the steps illustrated in Figure 2 are performed separately per target language except for the discriminative HMM training where all phoneme recognizers are optimizes together.

In figure 3 and 4 the changes in language identification error rates during MLIDE training on the training and the test set are plotted. All results shown in the plots are without the optional ANN. The error rates at iteration index 0 represent the three different non discriminatively trained setups: supervised HMM and LM Training (left in Fig. 2), unsupervised HMM adaptation and unsupervised LM training (center in Fig. 2) and the multilingual HMM together with unsupervised LM training (right in Fig. 2). As for all experiments the number of MLIDE iterations was set to 8 the numbers for this index give the final error rates with additional discriminative parameter optimization.

Looking at the error rates on the training set it can be seen that the large absolute differences in error rates before MLIDE training are greatly reduced: while the error rate for the optimized multilingual HMM reaches about 3%, the error rates for the language specific (SuLM) HMM and for language adapted HMM (UnAd) both get down below 1%. It should be noted here that the convergence of MLIDE training of the multilingual HMM was somehow difficult — the normalized step width had to be heavily reduced in comparison to the two other starting models. Maybe some additional gain would be possible if the step width was adjusted in every iteration to avoid divergence as in [6].

Similar trends as for the training set can be observed on the (fixed network) test data during MLIDE training using 3 different starting models according to different non discriminative training methods.
lot smaller after MLIDE optimization. While the relative (absolute) increase of error rate due to the unsupervised steps was 61% (4.1%) without MLIDE optimization it is only 26% (0.8%) after MLIDE. Similar observations can be found for the equal error rates. It can be seen that the ANN significantly reduces the error rates.

Table 2 gives results with ANN for the mobile telephone test data. As found in [3] MLIDE optimization shows robustness against database mismatch between training and recognition. Other than for for the fixed network test data MLIDE slightly increases the loss (no MLIDE: 17%, with MLIDE: 26%) of recognition performance introduced by unsupervised learning. On the other hand the result for unsupervised learning with additional MLIDE is much better than even for the supervised learning without MLIDE.

The recording conditions of the mobile databases are of course very different from the ones of the fixed network databases. Therefore the performance gain in case of the database mismatch indicates that the proposed system is really able to discriminate between languages and not between recording conditions.

### 10. Summary

An unsupervised and discriminative learning scheme for Hidden Markov Models (HMMs) in acoustic Language Identification (LID) based on Parallel Phoneme Recognizers (PPR) was presented. It combines recently proposed unsupervised adaptation ([2]) and Minimum Language Identification Error training (MLIDE, [3]). A series of experiments was performed on large telephone speech databases with and without database mismatch. Non discriminative supervised HMM training and unsupervised language adaptation of a multilingual HMM were compared. Both possibilities were investigated with and without additional MLIDE optimization. As an alternative a third combination with MLIDE training starting immediately from the multilingual model was tested.

Experimental results showed that the increase of error rates introduced by the unsupervised learning step is significantly smaller when discriminative training is applied. The best result on a four languages fixed network task is 3.0% LID error rate that is obtained with supervised learning and additional MLIDE optimization. In comparison the result for the new combined unsupervised and discriminative method on the same task is 3.8% LID error rate. It was also found that starting MLIDE directly from the multilingual HMM — omitting the unsupervised HMM adaptation — unnecessarily decreases the system performance.

### 11. References


