Minimum Classification Error Training of Hidden Markov Models for Acoustic Language Identification

Josef G. Bauer *Ekaterina Timoshenko*

Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81730 Munich, Germany

{Josef.Bauer,Ekaterina.Timoshenko.ext}@siemens.com

Abstract

The goal of acoustic Language Identification (LID) is to identify the language of spoken utterances. The described system is based on parallel Hidden Markov Model (HMM) phoneme recognizers. The standard approach for parameter learning of Hidden Markov Model parameters is Maximum Likelihood (ML) estimation which is not directly related to the classification error rate. Based on the Minimum Classification Error (MCE) parameter estimation scheme we introduce Minimum Language Identification Error (MLIDE) training that results in HMM model parameters (mean vectors) that give minimum classification error on the training data. Using a large telephone speech corpus with 7 languages achieve a language classification error rate of 4.7% which is a 40% reduction of error rate compared with a baseline system using ML trained HMMs. Even if the system trained on fixed network telephone speech is applied to mobile network speech data MLIDE can greatly improve the system performance.

Index Terms: Automatic Language Identification, Hidden Markov Model training, discriminative training, Minimum Classification Error

1. Introduction

Automatic acoustic Language Identification (LID) aims to determine the language of spoken utterances. One of the most successful approach to LID is the use of Parallel Phoneme Recognizers (PPR, [1]) based on Hidden Markov Models (HMMs). Usually the parameters of such HMMs are based on an Maximum Likelihood (ML) objective which is not directly related to the language identification error rate.

In automatic speech recognition so called discriminative training criteria like Maximum Mutual Information (MMI) and Minimum Classification Error (MCE, [2]) have become popular for estimation of model parameters ([3]). The key issue in discriminative training is the fact that for estimation of a model not only the data and the specific model are considered but also competitive models and patterns from competitive classes.

In [4] parameters of Gaussian Mixture Models (GMMs) for Language Identification were optimized based on Minimum Classification Error (MCE) objective. In [4] the OGI database was used which contains only a rather small amount of speech data.

In the described work we apply discriminative parameter optimization based on MCE to mean vectors of Hidden Markov Models (HMMs) use for parallel phoneme recognizers in combination with an Artificial Neural Network (ANN). Minimum Language Identification Error (MLIDE) training aims to find a set of model parameters with minimum language identification error rate on the set of training patterns. For experimental investigations we use 7 languages from the SpeechDat II Corpus ([5]) which offers a large amount of speech data (about 30 hours per language) which is crucial for discriminative training methods ([6]).

2. General LID System Description

The described language identification system consists of several language dependent phoneme recognizers optionally using an integrated language specific phoneme bigram model and one common Artificial Neural Network (ANN). Figure 1 illustrates the system architecture.

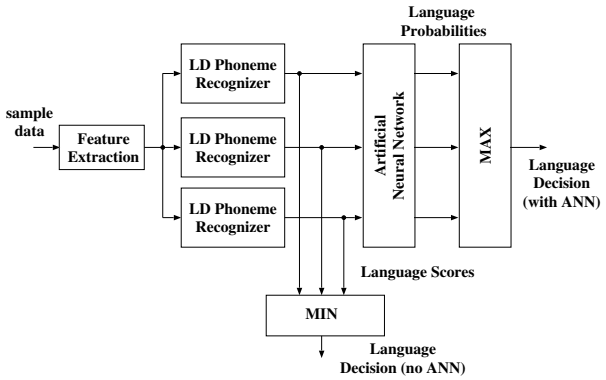

Figure 1: Example LID System for 3 languages with 3 Language Dependent (LD) phoneme recognizers.

Each language specific phoneme recognizer computes

the negative log likelihood for the feature vector sequence based on first best Viterbi decoding. In oder to roughly approximate a-posteriori probabilities we use a very simple normalization technique where we subtract the sum of the minimal neg-log state specific likelihoods and divide by the number of frames (see [7]). To make immediate use of this normalized scores we can use a minimum detector to find the most probable language. This language decision or the neg-log scores respectively are the system output when no ANN is in use.

In order to improve the classification performance and to better approximate the a-posteriori language probabilities an Artificial Neural Network (ANN) is employed. We are using a two layer perceptron with both the number of input nodes and output nodes being the number of considered languages. The ANN is trained using the neg-log scores from the language dependent phoneme recognizers as input an a binary pattern as output — the output value for the spoken value is always 1. Trained with minimum square error objective the output nodes should well approximate the a-posteriori probabilities for the considered languages. For a language decision a maximum detector after the ANN must be applied.

3. MCE Training of HMMs for LID

The basic idea of Minimum Classification Error (MCE) Training is a differentiable objective function l_{MCE} that approximates the classification error rate. The goal is to find a set of parameters Λ that minimizes the objective function for a set of training patterns $\{S_r\}$:

$$\Lambda_{\mathrm{MCE}} = \underset{\Lambda}{\operatorname{argmin}}\, l_{\mathrm{MCE}}(\Lambda, \{S_r\}) \tag{1}$$

In case of acoustic Language Identification (LID) the objective function for Minimum Language Identification Error (MLIDE) training approximates the language classification error rate. In the described work only the mean vectors of the Hidden Markov Models as the parameters Λ were optimized. For the optimization of the model parameters we apply a simple iterative gradient algorithm with constant learning rate ϵ:

$$\Lambda_{k+1} = \Lambda_k - \epsilon \cdot \nabla l(\{S_r\}, \Lambda_k) \tag{2}$$

l_{MCE} is the mean value of $l(S_r, \Lambda)$ for all training patterns S_r. $l(S, \Lambda)$ is defined via a sigmoid function based of the discriminant function d:

$$l(S, \Lambda) = \frac{1}{1 + e^{-\gamma d(S,\Lambda)}} \tag{3}$$

One important aspect of real world MCE training is to find an appropriate value for γ controlling the slope of the sigmoid function. For the definition of the MCE discrimination function d the logarithmic model probabilities

$$g(S, \lambda) = -logP(S|\lambda) \tag{4}$$

are essential:

$$d(S, \Lambda) = g(S, \lambda_i) + \log\left(\frac{1}{J-1}\sum_{j\neq i} e^{-g(S,\lambda_j)\eta}\right)^{\frac{1}{\eta}} \tag{5}$$

The exponent η hereby determines how many models (the total number of models is J) contribute to the objective function. For $\eta \to \infty$ only the correct model λ_i and the best matching competitive model for a pattern contribute to the objective function.

In case of an acoustic LID system based on Parallel Phoneme Recognition (PPR) the models λ consist of phoneme recognizers with language models for phonemes. Here the models for the correct (spoken) language $i = \Omega(S)$ as well as all other models are taken into account. Note that the phoneme language models will contribute to the objective function even if we only optimize the HMM mean vectors.

In order to implement MLIDE training of HMMs phoneme recognizers delivering the model probabilities are needed. In the described work we use a first-best continuous phoneme recognizer based on Viterbi decoding. State level alignment that is needed for parameter re-estimation is performed by an extra forced Viterbi step.

To illustrate the outcome of MLIDE optimization of HMM mean vectors let us consider the case $\eta \to \infty$. Then a positive value of $d(S, \Lambda)$ corresponds to a case where the PPR system was not able to correctly classify the pattern. Vice versa a negative value of d occurs in case of a correct classification. In any case the mean vectors from the correct model will be drawn towards the aligned feature vectors and the mean vectors of the competitive model will be drawn away from the aligned feature vector. The absolute value of d determines how strongly the mean vectors are shifted. The shift will be highest for small values of $|d|$ and become 0 for large values of $|d|$. In this way the MLIDE training is dominated by patterns at the classification boundaries.

4. Experiments and Results

4.1. Databases and System Setup

For training and evaluation we are using the SpeechDat II, Polyphone and SpeechDat II Mobile databases. From SpeechDat II and Polyphone we use Italian, Spanish, French, German, Polish, English and Dutch languages. For HMM parameter estimation the official set of training speakers from SpeechDat II are employed (17352 speakers). The official set of test speakers from SpeechDat II was divided in a development set (2819 speakers) used for optimizations of the ANNs and an evaluation set (955 speakers) for LID tests. A set of 813 speakers from SpeechDat

II Mobile was used for LID evaluation on mobile data. For the mobile data we are using Italian, English, German and Dutch language.

For training and evaluation phonetically rich sentences are used. Utterances with the exact wordings apearing in the test utterances were removed from the training set as in [8]. Phonetically rich sentences in SpeechDat II have a mean length of 7 seconds. Phonetically rich sentences in SpeechDat II Mobile have a mean length of 8 seconds.

The underlying CDHMM system was originally designed and developed for automatic speech recognition. The Maximum Likelihood (ML) HMM parameters are indeed the same as used for simple language dependent speech recognition. We are using 3-state mono-phone models in Bakis-topology with fixed transition penalties. For each language a set of 2048 Gaussian densities with diagonal covariance matrices and only one global variance parameter is applied. Feature extraction is based on MFCCs together with a linear transformation from Linear Discriminant Analysis (LDA) with multilingual mono-phone state being the classes. In future systems we want to use language dependent mono-phone models as classes for the LDA matrix estimation.

The phoneme bigram models are estimated on the transcriptions of phonetically rich sentences in the databases used for HMM training. The parameter of this language models are estimated with Maximum Likelihood objective. Rarely seen or unseen probabilities are floored.

The crucial parameter γ for MCE training is adjusted using the histogram method for d with $\eta \rightarrow \infty$ as described in [6]. We are also using the gradient normalization technique described there. The parameter η is adjusted in a way that most of the time only the most competitive model contributes to parameter optimization. Maximum Likelihood trained models serve as starting point for MCE parameter re-estimation. The number of MCE iterations was always set to 6.

4.2. Performance Measurement

The first scenario considered here is classification where one out of the set of investigated languages is the result of the LID system. In this case we are using the LID Error Rate (ER) in % as the performance measure. Note that the LID Error Rate (ER) is most closely related to the objective function for MLIDE HMM training as described above.

In a detection scenario the result of the LID system is a set of measures (usually approximated a-posteriori probabilities). In an extra processing step a decision for each language whether the language is detected as spoken is taken. This is done based on a threshold. Varying the threshold we can get the Equal Error Rate (EER) in % where the false acceptance rate is equal to the false rejection rate. Note that this measure is not directly related to the objective of MLIDE as describe in this paper.

4.3. Results without Database-Mismatch

In a first series of experiments we evaluate different setups on the SpeechDat II fixed telephone network database. This means that different subsets of the *same database* were employed for HMM training, ANN parameter optimization, and system evaluation. In this case we don't have a database mismatch.

Although we do not optimize the language models with a discriminative training procedure we do consider them in the MLIDE HMM parameter optimization scheme. But we have the choice to use the bigrams as in the LID system or to employ zerograms which corresponds to not considering the language models in MLIDE HMM training.

Figure 2 shows convergence of MLIDE training of HMMs with bigrams models in training. Results are with LMs but without ANN. Table 1 gives results for Maximum Likelihood and Minimum Language Identification Error training of HMM mean vectors. For MLID we consider two cases: with or without language models in MLIDE training. First of all the use of bigram phoneme language models and artificial neural network greatly improves the error rates as well as the equal error rates. The improvement on the equal error rates by the ANN is enormous which can be explained by the poor approximation of the a-posteriori probabilities through the simple score normalization.

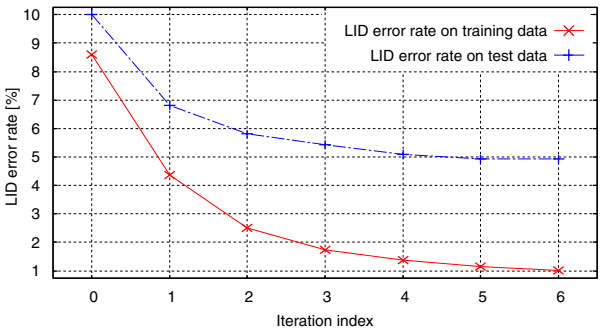

Figure 2: Convergence of MLIDE training with bigrams, without ANN, fixed network training and test data

It can be seen that in all cases MLIDE HMMs significantly outperform ML HMM concerning the classification error rates as well as concerning the equal error rates. The use of zerograms instead of bigrams (as used for LID) in MLIDE parameter estimation degrades results without ANN but when the ANN is in use the overall results are very similar.

HMM-Training	LMs	ANN	ER	EER
ML	-	-	14.6	23.4
ML	-	+	10.2	4.8
ML	+	-	10.0	26.7
ML	+	+	7.9	3.7
MLIDE using Bigrams	-	-	11.7	23.8
MLIDE using Bigrams	+	-	4.9	26.8
MLIDE using Bigrams	+	+	4.7	2.7
MLIDE using Zerogr.	-	-	7.6	23.5
MLIDE using Zerogr.	+	-	5.4	26.8
MLIDE using Zerogr.	+	+	4.8	2.6

Table 1: Comparison of HMM training criteria Maximum Likelihood (ML) and Minimum Language Identification Error (MLIDE) with Bigrams or Zerograms, with / without language models (LM) in LID, with / without Artificial Neural Network (ANN), language Error Rates (ER) and Equal Error Rates (EER) in % on *fixed* telephone test data

4.4. Results with Database-Mismatch

This section presents results of the system trained on fixed network telephone speech (SpeechDat II) tested on mobile network telephone speech (SpeechDat II Mobile). Because of the GSM channel and the different environmental conditions (e.g. car noise) we can speak of a severe database mismatch in this case.

Table 2 shows results on the mobile test data featuring 4 languages. The setup is using the ANN and language models. The error rates on the mobile data are higher than

HMM-Training	ER	EER
ML	19.5	12.0
MLIDE using Bigrams	13.2	8.4
MLIDE using Zerogr.	14.4	9.0

Table 2: Comparison of HMM training criteria Maximum Likelihood (ML) and Minimum Language Identification Error (MLIDE) with Bigrams or Zerograms, with language models (LM) in LID, with Artificial Neural Network (ANN), language Error Rates (ER) and Equal Error Rates (EER) in % on *mobile* telephone test data (database mismatch)

those for fixed network test data although the number of languages is only 4 compared to 7. This might be caused by the higher noise level in the mobile speech data and — maybe more important — by the general database mismatch. Obviously MLIDE HMM training can greatly improve classification results in spite of this mismatch problem.

5. Conclusions and Future Work

We have successfully applied Minimum Language Identification Error (MLIDE) training of Hidden Markov Model mean vectors based on the Minimum Classification Error (MCE) approach. We could show that for an experimental setup with large amount of training data it is possible to reduce the language identification error rate by as much as 40% relatively. Even given a severe database mismatch between training and test the use of MLIDE parameter estimation for HMM mean vectors can heavily reduce error rates.

In our future work we are planning to optimize the other components — phoneme bigram language models and ANN parameters — with MLIDE objective.

6. References

[1] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," in *IEEE Transactions on Speech and Audio Processing*, January 1996, vol. 4, pp. 31–44.

[2] Biing-Hwang Juang and Shigeru Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.

[3] Ralf Schlüter, W. Macherey, S. Kanathak, H. Ney, and L. Welling, "Comparison of optimization methods for discriminative training criteria," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1997, pp. 15–18.

[4] Dan Qu and Bingxi Wang, "Discriminative training of GMM for language identification," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003.

[5] "Elra web site," http://www.elra.info.

[6] Josef G. Bauer, "Enhanced control and estimation of parameters for a telephone based isolated digit recognizer," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 1531–1534.

[7] Jochen Junkawitsch and Harald Höge, "Keyword verification considering the correlation of succeeding feature vectors," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, vol. 1, pp. 221–224.

[8] Diamantino Caseiro and Isabel M. Trancoso, "Spoken language identification using the speechdat corpus," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 1998.