DISCRIMINATIVE TRAINING TECHNIQUES FOR ACOUSTIC LANGUAGE IDENTIFICATION

Lukáš Burget, Pavel Matějka and Jan Černocký

Speech@FIT group, Brno University of Technology, Czech Republic
{burget,matejkap,cernocky}@fit.vutbr.cz

ABSTRACT

This paper presents comparison of Maximum Likelihood (ML) and discriminative Maximum Mutual Information (MMI) training for acoustic modeling in language identification (LID). Both approaches are compared on state-of-the-art shifted delta-cepstra features, the results are reported on data from NIST 2003 evaluations. Clear advantage of MMI over ML training is shown. Further improvements of acoustic LID are discussed: Heteroscedastic Linear Discriminant Analysis (HLDA) for feature de-correlation and dimensionality reduction and Ergodic Hidden Markov models (EHMM) for better modeling of dynamics in the acoustic space. The final error rate compares favorably to other results published on NIST 2003 data.

1. INTRODUCTION

Automatic language identification (LID) has increasing importance among speech processing applications. It can be used to route calls to human operators (commerce, emergency), pre-select suitable speech recognition system (information systems) and has many uses in security applications.

The goal for Language Identification is to determine the language a particular speech segment was spoken. The algorithms for LID can be roughly divided (see for example [2]) into two groups. In *phonotactic modeling*, a tokenizer transcribes the input speech into phoneme strings or lattices and a language model (LM) scores these structures. This approach is mostly referred to as PRLM (Phoneme recognizer followed by language model) or PPRLM (Parallel PRLM). In *acoustic modeling*, input features are modeled directly by Gaussian mixture models (GMM), artificial neural networks, support vector machines, or other techniques [3].

This work concentrates on acoustic modeling using GMM and complements our successful PPRLM based on robust phoneme recognizer [4, 18]. In acoustic modeling, we were inspired by the advantages brought by discriminative training into large vocabulary continuous speech recognition (LVCSR) systems. The performance of acoustic modeling was also increased by the use of Heteroscedastic linear discriminant analysis (HLDA), also in common use in LVCSR.

The paper is organized as follows: in section 2 we briefly review feature extraction used for LID. In section 3, we compare maximum likelihood and discriminative training approaches for acoustic modeling. Section 4 presents the experimental data. Section 5 compares the results obtained with the two mentioned training paradigms and two feature extractions. The following section 6 concentrates on the use of HLDA in our system and section 7 describes experiments conducted with Ergodic HMM (EHMM) instead of GMM. As we will see, all experimental results were obtained on a subset of training data. Section 8 therefore presents the “ultimate” results with full training set.

2. FEATURES

The most widely used features for LID (as well as for other speech processing techniques) are Mel-Frequency Cepstral Coefficients (MFCC). The works of Torres-Carasquillo [6] and others have however shown the importance of broader temporal information for LID. The shifted delta cepstra (SDC) features are created by stacking delta-cepstra computed across multiple speech frames. The SDC features are specified by a set of 4 parameters: N, d, P and k, where N is the number of cepstral coefficients, d is the advance and delay for the delta-computation, k is the number of blocks whose delta-coefficients are concatenated to form the final feature vector and P is the time shift between consecutive blocks. In case we denote the original features $o_h(t)$[1], shifted deltas are defined:

$$\Delta o_h(t) = o_h(t + iP + d) - o_h(t + iP - d)$$

for $i = 0, P, 2P, \dots, (k-1)P$. Obviously, these feature vectors are heavily correlated (most of elements are merely copied from one vector to another when we go from t to $t+1$).

Two widely used enhancements of features for LID are RASTA filtering of cepstral trajectories ensuring channel normalization [2] and vocal-tract length normalization (VTLN) [1] which is a simple speaker adaptation.

3. ACOUSTIC MODELING

Language recognition can be seen as a classification problem with each language representing a class. The most straightforward way to model class s is to construct a Gaussian mixture model that represents feature vectors by a weighted sum of multivariate Gaussian distributions:

$$p_\lambda(\mathbf{o}(t)|s) = \sum_{m=1}^{M} c_{sm}\mathcal{N}(\mathbf{o}(t); \boldsymbol{\mu}_{sm}, \boldsymbol{\sigma}^2_{sm}),$$

where $\mathbf{o}(t)$ is the input feature vector and the parameters λ of model of s-th class are c_{sm}, $\boldsymbol{\mu}_{sm}$ and $\boldsymbol{\sigma}^2_{sm}$: mixture weight, mean vector and variance[2] vector respectively. The log likelihood of utterance

This work was partially supported by EC project Augmented Multi-party Interaction (AMI), No. 506811, Grant Agency of Czech Republic under project No. 102/05/0278 and by Ministry of Defence of Czech Republic.

[1] $o_h(t)$ denotes the h-th element of feature vector $\mathbf{o}(t)$

[2] we assume diagonal covariance matrices that can be represented by variances.

$\mathcal{O}_r$ given class s is then defined as:

$$\log p_\lambda(\mathcal{O}_r|s) = \sum_{t=1}^{T_r} \log p_\lambda(\mathbf{o}(t)|s),$$

where T_r is the number of feature vectors in $\mathcal{O}_r$.

In the standard *Maximum Likelihood* (ML) training framework, the objective function to maximize is the total (log) likelihood of training data given their correct transcriptions:

$$\mathcal{F}_{ML}(\lambda) = \sum_{r=1}^{R} \log p(\mathcal{O}_r|s_r) \quad (1)$$

where λ denotes the set of model parameters, $\mathcal{O}_r$ is r-th training utterance, R is the number of training utterances and s_r is the correct transcription (in our case the correct language identity) of the r-th training utterance. To increase the objective function, the GMM parameters are iteratively estimated using well known EM algorithm reestimation formulae (see for example [16]).

In *discriminative training*, the objective function is designed in such a way that it is (or is believed to be) better connected to the recognition performance. One of the most popular discriminative training technique nowadays is Maximum Mutual Information (MMI) training where the objective function is posterior probability of correctly recognizing all training utterances:

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\mathcal{O}_r|s_r)\mathcal{P}(s_r)}{\sum_{\forall s} p_\lambda(\mathcal{O}_r|s)\mathcal{P}(s)}. \quad (2)$$

We consider the prior probabilities of all classes equal and drop the prior terms $\mathcal{P}(s_r)$ and $\mathcal{P}(s)$. The denominator $\sum_{\forall s} p(\mathcal{O}_r|s)$ is the likelihood of utterance $\mathcal{O}_r$ given the "competing" model representing all possible transcriptions (in our case all language labels). The derivation of parameter update formulae is described in detail for example in [12].

Discriminative training techniques lead to consistent improvement in accuracy of LVCSR systems [11, 12]. To our knowledge, MMI training of GMMs has not been tested in LID so far. Dan and Bingxi [17] report results with Minimum classification error (MCE) criterion for the training, but the improvement they obtained was less than reported in our paper. We have tested MCE training too, but compared to MMI, the improvement was only about a half.

Our work on MMI training for LID was facilitated by the experience with discriminative training applied in AMI-LVCSR system[3] [5]. We could also rely on our HMM toolkit STK[4] that implements MMI and other discriminative training techniques.

4. EXPERIMENTAL DATA

Acoustic models were trained on the CallFriend Corpus [8]. There are 12 target languages: Arabic (Egyptian), Japanese, Farsi, French (Canadian French), German, Hindi, English (American), Korean, Mandarin, Spanish (Latin American), Tamil, and Vietnamese. The data of each target language contains 20 complete half-hour conversations. All results except for these reported in section 8 were obtained on a *small* training set: at first, all the training data were end-pointed by our phoneme recognizer [9] with all phoneme classes except `sil` linked to 'speech'. Then, 1 hour of data was selected from each of target languages by taking only segments longer than 2 seconds and balancing the amounts of data among speakers in each language.

Test data comes from NIST 2003 LID evaluation [10]. This data set consists of 80 segments with duration of 3, 10 and 30 second duration in each of 12 target languages. Unless stated otherwise, all results in this paper are reported for 30s segments. This data comes from conversations collected for the CallFriend Corpus but not included in its publicly released version. In addition, there are four additional sets of 80 segments of each duration selected from other LDC[5] supplied conversational speech sources, namely Russian, Japanese, English, and cellular English.

The *evaluation* is done according to NIST [10] per-language, considering each system is a language *detector* rather than recognizer. A standard detection error trade-off (DET) curve is evaluated as a plot of probability of false alarms against the probability of misses with the detection threshold as parameter and equal priors for target and non-target languages. Equal error rate (EER) is the point where these probabilities are equal.

5. ML AND MMI SYSTEMS - RESULTS

The following two feature extractions were tested. VTLN was applied for both of them.

- For the `MFCC38` system, 38 coefficients were used: standard setup of 13 direct coefficients, Δ and $\Delta\Delta$, without c_0 (we found that for this baseline, c_0 hurts).
- In the SDC setup, the cepstral coefficients were processed by RASTA filters. Several experiments were done with the parameters of shifted-delta computation but we ended up with the same setup as reported in [3]: 7,1,3,7 producing 49-dimensional feature vectors. The influence of adding direct coefficients was studied for $M = 512$ Gaussians. Without them (SDC only), the EER was 11.6% while with these coefficients, we obtained 8.9%. Therefore, the direct coefficients were added to the feature vector in all SDC experiments making it 56-dimensional. We experimented also with the c_0 coefficient. Unlike in MFCC, adding c_0 in SDC does not hurt so that c_0 (as well as all its shifted-deltas) was kept in feature vectors. This system is denoted `MFCC-SDC`.

For ML-training, the number of Gaussian components M was varied from 128 till 2048. For more computationally expensive MMI-training, the numbers were only 128, 256 and 512.

Upper part of Table 1 summarizes the results for ML-training. It is obvious that SDC clearly outperform MFCC which is coherent with results of other groups. We see also decreasing EER for increasing number of Gaussian components.

Lower part of Table 1 shows the results for MMI-training. We see that systems with discriminatively trained models clearly outperform the standard ML-trained ones by several percent *absolute*. In LVCSR, discriminative training usually yields only between 6-15% *relative* improvement [12], we have therefore tried to explain such a dramatic improvement in LID: In our opinion, the LVCSR-classes (phone states) that are modeled by Gaussian mixtures are already relatively well separated in the acoustic space, so standard ML training does already well enough. In LID however, the classes are heavily overlapped and Gaussians occupy mostly the same acoustic space. The models need to concentrate on "tiny details" that help to separate

[3] AMI is EC-sponsored project Augmented Multi-Party Interaction, http://www.amiproject.org

[4] http://www.fit.vutbr.cz/speech/sw/stk.html

[5] Linguistic Data Consortium, http://www.ldc.upenn.edu

system	128	256	512	1024	2048
ML-TRAINING					
MFCC38	18.8	17.3	16.2	14.8	14.5
MFCC-SDC	11.8	10.5	8.9	7.3	6.8
MMI-TRAINING					
MFCC38	7.6	7.3	7.5	-	-
MFCC-SDC	4.7	4.6	4.3	-	-

Table 1. Equal error rates in % for ML and MMI training

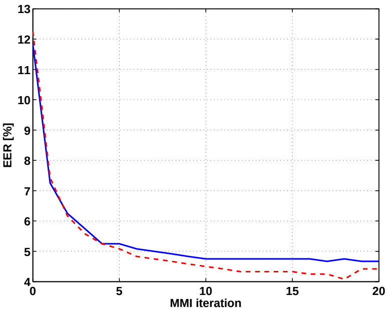

Fig. 1. Convergence of MMI-training for non-HLDA (solid line) and HLDA (dashed line) SDC features. Zeroth iteration of MMI is equivalent to ML-estimates.

the languages. With ML-training, the only possibility is the brute-force approach – increasing the number of Gaussians. On the other hand, discriminative training populates well these "tiny details" by definition.

Based on the results of this sections, we have continued with shifted delta-cepstra feature extraction and discriminatively trained models. MMI is iterative, but usually requires only a few iterations to converge (Fig. 1). Increasing the number of Gaussians in discriminative training improves the results only slightly (Tab. 1) at the price of very high computational load during the training, therefore, we stuck with $M = 128$ in the following experiments.

6. HLDA IN ACOUSTIC LID

As the next step in the development of our LID system, we have employed Heteroscedastic linear discriminant analysis (HLDA), which is also in common use in LVCSR. The reasons are obvious: our features are too highly-dimensional and (as it comes clearly from the nature of SDC) too correlated. HLDA provides a linear transformation that can de-correlate the features and reduce the dimensionality while preserving the discriminative power of features. In our previous works in small- and large-vocabulary speech recognition [14, 15], HLDA consistently improved the recognition performance.

The theory of HLDA is described in detail in [13] and [14]. HLDA needs classes to estimate its class-covariance statistics (which are then used to estimate the transform matrix). In our case, these classes were individual Gaussian components. HLDA is estimated for the ML-trained model and its parameters are fixed in the following MMI training.

Results with HLDA are summarized in Figure 2 again for MFCC-SDC feature extraction and $M = 128$ Gaussians. Different lengths of the output feature vector were tested with MMI training. Compared to non-HLDA result (4.7%), we see an improvement for wide range of these lengths; HLDA helps even in case we use it only to de-correlate, not to reduce the dimensionality (56 to 56 features). When the output dimensionality is tuned, we obtain EER of 4.1% which is a nice 0.6% absolute improvement over MMI-training only.

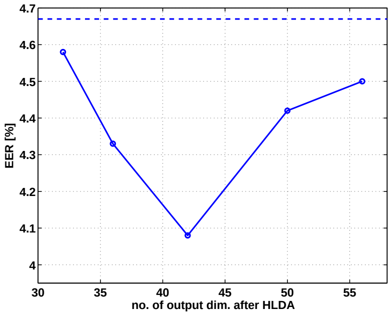

Fig. 2. HLDA processing of features. The dashed line shows the MMI result without HLDA.

7. FROM GMM TO EHMM

The final improvement we brought to our acoustic LID system was replacing the GMM by Ergodic HMM. By definition, one-state GMM can not model any dynamics in speech (some dynamics is of course represented by the delta or shifted-delta features). We felt that splitting the model of each language into several states could bring additional improvement. We therefore converted each MMI-trained GMM into a fully connected (ergodic) HMM simply by copying each individual mixture component into one state. We retrained the transition probabilities using simple ML and have not touched the mixture parameters.

The resulting likelihood of utterance can be evaluated in "Viterbi" (taking only the maximum path in EHMM into account) or "full model evaluation" styles. We have tested both without significant difference. The results are reported for Viterbi computation of the likelihood which is also about 10 times faster than full evaluation.

Table 2 summarizes the results. EHMM on the top of discriminative training and HLDA provided another 0.2% absolute improvement, which confirmed our assumption that better modeling of dynamics is advantageous.

8. FULL TRAINING DATA

All training so far was done on a small sub-set (12 out of about 275 hours) of the training data. The last step is therefore to use all the available data and report the results. This is done in the 2nd column of Table 2, with the usual feature extraction MFCC-SDC. While EHMM helped similarly as we have seen for small training data, HLDA does not perform as well — the number of output dimensions (42) was tuned on the small set, and for full training data, this reduction in dimensionality already seems to suppress useful information. This could be fixed by more careful tuning of the output dimensionality on the full set.

Compared to our best EER of 1.8% from our phonotactic system [18], we see that the best result obtained from acoustic modeling, also 1.8%, is very competitive. It also compares favorably to the

system	small	full
ML 2048	6.8	4.8
MMI 128	4.6	2.0
MMI 128 + HLDA	4.1	2.1
MMI 128 + EHMM	4.3	**1.8**
MMI 128 + HLDA + EHMM	**3.9**	-

Table 2. Equal error rates in % for MMI-trained system, completed by HLDA and EHMM. The results in the right column are for the full training set. For comparison, results of ML-training are presented for M=2048 Gaussians.

system	**30s**	**10s**	**3s**
ML 2048	4.7	7.9	16.3
MMI 128	2.1	5.5	14.8
MMI+PPRLM	0.8	3.0	11.8

Table 3. Equal error rates in % for ML and MMI systems depending on duration of test segments. MMI+PPRLM represents our final system – a fusion of acoustic and phonotactic approaches.

best published GMM-based results obtained by MIT: 4.8% (Table 2 in [3], note that with ML-training, we obtained exactly the same result).

So far, all the results were reported for test segments of duration 30 seconds. As mentioned in Section 4, the test data include also 10 and 3 second segments. The results obtained for all the three conditions are presented in Table 3. The MMI-system with only 128 Gaussians clearly outperforms the ML-system (2048 Gaussians) in all conditions, though improvement is not that prominent for shorter segments. The last line of the table presents results obtained with our final system, which is a fusion [4] of MMI-trained acoustic system, and PPRLM system [18]. Similar system was also very successful in 2005 NIST language recognition evaluations.

9. CONCLUSIONS

This paper deals with acoustic modeling for language identification. We have verified that the results of other labs obtained with shifted delta-cepstra (SDC) features are valid and that these features are good for this task. We have concentrated on discriminative training methods and have shown, that MMI-based training of models for LID clearly outperforms widely used ML-training. This verified the assumption that for LID, discriminative training would bring more significant improvement than to LVCSR systems due to high overlap of classes in the feature space. We have also studied HLDA applied to de-correlate feature vectors and reduce their dimensionality, and EHMMs to model the dynamics of features. Both methods bring further improvement.

10. REFERENCES

[1] Jordan Cohen, Terri Kamm and Andreas G. Andreou, “Vocal tract normalization in speech recognition: Compensating for systematic speaker variability”, *J. Acoust. Soc. Am.*, 97, 3246 (1995).

[2] M.A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech.,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[3] E. Singer, P.A. Torres-Carrasquillo, et al., “Acoustic, phonetic, and discriminative approaches to automatic language identification,” in *Proc. Eurospeech*, Sept. 2003, pp. 1345–1348.

[4] Pavel Matějka, Petr Schwarz, Jan Černocký, Pavel Chytil, “Phonotactic Language Identification using High Quality Phoneme Recognition” in *Proc. Eurospeech*, Sept. 2005.

[5] T. Hain, et al.: “The 2005 AMI System for the Transcription of Speech in Meetings”, in *Proc. NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh, July 2005.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., “Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features”, in *Proc. ICSLP* Denver, CO, September 2002.

[7] H. Hermansky and N. Morgan. RASTA processing of speech. *Trans. on Speech & Audio Processing*, 2(4):578–589, 1994.

[8] “Callfriend corpus, telephone speech of 15 different languages or dialects,” www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone.

[9] P. Schwarz, P. Matějka, and J. Černocký, “Towards lower error rates in phoneme recognition,” in *Proc. International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, Sept. 2004, pp. 465–472.

[10] A.F. Martin and M.A. Przybocki, “NIST 2003 language recognition evaluation,” in *Proc. Eurospeech*, Sept. 2003, pp. 1341–1344.

[11] R. Schluter, W. Macherey, B. Muller and H. Ney, “Comparison of discriminative training criteria and optimization methods for speech recognition”, *Speech Communication*, 34(2001):287–310, 2001.

[12] D. Povey, “Discriminative Training for Large Vocabulary Speech Recognition”, PhD. Thesis, Cambridge University, July, 2004.

[13] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, John Hopkins University, Baltimore, 1997.

[14] L. Burget, “Complementarity of Speech Recognition Systems and System Combination”, PhD. Thesis, Brno University of Technology, 2004.

[15] L. Burget, “Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis”, in *Proc. 8th International Conference on Spoken Language Processing - ICSLP*, Jeju Island, Korea, 2004.

[16] S. Young et al.: *The HTK Book (for HTK Version 3.3)*, Cambridge University Engineering Department, 2005, http://htk.eng.cam.ac.uk/.

[17] Q. Dan and W. Bingxi, “Discriminative training of GMM for language identification”, in *Proc. ICSA and IEEE Workshop on spontanesous speech processing and recognition*, Tokyo, Japan, April 2003.

[18] P. Matějka, L. Burget, P. Schwarz, and J. Černocký, “Use of anti-models to further improve state-of-the-art PRLM Language Recognition System”, accepted to ICASSP 2006, Toulouse, France, 2006.