LANGUAGE RECOGNITION USING PHONE LATTICES

J.L. Gauvain, A. Messaoudi, and H. Schwenk

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,abdel,holger}@limsi.fr

ABSTRACT

This paper proposes a new phone lattice based method for automatic language recognition from speech data. By using phone lattices some approximations usually made by language identification (LID) systems relying on phonotactic constraints to simplify the training and decoding processes can be avoided. We demonstrate the use of phone lattices both in training and testing significantly improves the accuracy of a phonotactically based LID system. Performance is further enhanced by using a neural network to combine the results of multiple phone recognizers. Using three phone recognizers with context independent phone models, the system achieves an equal error rate of 2.7% on the Eval03 NIST detection test (30s segment, primary condition) with an overall decoding process that runs faster than real-time (0.5xRT).

1. INTRODUCTION

Some of the most efficient approaches to language identification (LID) rely on language dependent phone N-gram models which are based on the assumption that phonotactic constraints contain enough information to identify the language[1]. Various implementations of this basic idea have been proposed, starting with the use of N-gram based phone decoders (one per language) [2, 3, 5], the use of a single phone recognizer followed by the computation of language dependent phone N-gram likelihoods [6], and the use of multiple phone recognizers followed a phone N-gram classifier [7] (approach denoted PPRLM for "parallel phone recognition followed by language dependent modeling"). For these three approaches, the acoustic and N-gram model can be estimated on either unlabeled or orthographically/phonetically labeled data, however the phone N-gram models are usually estimated on unlabeled data as this corresponds most closely to anticipated usage. The PPRLM method proposed by Zissman [8] is a very effective solution both in terms of decoding time and needed resources, as acoustic decoding can be limited to a few languages, and phone models can be trained for each language of interest as long as unlabeled data are available to estimate the phone N-gram probabilities.

One question raised by the phonotactic approach is how the posterior probabilities, $\Pr(L|X)$ for each considered language L given the speech segment X, are estimated from the phone N-gram probabilities and from the phone acoustic scores. Another related issue concerns the estimation of the phone N-gram probabilities for each language given some unlabeled data.

This paper demonstrates that phone lattices provide a very effective solution to these two problems and significantly improve the accuracy of the language identification system in comparison to previously investigated methods. In addition it is shown that a neural-network decision module can further improve the estimation of the posterior probabilities leading to better identification accuracy.

2. PHONOTACTIC APPROACH

The language identification problem can be described as finding the language with the highest posterior probability $\Pr(L|X,\Lambda,\Phi)$ given a speech segment X, a set of phone acoustic models Λ and a set of phonotactic models Φ. Assuming that the languages under consideration are equiprobable, the LID problem can be formulated as follows:

$$L^* = \operatorname*{argmax}_L \sum_H f(X|H,L,\Lambda)P(H|L) \qquad (1)$$

where L^* denotes the hypothesized language and $f(X|H,L,\Lambda)$ is the likelihood of the speech segment X given a phone sequence H and the language L, and is usually estimated with HMM phone models. The prior probability of H is estimated with a phone N-gram model $P(H|L) = \prod_i P(h_i|h_{i-N+1},...,h_{i-1},L)$.

The formulation (1) can be approximated by keeping only the dominant term of the summation, i.e.,

$$L^* \simeq \operatorname*{argmax}_L \max_H f(X|H,L,\Lambda)P(H|L). \qquad (2)$$

This implementation corresponds to the use of language dependent phone recognizers in parallel as proposed in [3] and is referred to as parallel phone recognition

(PPR) in [8]. Another approximation that can be used at this level is the assumption that the phone models are language independent [9], i.e., replacing $f(X|H,L,\Lambda)$ by $f(X|H,\Lambda)$ in equations (1) and (2). In the usual PRLM implementation [8], this is further approximated by not using any phonotactic constraints for phone decoding. The resulting formulation is

$$L^* \simeq \operatorname*{argmax}_L P(H^*|L) \qquad (3)$$

where H^* is the most likely phone sequence, i.e. $H^* = \mathrm{argmax}_H\, f(X|H,\Lambda)$. This formulation is also used in the PPRLM method, in which case phone decoding is done using multiple phone recognizers in parallel (for a few languages) and the corresponding phonotactic scores are combined for the final decision.

Finally, a better alternative solution to (3) is obtained by maximizing the expectation of $\log P(H|L)$ over L with respect to H given the observed speech X, i.e.

$$L^* \simeq \operatorname*{argmax}_L E_H[\log P(H|L)\,|\,X,\Lambda,L]. \qquad (4)$$

As for (3), the acoustic likelihood $f(X|H,\Lambda)$ is not directly included in the decision score, however the formulation (4) does take into account the hidden nature of the phone sequence.

For a language detection task (like for the NIST evaluation [13]) the posterior probability can be estimated by taking the likelihood ratio $\mathcal{F}(\cdot|L^*)/\sum_L \mathcal{F}(\cdot|L)$, where $\mathcal{F}(\cdot|L)$ denotes one of the 4 likelihoods which are maximize over L in equations (1) to (4). This likelihood ratio is also used here to combine results obtained from multiple phone recognizers.

3. BASELINE SYSTEM

For this work we adopt the PPRLM approach and make an attempt to remove some of the approximations of the original method with the goal of increasing the accuracy of the LID system.

Three phone recognizers for Arabic, American English, and Spanish are used in the experiments reported here. Each phone recognizer uses a language dependent set of context-independent phone models (primarily for efficiency reasons). Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities and 32 Gaussians per state. The acoustic feature vector has 39-components comprised of 12 PLP cepstrum coefficients and the log energy, along with their first and second order derivatives. The 42 Arabic phone models and the 27 Spanish phone models were trained on 80 conversations from the LDC CallHome family [10] of corpora (about 12 hours of Egyptian Arabic and 10 hours of Spanish data). The 48 English phone models were trained on 160 hours of conversations from the LDC Switchboard corpus [11].

Our experimental setup is in accordance with the NIST 2003 language recognition evaluation [13], where the task is to recognize the language spoken in segments of conversational speech with three nominal durations (3s, 10s, and 30s) extracted from LDC conversational speech corpora (CallFriend, Switchboard, Callhome). There are twelve target languages for this task: Egyptian Arabic, American English, Farsi, Canadian French, German, Hindi, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese.

For the baseline PPRLM system, the language dependent phone N-gram models were obtained by decoding the CallFriend [12] training data (12 languages) with each of the three phone recognizers, and by estimating 12 backoff 3-gram phone models from each phone transcription. The training data consist of 20 conversations of 30 minutes for each language, with the exception of English, Mandarin and Spanish, for which twice this amount of data was used since conversations are available for two dialects.

The decison score for the baseline system is obtained by computing the average of the three posterior probabilities. In Section 5 a much better combination method is proposed.

The NIST LID Dev'96, Eval'96 and Eval'03 data sets were used for development and testing purposes. The data sets respectively contain 1200, 1500 and 1280 test speech segments for each norminal duration (3s, 10s, 30s).

Results with our baseline PPRLM system are given in Table 1 in terms of equal error rate (EER) as reported by the NIST scoring software. These results are in the range of the best reported results with a PPRLM system. For comparison, the EER reported in [16] with the PPRLM system on the Eval'96 and Eval'03 test sets for the 30s segments are 5.6% and 6.6% respectively . It should be noted that for this work we use three phone recognizers trained on orthographically transcribed Switchboard and Callhome data, whereas the PPRLM system reported in [16] uses six phone recognizers trained on the phonetically transcribed OGI-TS corpus [15].

NIST Dev'96			
Method	3s	10s	30s
Baseline	25.7	13.3	7.2
Lattice	22.0	10.2	4.9
NIST Eval'96			
Baseline	20.5	9.7	4.9
Lattice	15.6	7.0	3.2
NIST Eval'03			
Baseline	23.7	12.6	6.8
Lattice	18.3	8.3	4.0

Table 1: Equal error rates on three NIST test sets (Dev'96, Eval'96, Eval'03 primary condition).

4. USING PHONE LATTICES

Phone lattices are graphs where nodes correspond to particular frames and where edges represent the phone hypotheses and have associated acoustic scores. The lattices are generated by the phone decoder using the acoustic models described in Section 3 without any phonotactic constraints. A typical phone lattice for a speech segment of 30s has on average 3700 nodes and 15000 edges.

The idea behind the use of phone lattices is to avoid some of the approximations made in the baseline system. To identify the language of a given speech segment, equation (1) can be better approximated by taking the summation over the phone sequences present in the phone lattice instead of just using the most likely one.

Similarly when training the phonotactic N-gram models from the unlabeled training data, X, the phone lattices can be used to obtain better maximum likelihood (ML) estimates. As a matter of fact, ML training of the phone N-grams from X, consists of finding $P(\cdot|L)$ that maximizes $\sum_H f(X|H,\Lambda)P(H|L)$ for a given set of acoustic models Λ. Therefore using only the 1-best hypothesis from the phone decoder to estimate the N-gram probabilities appears to be a crude approximation. We make the assumption here that this approximation can be overcome by summing the likelihoods over all paths in the phone lattice.

Finding estimates of the N-gram probabilities that maximize $\sum_H f(X|H,\Lambda)P(H|L)$ can be done iteratively by using the EM algorithm. Given the current estimates of these probabilities (denoted M') for one target language, the next EM estimates are obtained by computing the expectation of the N-gram frequencies $C(h_1,...,h_n)$ which can be approximated by taking the expected frequencies given the phone lattice $\mathcal{L}$. This gives us

$$E[C(h_1,...,h_N)\,|\,X,\Lambda,M'] \simeq \sum_{h(e_i)=h_i} P(e_1,...,e_N|\mathcal{L}) \tag{5}$$

where in the right hand part we compute the sum of the lattice posterior probabilities of all sequences of N edges corresponding to the phone N-gram $(h_1,...,h_N)$. The lattice posterior probabilities in (5) are computed by means of the forward-backward algorithm which gives us

$$P(e_1,...,e_N|\mathcal{L}) = \alpha(e_1)\beta(e_N)\prod_i \xi(e_i) \tag{6}$$

where $\alpha(e)$ is the forward probability of the starting node of the edge e, $\xi(e)$ is the posterior probability of the edge e, and $\beta(e)$ is the backward probability of the ending node of edge e. The new estimates of the N-gram probabilities can then be used to recompute the posterior scores with the same lattice (i.e. only the phonotactic scores are changed) for the next EM iteration. The EM procedure can be initialized with a uniform distribution.

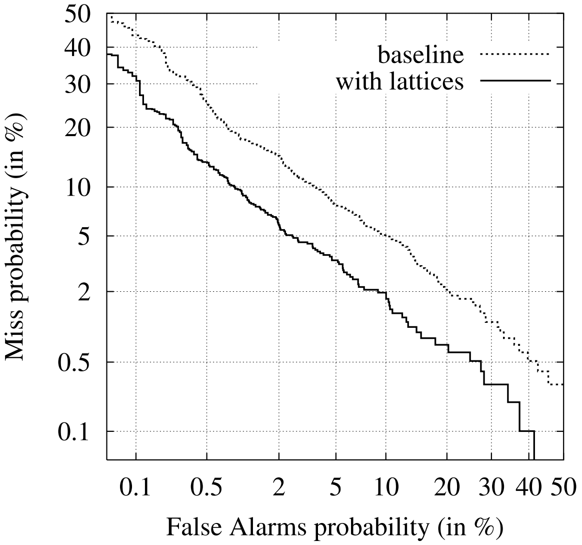

Figure 1: NIST DET curves for the 30s segments of the Eval03 test set with and without phone lattices.

During decoding, the computation of the sum of the terms $f(X|H,\Lambda)P(H|L)$ over the lattice for each target language, can be obtained by means of the forward algorithm. Alternatively, in keeping with the PRLM scheme we can use the formulation given in equation (4) which can be solved by computing the expected N-gram frequencies for each target language. This solution was adopted for the experiments reported in this paper. The results (ERR) with lattices are given in Table 1 where they can be compared to the baseline results. The error rate reduction is very significant on the three test sets for the three test durations. However, the relative error reduction is largest for long speech segments. Figure 1 plots the NIST DET curves with and without phone lattices for the 30s test segments. The lattice based method outperforms the baseline system at all operating points.

For a speech segment of 30s, computing the expected N-gram count from the lattice requires only a fraction of a second for N=3. Overall this lattice-based language identification system runs in about 0.5xRT (i.e. faster than real-time) on a Pentium4 at 2.4GHz. The measured elapsed time includes feature extraction, lattice generation for the three reference languages (about 0.4xRT), N-gram count estimation and likelihood computation.

5. NEURAL NETWORK COMBINATION

The simple method of taking the average of the posterior probabilities estimated for each phone recognizer to take the final decision can be replace by a more effective combination function. In this work a fully connected multi-layer neural network trained by stochastic back propagation, similar for instance to [14] is used. The network has 36 inputs, corresponding to three times the 12 phonotactic scores, one hidden layer with sigmoidal activation functions and 12 softmax outputs. In order to achieve good generalization behavior the Dev'96 and Eval'96 data sets have been joined together, and 60% was used for training and 40% for development and parameter optimization. In the Eval'96 data set there are approximately six times as many examples for the English language as there are for the other ones, which would result in an biased decision module. To avoid this bias only the 80 first speech segments for each language were used.

The best results are obtained by building duration dependent classifiers using 24 hidden units for the 30s test condition and 15 hidden units for the 10s condition. Table 2 reports the equal error rates for the two combination methods for the three test durations. Results are only reported for the NIST Eval'03 test set since the results on two other data sets are biased due to their use for training the neural networks (the values are very low).

Combination	3s	10s	30s
Average	18.3	8.3	4.0
NN fusion	18.3	7.9	2.7

Table 2: Equal error rates on the NIST Eval'03 test set for the two combination methods.

The neural network approach is most effective for the 30s condition, and does not improve the error rate for the 3s condition. This is in agreement with observations of other authors [16]. Using the neural network fusion the lattice based PPRLM LID system achieves state-of-the-art results [13] with a real-time factor of 0.5.

6. CONCLUSIONS

In this paper we have described our recent work in developing a language recognition system for conversational data relying on N-gram phonotactic models. The original parallel phonotactic method has been extended to use phone lattices both in training and testing instead of being limited to only the most likely phone sequence. Decoding is done by maximizing the expectation of the phonotactic likelihood for each language. The phone lattices offer much more accurate estimates of the N-gram frequencies given the hidden nature of the phone sequence in an LID system based on phonotactic constraints.

In this work three phone recognizers were used to produce phone lattices for each training and test segment. On the NIST Eval03 language recognition test set, the lattice based method reduces the equal error rate from 6.8% to 4.0% for the 30s segments, with smaller gains for the shorter segments. When the scores corresponding to the 3 phone recognizers are combined with a neural network, the equal error rate for the 30s segments is further reduced to 2.7%. This makes a very competitive language recognition system running in about 0.5xRT.

REFERENCES

[1] A.S. House, E.P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *JASA*, **62**(3):708-713, 1977.

[2] J.L. Gauvain and L.F. Lamel, "Identification of Non-Linguistic Speech Features," *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ, 96-101, Mar. 1993.

[3] L.F. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Proceedings of Eurospeech*, Berlin, Germany, **I**, 23-28, Sep. 1993.

[4] L. Lamel, J.L. Gauvain, "Language identification using phone-based acoustic likelihoods," *Proc. ICASSP*, Adelaide, vol. 1, pp. 292-295, April 1994.

[5] Y.K. Muthusamy et al. "A comparison of approaches to automatic language identification using telephone speech," *Proc. Eurospeech'93*, vol. 2, pp.1307-1310, Sep. 1993.

[6] T.J. Hazen, and V.W. Zue, "Automatic language identification using a segment-based approach," *ProcĖurospeech*, vol. 2, pp. 1303-1306, Sep. 1993.

[7] M.A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," *Proc. ICASSP*, vol. 1, pp. 305-308, Apr. 1994.

[8] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech and Audio Proc*, SAP-4(1), pp. 31-44, Jan. 1996.

[9] C. Corredor Ardoy, J.L. Gauvain, M. Adda-Decker, and L. Lamel, "Language identification with language-independent acoustic models," *Proc. Eurospeech*, vol. 1, pp. 5-8, Rhodes, Sep. 97.

[10] Linguistic Data Consortium, "The CallHome corpora," Philadelphia, 1996 and 1997.

[11] Linguistic Data Consortium, "The Switchboard corpus - LDC97S62," Philadelphia, 1997.

[12] Linguistic Data Consortium, "The CallFriend corpora, LDC96S46 through LDC96S60," Philadelphia, 1996.

[13] A. Martin, M. Przybocki,"NIST 2003 Language Recognition Evaluation," *Proc. Eurospeech'03*, pp. 1341-1344, Geneva, 2003.

[14] Y. Yan, E. Barnard, "An Approach to Automatic Language Identification Based on Language-Dependent Phone Recognition," *Proc. ICASSP*, vol. 5, pp. 3511-3514, 1995.

[15] Y.K. Muthusamy, R. Cole and B. Oshika, "The OGI Multi-Language Telephone Speech Corpus," *Proc. of ICSLP*, vol. 2, pp. 895-898, Banff, 1992.

[16] E. Singer, P. Torres-Carrasquillo, T. Gleason, W. Campbell, and D. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification," *Proc. Eurospeech'03*, pp. 1345-1348, Geneva, 2003.