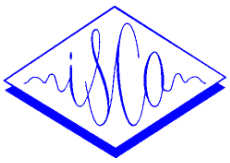

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

AUTOMATIC LANGUAGE IDENTIFICATION USING MIXED-ORDER HMMS AND UNTRANSCRIBED CORPORA

Ludwig Schwardt, Johan du Preez

Department of Electrical and Electronic Engineering,
University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa
E-mail: schwardt@ing.sun.ac.za

ABSTRACT

The state-of-the-art language identification (LID) systems are based on phone recognisers and n-gram language models, which require the use of transcribed speech databases for training. An alternate solution to the LID problem directly applies mixed-order hidden Markov models (HMMs) to untranscribed speech. The competitive performance of these mixed-order HMMs on the NIST 1996 evaluation set is very promising, considering the ease of implementation and many possible improvements. This validates a novel mixed-order HMM training procedure and extends previous results obtained with high-order HMMs to take advantage of larger datasets.

1. INTRODUCTION

An important category of automatic language identification (LID) systems is based on distinctive phonotactic information contained in the different languages. In this grouping a distinction is made between systems based on explicit phoneme recognition followed by n-gram modelling [1] [2], and the much larger and more complex systems that are based on Large Vocabulary Continuous Speech Recognition (LVCSR) systems [3]. For good performance both these approaches require the availability of aligned transcriptions for all or most of the corpora on which these systems are trained. This is an obstacle for deploying this technology in situations where such transcribed databases are not yet available. (South Africa has 11 official languages, most of them with no speech corpora available.) Being able to train phonotactic-based LID systems on untranscribed corpora can provide a major advantage in such circumstances.

This paper examines the use of ergodic high-order HMMs for phonotactic language recognition without requiring transcriptions. The rationale behind phonotactic methods is that languages have access to a largely common set of basic sounds or phones that are combined in a language-specific fashion to form sentences. It is well known that these sound dependencies are high-order in nature, but initial approaches to exploit this high-order language structure directly met with intractable results [4].

Du Preez [5] [6] recently showed that high-order HMMs can be used efficiently by converting them to equivalent first-order HMMs that represent context by states instead of links. This led to the fast incremental training (FIT) algorithm that incrementally expands the HMM order during Expectation-Maximisation (EM) training to alleviate the problems of excessive computational requirements and specialisation due to local optima. The high-order models outperform first-order models on language recognition by capturing additional language structure. These models rapidly grow with increasing order and only contain fixed-order links. This leads to an over-abundance of model parameters if the true source is mixed-order in nature.

In this paper a new mixed-order ergodic HMM training algorithm is evaluated for language identification. More details on the technique can be found in a companion paper [7]. Results show that models trained with this algorithm are smaller than comparable fixed-order models and also train faster. Performance on unseen test data remains similar to that previously obtained with (much larger) fixed-order models.

Previous LID tests with high-order HMMs [8] were hampered by a lack of training data. This is especially severe in the case of data-driven methods that depend on longer contexts, since the number of contexts to be considered grows exponentially with the model order. The current experiments use the full CALLFRIEND corpus, which allows the high-order models to reach their full potential. The models are evaluated on the NIST 1996 test set for comparison with state-of-the-art LID systems.

The results are very promising, considering the fact that several obvious improvements have not been implemented and that the entire training process requires little effort compared to phone-based systems.

2. EXPERIMENTAL SETUP AND SIGNAL PROCESSING

The LID experiments were performed on the CALLFRIEND speech corpus[1], a large untranscribed database of conversational telephone speech. The database size promotes the use of longer contexts. The corpus contains twelve languages, of which three (English, Mandarin and Spanish) have two dialects each. A language model was trained for each of the fifteen dialects on the full training and development set. This resulted in approximately 15 hours of training data per model after silences and crosstalk were removed.

After pre-emphasis and normalisation of the power in the speech signal, tenth-order LPC cepstra and delta cepstra are calculated from 32ms speech frames at 16ms intervals. Cepstral mean subtraction is used to alleviate some of the adverse channel effects. These 20-dimensional feature vectors are directly fed to HMM

[1]The CALLFRIEND corpus was obtained from the Linguistic Data Corporation (LDC).

recognisers for language identification.

Evaluation was done on the National Institute of Standards and Technology (NIST) 1996 Language Identification Evaluation corpus. This dataset consists of telephone speech segments of various durations (3s, 10s and 30s), spoken in one of twelve target languages. The vast majority of segments are derived from the CALLFRIEND evaluation set, augmented with extra English conversations from the SWITCHBOARD, KING, OGI_TS and OGI_22 corpora. These extra segments were ignored during testing, in line with existing published results [9]. The evaluation set represents approximately one hour of test data per dialect.

The three languages with two dialects each are treated as single languages during evaluation. This is achieved by taking the maximum of the two related dialect scores as the score for the language they represent. The final system therefore considers 12 alternatives during classification. Additional tests evaluated pairwise classification, whereby every language was compared with every other language and the error rates averaged. In all cases the classifier is forced to choose a specific language from a closed set, thereby avoiding verification issues.

3. LANGUAGE MODELS AND ALGORITHMS

In each experiment the individual HMMs share a common acoustic alphabet of 32 diagonal Gaussian observation pdfs. This language-independent codebook makes the system more robust against adverse channel effects, a major concern when using telephone speech. The language models therefore only differ in their transition probabilities, resulting in a purely phonotactic approach to language identification. The observation pdfs are kept simple in form in order to allocate more degrees of freedom to high-order transition statistics.

All HMMs were trained with Viterbi reestimation [4] incorporating a beam that discards highly unlikely state sequences at each time instant. This cut computation times by between a factor two and four, with very little effect on the final likelihood scores.

The baseline system is a standard ergodic first-order HMM with 32 states (termed X1 in the following). Although not as powerful as the high-order models, it will serve as a useful reference and also as initialisation for the incremental training of second-order models.

The fixed-order models Fn of order n were trained with the FIT algorithm, which initialises the training of model Fn with the trained model F$(n-1)$ of one lower order.

The mixed-order training method replaces the traditional transition probability reestimation step in Viterbi HMM training with a flexible Markov structure inference step. The inference is done by smallest encoding context trees (SECT), which explicitly models context with a prediction suffix tree representation. The algorithm recursively grows the context tree to include the contexts that, based on minimum encoding inference, are judged well supported in the training data. The tree is finally converted to an equivalent ergodic Markov chain to form the new HMM topology for the next EM training iteration. More details on the mixed-order HMM training procedure can be found in a companion paper [7].

Model	Training time (hr/language)	Model size (links)	Test set error (%) 12-way	pairwise
X1	0.5	585	48.2	13.4
F2	1.6	2640	37.8	10.2
MF2	1.4	2313	38.5	10.3
C2	1.3	4177	35.0	8.9
F3	5.8	9074	33.9	9.1
MF3	2.0	4313	35.0	9.1
C3	7.6	18675	32.4	7.8
MFx	3.2	11280	38.1	8.7

Table 1: LID results obtained on the 30s CALLFRIEND segments of the NIST 1996 test set. X1 is a standard first-order model. The Fn models were trained incrementally via FIT, with n indicating the fixed model order. The mixed-order MFn models were trained incrementally with SECT inference, where n indicates the maximum order, while MFx was trained directly with no order limiting. The Cn models describe contexts with n distinct states and arbitrary repetitions. The training times are measured on a Celeron 400 MHz processor.

The SECT representation is closely related to the multigram approach to language modelling [10] [11]. Both are based on variable-length acoustic units that are inferred in an unsupervised fashion from raw speech data using the MDL principle [12]. While multigrams follow a lexicon approach, SECT uses a context tree for inference and an equivalent finite-state machine representation for scoring.

The mixed-order approach also allows for FIT training, by incrementally lifting a constraint on the maximum order of the model. The MFn models were trained in this way, whereby n is the maximum order and the training of model MFn is initialised by the trained model MF$(n-1)$. The MFx model is trained directly from the initial vector codebook, without any limit on its context lengths.

The Cn models are representatives of a special type of mixed-order model that focuses on sequences of distinct states, while disregarding any state repetitions in this sequence [6]. The context order n refers to the number of distinct states in the state history that influences the choice of next state. This allows the high-order contexts to operate closer to the phone-level language structure instead of being tied up doing duration modelling. The Cn models are trained incrementally as well.

4. NIST 1996 RESULTS

Table 1 and Figure 1 summarise the results obtained on the 30s CALLFRIEND segments of the NIST 1996 test set. As expected, performance increases with increasing model order. The mixed-order models MFn achieve similar performance to their fixed-order Fn counterparts, while training faster and ending up smaller. The advantage of incremental (FIT) training can be seen from the fact that MF3 outperformed the directly trained MFx model. When the extra English segments from non-CALLFRIEND corpora are added to the evaluation set, the classification errors drop uniformly, as can be seen in Figure 2.

The best model was C3, a third-context-order HMM that describes sequences containing three distinct states. Its perfor-

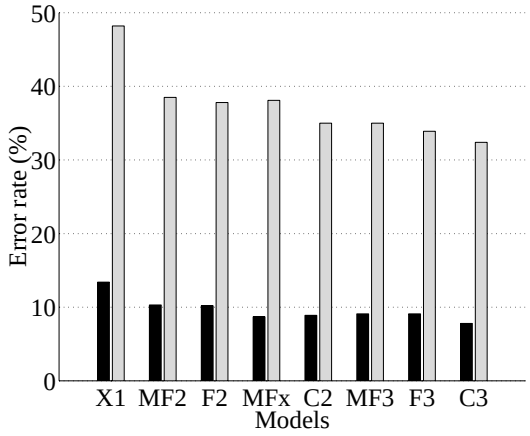

Figure 1: Average error rates on the 30s CALLFRIEND segments of the NIST 1996 evaluation set. The left-hand bars represent pairwise classification, and the right-hand bars 12-way classification.

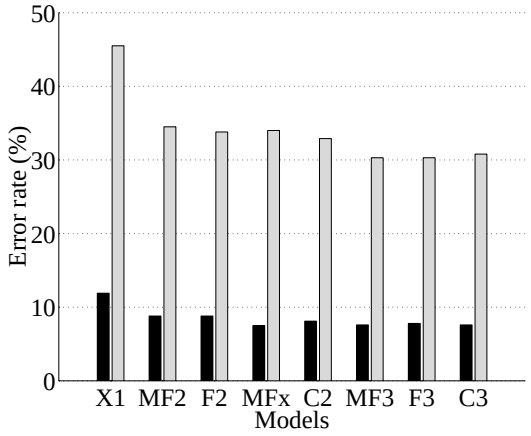

Figure 2: Average error rates on all 30s segments in the NIST 1996 evaluation set. The left-hand bars represent pairwise classification, and the right-hand bars 12-way classification.

Classification	3s	10s	30s
12-way	68.4	52.0	32.4
pairwise	24.2	14.3	7.8

Table 2: Average error rates of best model (C3) on CALLFRIEND segments of NIST 1996 evaluation set.

mance on all the CALLFRIEND segments of the NIST 1996 set is shown in Table 2. This compares favourably to results reported for an optimised phone-based system [9], which achieved 65.2% on the 3s segments, 46.6% on the 10s sections and 25.7% on the 30s segments for 12-way classification. The performance of the C3 model is on par with systems that participated in the official NIST 1996 evaluation, thereby vindicating the use of direct HMM modelling for language identification.

The performance of the Cn models emphasises the advantage of longer distinct contexts. In previous tests [8] the C3 model failed to improve on F3. Its current superiority confirm that the previous test suffered from a lack of training data. This illustrates the considerable data requirements of data-driven high-order methods.

When the full NIST 1996 set is used for testing, however, all the third-order models (MF3, F3 and C3) show very little difference in accuracy. This probably reflects differences between the statistics of CALLFRIEND versus the other corpora (we only trained on CALLFRIEND data).

5. FUTURE IMPROVEMENTS

The high-order HMM methods described so far are still considered to be in its early stages of development, and several improvements are immediately apparent. These modifications have the potential to bring these algorithms on par with the state-of-the-art in language identification today.

- It is straightforward to include a Gaussian back-end classifier, which treats the combined likelihood scores of the language models as a feature vector in itself. This captures correlations between the scores of the different language models on segments of a specific language, and transforms the separate classifiers into a mixture of experts. Its inclusion has resulted in an error rate reduction of about a factor two in other LID systems [9].

- The shared alphabet of 32 densities can be expanded to allow finer-grained acoustic modelling, as natural languages typically contain more than 32 identifiable phones.

- The explicit removal of the acoustic contribution to the likelihood scores during testing was found to be useful in a previous study [3]. This improves the rejection of adverse channel effects. The final score for each language therefore only depends on the transition probabilities. The effect on classification accuracy should be investigated.

- The addition of explicit duration modelling, as in the CnDm models of [8], is expected to increase classification accuracy. This complements the exclusive context modelling of the Cn models by concentrating on repetitions of the same state. The combination of context and duration modelling has been shown to outperform the separate approaches [8].

- As pointed out in previous work [11], the direct phonotactic modelling of codebook classes or "soft" symbols is problematic due to the short length of these units (32ms in this study). This demands large context lengths and model orders to capture the phone-level statistical structure associated with the specific language, making the algorithms data hungry. By concentrating on modelling distinct symbols as in the Cn models,

this problem can be reduced.

- The context tree framework of the mixed-order HMMs allows for simple discriminative training, which concentrates on those contexts that differentiate between languages. This will allow the modelling of longer, more language-specific contexts.

- Further improvements to the mixed-order training procedure include Baum-Welch-based reestimation to take account of symbol ambiguity, explicit context/duration modelling and improved encoding schemes. All these changes are expected to increase classification accuracy.

6. CONCLUSIONS

The LID results on the NIST 1996 evaluation set indicate that direct HMM methods are useful for language identification, in contrast with previous reports [13]. These methods do not require transcribed speech for training and are much simpler to implement than phone-based approaches. The SECT algorithm decreases training times and model sizes of high-order HMMs. Although mixed-order SECT models still show modest classification performance compared to fixed-order methods, it is expected to become more powerful as longer contexts are modelled and training data becomes scarcer. Several improvements are apparent, which promise to make mixed-order HMM classifiers even more competitive with phone-based LID systems.

7. REFERENCES

1. Yan, Y. and Barnard, E., "Experiments for an approach to language identification with conversational telephone speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, USA, May 1996, vol. 2, pp. 789–792.

2. Zissman, M. A., "Language identification using phoneme recognition and phonotactic language modelling," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, USA, 1995, pp. 3503–3506.

3. Mendoza, S., Gillick, L., Ito, Y., Lowe, S., and Newman, M., "Automatic language identification using large vocabulary continuous speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, USA, 1996, pp. 785–788.

4. Deller, J. R., Proakis, J. G., and Hansen, J. H. L., *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, Englewood Cliffs, New Jersey, USA, 1993.

5. Du Preez, J. A., *Efficient high-order hidden Markov modelling*, Ph.D. thesis, University of Stellenbosch, Stellenbosch, South Africa, 1998.

6. Du Preez, J. A. and Weber, D. M., "Automatic language recognition using high-order HMMs," in *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998, vol. 2, pp. 117–120.

7. Schwardt, L. C. and Du Preez, J. A., "Efficient mixed-order hidden Markov model inference," in *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

8. Du Preez, J. A. and Weber, D. M., "Efficient high-order hidden Markov modelling," in *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998, vol. 7, pp. 2911–2914.

9. Zissman, M. A., "Predicting, diagnosing and improving automatic language identification performance," in *Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997, vol. 1, pp. 51–54.

10. De Marcken, C. G., *Unsupervised language acquisition*, Ph.D. thesis, MIT, 1996.

11. Harbeck, S. and Ohler, U., "Multigrams for language identification," in *Proceedings of European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 375–378.

12. Rissanen, J., *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

13. Zissman, M. A., "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1993, vol. 2, pp. 399–402.