MIXED-MEMORY MARKOV MODELS FOR AUTOMATIC LANGUAGE IDENTIFICATION

Katrin Kirchhoff, Sonia Parandekar, Jeff Bilmes

Department of Electrical Engineering
University of Washington, Seattle, WA, USA

ABSTRACT

Automatic language identification (LID) continues to play an integral part in many multilingual speech applications. The most widespread approach to LID is the phonotactic approach, which performs language classification based on the probabilities of phone sequences extracted from the test signal. These probabilities are typically computed using statistical phone n-gram models. In this paper we investigate the approximation of these standard n-gram models by mixed-memory Markov models with application to both a phone-based and an articulatory feature-based LID system. We demonstrate significant improvements in accuracy with a substantially reduced set of parameters on a 10-way language identification task.

1. INTRODUCTION

Automatic Language Identification (LID) continues to play an integral role in many multilingual speech-based system. Various approaches to LID have been proposed in the past, ranging from simple distance measures applied to acoustic feature vectors to integrated LID and large-vocabulary speech recognition (LVCSR). The most widespread technique is the phone-based approach [1], which classifies languages based on the statistical characteristics of their phone sequences. More recently we have developed an alternative approach based on multiple sequences of articulatory features.

1.1. Phone-based Language Identification

Phone-based systems typically consist of a phone recognition front end which extracts a sequence of phone symbols from the acoustic signal, followed by a set of language-specific phone n-gram models, one for each language in the system. The n-gram models compute the probability of the phone sequence given the language. The model obtaining the highest score identifies the language in question. Formally this can be expressed as

$$L^* \rightarrow argmax_L P(\phi_1, \phi_2, ..., \phi_N | L) \qquad (1)$$

where L is a language and $\phi_1, \phi_2, ..., \phi_N$ is a phone sequence of length N. The statistical phone n-gram model approximates the probability of the phone sequence as follows:

$$P(\phi_1, \phi_2, ..., \phi_N) = \prod_{i=n}^{N} P(\phi_i | \phi_{i-1}, ..., \phi_{i-n+1}) \qquad (2)$$

Here, n is the order of the n-gram model. Typically, an order of 2 or 3 is used. The phone recognition front end may contain either a global set of acoustic models, or it may consist of a set of recognizers, each of which uses a different (e.g. language-specific) set of acoustic models. Phone-based language identification systems have the advantage of easy training and scoring procedures in comparison to e.g. integrated LID and LVCSR. However, they suffer from certain drawbacks: first, their performance on very short test signals (3 seconds or less) is often unsatisfactory, presumably because the time span is too short to provide a reliable phone n-gram context. Second, problems may arise from previously unseen phones and phone combinations when porting a phone-based LID system to new languages. For these reasons we have developed an alternative approach to LID, which is based on units below the phone level, viz. articulatory features.

1.2. Feature-based Language Identification

The articulatory-feature based approach [2], uses not just a single but multiple sources of information. Instead of extracting a phone sequence from the acoustic signal, a feature-based system extracts multiple parallel sequences of articulatory features and then scores each sequence with a separate feature n-gram model. We use articulatory features belonging to five different categories: *manner* of articulation, *consonantal place* of articulation, *vowel place* of articulation, *front-back* tongue position and lip *rounding*. Acoustic models are trained for each individual feature and are assigned to separate recognition networks for the different feature streams. N-gram modeling of a single feature sequence is performed analogous to phone n-gram modeling, i.e.

$$P(f_1, f_2, ..., f_N) = \prod_{i=n}^{N} P(f_i|f_{i-1}, ..., f_{i-n+1}) \qquad (3)$$

where $f_1, f_2, ..., f_N$ is a sequence of feature symbols of length N produced by an acoustic feature recognition front end. The individual scores from all K feature streams are subsequently combined by some combination function to produce the overall LID score for a given language. In our baseline system we use a simple product as a combination function:

$$P(\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_K|L) = \prod_{k=1}^{K} P(\mathcal{F}_k|L) \qquad (4)$$

The final combined score is used in the decision rule to identify the most probable language:

$$L^* = argmax_L P(\mathcal{F}_1, ..., \mathcal{F}_K|L) \qquad (5)$$

There are several advantages to the feature-based approach: first, the number of articulatory features needed to encode sounds is typically smaller than the set of phones. Since features are shared across phones, more training material is available for them, which means that both the acoustic models and the n-gram models can be trained more robustly. Second, the number of potential feature n-gram contexts is smaller, which reduces the possibility of encountering unseen contexts when porting the system to new languages. Third, the feature-based model provides an easy way of modeling language differences arising from subtle articulatory timing, such as aspiration, vowel nasalization, etc. without having to enlarge the set of basic units in the system.

2. MIXED-MEMORY MARKOV MODELS

Mixed-memory Markov models (MMMs) were proposed by Saul and Jordan [3, 4] for the analysis of time series. The basic idea is to reduce Markov models with large state spaces to a combination of simpler Markov models with smaller state spaces. This done by representing the transition matrix of the larger model as a mixture of the transition matrices of smaller models: let S be a random variable with i possible values. An $n'th$ order Markov model over S is specified by the transition matrix

$$P(s_t|s_{t-1}, s_{t-2}, ..., s_{t-n}) \qquad (6)$$

In a mixed-memory Markov model, the transition matrix is decomposed as follows:

$$P(s_t|s_{t-1}, s_{t-2}, ..., s_{t-n}) = \sum_{\mu=1}^{n} \phi(\mu) a^{\mu}(s_t|s_{t-\mu}) \qquad (7)$$

where $a^{\mu}(s_t|s_{t-\mu})$ is a $i \times i$ transition matrix modeling the probability of the model state at time t given its value at state $t - \mu$. Naturally, $0 \leq \phi(\mu) \leq 1$ for all μ and $\sum_{\mu} \phi(\mu) = 1$. The term $\phi(\mu)$ can be viewed as the probability that a hidden variable X assumes value μ at time t, i.e. the prior probability $P(x_t = \mu)$. Marginalizing over all possible assignments to X yields the model in Equation 7. The relative contributions of the individual transition matrices are thus controlled by a hidden variable representing the mixture coefficients, whose values can be estimated using the EM algorithm.

When the mixed-memory model is applied to a single time series, the states $t-\mu$ are states μ steps into the past. In the case of coupled time series, where the observations can be described by a vector of values, the decomposition applies to the transition matrices between the individual components of the current and the previous vector(s). Suppose that there are m different random variables, each of which is associated with its own observation sequence and can take on i values. Let us assume further that the m different components of the observation vector V at time t, $v_t^1, ..., v_t^m$, are conditionally independent given the previous observation vector V_{t-1}, i.e.

$$P(V_t|V_{t-1}) = \prod_{j=1}^{m} P(v_t^j|V_{t-1}) \qquad (8)$$

The probability of an individual vector component given the previous complete observation vector can then be approximated by a mixture of transition probabilities between individual components:

$$P(v_t^j|V_{t-1}) = \sum_{l=1}^{m} \phi^j(l) a^{jl}(v_t^j|v_{t-1}^l) \qquad (9)$$

Here the $a^{lj}(v_t^j|v_{t-1}^l)$ are individual $i \times i$ transition matrices between two vector components l at time $t-1$ and and j at time t.

The full-memory model specified in Equation 6 requires $O(n^{i+1})$ parameters, whereas the model in Equation 7 only requires $O(ni^2)$ parameters. In the case of multiple Markov chains, the complexity is reduced from $O(i^{2m})$ to $(O(m^2i^2)$. MMMs are thus particularly suitable for modeling complex dynamic processes where the addition of conditioning variables increases the state space of the model exponentially. This includes processes where the current model state is conditioned on a large number of previous states in time as well as those processes where a given state is conditioned on a large number of simultaneous states in space. Although the mixed-memory model cannot model the complete set of conditional dependencies represented in the full-memory model, it may be advantageous in cases where the number of parameters in the full-memory model greatly exceeds the available training material or where the process is truly composed of different sources. Previous applications of mixed-memory Markov models to speech and language processing include large-vocabulary statistical language modeling [3] and acoustic modeling for speech recognition using frequency subbands [5]. In our case, MMMs can be employed in two different ways. In both the phone-based and the feature-based system, they can be used to approximate the full-memory n-gram models for the phone or feature sequences. In the feature-based system, they can additionally be used to model dependencies between different feature streams. In this paper, we focus exclusively on the first application in order to determine if the mixed-memory approximation to standard n-gram models is advantageous when n-gram models are used as classifiers.

3. CORPUS AND BASELINE SYSTEMS

Our experiments were performed on the OGI-TS multilingual telephone speech database [6]. We use the training, development and evaluation set definitions proposed in [7], which include data from ten different languages (English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese). The training, development and evaluation sets are comprised of 4650, 1898 and 1848 utterances, respectively. We have grouped the test signals into three different categories according to their length, viz. very short ($<$ 3s), short (3-15s) and long ($>$ 15s).

We have developed both a phone-based and a feature-based baseline system, which were first described in [2]. For both systems, language-independent acoustic front-ends were used, i.e. a generic set of acoustic models was used in each system to extract symbol sequences from the speech signal. The acoustic models are HMMs with a variable number of states (tuned to model the average duration of each unit) and multiple Gaussian mixture components. The phone-based system contains 133 context-independent phone models with 1089 states in total and has 2 mixture components per state. The n-gram modeling component consists of language-specific bigram models with Witten-Bell smoothing. The feature-based system has 68 models and a total of 761 states with two Gaussian mixture components each. Prior to n-gram model training, all instances of manner and rounding features were relabeled as either short or long, depending on their length in terms of the number of acoustic frames. This increased the number of units for the purpose of n-gram modeling to 96 (consonantal place: 18, manner: 32, vowel place and front-back: 12, rounding: 20); however, no additional acoustic models are created for the different duration categories. The sequence of feature symbols in each group is modeled by a 4-gram with Witten-Bell smoothing. The combination of individual feature-stream scores is performed by Equation 4. The baseline accuracy rates for both the phone-based and the feature-based system are shown in Table 1. The overall performance of both systems is roughly equal; however, the feature-based system shows a markedly better performance on very short signals whereas the phone-based system is superior on long signals.

	very short	short	long	average
phone-dev	35.6	57.4	65.2	52.7
feature-dev	42.1	55.7	63.9	53.0
phone-eval	30.6	53.1	69.7	48.8
feature-eval	44.8	57.8	60.0	54.6

Table 1. LID accuracy (in %) of the baseline feature and phone-based systems on the development (dev) and evaluation (eval) sets, broken down by test signal length.

4. EXPERIMENTS WITH MIXED-MEMORY MARKOV MODELS

The application of MMMs to standard phone-based n-gram models is quite straightforward: Equation 7 can be used to approximate a higher-order n-gram, such as a 3-gram or a 4-gram, by a mixture of bigrams. For the implementation of the mixed-memory systems we used the Graphical Models Toolkit GMTK [8]. This toolkit provides ways of specifying a wide range of graphical models for the purpose of speech recognition as well as a general inference mechanism to maximize the likelihood of the model given the data. A MMM is a special case of a graphical model, where the weights for each transition matrix are encoded as hidden variables and the variables in the Markov chains are observed. For encoding the hidden variables we used the switching-parents feature of GMTK, which greatly facilitates specifying of mixtures of probability distributions and the data-driven learning of mixture weights using Expectation-Maximization. In order to incorporate standard n-gram smoothing methods, the basic n-gram probabilities were computed using the CMU-Cambridge Language Modeling toolkit [9]. The GMTK toolkit was subsequently used to estimate the parameters of the hidden variables, i.e. the mixing coefficients.

Table 2 shows the comparison between phone N-grams and their mixed-memory equivalents, e.g. a trigram compared to a mixture of two bigrams. We can see that in all cases the accuracy of the MMM is superior to that of the N-gram. The best-performing system, MMM-2, was then applied to the evaluation set. Table 3 shows results on both sets for different signal lengths. The improvement of the overall accuracy on the evaluation set compared to the baseline system (Table 1) is statistically significant (significance level of 0.05 using a difference of proportions significance test). It is particularly noteworthy that performance improves for very short signals whereas it declines

	2	3	4	5
N-gram	52.7	52.7	47.9	47.6
MMM	-	54.1	53.6	52.7

Table 2. LID results (accuracy in %, dev set) for standard phone n-grams and phone MMMs for orders ranging from 2 to 5. For order 2, the mixed-memory model is equivalent to a standard bigram.

	very short	short	long	average
dev	38.8	58.2	64.6	54.1
eval	35.7	56.2	65.7	51.8

Table 3. LID accuracy (in %) on evaluation and development sets of the best phone MMM, broken down by signal length.

slightly on long signals. Next we applied MMMs to each feature stream in the feature-based system, approximating the feature 4-gram models with mixtures of four bigrams. In order to determine the effect on different feature streams we measured the performance on the development set independently for each stream. We found that an increase in accuracy was only obtained for the manner stream, which is the stream with the largest set of models (34). All other streams showed slight losses in accuracy. A combination of the mixed-memory manner stream model with the standard 4-gram models for the remaining streams increase the accuracy on the development set to 54.1%. Table 4 shows the detailed results of this system. Again we see a marked improvement on very short test signals, albeit a slight decrease in overall accuracy on the evaluation set.

	very short	short	long	average
dev	48.6	54.9	63.0	54.1
eval	47.7	55.3	60.0	53.7

Table 4. Accuracy (in %) on development and evaluation sets for a combined MMM (manner stream) and N-gram (other streams) feature-based model.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated the application of mixed-memory Markov models to automatic language identification. MMMs were used to approximate standard statistical n-gram models ($n > 2$) by a mixture of bigram models. This was done both for a phone-based and an articulatory feature-based system. It was found that the mixed-memory approximation improves accuracy when the model set is large compared to the available training data, as in the phone-based system. By contrast, n-grams involving smaller symbol sets, like most of the feature n-gram models, benefit from a full-memory model. On our present LID task, MMMs have shown the largest improvements on very short test signals, which indicates that the mixed-memory approximation may be a useful technique for real-world, real-time LID applications. In the future we intend to investigate the use of MMMs to model dependencies between different streams in the feature-based system.

Acknowledgements
The work presented in this paper was funded by US DoD.

6. REFERENCES

[1] M.A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech and Audio Processing*, vol. 4(1), pp. 31–44, 1996.

[2] K. Kirchhoff and S. Parandekar, “Multi-stream statistical language modeling with application to automatic language identification,” in *Proceedings of Eurospeech-01*, 2001.

[3] L. Saul and F. Pereira, “Aggregate and mixed-order markov models for statistical language processing,” in *Proceedings of the 2nd Conference on Empirical Methods in NLP*, Somerset, NJ, 1997, pp. 81–89.

[4] L. Saul and M. Jordan, “Mixed memory Markov models,” *Machine Learning*, vol. 37, pp. 37–87, 1999.

[5] H.J. Nock and S.J. Young, “Loosely coupled HMMs for ASR,” in *Proceedings of ICSLP-00*, 2000.

[6] Y.K. Muthusamy et al., “The OGI multi-language telephone speech corpus,” in *Proceedings of ICSLP-92*, 1992.

[7] Y.K. Muthusamy, *A Segmental Approach to Automatic Language Identification*, Ph.D. thesis, Oregon Graduate Institute, 1993.

[8] J. Bilmes & G. Zweig, “The Graphical Models Toolkit: An open source software system for speech and time-series processing,” in submission.

[9] P. Clarkson and R. Rosenfeld, “Statistical language modeling using the CMU-Cambridge toolkit,” in *Proceedings Eurospeech-97*, Rhodes, Greece, 1997.