AUTOMATIC LANGUAGE IDENTIFICATION USING ERGODIC-HMM

S. A. SantoshKumar *V. Ramasubramanian*

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore 560 012, India
vram@ece.iisc.ernet.in

ABSTRACT

Recently, we established the equivalence of an ergodic HMM (EHMM) to a parallel sub-word recognition (PSWR) framework for language identification (LID). The states of EHMM correspond to acoustic units of a language and its state-transitions represent the bigram language model of unit sequences. We consider two alternatives to represent the state-observation densities of EHMM, namely, the Gaussian mixture model (GMM) and hidden Markov model (HMM). We present a segmental K-means algorithm for the training of both these types of EHMM (EHMM of GMMs and EHMM of HMMs) and compare their performances on a 6 language LID task in the OGI-TS database. EHMM of GMMs has a performance comparable to PSWR and superior than EHMM of HMMs; we provide reasons for the performance difference between EHMM(G) and EHMM(H), and identify ways of enhancing the performance of EHMM(H) which is a novel and powerful architecture, ideal for spoken language modeling.

1. INTRODUCTION

Automatic language identification (LID) has become an important research problem over the last decade with several promising solutions [1], [2]. One of the earliest work in LID by House and Neuberg [3] was based on the now popular hidden Markov model (HMM); here, they exploited the potential of the discrete ergodic HMM to model sequential characteristics of broad phonetic labels derived from texts of different languages. Following this, there have been a few other attempts to use HMMs for LID [4], [5]. However, the moderate results of these work only raised doubts on the modeling capability of HMMs [1]; in general, it was concluded that multi-state HMMs cannot perform any better than static models like GMM [4], [1]. However, more recently [6], we established the equivalence of the 'parallel sub word recognition' framework to an ergodic HMM (EHMM) along with clear experimental validation which showed that ergodic HMMs can offer as good a performance as PSWR, which is essentially a sub-word unit based 'parallel phone recognition' (PPR) system – one of the popular phone recognition frameworks for LID till date [7], [1], [2]. The equivalence is based on the following correspondences between PSWR and EHMM:

1. The states (observation densities) of the EHMM correspond to sub-word units (with associated sub-word HMMs) which constitute the front-end sub-word recognizer (SWR) in PSWR.

2. The state-transition probabilities of EHMM represent the bigram statistics of sub-word units in sequences obtained by the front-end SWR decoding. This is equivalent to the bigram back-end language model (LM) of PSWR, which models the phonotactics of a language.

3. Evaluation of the Viterbi likelihood by EHMM exactly corresponds to the joint-decoding in PSWR using both the front-end SWR and back-end (bigram) LM.

In this paper, we deal with two types of EHMM, based on the type of model used to represent the state observation densities of the EHMM, namely, Gaussian mixture model (GMM) and hidden Markov model (HMM). We refer to the EHMM with GMM observation density as EHMM(G) and the EHMM with HMM observation density as EHMM(H). We present a segmental K-means algorithm for training of both these types of EHMM (EHMM(G) and EHMM(H)). We report their performances on a 6 language LID task in the Oregon Graduate Institute – Telephone Speech (OGI-TS) database [7].

EHMM(G) is equivalent to a PSWR system, where an acoustic sub-word unit (SWU) is modeled by a GMM. Likewise, EHMM(H) corresponds to the case when a SWU in PSWR is modeled by a HMM. Interestingly, the SWU GMM of EHMM(G) proves to be a more appropriate model of the SWUs of a language (states of EHMM(G)); the SWU GMM, despite being a static model of the SWU, enjoys the advantage of being insensitive to context-dependencies of the SWUs, thereby generalizing to acoustic segments of all possible contexts. In contrast, in the case of EHMM of HMMs, a sub-word unit HMM (state of EHMM(H)), being a temporal model of a SWU, becomes specialized to specific contexts and suffers from poor context generalizability to other contexts of the SWUs as may occur in unseen data. Since EHMM(H) is a novel and powerful architecture in the theory of HMMs, ideally suited to spoken language modeling, we identify ways of enhancing its performance.

2. ERGODIC-HMM (EHMM) BASED LID

Fig. 1 shows a typical EHMM based LID system for N languages. An N - language LID task is to classify an input speech utterance (of any speaker and any text), as belonging to one of N languages $\mathcal{L}_1, \ldots, \mathcal{L}_N$. The EHMM system has N paths for a N language LID task. A path 'i' ($i = 1, \ldots, N$), has an EHMM $\mathcal{E}_i$ of language $\mathcal{L}_i$. For a given input utterance, EHMM yields N 'Viterbi likelihood' scores ($\mathbf{V}_i$ in Fig. 1), one for each language $\mathcal{L}_i$, obtained by a Viterbi decoding of the input utterance $\mathbf{O}$ by the EHMM $\mathcal{E}_i$ of language $\mathcal{L}_i$. The maximum - likelihood (ML) classifier identifies the language of the input utterance as $\mathcal{L}_{i^*}$ which has the highest likelihood (score) $\mathbf{V}_i$, i.e., $i^* = \arg\max_{i=1,\ldots,N} \mathbf{V}_i$.

3. EHMM PARAMETERS

An M-state ergodic HMM $\mathcal{E}_i$ of language $\mathcal{L}_i$ is specified as $\mathcal{E}_i = (A_i, B_i, \pi_i)$; we refer to this as the 'primary HMM'. These three parameters of $\mathcal{E}_i$ are as follows (Fig. 2):

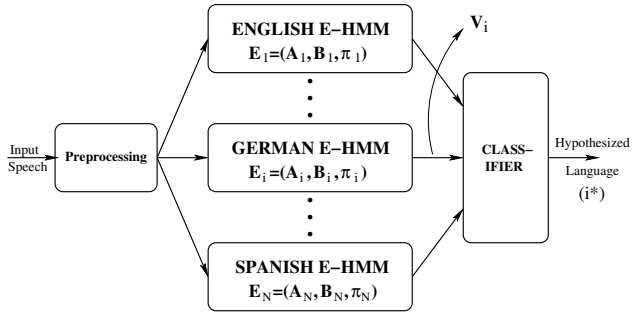

Fig. 1. *LID by ergodic HMM*

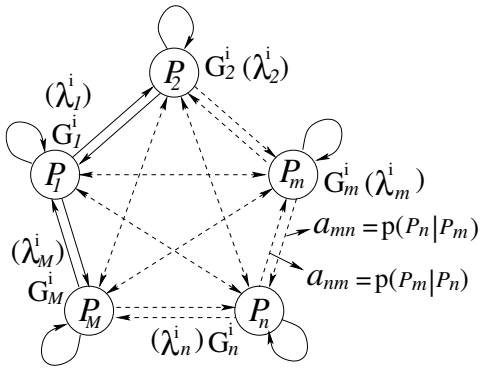

Fig. 2. *Ergodic-HMM of sub-word GMMs / sub-word HMMs*

1. Observation density B_i: B_i is the set of M observation densities of the $\mathcal{E}_i$.

- In the case of EHMM(G), each state of the primary HMM is modeled by GMMs and is given by $B_i = \{b_m^i(\mathbf{o}_t)\}_{m=1}^M = (G_1^i, G_2^i, \ldots, G_M^i)$. The GMM of state m, G_m^i is given by $G_m^i = (c_{ml}, \mu_{ml}, \Sigma_{ml})_{l=1}^L$ where L = 9, i.e., each G_m^i has 9 mixtures.
- In the case of EHMM(H), each state of the primary HMM is modeled by a secondary HMM and is given by $B_i = \{b_m^i(\mathbf{o})\}_{m=1}^M = (\lambda_1^i, \lambda_2^i, \ldots, \lambda_M^i)$. A secondary HMM, λ_m^i, of state m is typically a 3 state left-to-right HMM with 3 mixture Gaussian per state.

2. Transition matrix A_i: $A_i = \{a_{mn}\}, m, n = 1, \ldots, M$ specifies the transition probabilities a_{mn} for a sub-word unit modeled by λ_m^i (or G_m^i) to transit to another unit modeled by λ_n^i (or G_n^i).

3. Initial state distribution π_i : $\pi_i = \{\pi_{im}\}, m = 1, \ldots, M$ specifies π_{im} – the probability that sub-word unit λ_m^i (or G_m^i) is the starting state.

In the case of EHMM(H), a state m in $\mathcal{E}_i$ is characterized by an observation density $b_m^i(\mathbf{o}) = p(\mathbf{o}|\lambda_m^i)$, which is a 'segmental' density in that, it yields the likelihood of a segment $\mathbf{o}$ given the sub-word HMM λ_m^i. Thus $\mathcal{E}_i$ is an HMM of HMMs; i.e., each state of EHMM is itself another HMM. In the case of EHMM(G), each state corresponds to a frame (a feature vector $\mathbf{o}_t$). (A_i, π_i) model the 'phonotactics' of language $\mathcal{L}_i$. Such an ergodic HMM is ideally suited to model a language at two levels: phonotactics is modeled by A_i which governs the sequence of SWUs (states) that can be realized and the acoustic manifestation of each sub-word so realized is modeled through B_i.

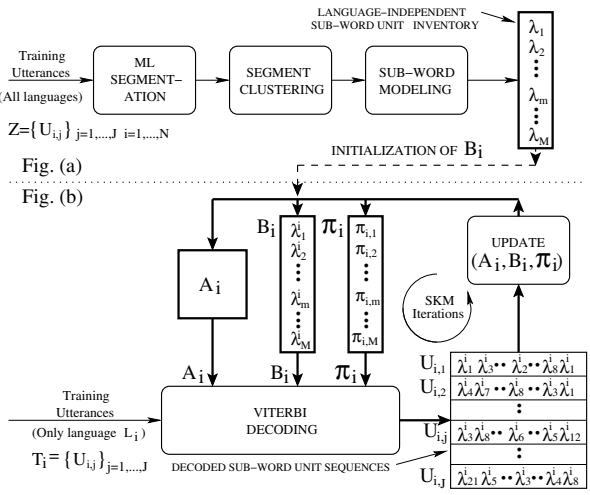

Fig. 3. *EHMM training by SKM: (a) Generation of language-independent sub-word unit inventory, (b) SKM iterations for EHMM parameter estimation of language $\mathcal{L}_i$.*

3.1. Viterbi likelihood

Given an input utterance $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, the ergodic HMM $\mathcal{E}_i$ of language $\mathcal{L}_i$ can evaluate the Viterbi likelihood (score) $\mathbf{V}_i$ as,

$$\begin{aligned} \mathbf{V}_i &= P^*(\mathbf{O}, \mathbf{q}|\mathcal{E}_i) = \max_{\mathbf{q}} P(\mathbf{O}, \mathbf{q}|\mathcal{E}_i) \\ &= \max_{\mathbf{q}} \{P(\mathbf{O}|\mathbf{q}, \mathcal{E}_i) \cdot P(\mathbf{q}|\mathcal{E}_i)\} \end{aligned} \quad (1)$$

In the case of EMM(G), Eqn. (1) is given by

$$\mathbf{V}_i = \max_{\mathbf{q}} \left\{ \pi_{iq_1} b_{q_1}^i(\mathbf{o}_1) \cdot \prod_{t=2}^{T} \left[a_{q_{t-1}q_t} \cdot b_{q_t}^i(\mathbf{o}_t) \right] \right\} \quad (2)$$

In the case of EHMM(H), Eqn. (1) is given by

$$\mathbf{V}_i = \max_{\mathbf{q},\mathbf{B},K} \left\{ \pi_{iq_1} p(\mathbf{s}_1|\lambda_{q_1}^i) \cdot \prod_{k=2}^{K} \left[a_{q_{k-1}q_k} \cdot p(\mathbf{s}_k|\lambda_{q_k}^i) \right] \right\} \quad (3)$$

where, $\mathbf{B} = (b_0, b_1, \ldots, b_K)$, with $b_0 = 0$ and $b_K = T$, are the segment boundaries which segments $\mathbf{O} = (\mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_T)$ into K segments $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$, where, segment $\mathbf{s}_k = (\mathbf{o}_{b_{k-1}+1}, \ldots, \mathbf{o}_{b_k})$. $\mathbf{q}$ is any arbitrary state sequence of $\mathcal{E}_i$ given by $\mathbf{q} = (q_1 q_2 \ldots q_{k-1} q_k \ldots q_K)$, where state $q_k \in \{1, 2, \ldots, M\}$. The corresponding observation density is $\lambda_{q_k}^i$ (drawn from $B_i = (\lambda_1^i, \lambda_2^i, \ldots, \lambda_M^i)$), which evaluates the probability of the 'observation segment' $\mathbf{s}_k$, $b_{q_k}(\mathbf{s}_k)$, as the Viterbi likelihood $p(\mathbf{s}_k|\lambda_{q_k}^i)$.

Eqn. (3) maximizes $\mathbf{V}_i$ over the variables $(\mathbf{q}, \mathbf{B}, K)$. Evaluation of $\mathbf{V}_i$ by Eqn. (2) or Eqn. (3) in EHMM optimally combines both the acoustic likelihood $P(\mathbf{O}|\mathbf{q}, \mathcal{E}_i)$ and the language model likelihood $P(\mathbf{q}|\mathcal{E}_i)$ as in the joint decoding for PSWR [6].

4. SEGMENTAL K-MEANS TRAINING OF EHMM

The parameters (A_i, B_i, π_i) of EHMM $\mathcal{E}_i$ are learnt from the training utterances $\mathcal{T}_i = \{U_{ij}\}_{j=1}^J$ of language $\mathcal{L}_i$ using a segmental K-means (SKM) algorithm. Through the SKM, we jointly optimize the state observation densities and the state-transitions (bigram language model). Fig. 3 illustrates this procedure which is as follows:

Step 1: Initialization

Set iteration count $r = 1$. Initialize $\mathcal{E}_i(r) = (A_i, B_i, \pi_i)$ with

1. A_i as equiprobable, i.e., $a_{mn} = 1/M, \forall m, n$.
2. π_i as equiprobable, i.e., $\pi_{im} = 1/M, m = 1, \ldots, M$.
3. Initialization of the state observation densities B_i:

   • **For EHMM(G):** The m state observation densities $B_i = \{G^i_m\}_{m=1}^M$ are initialized as follows:

   A vector quantization (VQ) codebook is designed using the training feature vectors of language $\mathcal{L}_i$. Let this VQ codebook be $\mathbf{C}_i = \{\mathbf{c}_{i1}, \mathbf{c}_{i2}, \ldots, \mathbf{c}_{iM}\}$. All training vectors quantized to codeword $\mathbf{c}_{im}$ are used to initialize the GMM G^i_m of state 'm' of $\mathcal{E}_i$, i.e., the parameters of G^i_m, $(c_{ml}, \mu_{ml}, \Sigma_{ml})_{l=1}^L$ are estimated from $\{\mathbf{o}_t : q(\mathbf{o}_t) = \mathbf{c}_{im}\}$ (where $q(\mathbf{o}_t) = \mathbf{c}_{im} : d(\mathbf{o}_t, \mathbf{c}_{im}) \leq d(\mathbf{o}_t, \mathbf{c}_{in}), n = 1, \ldots, M$) using the standard K-means algorithm.

   • **For EHMM(H):** $B_i = \{b^i_m(\mathbf{o})\}_{m=1}^M$ is initialized by the language-independent sub-word unit inventory, $\{b^i_m(\mathbf{o})\}_{m=1}^M = (\lambda_1, \lambda_2, \ldots, \lambda_M)$ (Fig. 3(a)) obtained as follows:
   i) **Automatic segmentation:** The training utterances (in the form of MFCC vector sequence) are segmented into acoustic segments using the maximum-likelihood (ML) segmentation technique. ii) **Segment clustering:** The resulting acoustic segments are clustered into M clusters by applying the K-means algorithm on the centroids of the acoustic segments. iii) **Segment modeling:** The segments belonging to each of the M clusters are modeled by a 3-state left-to-right HMM resulting in an inventory of M sub-word HMMs, $(\lambda_1, \lambda_2, \ldots, \lambda_M)$.

Step 2: Viterbi decoding

Given a set of training utterances of language $\mathcal{L}_i$, $\mathcal{T}_i = \{U_{ij}\}_{j=1}^J$, let $\mathbf{O} = (\mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_T)$ be the observation feature vector sequence (of T frames) of a typical utterance U_{ij}. The Viterbi decoding of utterance U_{ij} (or simply, $\mathbf{O}$) by $\mathcal{E}_i(r)$ yields the Viterbi likelihood $P^*_{ij}(r)$ given by,

$$\begin{aligned} P^*_{ij}(r) &= P^*(\mathbf{O}, \mathbf{q}|\mathcal{E}_i(r)) \\ &= \max_{\mathbf{q}} \{P(\mathbf{O}|\mathbf{q}, \mathcal{E}_i(r)) P(\mathbf{q}|\mathcal{E}_i(r))\} \end{aligned} \quad (4)$$

which is evaluated as in Eqn. (2) for EHMM(G) or Eqn. (3) for EHMM(H).

Step 3: Parameter update

Let $\mathcal{S}_i(r) = \{S_{ij}(r)\}_{j=1}^J$ be the set of optimal sub-word unit (state) sequences obtained by Viterbi decoding of the training utterances $\mathcal{T}_i = \{U_{ij}\}_{j=1}^J$ at iteration r of the SKM algorithm, i.e., $S_{ij}(r)$ is the optimal state sequence $\mathbf{q}^* = (q^*_1 q^*_2 \ldots q^*_{K*})$ obtained by Viterbi decoding of utterance U_{ij} using $\mathcal{E}_i(r)$ (the ergodic-HMM parameters at iteration (r)) as given by Eqn. (4). At iteration $(r+1)$, the parameters (A_i, B_i, π_i) of $\mathcal{E}_i(r+1)$ are updated using *both* $\mathcal{T}_i = \{U_{ij}\}_{j=1}^J$ and $\{S_{ij}(r)\}_{j=1}^J$ as follows:

• **Update of A_i and π_i:**

$$a_{mn} = \frac{n(q^*_{k-1} = m, q^*_k = n)}{n(q^*_{k-1} = m)} \quad m, n = 1, \ldots, M \quad (5)$$

$$\pi_{im} = \frac{n(q^*_1 = m)}{J} \quad m = 1, \ldots, M \quad (6)$$

where, the occurrence counts $n(.,.)$ and $n(.)$ are measured over all the J optimal state-sequences $\mathbf{q}^*$ of the J training utterances, i.e., over all the J SWU sequences $\{S_{ij}(r)\}_{j=1}^J$.

• **Update of B_i:**

∘ **For EHMM(G):** Let $\mathbf{S}_m = \{\mathbf{o}_t : q^*_t = m\}$ be the set of all observation feature vectors of $\{U_{ij}\}_{j=1}^J$ which have been assigned to state 'm' by the optimal Viterbi decoding of Eqn. (4). Update the GMM parameters $(c_{ml}, \mu_{ml}, \Sigma_{ml})_{l=1}^L$ of state 'm' as follows: Perform K-means clustering of feature vectors in $\mathbf{S}_m$ into L clusters. Let the resultant clusters be given by $\mathbf{O}^l_m = \{\mathbf{o}_t : \mathbf{o}_t \in$ cluster l of $\mathbf{S}_m\}$. The updated GMM parameters are given by :

$$\mu_{ml} = \frac{1}{n_{ml}} \sum_{\mathbf{o}_t \in \mathbf{O}^l_m} \mathbf{o}_t \quad (7)$$

$$\Sigma_{ml} = \frac{1}{n_{ml}} \sum_{\mathbf{o}_t \in \mathbf{O}^l_m} [\mathbf{o}_t - \mu_{ml}][\mathbf{o}_t - \mu_{ml}]' \quad (8)$$

$$c_{ml} = \frac{n_{ml}}{n_m} \quad (9)$$

where, $1 \leq l \leq L$; n_{ml} is the number of observation vectors in cluster $\mathbf{O}^l_m$ and n_m is the total number of observation vectors in $\mathbf{S}_m$.

∘ **For EHMM(H):** Let $\mathbf{S}_m = \{\mathbf{s}_k : q^*_k = m\}$ be the set of all segments of $\{U_{ij}\}_{j=1}^J$ which have been assigned to state m (sub-word unit model λ^i_m) by the optimal Viterbi decoding of Eqn. (4). Update sub-word unit model λ^i_m using the segments in $\mathbf{S}_m$, i.e., build a new HMM λ^i_m from these segments:

$$\lambda^i_m = HMM\left(\mathbf{S}_m = \{\mathbf{s}_k : q^*_k = m\}\right), \; m = 1, \ldots, M \quad (10)$$

Step 4: Convergence

$P^*_i(r) = \frac{1}{J} \sum_{j=1}^J P^*_{ij}(r)/T_j$ is the average Viterbi likelihood over all the J training utterances of language $\mathcal{L}_i$, after Eqn. (4) of iteration r. T_j is the number of frames in observation vector sequence of utterance U_{ij} and $P^*_{ij}(r)$ is the Viterbi likelihood as per Eqn. (4) using $\mathcal{E}_i(r)$ at iteration r.

Terminate SKM iteration if $|P^*_i(r) - P^*_i(r-1)| < \epsilon$; otherwise continue with 'Step 2: Viterbi decoding' with $r = r + 1$. ϵ is a suitable threshold to ensure a good convergence.

5. EXPERIMENTS AND RESULTS

We present here experimental results of LID performance using EHMM(G) and EHMM(H) systems and compare the results [1].

5.1. Database

EHMM(G) and EHMM(H) are evaluated on 6 languages of the OGI-TS corpus [7] - English, German, Hindi, Japanese, Spanish and Mandarin. The EHMM(G) and EHMM(H) systems are trained on 50 'story-bt'(story-before-the-tone) utterances per language spoken by 50 different speakers. Both the systems are tested using 20 'story-bt' utterances per language outside the training data; the training and test utterances are each 45 seconds long. Both the systems use a 26-dimensional parameter vector of 12 MFCC, 12 delta-MFCC, energy and delta energy.

[1] We thank P. Srinivas and H. V. Sharada for their code and data.

5.2. Parameters of EHMM(G)/(H) systems

In EHMM(G), the main parameters are the number of states 'M' and the number of Gaussian mixtures/state 'L'. For each language, EHMM(G) system is designed for M = 8, 16, 32, 64 with L = 9 for each M.

In EHMM(H) system, the main parameters are the number of states of primary HMM 'M', the number of states of secondary HMM and the number of Gaussian mixture/state of secondary HMM. In our experiment, for each language, EHMM(H) systems were designed for M = 8, 16, 32, 64, and each state is modeled by 3 state left to right HMM with 3 Gaussian mixture/state.

5.3. Results

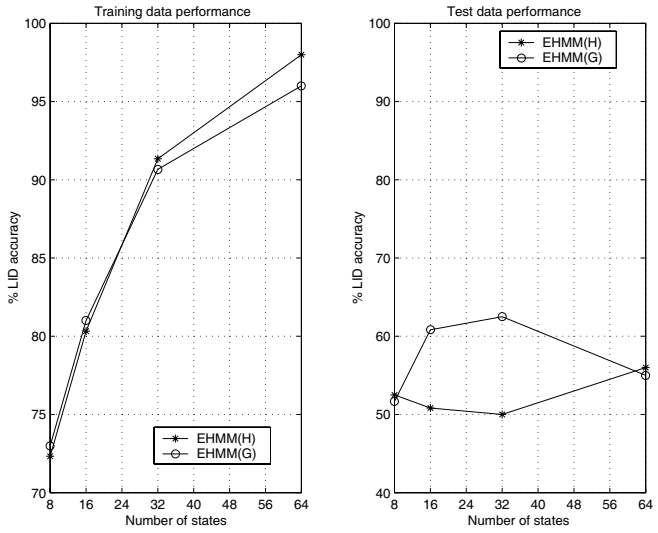

Fig. 4. *LID performance of EHMM(G) and EHMM(H)*

Fig. 4 shows the % LID accuracy of EHMM(G) and EHMM(H) for training and test data on the 6-language task in OGI-TS database for number of states 'M' ranging from 8 to 64. The following can be observed from this figure:

The training data performance shows the potential of EHMM(G) and EHMM(H) to achieve very high LID classification accuracy. The recognition performance increases significantly with increase in the number of states M for both systems. Both EHMM(G) and EHMM(H) have comparable performance on training data (96% for EHMM(G) and 98% for EHMM(H) for M = 64). On test data, the best performance for EHMM(G) is 62.5% for M=32 and that of EHMM(H) is 55.83% for M = 64. EHMM(G) has a performance comparable to that of the PSWR system [6] (which is an acoustic sub-word equivalent of the PPR system [7]), which has an LID accuracy of 98.3% on training data and 65% on test data with a sub-word unit inventory size of 50 (equivalent to the number of states M in EHMM here).

The above results show that EHMM(G) has better (or comparable) performance to EHMM(H). An important factor contributing to this performance difference between EHMM(G) and EHMM(H) is the extent to which their state observation densities (GMM or HMM) can generalize to various contexts of the acoustic segments associated with a state of the respective EHMM. A secondary HMM associated with each state in EHMM(H) is left-to-right, which allows for modeling temporal dynamics of the segments of a state (sub-word unit). However, it can atmost represent optimally only one kind of left-context and right-context of the various segments associated with a state. Thus, it fails to generalize to acoustic segments of other contexts.

A GMM model of a SWU (state) in EHMM(G) is only a static model of the observation vectors belonging to that state; while this may appear as a poor modeling of the dynamics of the acoustic segments spanning the state, it provides the incidental advantage of being able to give high likelihoods equally well to segments of various contexts; i.e., the GMM is not fine tuned to any one particular context, and generalizes to other contexts; whereas, the left-to-right secondary HMM of a state becomes fine tuned to a particular context and fails to adequately model other contexts.

Therefore, considering means of enabling the secondary HMM to model context dependencies may be one way to improve the performance of EHMM(H). One possible way is to have context-dependent secondary HMMs as sub-states of a state of EHMM(H); another possibility is to consider the use of ergodic HMMs as the secondary HMMs (rather than left-to-right HMMs) which allow for entry and exit from any state thereby providing means of context-dependent modeling. The success of such an ergodic-HMM to model context-dependency can have implications to similar context-dependent acoustic modeling in large vocabulary continuous speech recognition (in the place of now commonly used left-to-right triphone models). Moreover, we also note that the training of EHMM(G) and EHMM(H) with Baum-Welch (BW) algorithm can also be expected to improve their generalizability.

6. CONCLUSIONS

We have proposed two types of ergodic HMMs (EHMM) for automatic spoken language identification, namely, EHMM(G) and EHMM(H), based on the modeling of the state observation density, either by a GMM or an HMM, respectively. We have presented a segmental K-means algorithm for the training of both these types of EHMM and compared their performance on a 6-language LID task using the OGI-TS database. We have provided reasons for the performance difference between EHMM(G) and EHMM(H), and identified ways of enhancing the performance of EHMM(H).

7. REFERENCES

[1] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, Oct 1994.

[2] M. A. Zissman and K. M. Berkling. Automatic language identification. *Speech Commun.*, 35(1-2):115–124, Aug 2001.

[3] A. S. House and E. P. Neuberg. Toward automatic identification of the language of an utterance. *Journal of Acoustic Society of America*, 62(3):708–713, Sep 1977.

[4] M. A. Zissman. Automatic language identification using Gaussian mixture and hidden Markov models. In *Proc. ICASSP*, pages 399–402, Apr 1993.

[5] M. Savic, E. Acosta, and S. K. Gupta. An automatic language identification system. In *Proc. ICASSP*, pages 817–820, 1991.

[6] V. Ramasubramanian, A. K. V. Sai Jayram, and T. V. Sreenivas. Language identification using parallel sub-word recognition - an ergodic HMM equivalence. In *Proc. Eurospeech*, pages 1357–1360, Geneva, Switzerland, Sep 2003.

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