LANGUAGE IDENTIFICATION WITH DYNAMIC HIDDEN MARKOV NETWORK

Konstantin Markov[1,2], *Satoshi Nakamura*[1,2]

[1]Spoken Language Communication Research Labs, ATR, Japan
[2]Spoken Language Communication Group, NICT, Japan
konstantin.markov@atr.jp, satoshi.nakamura@atr.jp

ABSTRACT

In this paper, we describe new language identification system based on the recently developed Dynamic Hidden Markov network (DHMnet). The DHMnet is a never-ending learning system and provides high resolution model of the speech space. Speech patterns are represented by paths through the network, and these paths when properly labeled with language IDs provide efficient means to discriminate between languages. First experiments indicated that our system can work on-line and is able to deliver relatively high performance with low latency. Evaluated on three language (English, Japanese and Chinese) identification task, the system achieved identification rates of 87.3% and 89.3% for 3 and 5 seconds long speech segments respectively.

Index Terms— Language identification, never-ending learning, dynamic hidden markov network, on-line learning, bio-inspired algorithms.

1. INTRODUCTION

Increased globalization and expanding practical adoption of the human language technologies require automatic speech processing systems, such as dialog or speech-to-speech translation systems, to operate in a multilingual environment. This, for its part, requires the availability of high performance, on-line language identification (LID) systems with minimum latency.

Research on LID has been focused mainly on two approaches. In the first one, the idea is to use the phonotactic content of the speech signal for language discrimination. Typically, language dependent phone n-gram models are trained from labeled data produced by single or multiple phone recognizers[1, 2]. The most popular technique is called Parallel Phone Recognition and Language Modeling (PPRLM) [3]. Input utterance is transformed into phone sequence by a set of recognizers and the probability of occurrence of this sequence in each language is obtained from the language n-grams. It has been shown that this approach is effective and gives good performance. An essential drawback, however, is the need for fine phone labels during training, but such labels may be too expensive or even impossible to obtain for some languages. The other approach to LID exploits the acoustical differences between languages. In a manner similar to the text-independent speaker identification, each language is represented by a Gaussian mixture model (GMM), trained on some language specific data. Initially, this method did not show good results, but recent improvements have increased its performance to match that of the phone-based systems [4, 5].

The goal of this study is to develop a LID front-end for our multilingual speech-to-speech translation system which will allow it to operate in a fully automatic mode. This requires the LID system to work on-line, in real time and to introduce minimum delay, i.e. to have as small as possible latency. Unfortunately, non of the widely used methods satisfies these conditions. That is why, we decided to experiment with the newly developed Dynamic Hidden Markov network (DHMnet)[6]. It is a never-ending learning model which when presented with sufficient data is capable of representing the speech manifold embedded in the feature space. The number of states and the network structure are automatically determined and can change in time depending on the data distribution. Speech patterns form paths through the network whose output is the best state sequence. For language identification, initially, the DHMnet is learned with data sequences, where every feature vector is labeled with silence or the corresponding language label. Each DHMnet state and transition are assigned discrete label probability distribution. Probabilities are estimated incrementally as the data come in. During the test, for speech segments only, the language labels probabilities are accumulated along the best state sequence and the language is identified as the label with the highest probability score. This approach is somewhat similar to the one from recently presented study and based on the self-organizing map (SOM)[7]. The difference is that, in contrast to the DHMnet, SOM topology and the number of nodes needs to be selected manually, which cannot be optimal and always introduces distortions of the speech space. Also, instead of probability distributions, SOM nodes are assigned single language label and the identification decision is made by simple label counting.

2. THE DHMNET

The DHMnet consists of hidden Markov states with self-loops and transitions between them. Additionally, neigh-

boring states are connected with lateral connections. Each state represents a part of the input feature space modeled by a multivariate Gaussian function. State sequences or paths through the network correspond to learned speech patterns or classes of patterns. New states and transitions are added to the network whenever new pattern is encountered. The practical problem is to define what should be considered as a "new" pattern and how to detect it. Inevitably, spurious events and noises would allocate states that may never be visited again. Such states (and paths) are considered "dead" and will be gradually removed from the network. The schematic structure of the DHMnet is shown in Fig.1, where transitions of a learned path are represented by directed solid lines, new paths with directed short dashed lines, and "dead" paths with directed long dashed lines. Undirected dashed lines represent lateral connections between states. Generally, any pattern

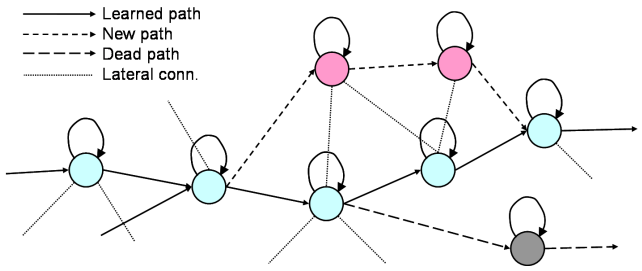

Fig. 1. Schematic structure of a Dynamic Hidden Markov network.

that is sufficiently different from those that have been already learned can be considered a new pattern. In the DHMnet, we use single multivariate Gaussian function with fixed diagonal covariance matrix for all the state PDFs and apply a threshold to the likelihood value, or distance from the mean, for "novelty" detection. Since the DHMnet is a first-order Markov chain where input vectors are presumed conditionally independent, the pattern-level novelty detection can be substituted by multiple frame-level novelty detections. Thus, any given input vector x will be considered "new" if $|x-\mu_b| > \theta = k\sigma$, where μ_b is the mean of the best matching state and the θ is the so-called *vigilance* threshold. Thus, the only two parameters that need to be set for the DHMnet are the Gaussian variance σ^2 and θ.

For the DHMnet state PDF learning, we consider the sequential version of the Maximum Likelihood estimation algorithm. In this case, the Gaussian mean update $\Delta\mu_n$ after input vector x_n will be:

$$\begin{aligned}\Delta\mu_n &= \mu_n - \mu_{n-1} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{n}{n}\mu_{n-1} \qquad (1)\\ &= \frac{(n-1)\mu_{n-1} + x_n - n\mu_{n-1}}{n} = \frac{1}{n}(x_n - \mu_{n-1})\end{aligned}$$

Since the DHMnet states represent different regions of the input feature space, it is important that neighboring states correspond to neighboring regions. That is, the state network should be a topology representing network. It has been shown that if lateral connections between neural network nodes (states in the DHMnet case) are built using the competitive Hebbian rule [8], the resulting network is a perfect topology representing network. The competitive Hebbian rule can be described as: for each input vector, connect the two closest nodes by an edge. Such networks have two very useful properties: 1) vectors that are neighbors in the input space will be represented by neighboring nodes; 2) if there is a path in the input space between two vectors, there will be a path connecting the two nodes that represent those vectors.

When a network dynamically changes its structure, the state neighborhood relations also change. To account for these changes, each lateral connection is given an age that is set to zero when a connection is made or refreshed. Otherwise, the connection age is increased every time one of the connection's states is visited. This way, connections that reach a certain age, i.e. ones that have not been refreshed for some time, are removed. The DHMnet states can have many lateral connections and if for some state all connections are removed, this state is pronounced "dead" and is removed along with all transitions to and from it.

For any input speech pattern represented by a sequence of feature vectors we are interested in finding the best state sequence or path through the network. Formally, this can be stated as follows:

$$\overline{S} = \max_{S} P(S|X), \;\; X = \{x_i\}_i^T, \; S = \{s_i\}_1^T \qquad (2)$$

The neighborhood and path preserving properties of the network ensure that each current state s_t is the best state given the current vector x_t. The best state sequence can be found by using a recursive procedure. Suppose that S_1^t is the best path until time t. Then

$$\begin{aligned}&P(S_1^{t+1}|X_1^{t+1}) = \qquad (3)\\ &= \left[\max_{s_j \subset Succ(s_t)} P(s_j|s_t)P(x_{t+1}|s_j)\right] P(S_1^t|X_1^t)\end{aligned}$$

where $Succ(s_t)$ is the set of succeeding states for state s_t.

We summarize the complete DHMnet algorithm as follows:

(1) Start with an empty network.
(2) For the next input vector x_t, given the current state s_{curr}, find the best matching succeeding state s_c. If it passes the vigilance test, set it as the next state, i.e. $s_{next} = s_c$, and go to (5).
(3) Find the best state, s_a, from all other states. If it passes the vigilance test, $s_{next} = s_a$, and go to (5).
(4) Add a new state, s_t, i.e. $s_{next} = s_t$, and set its mean to x_t.
(5) Make (update) the transition from the current state s_{curr} to s_{next}.

(6) Update the means of s_{next} and all its neighbors (Eq.1).

(7) Make (refresh) the connection between s_{next} and the second best state. Increase the ages of all s_{next} connections.

(8) If any connection age has reached the age threshold, remove this connection. Remove states with no connections.

(9) Add s_{next} to the best state sequence. Set the current state $s_{curr} = s_{next}$, and go to (2).

3. LANGUAGE INFORMATION LEARNING

The DHMnet is capable of learning the input data characteristics in an unsupervised manner, but it cannot acquire the higher level abstract knowledge about words or languages because such information is not presented in the acoustic signal. However, in a way similar to how humans learn, such knowledge can be learned by associating speech signals, or more precisely, the paths through the network, with the corresponding abstract notions. And this can only be achieved by a supervised training.

The simplest and most straightforward association is the labeling. Assuming that our data are segmented into speech and non-speech segments, we first assign labels to each feature vector. These labels, l_k, are either sil for silence or $lang_n$ for the n^{th} language to be identified. On the other hand, discrete label probability distributions (PD), $P_i(l_k)$ and $P_{i\rightarrow j}(l_k)$, are assigned to each DHMnet state s_i and to each transition from state s_i to state s_j whenever the state is created or the transition is established. During the DHMnet learning, the PDs of the states along the best path are incrementally updated with the corresponding labels according to the following rule:

$$P_i^n(l_k) = P_i^{n-1}(l_k) + \frac{1}{n}(C - P_i^{n-1}(l_k)) \qquad (4)$$

where

$$C = \begin{cases} 1, & if \quad l^n = l_k \\ 0, & if \quad l^n \neq l_k \end{cases} \qquad (5)$$

and n is the number of times the state has been visited and the l^n is the current label. The transition PDs, $P_{i\rightarrow j}(l_k)$, are updated in the same manner.

To identify the language of an unknown utterance X of length T[1], we accumulate the probabilities of each label l_k along the DHMnet's best state sequence. The decision is made by maximum probability principle, i.e.:

$$L(X) = \arg\max_{l_k} \sum_{t=1}^{T} \log P_t(l_k) \qquad (6)$$

We can use only state distributions, $P_t(l_k) = P_{s_t}(l_k)$, or transition distributions, $P_t(l_k) = P_{s_{t-1}\rightarrow s_t}(l_k)$, or both, $P_t(l_k) = P_{s_t}(l_k)P_{s_{t-1}\rightarrow s_t}(l_k)$. In the first case, only static, acoustical language dissimilarities are exploited, which corresponds to the idea of the GMM based LID approach. The second one, is somehow similar to the phone-based approach since it captures the differences in the acoustic dynamics. Naturally, we can expect that the last case will be the best option, because it combines both the static and dynamic information.

It may happen that some of the states in the best states sequence are newly added states to the DHMnet, and for them PDs do not exist yet. In such cases, we take the PD of the nearest old state. Common sense suggests that these probabilities should be weighted depending on the distance between the new and the old states, but in this study we did not use such weighting.

4. EXPERIMENTS

For experimental evaluation of the LID system, we used part of the ATR multilingual travel domain speech database [9]. It consists of read style studio recordings in three languages - English, Japanese and Chinese, from many speakers. For the initial DHMnet learning and supervised label training, we randomly selected 1000 utterances per language with the condition that the minimum speech duration of each utterance is 2 sec. This makes the total amount of speech about 75 min. per language. Half of the data are from male speakers and the other half is from female speakers. We denote this data set as "Train" set. For testing, two different data sets were selected, both consisting of 200 utterances (100 male and 100 female utt.) per language. The difference is that for the first set, called "Test1", there was no lower limit on the speech duration, but for the other one, "Test2", the limit was set at 5 sec. Note, that in the "Test1" data set, there were some utterances with one or two words only. The average speech length for this set is about 2.6 sec.

The speech pre-processing is the same as in our previous study [6], i.e. we use 24 Filter Bank log energies from 20ms long windows taken at 10ms. rate. All data are segmented into speech and silence regions using forced alignment by our conventional speech recognition system.

As explained in Section 2, the only two DHMnet parameters that need to be set are the Gaussians variance σ^2 and the novelty threshold θ. In this experiments we kept $\sigma = 1$, but θ was set to 1.4σ, 1.5σ and 1.6σ. The bigger is the threshold, the fewer states are created during the learning and, consequently, the lower is the resulting speech space resolution. The number of DHMnet parameters after the initial learning with the data set "Train" for each value of θ are summarized in Table 1. In these experiments, removing of "dead" DHMnet states was disabled. Label probability distributions were incrementally estimated during learning with the "Train" data set. There were four labels: *sil, en, jp, ch* for the silence and the three languages. Due to the label data sparseness, especially for the transition PDs, there were many distributions with unseen label values. In order to avoid numerical prob-

[1] Here we assume that X consists of speech frames only.

Table 1. DHMnet parameters after initial learning with the data set "Train".

Threshold	# states	# transitions
1.6	5703	282772
1.5	8825	358445
1.4	14177	449000

lems during scoring, probability value for the missing labels was set to a small fixed floor.

First, we tested the LID system with the data set "Test1" and the results are presented in Table 2 where the columns "State PDs", "Trans.PDs" and "Both PDs" show the LID rates corresponding to the three different cases of PDs usage for scoring described in Section 3. As we expected, DHMnets with more states performed better. But, in contrast to the common HMM sense, the information captured by DHMnet transitions shows same or even better discriminative abilities than that of the states. This maybe due to the fact that transitions implicitly hold some static information because they are state dependent and there is no PD sharing between them. In order

Table 2. LID rates (%) for different DHMnet novelty thresholds and data set "Test1".

Threshold	State PDs	Trans. PDs	Both PDs
1.6	82.2	84.6	85.3
1.5	84.5	84.5	84.6
1.4	85.3	87.5	86.3

to work on-line together with a multilingual speech recognition system, the LID system should be able to make decisions about the language not at the end of an utterance, but as soon as possible after the beginning of the speech segment. To check our system performance in this mode, we used the data set "Test2" and forced LID decision after a fixed number of speech frames have been processed. This number was set to correspond to 1, 2, 3, 4 and 5 seconds. The results of this experiment are shown in Table 3. Naturally, the performance for longer speech segments is better, but even for 3 seconds, which can be considered acceptable for on-line operation, the LID rate is more than 85% for all types of DHMnet.

5. CONCLUSIONS

We presented new LID system based on the Dynamic Hidden Markov network. The experiments showed that this system is able to achieve good performance even with short speech segments and could be used on-line as a front-end for a multilingual speech recognition system.

This is the first study of such kind of LID system and, actually, the first application of the DHMnet as a speech model in a real task, so there are much more investigations to be done

Table 3. LID rates (%) for different DHMnet novelty thresholds and different speech segment lengths from data set "Test2".

Length	State PDs	Trans. PDs	Both PDs
Threshold 1.6			
1 sec.	77.5	78.8	80.5
2 sec.	81.0	83.5	83.3
3 sec.	82.3	85.2	85.7
4 sec.	82.8	85.5	87.0
5 sec.	83.5	86.5	86.0
Threshold 1.5			
1 sec.	81.0	81.6	82.0
2 sec.	84.5	84.2	84.8
3 sec.	85.3	85.5	85.6
4 sec.	87.2	87.5	87.3
5 sec.	86.6	88.0	88.2
Threshold 1.4			
1 sec.	82.0	79.0	81.8
2 sec.	85.2	84.2	86.0
3 sec.	85.8	86.3	87.3
4 sec.	86.8	86.8	88.3
5 sec.	88.3	88.8	89.3

including the effect of the amount of training data, possibility of automatic abstract knowledge acquisition, error recovery, etc.

6. REFERENCES

[1] Y. Yan and E Bernard, "An approach to automatic language identification based on language-dependent phone recogntion," in *Proc. ICASSP*, 1995, pp. 3511–3514.

[2] M. Zissman, "Predicting, diagnosing, and improving automatic language identification performance," in *Proc. Eurospeech*, 1997, pp. 51–54.

[3] M. Zissman, "Comparisaon for four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 31–44, 1996.

[4] E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system," in *Proc. ICSLP*, 2002, pp. 93–96.

[5] E. Singer, P. Torres-Carrasquillo, T. Gleason, W. Campbell, and D. Reynolds, "Acoustic, Phonetic, and Discriminative approaches to automatic language identification," in *Proc. Eurospeech*, 2003, pp. 1345–1348.

[6] K. Markov and S. Nakamura, "Never-Ending Learning with Dynamic Hidden Markov Network," in *Proc. Interspeech*, 2007, pp. 1437–1440.

[7] L. Wang, E. Ambikairajah, and E. Choi, "Multi-layer Kohonen self-organizing feature map for language identification," in *Proc. Interspeech*, 2007, pp. 174–177.

[8] T. Martinetz and K. Schulten, "Topology representing networks," *Neural Networks*, vol. 7, no. 3, pp. 507–522, 1994.

[9] S. Nakamura, K. Markov, et al., "The ATR Multilingual Speech-to-Speech Translation System," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, Mar. 2006.