GENERALIZED MIXTURE OF HMMS FOR CONTINUOUS SPEECH RECOGNITION

Filipp Korkmazskiy, Biing-Hwang Juang and Frank Soong

Lucent Technologies Bell Laboratories, Murray Hill, NJ 07974, USA
yelena@research.bell-labs.com

ABSTRACT

This paper presents a new technique for modeling heterogeneous data sources such as speech signals received via distinctly different channels. Such a scenario arises when an automatic speech recognition system is deployed in wireless telephony in which highly heterogeneous channels coexist and interoperate. The problem is that a simple model may become inadequate to describe accurately the diversity of the signal, resulting in an unsatisfactory recognition performance. To deal with such a problem, we propose a Generalized Mixture Model (GMM) approach. For speech signals, in particular, we use mixtures of hidden Markov models (i.e., GMHMM, Generalized Mixture of HMM's). By applying discriminative training for GMHMM we obtained 1.0% word error rate for the recognition of the digits strings from the wireless database, comparing to 1.4% word error rate for the conventional HMM based discriminative technique.

1. INTRODUCTION

The objective of this study is to find an efficient speech modeling technique, allowing to cope with environmentally adverse conditions([1]) for robust recognition performance. Most existing approaches in dealing with the robustness issue use transformation of the speech signals or their models. Signal bias removal([2], [3]) and linear regression([4]) techniques may serve as examples of such transformations. It is very difficult, however, to find a universal transformation, applicable in the adaptation to different speech data, coming from highly heterogeneous sources.

In our modeling approach we assume that a Hidden Markov Model is only sufficient in characterizing the behavior of the speech signal from a known homogeneous source. When the signal source becomes heterogeneous due to the variety of the channels or noisy conditions, a natural extension to a single HMM is a mixture of HMM's. Such a mixture of HMM's may be obtained by clustering the speech patterns, that represent the same speech unit. In a conventional modeling framework, each speech unit is represented by one single HMM. Using clusters of samples belonging to the same speech unit, we can create multiple representations of the unit in the form of several HMM's, each of them being produced by the corresponding training procedure according to some universal criteria.

2. GENERALIZED MIXTURE OF HMMS FORMULATION

Let us consider all speech tokens in training the HMM for some particular segment of the speech(e.q., a word or subword unit). A conventional HMM for the speech unit can describe the behavior of the real speech signals only approximately, because it does not take into account the correlation between the parts of the speech signal assigned to the different HMM states. In order to make speech modeling more accurate we propose to expand single HMM to a mixture of HMMs, thereby delivering more precise description for the different groups of the speech signals. To implement that, we have to split all speech samples involved in the single HMM construction into several groups(clusters) and then build individual HMMs for each of the groups of the speech samples. The first problem that we come across in connection with speech samples clustering is the variable length of the speech samples. This obstacle does not allow direct application of the traditional clustering methods used for fixed dimensional vectors. We apply a conversion procedure to overcome this problem. According to this procedure the position of a speech sample is characterized by the set of distances from the speech sample to the corresponding elements(states or mixtures) of the HMM. Let us consider a speech sample $O^{(i)}$, which is a representative of the samples of the speech unit i. We define a distance from $O^{(i)}$ to the corresponding HMM $\lambda^{(i)}$ as follows. For each frame $O_f^{(i)}, (1 \leq f \leq F)$ of $O^{(i)}$ let $d_j(O_f^{(i)}|\lambda^{(i)})$ be the distance(log probability) from $O_f^{(i)}$

to the j-th state. Usually, this distance is a byproduct of the Viterbi segmentation procedure. By averaging the corresponding distances over all the frames assigned to the same j-th state of HMM $\lambda^{(i)}$ after the segmentation we obtain fixed number of the vector components describing a speech sample $O^{(i)}$:

$$d_j(O^{(i)}|\lambda^{(i)}) = \frac{1}{F_j} \cdot \sum_{f_j=1}^{F_j} d_j(O_{f_j}^{(i)}|\lambda^{(i)}) \qquad (1)$$

Here $d_j(O^{(i)}|\lambda^{(i)})$ is the average distance for all F_j frames f_j from the speech sample $O^{(i)}$, assigned to the j-th state($1 \leq j \leq N^{(i)}$) of the HMM $\lambda^{(i)}$. So, we've got a fixed number $N^{(i)}$ of the vector components $d_j(O^{(i)}|\lambda^{(i)})$ representing the speech sample $O^{(i)}$.
In a similar manner we can define a more detailed representation of the speech sample by evaluating corresponding distances to the states mixtures. For each frame $O_f^{(i)}$, ($1 \leq f \leq F$) of $O^{(i)}$ let $d_{jm}(O_f^{(i)}|\lambda^{(i)})$ be the corresponding distances(log probabilities) to the all state mixtures M_j, representing the j-th state ($1 \leq m \leq M_j$; $1 \leq j \leq N^{(i)}$) of the HMM $\lambda^{(i)}$. By averaging the corresponding vector components over all the frames assigned to the same j-th state of HMM $\lambda^{(i)}$ we get the new representation for the sample $O^{(i)}$:

$$d_{jm}(O^{(i)}|\lambda^{(i)}) = \frac{1}{F_j} \cdot \sum_{f_j=1}^{F_j} d_{jm}(O_{f_j}^{(i)}|\lambda^{(i)}) \qquad (2)$$

By introducing some distance measure between vectors of the fixed dimensionality we can easily apply traditional(like K-means) clustering procedure in order to distribute all samples, associating the i-th unit, into $K^{(i)}$ subsets and then create individual HMM $\lambda_k^{(i)}$ ($1 \leq k \leq K^{(i)}$) for each of the $K^{(i)}$ groups of the speech samples. Such groups of cluster HMM's, representing the same speech unit, we'll refer to as *Generalized Mixture of HMMs*. So, during decoding we consider a mixture of HMM's for each speech unit rather than a single HMM for a speech unit.
Creating HMM mixtures for connected word or continuous speech recognition involves the notion of 'decoded unit trajectory'. The following diagrams illustrate the difference between generalized mixture of HMMs and the conventional mixture density HMM. The initial parameters for the mixture density in an HMM state may be evaluated by clustering the speech frames assigned to the particular HMM state after segmentation(Figure 2). In a similar manner, we cluster the speech samples based on trajectories, associated with a specific unit(Figure 2).
Clusters of the speech sample trajectories are be used

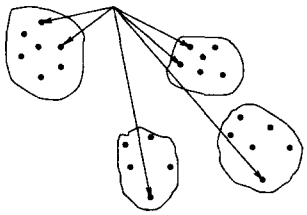

Figure 1: HMM state mixtures, where speech frames are assigned to an HMM state

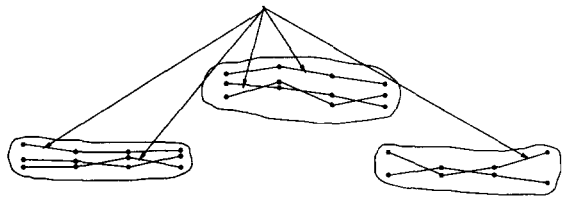

Figure 2: Generalized mixture of HMM's, where speech samples trajectories are assigned to a speech unit

to create a new set of HMM's, each of them delivering more accurate modeling for the different groups of the speech samples. The conventional mixture density HMM, being inadequate to characterize the interrelations between adjacent speech frames, may not give the best possible recognition performance. With a GMHMM set, an N-best algorithm is used to produce N different cluster trajectories rather than N different word strings. That is, different cluster trajectories may be of the same lexical content (i.e., the same sequence of words). The example below illustrates this point.

The output of the HMM-based N-best strings

(1 2 3) – the best digit string
(3 4 2) – the 2nd best digit string
(2 3) – the 3rd best digit string

The output of the GMHMM-based N-best strings

(1-b 2-a 3-a) – the best cluster trajectory
(1-a 2-a 3-a) – the 2nd best cluster trajectory
(3-b 4-c 2-b) – the 3rd best cluster trajectory
(1-a 2-b 3-b) – the 4th best cluster trajectory
(2-b 3-a) – the 5th best cluster trajectory
(3-a 4-b 2-a) – the 6th best cluster trajectory
1-b – denotes cluster b for digit 1.

In order to produce the best cluster trajectories at each grammar node while runnig the N-best algorithm we have to consider the alternative junctions between different clusters representing the units rather than the junctions between the units themselves.
In order to get scores for the competing lexical strings we proposed a formula, which allows a combination of

scores for different trajectories of the same lexical content:

$$G(O|\Lambda, W_r) = \frac{1}{\mu} \cdot \log \left[\frac{1}{M_{W_r}} \cdot \sum_{m=1}^{M_{W_r}} \exp(\mu \cdot g(O|\Lambda, W_r^{(m)})) \right] \quad (3)$$

Here, $\mu > 0$, W_r is the r-th lexical string after N-best decoding($1 \leq r \leq R$), $W_r^{(m)}$ is the m-th cluster trajectory, corresponding to the r-th lexical string($1 \leq m \leq M_{W_r}$), $g(O|\Lambda, W_r^{(m)})$ is the score for the cluster trajectory $W_r^{(m)}$, Λ is the set of the cluster HMMs. It should be clear, that

$$\sum_{r=1}^{R} M_{w_r} = N, \quad (4)$$

where N is the total number of the best cluster trajectories produced by the N-best algorithm.

3. DISCRIMINATIVE TRAINING FOR GENERALIZED MIXTURE OF HMMS

We further cast the GMHMM modeling technique in the framework of discriminative training. Proper use of the GPD method([5]), applied to the component HMM's, combines the advantage of the more accurate modeling, obtained by the GMHMM approach, and the discriminative power of the GPD method. According to the GPD formulation we use a gradient-descent technique to minimize the expectation of the classification error defined according to the formula:

$$\ell(O|\Lambda) = \frac{1}{1 + \exp(-\gamma D(O|\Lambda))}, \ \gamma > 0 \quad (5)$$

The value of $D(O|\Lambda)$ for the GMHMM is defined as follows:

$$D(O|\Lambda) = -G(O|\Lambda, W_i) + \frac{1}{\eta} \log \left[\frac{1}{R-1} \cdot \sum_{r \neq i}^{R} \exp(\eta G(O|\Lambda, W_r)) \right], \ \eta > 0 \quad (6)$$

Here, W_i is the correct lexical sequence.

In applying the gradient-descent technique we have to evaluate partial derivatives for all parameters of the GMHMM. For some arbitrary HMM parameter α a corresponding estimate takes on such a form:

$$\frac{\partial \ell(O|\Lambda)}{\partial \alpha} = \sum_{r=1}^{R} \frac{\partial \ell(O|\Lambda)}{\partial G(O|\Lambda, W_r)} \cdot \frac{\partial G(O|\Lambda, W_r)}{\partial \alpha} \quad (7)$$

The first term in the last formula may be evaluated as follows:

$$\frac{\partial \ell(O|\Lambda)}{\partial G(O|\Lambda, W_r)} = \frac{\partial \ell(O|\Lambda)}{\partial D(O|\Lambda)} \cdot \frac{\partial D(O|\Lambda)}{\partial G(O|\Lambda, W_r)} \quad (8)$$

The expressions for the derivatives in the formula (8) are well known([6]).

In turn, the second term in the formula(7) may be evaluated according to the formula:

$$\frac{\partial G(O|\Lambda, W_r)}{\partial \alpha} = \sum_{m=1}^{M_{w_r}} \frac{\partial G(O|\Lambda, W_r)}{\partial g(O|\Lambda, W_r^{(m)})} \cdot \frac{\partial g(O|\Lambda, W_r^{(m)})}{\partial \alpha} \quad (9)$$

The derivative $\frac{\partial G(O|\Lambda, W_r)}{\partial g(O|\Lambda, W_r^{(m)})}$ may be evaluated as follows:

$$\frac{\partial G(O|\Lambda, W_r)}{\partial g(O|\Lambda, W_r^{(m)})} = \exp(g(O|\Lambda, W_r^{(m)})) \cdot \left[\sum_{m=1}^{M_{w_r}} \exp(\mu \cdot g(O|\Lambda, W_r^{(m)})) \right]^{-1} \quad (10)$$

The above formulas allow to apply the GPD method for optimization of the GMHMM parameters.

4. EXPERIMENTAL RESULTS

We conducted experiments to verify the effectiveness of the proposed GMHMM approach. Wireless data of connected digit strings recorded over analog AMPS and digital cellular (TDMA with IS-54 coding) channels were used in the experiments. The collected data include different channel and noise conditions(from clean speech to hardly audible speech, contaminated mainly by environmental car noise). The digit string length in the database ranges from 1 to 30 digits. In the experiments we used context-dependent subword units. Each digit was represented by a concatenation of its head, body and tail models. Altogether 274 context dependent such units were employed. For GMHMM approach K-means clustering technique([7]) was applied to obtain subsets of the training data used to build corresponding cluster HMM's(GMHMM). First, we conducted comparative experiments using 2 different distance measures(eqs. (1) and (2)) For MLE training(MLE stands for speech models obtained by the maximum likelihood estimation) of the GMHMM, we obtained 2.4% word error rate for the states based distance measure(eq. (1)) and 1.9% word error rate for the state mixtures based distance measure(eq. (2)). So, in the experiments, based on the discriminative training of GMHMM, formula (2) appears to be a more

efficient distance measure. The total number of the clusters obtained via K-means($K = 5$) algorithm, after outliers were removed subsequently, was 678. In both representations(HMM and GMHMM) 8 mixtures per a statewere used. The recognition results are summarized in the following table, where word error rates are tabulated for the traditional HMM approach and GMHMM approach:

Wireless Data Recognition Performance

Model	MLE	GPD
HMM	2.6%	1.4%
GMHMM	1.9%	**1.0%**

From the performance table it is observed that the proposed GMHMM technique outperforms the traditional single HMM approach by a significant margin. We also conducted the experiment of MLE training for the conventional HMM's using approximately the same total number of parameters as for GMHMM approach. To implement that, we used 20 mixtures per a state for HMM(versus 8 mixtures per a state for GMHMM). The word error rate obtained in this experiment was only 2.4%(versus 1.9% for GMHMM).

5. CONCLUSIONS

The proposed GMHMM speech representation allows us to achieve more accurate modeling of heterogeneous data by using several clustered HMM's per speech unit rather than a single HMM. Experiment showed that the use of the GMHMM reduced word error rate in a continious speech recognition task, using either MLE or GPD training technique. It was found, that the quality of GMHMM depends on the distance measure, chosen for the speech sample clustering. Further improvements of the proposed method may be achieved by finding a more relevant cluster distance measure and by modifying the corresponding clustering technique. In order to reduce the computational complexity, caused by the increased number of the HMMs, the possibility of tying across clusters for different speech units may be explored. Additionally, such procedure may yield better recognition performance because of the more efficient use of training data. Also, in order to achieve faster implementation it is worth to consider the possibility to use GMHMM representation in the second pass of the recognition after obtaining N-best candidates in its the first pass with a traditional HMM representation.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. C.H. Lee and Dr. Wu Chou for their suggestions and support during the process of this research.

6. REFERENCES

[1] B.-H. Juang "Speech recognition in adverse environments," *Computer Speech and Language*, **5**, pp. 275–294, 1991.

[2] M.G. Rahim, B.-H. Juang "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, **4**(1), pp. 19–30, January 1996.

[3] M.G. Rahim, B.-H. Juang, W. Chou and E.Buhrke "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters*, **3**(4), pp. 107–109, April 1996.

[4] O. Siohan, Y. Gong, J.-P. Haton "Noise adaptation using linear regression for continuous noisy speech recognition," *Eurospeech'95'*, **1**, pp. 465–468, 1995

[5] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, **40**(12), pp. 3043–3054, December 1992.

[6] C.-S. Liu, C.-H. Lee, W. Chou, B.-H. Juang and A. Rosenberg "A study on minimum error discriminative training for speaker recognition," *J. Acoust. Soc. Am.*, **97**(1), pp. 637–648, January 1995.

[7] J. G. Wilpon, L.R. Rabiner "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. on Acoust., Speech, Signal Processing*, **33**(3), pp. 587–594, June 1985.