Spoken Web Search using an Ergodic Hidden Markov Model of Speech

Asif Ali
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250 USA
asif.ali@gatech.edu

Mark A. Clements
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250 USA
clements@ece.gatech.edu

ABSTRACT

An ergodic hidden Markov model (EHMM) of speech can be trained in an unsupervised manner using unlabeled speech. A keyword spotting system has been developed where the queries and test observations are represented as sequences of states of the EHMM. A graphical keyword model is built by aggregating multiple instances of a query or by using mappings between phonemes and states of the EHMM. A modified Viterbi algorithm with a 3D lattice structure has been used to score the observations.

1. INTRODUCTION

Traditional statistical approaches model speech at word or subword level using left-right HMMs. These schemes require knowledge of the linguistic structure of a language and the availability of large amount of labeled data for training. This methodology of speech recognition, therefore, cannot be adopted for a large number of resource-limited languages.

An alternate statistical approach that eliminates reliance on time-aligned labeled training data, is to built a single model of speech. While phonemes always occur in a particular sequence in an utterance, the set of phonemes, in different combinations, form all the utterances in a language. Modeling the entire signal space of a language would require a model with a far more flexible structure than that of a left-right HMM.

In this work, a single, large EHMM has been used to model the entire speech. An EHMM can model non-linear observation distributions and capture the short-term correlations in speech. Furthermore, an EHMM can be trained without labeled data. These characteristics have been the motivation for the design of an EHMM-based keyword spotting system for the Spoken Web Search task of MediaEval 2013, described in detail in [2].

2. AN EHMM OF SPEECH

The design of an EHMM encompasses a number of factors including the size of the EHMM, the form of the observation distributions, initial values of the parameters, and the training methodology.

2.1 Training

The training involves the established Baum-Welch algorithm and can be carried out using the Hidden Markov Model Toolkit (HTK) [5]. Every frame in the training data contributes to the parameter estimates of the whole model. Mel-frequency cepstral coefficients (MFCC) with delta and delta-delta coefficients were used as a feature set. The MFCC features were generated at a frame rate of 100 Hz using a 25 ms window. The observation density for each state was modeled using a single Gaussian mixture with a diagonal covariance matrix. The mean value and covariance for each state was initially set to the global mean and covariance of the training data. Random perturbations were then added to the mean values before re-estimation using the Baum-Welch algorithm.

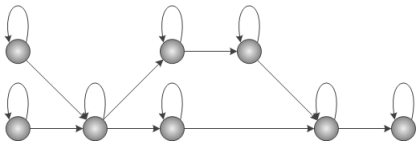

Figure 1: Graphical keyword model

2.2 Speech Retrieval using an EHMM

An EHMM stores sufficient detail about the original signal that perfectly intelligible speech can be synthesized from its state sequence [4]. In this research, an EHMM has been used to transform speech into a model-specific representation which is then used for the task of spoken web search. For each keyword, the corresponding most likely sequence of EHMM states can be calculated using the Viterbi algorithm, which can then be used as a left-right HMM of the query.

In absence of any time-aligned labeled data for training, observations are grouped into a state based on their proximity in the feature space and exhibition of similar temporal characteristics. Since most feature sets, including MFCCs, are not invariant to all speaker-dependent variations, speech segments with the same perceptual character may occupy distinct regions in the feature space, and consequently, mapped to different states of the EHMM.

A graphical keyword model [1], shown in Figure 1, can model an utterance using a network of states of the EHMM. A segment of speech can be modeled with multiple EHMM states, each corresponding to a different pronunciation and each state retaining its state transition probabilities.

3. EXPERIMENTS

A series of experiments were carried out to identify the configuration of system with the highest precision. Due to

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

Table 1: MTWV for the development set as a function of number of states in the model

States	64	128	192	256
MTWV	0.0565	0.0830	0.0868	0.0933

Table 2: MTWV for the development set as a function of number of clusters in the model

Clusters	160	192	224	256
MTWV	0.1026	0.1033	0.102	0.0993

time constraints, an exhaustive evaluation of every combination of the parameters was not possible.

3.1 Number of States in EHMM

The optimal number of states in an EHMM is a function of acoustic variety in the training data and may differ from one language to another. A small number of states may not fully capture the characteristics of the modeled data. Conversely, if the number of states in the EHMM are larger than that required then only a subset of the states will be used for modeling. For the SWS2013 data, a number of different EHMMs were tested but larger EHMMS, with number of states greater than 192, lost some of the states during training. Hence, the maximum size of the EHMM was limited to 256 states. The maximum term weighted value (MTWV) was observed for the EHMM with 256 states (Table 1).

3.2 Clustering the States of EHMM

A graphical keyword model can also be built from a single utterance by incorporating knowledge of similar states of the EHMM obtained through supervised or unsupervised clustering schemes. In this work, similar states in the 256-state EHMM were merged to form superstates or state clusters using the approach outlined in [1]. Beginning from a 256-state EHMM, states, with the least distance between them, were merged in succession to form a new model with smaller number of clusters.

A merger of similar states initially led to an increase in the value of MTWV (Table 2). The EHMM with 192 superstates yielded a higher MTWV than any of the EHMMs trained directly by the Baum-Welch algorithm.

3.3 Extended Trials

In these experiments, graphical keyword models were built from multiple instances of a query. Surprisingly, the highest gain in MTWV was observed after the addition of the second example and the precision decreased consistently (Table 3) after the inclusion of each additional example.

3.4 Score Normalization

The range of values for log-likelihoods is much larger than most similarity/distance measures. The histograms of log-likelihood scores of candidates for two different queries, after normalizing for length differences, is shown in Figure 2. The large variations in the candidate scores for different trials make the computed TWV, which uses a single scale for all trials, susceptible to score normalization parameters. The scores in Tables 1 and 2 were normalized using Z-Norm while the scores in Table 3 and the final submissions were normalized using a variant of T-Norm, as outlined in [3].

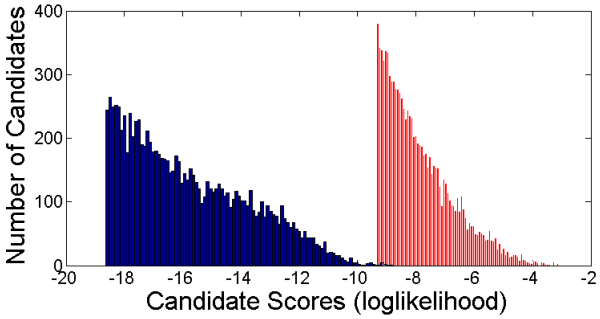

Figure 2: Histogram of length-normalized scores for two different trials

Table 3: MTWV for extended trials

Instances	2	3	5	7	9
MTWV	0.1259	0.1219	0.1151	0.1096	0.1081

3.5 Performance Metrics

The experiments were carried out on Georgia Tech's heterogeneous PACE cluster, which consists of multi-core (24-64 cores, 64-256 GB RAM) Altus-branded servers running RHL6.3. A typical run on the development set would take 40 hours of CPU time with an average memory footprint of 5MB leading to an estimated real-time factor of 0.003.

4. CONCLUSIONS

A zero-resource approach to spoken web search using a single large ergodic hidden Markov model of speech has been presented. The novel graphical keyword model represents a query as a network of states of the EHMM, and can incorporate multiple examples of a keyword or phoneme-to-state mappings. This approach can benefit from improved discovery of phoneme-to-state mappings and the scored TWV can further be increased by a better score normalization scheme.

5. REFERENCES

[1] A. Ali, S. Rustamov, and M. Clements. Speech retrieval using a single ergodic hidden Markov model. *AICT 2013*, forthcoming.

[2] X. Anguera, F. Metze, A. Buso, I. Szoke, and L. J. Rodriguez-Fuentes. The spoken web search task. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1):42–54, 2000.

[4] M. Lee, A. Durey, E. Moore, and M. Clements. Ultra low bit rate speech coding using an ergodic hidden Markov model. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 765 – 768, 18-23, 2005.

[5] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK book. *Cambridge University Engineering Department*, 2002.