Unsupervised Hidden Markov Modeling of Spoken Queries for Spoken Term Detection without Speech Recognition

Chun-an Chan and Lin-shan Lee

Graduate Institute of Communication Engineering,
National Taiwan University, Taipei, Taiwan

chunanchan@gmail.com, lslee@gate.sinica.edu.tw

Abstract

We propose an unsupervised technique to model the spoken query using hidden Markov model (HMM) for spoken term detection without speech recognition. By unsupervised segmentation, clustering and training, a set of HMMs, referred to as acoustic segment HMMs (ASHMMs), is generated from the spoken archive to model the signal variations and frame trajectories. An unsupervised technique is also designed for ASHMMs parameter training. A model-based approach for spoken term detection is then developed by constructing a query HMM from the ASHMMs, and then scoring the spoken documents using the query HMM. Experiments show that this model-based approach complements the feature-based dynamic time warping approach. A significant improvement on detection performance is achieved by integrating the two methods.

Index Terms: Spoken term detection, hidden Markov model, dynamic time warping, acoustic segment model

1. Introduction

Spoken term detection (STD) for multimedia content retrieval has gained much attention in recent years. Many successful approaches utilize automatic speech recognition (ASR) techniques to transcribe spoken utterances into lattices for searching, which relies on accurate recognition performance. However, it is practically difficult to train well matched acoustic/language models for the wide variety of content available over the Internet with very limited annotated training data. The unsupervised STD approaches using frame-based dynamic time warping (DTW) were thus proposed [1, 2], which do not require speech recognition, but at the cost of relatively high computation requirements. The segment-based DTW was thus proposed for fast selection of spoken term hypothesis [3]. In the segment-based DTW approach, the feature frames are decomposed into segments containing acoustically similar frames. Then the dynamic programming algorithm is applied on the segments. In this way, the distance calculations of highly redundant feature vectors in frame-based DTW can be reduced by using representing vectors in each segment. The hypothesize regions, or the signal parts in the spoken documents most similar to the query, are also located for frame-based rescoring and pseudo relevance feedback (PRF) [4]. The rescoring is achieved by applying frame-based DTW on the query and the hypothesized regions. Then a few most similar hypothesized regions are regarded as pseudo relevant regions. The distances between hypothesized regions and pseudo relevant regions are calculated using frame-based DTW, and integrated in the PRF approach.

One of the limitations of DTW-based approaches is the incapability of modeling signal variations. Gaussian posteriorgrams and phoneme posteriorgrams were hence used in DTW to address this issue [1, 2]. However, the posteriorgram is calculated from each frame individually without considering the trajectory of frames in the feature space. In contrast, hidden Markov model (HMM) with Gaussian mixtures is the most well-known solution for modeling the signal variations as well as signal trajectory in speech. Based on HMM, an acoustic segment model was proposed to incorporate temporal information in speech signal modeling [5], which was recently applied in spoken language recognition and speaker recognition [6, 7].

In this work, we model the speech signals using HMMs instead of Gaussians in an unsupervised fashion. The speech signals in the spoken archive are grouped into segments and applied a segment clustering algorithm [8]. The number of clusters can be automatically determined using minimum description length criterion. For each cluster, an HMM, referred to as acoustic segment HMM (ASHMM), is created to model all signal trajectories in the cluster. We develop an unsupervised training technique that simultaneously train the parameters in each ASHMM and the inter-ASHMM transition probabilities. We then propose a model-based STD method using these ASHMMs. A query HMM is constructed for each spoken query using the ASHMMs, and the similarity between the query and the spoken documents in the archive is calculated using Viterbi algorithm based on this query HMM. The previously proposed feature-based DTW approaches [3, 4] and the model-based HMM approach proposed here are further integrated. Experiments show a 1.7% absolute improvement of mean average precision over the DTW-based methods.

2. Unsupervised hidden Markov modeling of spoken archive

In this section, we describe the process for constructing a collection of ASHMMs that model the speech signals in the archive. This process is off-line performed without any spoken query.

2.1. Speech segmentation and segment feature vector

The speech signals in the utterances are first transformed into frames of Mel-filter bank outputs. The hierarchical agglomerative clustering algorithm is then performed on the sequences of feature vectors to construct a hierarchical tree for each utterance with the minimum total variance criterion [9]. Using a threshold selection method for tree height, each hierarchical tree is decomposed into subtrees, corresponding to segments in the utterance [3]. Each segment contains consecutive acoustically similar frames that approximately correspond to a vocal tract stage. A segment feature vector is then extracted by concatenating the mean vectors of the front, middle and end parts of each segment. Here, the corresponding subtree of a segment is divided into three smaller subtrees corresponding to the three parts of a segment [3]. Karhunen-Loève transform is then applied on all such concatenated mean vectors to reduce the correlation between dimensions while retaining 97% of total variance.

2.2. Segment clustering

Let $X = \{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^d$ be the set of segment feature vectors obtained above, where N is the total number of segments. An unsupervised clustering algorithm is applied to construct Gaussian mixtures in the segment feature space. A Gaussian mixture model $\Gamma = \{\pi_k, \mathcal{N}(\mu_k, \mathbf{\Sigma}_k)\}_{k=1}^{K}$ is used to model the segment feature distribution, where π_k is the weight of a Gaussian distribution and K is the number of Gaussians. The minimum description length (MDL) principle is used to select parameter set Γ including the number of Gaussians K [8]. The objective function is

$$f(\Gamma, K) = \log Pr(X|K, \Gamma) - \frac{1}{2} L \log(Nd), \qquad (1)$$

and the goal is to find Γ and K that maximize (1). The objective function is essentially the log likelihood function minus a penalty term, where

$$L = K\left(1 + d + \frac{(d+1)d}{2}\right) - 1 \qquad (2)$$

is the number of continuously valued real numbers to specify Γ. The penalty term prevents the model from overfitting to the training data. The iterative algorithm for MDL principle starts with an initial $K = K_0$. The means are randomly selected, and the covariance matrices are initialized as the covariance matrix of all segment feature vectors. At each iteration Γ was trained using EM algorithm with K fixed. At the end of each iteration, $f(\Gamma, K)$ is recorded, and then K is decreased by one by combining the two closest Gaussians [8]. After all Gaussians are merged into one, the parameters Γ and K that maximize (1) are returned. Each segment x_i is then labeled with a cluster label y_i,

$$y_i = \arg \max_{1 \le k \le K} \pi_k \cdot \mathcal{N}(x_i|\mu_k, \mathbf{\Sigma}_k). \qquad (3)$$

The segments and cluster labels are used for initializing the HMM training below.

2.3. HMM training

Using the segment clustering approach, the segments in the spoken archive are now labeled with $Y = \{y_n\}_{n=1}^{N}$. For each cluster of segments, an ASHMM is constructed with left to right state transitions without state skipping. The observations of these ASHMMs are the frame-level Mel-filter bank output vectors. The k-th ASHMM has states $q_1^k, \ldots, q_{M_k}^k$ and parameter set $\theta_k = \{\Pr(q_1^k), \Pr(q_{i+1}^k|q_i^k), \mu_i^k, \Sigma_i^k\}$ with $\Pr(q_i^k) = 0$ if $i \neq 1$. The emission probability is modeled by a single Gaussian in each state. The number of states $M_k = \lceil \frac{l_k}{10} \rceil$ in θ_k where l_k is the average length of segments labeled k. The transition probabilities from k-th to l-th ASHMM is $\Pr(q_1^l|q_{M_k}^k)$. Hence $\Pr(q_i^l|q_j^k) > 0$ only if $i = 1$ and $j = M_k$ when $l \neq k$. Fig. 1 shows an example of all parameters related to k-th ASHMM. These K ASHMMs including the transition between them jointly model the distribution and trajectories of speech signals in the spoken archive.

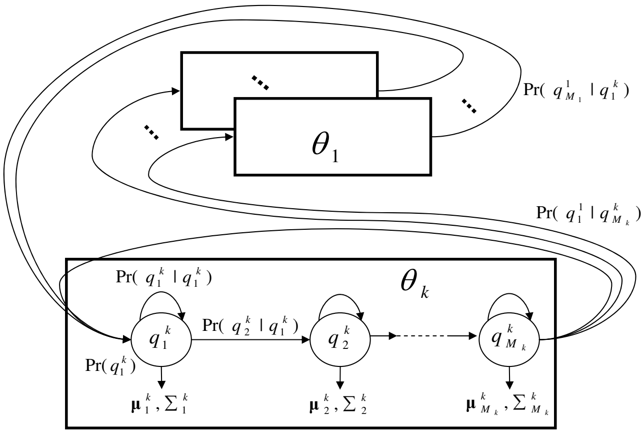

Figure 1: All parameters related to k-th ASHMM.

Using the segment labels in Y, the standard HMM initialization techniques are applied to the ASHMMs. The models are then trained with maximum likelihood estimation using EM algorithm. In conventional phone HMM training, the phoneme sequences are specified in the training transcriptions and only left-to-right state transitions are allowed with unconstrained phoneme boundaries. However in ASHMMs training here, since the labels are obtained in an unsupervised way, labels in Y are not specified for the training utterances. With unspecified labels, the number of transitions between ASHMMs will be unbounded. This could ruin the model because the likelihood function is dominated by emission probability density functions modeled by Gaussians whereas the transition probabilities are probability mass functions. The training will act like Gaussian mixture training and the temporal information carried by transition probabilities will be lost. To avoid this, transitions between ASHMMs are only allowed at segment boundaries. In other words, any segment will only belong to a single ASHMM in a state sequence during training.

3. Spoken query HMM for spoken term detection

In this section, we first briefly summarize the previously proposed feature-based STD approaches using DTW, which will be used here. Then we elaborate on methods for creating a query HMM with ASHMMs and the scoring mechanisms.

3.1. STD with segment-based DTW and pseudo relevance feedback

Given the spoken query term Q, it is first transformed into frames of Mel-filter bank outputs and decomposed into segments using the same HAC segmentation algorithm described in section 2.1. In the segment-based DTW approach, the dynamic programming algorithm is applied on the segment sequences in the query and the spoken document to efficiently locate the signal parts most similar to the query, referred to as hypothesized regions [3]. Then a second-pass frame-based DTW is applied to the query and the located hypothesized regions [4]. In this way, the segment-based DTW reduces the search space from the whole archive to only several short hypothesized regions in each document for the time-consuming frame-based DTW. A pseudo relevance feedback (PRF) approach is then applied to expand

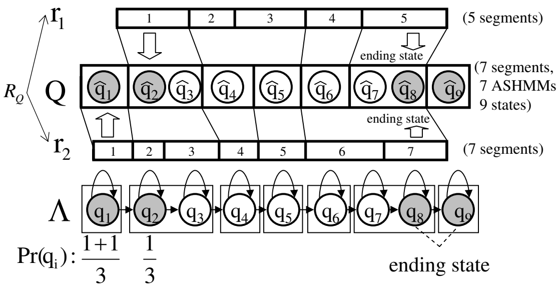

Figure 2: An example of initializing query HMM Λ.

the signal realizations of the desired query. After the frame-based rescoring, T hypothesized regions $R_Q = \{r_1, \ldots, r_T\}$ with the smallest distances to the query are regarded as pseudo relevant regions. The frame-based DTW is applied again on all hypothesized regions in the archive to all the pseudo relevant regions in R_Q. The relevance score $\text{score}_{\text{prf}}(u)$ of a hypothesized region u to the query Q using PRF approach is defined as

$$\text{score}_{\text{prf}}(u) = d(u, Q) + \sum_{t=1}^{T} e^{-\kappa t} \cdot d(u, r_t), \tag{4}$$

where $d(\cdot, \cdot)$ is the distance calculated with frame-based DTW and κ is the weight decay factor for t-th pseudo relevant region [4].

3.2. Query HMM initialization

Given the spoken query Q, a state sequence $(\widehat{q}_1, \widehat{q}_2, \ldots, \widehat{q}_M)$ is obtained by decoding Q using Viterbi algorithm with the ASHMMs described above. In the decoding procedures, transitions are allowed at segment boundaries if and only if they are transitions between ASHMMs. Hence the number of decoded ASHMMs is the same as the number of of segments in Q. A query HMM Λ with M states $(q_1, \ldots, q_M)$ is then constructed for the query. The Gaussian of state q_i is initialized to the same Gaussian of $\widehat{q}_i$. The state transition probabilities are initialized as

$$\Pr(q_j|q_i) = \begin{cases} \Pr(\widehat{q}_i|\widehat{q}_i) & \text{if } j = i \\ 1 - \Pr(\widehat{q}_i|\widehat{q}_i) & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

for $i < M$ and $\Pr(q_M|q_M) = 1$. Therefore Λ allows only left-to-right transitions without state skipping. All pseudo relevant regions in R_Q and their segment-based DTW warping functions are used to initialize $\Pr(q_i)$ and to mark the ending states. Let $h(s)$ be the number of pseudo relevant regions in R_Q whose first segment are matched to s in Q. The summation of $h(s)$ over all segment s in Q is T since there are T pseudo relevant regions. Assume the first decoded state of segment s is $q_{\text{first}}(s)$. Then

$$\Pr(q_{\text{first}}(s)) = \tfrac{h(s)+1}{T+1} \quad \text{if } q_{\text{first}}(s) = q_1, \tag{6}$$

$$\Pr(q_{\text{first}}(s)) = \tfrac{h(s)}{T+1} \quad \text{if } q_{\text{first}}(s) \neq q_1. \tag{7}$$

A state q_i is an ending state if either $i = M$ or $\widehat{q}_i$ is the last state in an ASHMM matched to an ending segment of some $r_j \in R_Q$. Fig. 2 shows an example of query Q with two pseudo relevant regions $R_Q = \{r_1, r_2\}$. Since the first segment in r_1 is matched to the second segment in Q, $h(2) = 1$. The first segment in r_2 is matched to the first segment in Q, hence $h(1) = 1$. Then $\Pr(q_1) = \frac{2}{3}$ by applying (6) and $\Pr(q_2) = \frac{1}{3}$ by (7). q_8 is an ending state because $\widehat{q}_8$ is the last state in the ASHMM matched to the last segments in r_1 and r_2. q_9 is an ending state because $M = 9$. At training and decoding stage, only hidden state sequences with trailing ending state are permissible.

3.3. Query HMM training

After the query HMM is initialized, the model is trained using the top T' most similar hypothesized regions ranked by PRF method. Since the hypothesized regions have different confidence of being relevant to the query, the training weight for them are different. In PRF method, all hypothesized regions are ranked by each of the pseudo relevant region $r_j \in R_Q$. Therefore there is a ranking list of hypothesized regions for each r_j. Assume hypothesized region u is ranked as $\text{rank}(u)$ by Q and $\text{rank}_j(u)$ by r_j for $j = 1, \ldots, T$. The training weight for u is

$$w(u) = e^{-\lambda \cdot \text{rank}(u)} + \sum_{j=1}^{T} e^{-\lambda (j + \text{rank}_j(u))}. \tag{8}$$

Unlike the training process for the ASHMMs as in section 2.3, all transitions are allowed at any frame since the total number of cross-state transitions in training query HMM is upper bounded by $M - 1$. The Gaussians updated in EM algorithm backoff to the Gaussians in the ASHMMs because the limited number of training instances may lead to overfitting and singular covariance matrix. For each Gaussian in Λ, assume that $\widehat{\mu}$ and $\widehat{\mathbf{\Sigma}}$ are the mean and covariance obtained at iteration t, then the update formula for μ_t and $\mathbf{\Sigma}_t$ are

$$\mu_t = \beta \cdot \mu_{t-1} + (1 - \beta) \cdot \widehat{\mu} \tag{9}$$

$$\mathbf{\Sigma}_t = \beta \cdot \mathbf{S}_{t-1} + (1 - \beta) \cdot \widehat{\mathbf{S}} - \mu_t \cdot \mu_t^T, \tag{10}$$

where $\mathbf{S}_{t-1} = \mathbf{\Sigma}_{t-1} + \mu_{t-1} \cdot \mu_{t-1}^T$ and $\widehat{\mathbf{S}} = \widehat{\mathbf{\Sigma}} + \widehat{\mu} \cdot \widehat{\mu}^T$. The parameter β controls the backoff level.

3.4. Integrating model-based and feature-based methods

After the query HMM Λ is trained, the relevance score $\text{score}_{\text{hmm}}(u)$ of a hypothesized region u is the log likelihood of u generated from Λ calculated with Viterbi algorithm. The concepts of the HMM and PRF approach are different: the HMM method relies on the statistical model that contains the variations and temporal information carried by the Q and R_Q, whereas the score calculated with PRF method is based on distances produced by $T + 1$ mutually unrelated DTW processes. Therefore integrating these two scores may further improve the detection performance. The integrated score is

$$\text{score}_{\text{int}}(u) = w \cdot \widehat{\text{score}}_{\text{hmm}}(u) + (1 - w) \cdot \widehat{\text{score}}_{\text{prf}}(u), \tag{11}$$

where $\widehat{\text{score}}_{\text{prf}}(u)$ and $\widehat{\text{score}}_{\text{prf}}(u)$ are mean and covariance normalized relevance scores, and w is the interpolation weight for HMM method.

4. Experiments

4.1. Experimental setup

We evaluated the proposed method with a spoken term detection task on broadcast news. The audio archive to be retrieved

Table 1: Performance of frame-based DTW, PRF method, HMM method and the integration.

Method	MAP (%)	P@5 (%)	P@10 (%)	p-value
(1)frame-based DTW	45.2	76.2	63.3	—
(2)PRF	48.8	78.6	66.2	baseline
(3)HMM	48.3	77.1	68.3	0.56
(4)integration	50.5	79.1	67.7	< 0.01

was the Mandarin broadcast news collected in Taiwan in August and September 2001, divided into 5034 documents. Ten spoken query terms were used as the development set for parameter selection and 42 for evaluation, all of which were collected from the speakers in the same archive. No document in the archive contained these spoken query instances. The percentage of relevant documents for each query ranged from 0.2% to 2.2%, averaged 0.5%, of the entire archive. The lengths of the test query terms were from 2 to 7 syllables, with majority of 2 to 3 syllables. The mean average precision (MAP), precision at 5 (P@5), and precision at 10 (P@10) were used to evaluate the detection performance. A hit or miss was evaluated on per document basis. That is, a hit was counted if the returned document contained a desired query. The initial number of Gaussians K_0 for segment features was 100, the number of pseudo relevant regions T was 7, and the number of training hypothesized regions T' was 15. The parameters were set as $\kappa = 0.4$ in (4) and $\lambda = 0.25$, $\beta = 0.15$ in (8)-(10).

4.2. Evaluation results

The results of the segment clustering experiment are shown in Fig. 3. The value of objective function is increased when K is decreased at each iteration, maximized at 44 Gaussians, and decreased afterwards. The optimal number of clusters (44) is close to the number of phonemes (36) in Mandarin, indicating the MDL criterion is a reasonable estimation for the number of clusters. There were on average 1.82 states in each ASHMM and total 80 states were generated and trained.

The evaluation results for frame-based DTW, PRF and HMM methods are listed in Table 1. The mean average precision (MAP) for the system using frame-based DTW is 45.2% (row(1) of Table 1). With the PRF method, MAP is improved to 48.8% (row(2)). This shows that including different signal realizations provides better detection performance. Using the proposed HMM method (row(3)), both MAP (48.3%) and P@5 (78.6%) are slightly lower than PRF method. However, P@10 is improved by 2.1%. This indicates that some relevant documents not similar to the query were included in the top 10 list by using the query HMM. The p-value of PRF and HMM methods is 0.56, showing that the performance of these two methods are in fact comparable. The results of integrating the two scores with (11) with different weight w are shown in Fig. 4. It can be seen that the integration brought benefit for w ranging from 0.1 to 0.9, with the best MAP of 50.5% achieved when $w = 0.6$. The p-value for this integration of $w = 0.6$ compared with PRF method is 9.2×10^{-3}, indicating the improvement is significant.

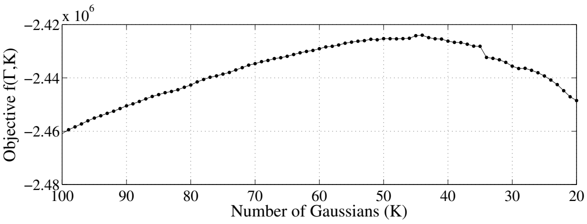

Figure 3: Cluster number selection: value of $f(\Gamma, K)$ after EM algorithm for each K.

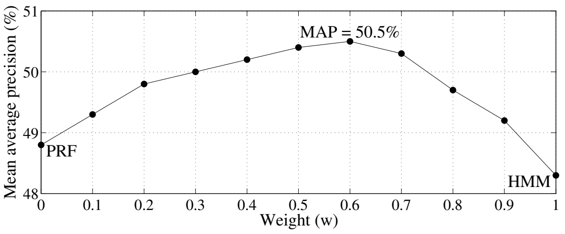

Figure 4: MAP for integrated PRF and HMM methods with different weight w.

5. Conclusions

The proposed unsupervised hidden Markov modeling technique for speech archive shows the ability of modeling signal variations and frame trajectories of speech signals in the spoken archive. The unsupervised segmentation and clustering method using MDL principal provide a reasonable estimation of total number of clusters. The query HMM constructed from ASHMMs and trained with pseudo relevant regions provides a model-based STD method that complements the feature-based DTW methods. The evaluation results show a comparable and complementing detection performance for PRF and HMM approaches, whereas a significant improvement is achieved by integrating the two.

6. References

[1] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *ASRU*, 2009.

[2] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *ASRU*, 2009.

[3] C.-a. Chan and L.-s. Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *Interspeech*, 2010.

[4] ——, "Integrating frame-based and segment-based dynamic time warping for unsupervised spoken term detection with spoken queries," in *ICASSP*, Prague, May 2011, to appear.

[5] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recoginition," in *ICASSP*, 1988.

[6] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, 2007.

[7] Y. Tsao, H. Sun, H. Li, and C.-H. Lee, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition," in *ICASSP*, 2010.

[8] C. A. Bouman, "Cluster: An unsupervised algorithm for modeling Gaussian mixtures," April 1997, available from http://www.ece.purdue.edu/˜bouman.

[9] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," in *ICASSP*, 2008.