ENHANCEMENT OF HMM THROUGH DISCRIMINATIVE ANALYSIS

Jianming Song

Department of Electrical and Computer Engineering
The University of Wollongong

ABSTRACT - Although the HMM approaches have been extensively studied in the context of speech recognition for nearly two decades, there still appears to be room for improving the discrimination capability within the HMM framework, from both training side and recognition side. The algorithm proposed in this paper integrates the concepts of variable frame rate and discriminative analysis to modify the conventional Viterbi algorithm, in such a way that the steady or stationary signal is compressed, while transitional or non-stationary signal is emphasized through the frame-by-frame searching process. The usefulness of each frame is decided entirely within the Viterbi process and needs not to be the same for different models. To evaluate this algorithm, we tested a speech database of 9 highly confusable E-set English letters. With 5 state and 6 mixture components, the conventional HMM baseline system only delivered the recognition accuracy of 73.9%. Through the use of the algorithm proposed in this paper, the recognition accuracy was increased to 82.5%.

1. INTRODUCTION

The hidden Markov modelling (HMM) has become a predominant technique for automatic speech recognition due to its modeling power to characterize the statistics of speech signal and its ability to integrate different sources of knowledge (acoustic, phonetic, lexicon, syntax, etc.) within a unified searching mechanism. Although it has been successful in dealing with a wide range of speech recognition tasks, there is a noticeable weakness inherently resided in the conventional HMM framework. In fact, due to its inadequacy in addressing the issues of discrimination and robustness, the HMM usually fails to attain high performance for difficult vocabulary. The reason for this limited discrimination capability is two-fold, i.e. the training algorithm based on the maximum likelihood estimation (MLE) and the uniform frame-to-frame searching mechanism.

Assuming we have N speech classes to model, $C = \{C_1, C_2, \ldots C_N\}$ and for each class we have a set of samples used as training data, it is well known from Bayesian decision theory that the MLE approach will lead to the asymptotically best recognition performance, if both the underlying structure (form) of the source models $P(O|\lambda_i)$ and the *prior* probability $P(C_i)$ are correct ones. However, the assumptions made by the HMM are generally not true in speech signal context. These incorrect assumptions are numerous, such as the first-order Markov chain, the form of HMM topology, the independent observation events (acoustic feature vectors), the form of probability density function on each state, and the diagonal covariance matrix in the Gaussian probability density function, etc. Due to the fact that MLE based HMM is estimated from a training database, the effect of the incorrect HMM assumptions imposed on speech may be emphasized through the training phase. Therefore the MLE approach dose not lead to the result of minimizing recognition error.

Secondly, the conventional frame-synchronous recognition performed by the Viterbi algorithm is also lack of discriminative power due to its linear and uniform process. The best global likelihood resulted from each model via dynamic programming (Viterbi alignment) is in fact simply determined by summing up the local likelihoods along the best state sequence. However, since this simple sum-up local likelihoods along the best state sequence treats the local likelihood from each frame equally and independently, it fails to make any use of distinct difference clearly demonstrated in speech signal, e.g. rapid and dynamic change in some regions, quasi periodic and stationary in other regions. Therefore, it may not provide a good discriminative capability for classifying words in an acoustically and phonetically confusable vocabulary, such as nine E-set English letters.

Thus there appears to be room for improving the discrimination capability within the HMM framework, from both training side and recognition side. In this paper we present a novel algorithm to modify the conventional Viterbi algorithm based on the likelihood distribution of each state of HMMs.

2. VARIABLE FRAME RATE ANALYSIS

The first module in any recognition system is a front-end, which extracts from an incoming utterance an adequate set of parameters, such as commonly adopted mel frequency cepstral coefficients (MFCC) and its first order time difference (ΔMFCC). Given the physical constraints of human vocal contract and phonetic structure of a language, it is clear that in speech signal, there exist some regions where changes in acoustic characteristics is dynamic and rapid, while in some other regions, the signal is much more stationary or quasi periodic. Therefore, it is expected that some successive feature vectors will be very similar, while some other successive feature vectors will be quite different, hence it is possible to utilize a similarity metric and then process each frame in the vector sequence in a discriminative way.

Conventional variable frame rate analysis (Bridle *et al* 1983, Ponting *et al* 1991) is to retain all the input feature vectors when they are changing most rapidly and omit the high proportion when they are relatively constant. The typical techniques used to achieve this is through calculation of some pre-defined similarity measure, such as Euclidean distance, between two feature vectors. Specifically, assuming a vector sequence is represented by $X = \{x_1, x_2, \ldots x_N\}$ and the distance between previously selected frame i and following frame j is represented by $D(i,j)$, the decision whether the frame j should be kept or drop is based on a pre-calculated threshold T in Figure 1.

Although the idea behind this approach is to enhance the important region of speech signal by removing out some highly stationary part of signal, it is *ad hoc* and lacks a clearly defined optimality criterion to perform the task. Furthermore it is not compatible with the stochastic framework of the HMM. Therefore, this variable frame rate analysis is not very useful in the current speech recognition systems. In fact, even the concept of the variable frame analysis is quite ignored by most researchers working on HMM based speech recognition.

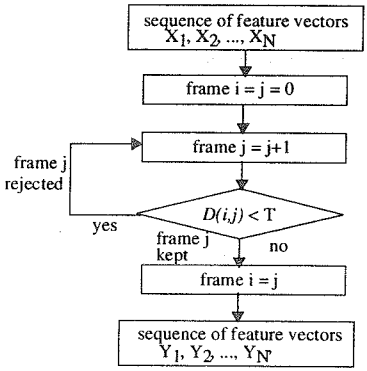

FIGURE 1. Flow diagram of the conventional frame compression approach

3. CONVENTIONAL VITERBI DECODING

The most popular mechanism adopted in speech recognition systems is based on the Viterbi algorithm. It calculates the accumulated likelihood across each state and each model in a frame synchronous manner. For each frame of signal (acoustic feature vector), the Viterbi algorithm updates and propagates all the accumulated likelihoods according to the assumption of the first-order Markov chain, the HMM topology, and the local probability values. Thus the Bayes decision to recognize an utterance within Viterbi searching context can be formulated as

$$\lambda_{best} = \arg\max_{\lambda_i} \{P(\lambda_i | X)\}$$

$$\log P\left(\lambda_i|X\right) = \log\left(\frac{P(\lambda_i)\,P(X|\lambda_i)}{P(X)}\right) \approx c + \max_{S} \log P(X, S|\lambda_i) \approx c + \sum_{t=1}^{T} \log b_{q_t}(x_t)$$

where the prior language model is assumed to be constant c (no grammar), and the best state sequence is represented by $S_{best} = \{q_1, q_2, \ldots q_T\}$ for model λ_i.

Thus the decision making in a non-grammar isolated word recognition system is approximated as

$$\lambda_{best} = \arg\max_{\lambda_i} \left\{ \sum_{t=1}^{T} \log b^{\lambda_i}_{q^{\lambda_i}_t}(x_t) \right\}$$

It is important to note that the above formulation is not conventional, in the way that the recognition score dose not depends on transition probabilities. The reason for that is because the transitional probabilities play an insignificant role due to their dynamic numerical range compared with local likelihood, thereby they can be ignored altogether without any notable effect on recognition accuracy (Song *et al* 1993).

Further examining the above formulation gives more insight of the recognition mechanism resided in the HMM framework. Assuming the local likelihood in each state of each model as an observation event, i.e. $Y = \{y_1, y_2, \ldots y_T\}$, it becomes clear that the discrimination function formulated after time-state warping through the Viterbi decoding is a simple sum-up discrimination function, i.e.$g(Y, \Lambda) = \sum y_i$, Λ is defined as an HMM space. This simple sum-up through the frame-to-frame and state-to-state likelihood propagation in the Viterbi algorithm has two problems, i.e. firstly, the estimation error from the non-discriminative states may offset the likelihood difference from discriminative states, secondly, the difference in likelihood generated from stationary frames may be more important than the difference generated from non-stationary frames. In fact, when tested on a highly confusable vocabulary consisting of nine English E-set letters, the conventional HMM approach with 5 Gaussian mixture components, 5 states and 24 dimensional feature vector (cepstrum + Δcepstrum) only produced word accuracy of 61.7% on telephone data (Kitagiri *et al*, 1993).

It seems to exist some room for us to reduce the recognition error rate by modifying the conventional recognition process presented above. In fact, if we view the HMM as a sophisticated mapping which transforms a sequence of observation events (acoustic feature vectors) to another sequence of observation events (local likelihood), it is possible to design a more powerful linear discrimination function in the forme of $g(Y, \Lambda) = \sum_i w_i y_i$, where $W = \{w_1, w_2, \ldots w_T\}$ is a weighting function. Recently a promising and powerful method called generalized probability decent (GPD) with an optimality criterion of directly reducing the recognition error rate on training data was under intensive investigation (Juang *et al*, 1993). Although the mathematics involved in the GPD is quite complex, it is not difficult to see that the optimized weighting function will emphasize the rapid changing signal resided on discriminative states and de-emphasize the stational signal from non-discriminative states.

4. VITERBI DECODING WITH DISCRIMINATIVE ANALYSIS

It is well known that speech signal is generally a non-stationary process which can be approximated by a short-term stationary chain. From the probability perspective, a segment of signal which is locally stationary will be characterized by a same probability distribution function. This characteristic is well represented within HMM, i.e., local stationarity of speech signal is assumed to be located on same state through self-loop transition, with non-stationarity being represented by state-to-state transition (Gurgen *et al*, 1993).

Although the degree of stationarity or redundancy embedded in a sequence of acoustic feature vectors $X = \{x_1, x_2, \ldots x_N\}$ is fixed and only depends on speech signal, it is very difficult, if not impossible,

to be measured through conventional distance metric, such as Euclidean distance, between two feature vectors. Thus frame compression manipulated in the feature extraction module might not be an effective solution to improve the discrimination capability of a recognition system. Much more insight can be gained if we analyze the sequence of the vectors within the HMM framework and evaluate the degree of stationarity based on each HMM model individually. In doing so, it is more reliable to decide whether x_t is a rapid changing signal or a stationary piece of the surrounding signal. Thus it is not difficult to see that a larger consistency is met with the stochastic pattern matching paradigm, which is the key aspect in applying this broad methodology.

To evaluate each frame of signal (vector) in a sensible way, we can align this vector to its most likely state and evaluate it from that point based on the transition pattern and the link to its previous frame. There are a number of ways to decide the most likely state on which x_t is located. To add the new feature of discriminative analysis into the well established Viterbi decoding algorithm, the most likely state is determined for each model at each frame, i.e.

$$s_t^{\lambda, best} = \arg \max_i \delta_t^{\lambda}(i) \quad \delta_t^{\lambda}(i) = \max_{s_1, s_2, s_{t-1}} P(x_1, x_2, \ldots x_t, s_1, s_2, \ldots s_t = i | \lambda)$$

It is not difficult to see that the index of the most likely state corresponds to the acoustic vector x_t for different model needs not to be the same. This gives a larger degree of freedom for the HMMs to process signal in a unified decision making criterion. Once the most likely state within a model λ is located, the effectiveness of a vector x_t evaluated by that model is decided by the following rules.

Rule 1 - transition type on the best state of each model

The stationarity of a vector is indicated by the self transition loop within a HMM as

$$\delta_t^{\lambda}(s_{best}) = \max_i \delta^{\lambda}_{t^\circ}(i)\, b^{\lambda}_{s_{best}}(x_t) = \delta^{\lambda}_{t^\circ}(s_{best})\, b^{\lambda}_{s_{best}}(x_t)$$

where t° denotes previous frame index which is an active frame and contributes to the accumulated likelihood.

x_{t° is referred to as reference frame with respect to x_t. It can be seen from the above equation that the transitional probabilities are not used. The reason for this modification has been mentioned earlier in this paper.

When the transition on the best state s_{best} is self-loop, the degree of stationarity with respect to its previous frame needs to be further examined. Only when the current frame is sufficiently close to the previous frame in a sense of likelihood, one can safely say that the information it carries is insignificant and therefore can be compressed.

Rule 2 - local log likelihood distance

The local likelihood $\log b^{\lambda}{}_i(x_t)$ is a measure of the probability that vector x_t is generated from state *i* of model λ. x_t is considered to be stationary with respect to previous reference frame x_{t°, if the best state determined by the rule 1 is of a self transition and the difference in magnitude between two local likelihoods is below a model-state-dependent threshold, i.e.

$$\left| \log b_{Sbest}{}^{\lambda}(t) - \log b_{Sbest}{}^{\lambda}(t^\circ) \right| < T^{\lambda}(s_{best})$$

where $T^{\lambda}(s_{best})$ is a model-state-dependent threshold (MSDT) estimated from HMM training stage.

The recognition process within the Viterbi algorithm needs to be modified. If any of the above-mentioned conditions (rules) are broken, the feature vector is considered active and therefore is used to update all the accumulated likelihoods within the model λ, the frame index t° is updated to be *t*. However, when both rule 1 and rule 2 are valid, x_t is not used to update the accumulated likelihood on any

state of the model λ. That is, all the internal likelihoods keep the values decided at the frame t°. In this case, the effective number of signal frames with respect to the model λ is decremented by 1 and the active reference frame index t° remained unchanged with respect to model λ. To determine the rank in the recognition, the global likelihood resulted from each HMM should be normalized with its effective length.

In the HMM training stage, the model and state dependent threshold $T^{\lambda}(i)$ is estimated. The basic assumption here is, once again, that frames of a stationary signal should exhibit similar probability characteristics within HMM, i.e. they should be located on the same state, as well as their local likelihood should be close to each other. The conventional Viterbi algorithm is used to align acoustic feature vectors to the best sequence of states based on a set of well trained HMM. Without loss of generality, assuming that number of vectors located on state i of model λ is N_i. The number of different absolute differences in logarithmic likelihood between any two vectors x_k, x_l is $N_i \times (N_i - 1)/2$. These local likelihood differences (in magnitude) in each state of each model can be viewed as a sequence of discrete observation events $Y = \{y_1, y_2, \ldots y_M\}$, and be modeled through non-parametric or parametric way. In our preliminary implementation, the mean and the standard deviation of each sequence is estimated individually for each state of each model, i.e.

$$\mu = \frac{1}{M}\sum_n y_n \qquad \sigma = \sqrt{\frac{1}{M}\sum_n (y_n - \mu)^2}$$

Since the mean value is close to zero due to the relation of $E\{y - z\} = \mu_y - \mu_z$, y and z are two random variable, the model and state dependent threshold is formulated as $T^{\lambda}(i) = \theta\sigma^{\lambda}(i)$, where q is a coefficients determined empirically from training data.

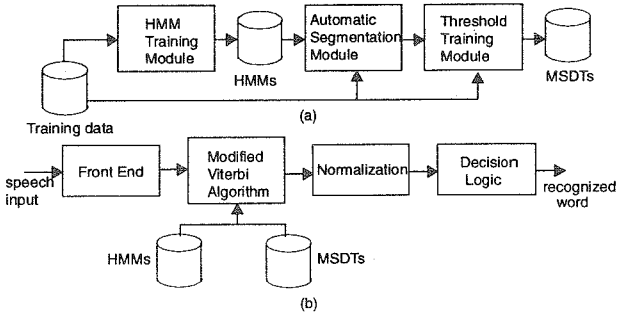

FIGURE 2. Block diagram for (a) Training phase (b) Recognition phase

5. EXPERIMENT SETUP

To evaluate the proposed algorithm, we chose a highly confusable vocabulary consisting of nine American English alphabet letters, i.e. *{b, c, d, e, g, p, t, v, z}*. The speech material used in our experiments was the standard database TI-46 distributed by NIST. It was constructed in multi speaker, isolated mode and contained 16 speakers: 8 males and 8 females. Each speaker produced 10 utterances of each letter as training data and 16 utterances of each letter as testing data. All speech tokens were recorded in a sound proof condition at sampling frequency of 12.5 KHz. To be compatible with telephone bandwidth, the original speech data were down-sampled to 8 kHz. The analysis condition and the system parameters for the based line system is presented as follows.

Front End:
- Telephone bandwidth
- Time window 30 ms with frame shift 10 ms
- MFCC derived from FFT + Mel filter bank + DCT + Cepstrum liftering

Acoustic feature vector: 12 MFCCs + 12 ΔMFCCs

HMM:

Type: Continuous density HMM

Topology: 5-state left-to-right.

PDF type: mixture of 6 Gaussian pdfs with diagonal covariance.

Segmental K-means clustering.

Additional parameters:

Model-dependent state-dependent threshold $T^{\lambda}(i)$ for each model and each state.

A conventional HMM recognition system is based on the one described in (Song *et al* 1993), with 5 left-to-right states and 6 mixture components per state being used to get a baseline performance. The recognition accuracy based on this set of models and conventional frame-synchronous Viterbi decoding algorithm was for the multi-speaker testing database. The comparison of the baseline performance with the result using the algorithm proposed in this paper in illustrated in Table 1.

TABLE 1. performance comparison for baseline HMM recognizer and discriminative HMM recognizer.

Approach	5 state 6 mixture component HMMs		
	correct samples	mis-recognized samples	recognition accuracy (%)
Baseline HMM	1692	598	73.9
Discriminative HMM	1889	401	82.5

6. CONCLUSION

The algorithm proposed in this paper is inspired by the concepts of the variable frame rate and discriminative analysis on the sequence of accumulated likelihood. The basic assumption of this algorithm is that the score contribution of each frame of signal matched with each HMM model should be treated discriminatively and depends on previous frames. Unlike either the conventional variable frame rate analysis which compresses signal frames based on a simple spectrum distance metric at the front-end stage, or the Viterbi algorithm which processes each frame of signal in an uniform way, this algorithm examines each frame of signal within the recognition module with enhanced discrimination capability. Whether this frame will contribute to the final likelihood score will be decided by each state of a model and the degree of stationarity. In this algorithm, a particular frame could be rejected by a model, while it is considered useful by another model. If a particular frame is considered redundant to a model to which it is evaluated, the effective length of the signal (number of frame) will be decremented by one from the initial length. Therefore, the final lengths of signal resulted from different models need not to be the same. This processing is more powerful and gives greater degree of freedom to explore the inherent stochastic modelling power within HMM.

The experiment presented in this paper was carried out by the author of this paper at the Speech Technology Research Group, Department of Electrical Engineering, University of Sydney.

REFERENCES

Bridle, J.S, Brown, M.D. (1982) *A Data-adaptive Frame Rate Technique And Its Use In Automatic Speech Recognition*, Proc. Institute Acoustics Autumn Conference, Bournemouth, UK, pp. C2.1-C2.6.

Chang, P.C, Juang, B.H. (1993) *Discriminative Training Of Dynamic Programming Based Speech Recognizers, IEEE Trans.* Speech and Audio Processing Vol. 1, No.2, April., pp. 135-143.

Gurgen, F, Song, J, R. King (1994) *A Continuous HMM Based Preprocessor For Modular Speech Recognition Neural Net Works*, Proc. of International Conference on Spoken Language Processing, Japan.

Kitagiri, S, Lee, C.H. (1993) *A New Hybrid Algorithm For Speech Recognition Based On HMM Segmentation And Learning Vector Quantization, IEEE Trans.* Speech and Audio Processing Vol. 1, No. 4, October., pp.421-430.

Ponting, K.M, Peeling, S.M. (1991) *The Use Of Variable Rate Analysis In Speech Recognition*, Computer Speech and Language, Vol. 5, pp. 169-179.

Song, J, Samouliean, A. (1993) *A Robust Speaker Independent Isolated Word HMM Recognizer For Operation Over The Telephone Network*, Speech Communication, Vol.13, December, pp. 287-295.