ENHANCEMENT OF DISCRIMINATIVE CAPABILITIES OF HMM BASED RECOGNIZER THROUGH MODIFICATION OF VITERBI ALGORITHM

Jianming Song

Department of Electncal and Computer Engineerıng
The University of Wollongong
NSW 2521, Australia

ABSTRACT

The algorithm proposed in this paper integrates the concepts of variable frame rate and discriminative analysis based on Tanimoto ratio 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%. In the preliminary experiment using 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 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 twofold, i.e. the training algorithm based on the maximum likelihood estimation (MLE) and the uniform frame-to-frame searching process

Assuming the class set of a model inventory is $C = \{C_1, C_2, \dots C_N\}$ 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, only if both the underlying structure (form and parameters) 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, the diagonal covariance matrix in the Gaussian probability density function, etc. Due to the fact that the MLE based HMM is estimated from a training database, the effect of the incorrect HMM assumptions imposed on speech may be emphasized through training phase. Therefore the MLE approach, even though it yields a converged (optimal or suboptimal) model in the sense of the maximum likelihood, dose not lead to the result of minimizing recognition error.

Furthermore, the conventional frame-synchronous recognition performed by the Viterbi algorithm is also lack of discriminative power due to its uniform process. The best global likelihood resulted from each model via dynamic programming (Viterbi alignment) is in fact well approximated 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 by selectively making use of speech frames in the searching process.

2. THE PROBLEM WITH 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 and the first order time differences. 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 are very similar, while other successive feature vectors exhibit dynamic change in characteristics. Thus it is possible to utilize a similarity metric and process each frame in the vector sequence in a discriminative way.

Conventional variable frame rate analysis [1][2] 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 predefined similarity measure, such as the Euclidean distance, between two feature vectors. Specifically, assuming a vector sequence is represented by $X = \{x_1, x_2, \dots x_N\}$ and the distance between previously selected frame i and following frame j is represented by

0-7803-2431-5/95 $4.00 © 1995 IEEE

$D(i, j)$, the decision whether the frame j should be kept or dropped is based on a predetermined threshold T illustrated in Figure 1.

Although the idea behind this approach is to enhance the important region of speech signal by removing out some highly stationary parts of signal, it 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 seldom used in the current speech recognition systems.

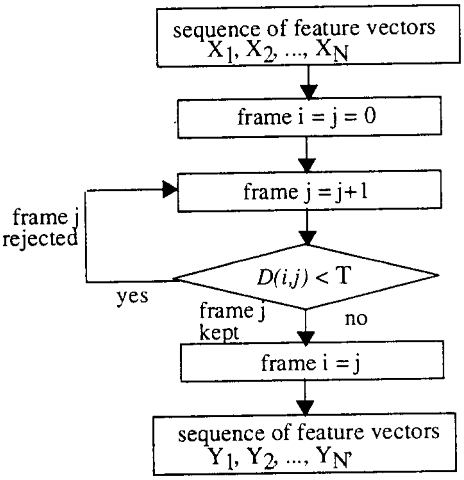

FIGURE 1. Flow diagram of the conventional frame compression approach

3. THE PROBLEM WITH CONVENTIONAL VITERBI DECODING

The most popular searching process adopted in the HMM based speech recognition systems is 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 accumulated likelihoods according to the assumption of the first-order Markov chain, the HMM topology, and the local probability values, etc. Thus the Bayes decision to recognize an utterance within Viterbi searching context can be formulated as

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

$$\log P(\lambda_i|X) = \log\left(\frac{P(\lambda_i)\,P(X|\lambda_i)}{P(X)}\right) \approx c + \max_S \log P(X, S|\lambda_i)$$

$$\approx c + \sum_{t=1}^{T} \log b_{s_t}^{\lambda_i}(x_t)$$

where the prior language model is assumed to be constant C (no grammar), and the best state sequence is represented by $S^{\lambda_i}_{best} = \{s_1, s_2, \ldots s_T\}$ for model λ_i.
Thus the decision making in a non-grammar isolated word recognition system is approximated as

$$\lambda_{best} = \arg\max_{\lambda_i} \left\{\sum_{t=1}^{T} \log b_{s_t}^{\lambda_i}(x_t)\right\}$$

Note that the above formulation is not conventional, in the way that transition probabilities are ignored. The reason for this is because transitional probabilities of left-to-right HMMs 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 [3].

This simple sum-up through the frame-to-frame and state-to-state likelihood propagation in the Viterbi algorithm possesses a very interesting characteristic, i.e. the matching process can be viewed as a sophisticated model dependent transformation, which transforms each acoustic vector x_t to a scalar quantity $\log b_{s_t}^{\lambda_i}(x_t)$. The rank of recognition candidates is determined by the sum of the transformed values. However, this paradigm has several problems, such as the estimation error from non-discriminative states may offset likelihood difference from discriminative states, the difference in likelihood generated from stationary frames may be more important than the difference generated from non-stationary frames. Some approaches based on discriminative analysis are proposed to improve the discriminatively of current recognizer, without modifying the existing framework significantly [4][5][6].

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 [7].

Although the degree of stationarity or redundancy embedded in a sequence of acoustic feature vectors $X = \{x_1, x_2, \ldots x_N\}$ is inherently fixed and only depends on speech signal, it is 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. However, 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.

To evaluate the stationarity degree of consecutive signal frames (vectors), we need to align each vector to its most likely state and to further evaluate it based on the transition pattern and the link to its previous frame from that stage. This frame-to-state alignment is performed by the Viterbi decoding algorithm, i.e.

$$s^{\lambda_i}_{best} = \arg\max_i \delta_t^{\lambda_i}(i)$$

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

The partial best state sequence with respect to model λ_i is linked with the sequence of local likelihood $Y_t = \{y_1, y_2, \ldots y_t\}$ where

$y_t = \log b_{s_t}^{\lambda_i}(x_t)$

To measure the degree of local stationarity of the consecutive frames aligned to the same state of a HMM, we use a metric called Tanimoto ratio defined as

$$d(Z_a, Z_b) = \frac{Z_a^T Z_b}{Z_a^T Z_a + Z_b^T Z_b - Z_a^T Z_b}$$

where Z_a and Z_b represent two vectors containing adjacent local likelihood aligned on the same state, $Z_a^T Z_b$ denotes the number of common attributes between Z_a and Z_b, $Z_a^T Z_a$ denotes the number of attributes possessed by Z_a. The denominator then gives the number of the attributes that are in Z_a or Z_b but not in both. The entire representation represents the ration of the number of common attributes between Z_a and Z_b to the number of attributes that are in either one of the vectors Z_a or Z_b but not in both. Thus it gives rise to a measure for the stationarity degree between local likelihood segments aligned on the same state. For each frame index t and each model λ, if the value of the Tanimoto ratio is larger than a state dependent threshold, the correlation between proceeding frame and current frame is assumed high. As a result, this frame is considered redundant with respect to the state concerned.

The advantage of this kind of processing is twofold. First, the incorrect assumption of the independent and identical distribution (IID) of frames located on the same state made by the conventional HMM is somehow mended by taking into account of the correlation between adjacent local likelihood on the same state. Second, in contrast to the single choice of the redundant stationary frame rejection made by the conventional variable frame analysis, the stationarity of adjacent frames is examined individually within each HMM model under consideration. This will therefore provide a larger degree of freedom for the HMMs to process signal frames.

The two-stage process of this algorithm is described as following:

Stage 1 - Transition type on the best state of each model

For each frame index t and each model λ, find out the most likely state, i.e.

$$s_{best} = \arg \max_s \delta_t^\lambda(s) = \arg\max_s \left\{ \max_i \delta^\lambda_{t^\circ}(i)\, b^\lambda_s(x_t) \right\}$$

where t° denotes the frame index of the active frame x_{t° which contributes to the accumulated likelihood.

If the frame x_t located on the best state s_{best} is of a self-loop transition, the stationarity of this vector is suggested and the degree of stationarity with respect to its previous frame needs to be further examined.

Stage 2 - Degree of local stationarity on the best state

The degree of local stationarity of consecutive frames located on the most likely state is quantified by a simplified Tanimoto ratio, with Z_a and Z_b representing previous active local likelihood $\log b_{s_{best}}^{\lambda}(t^\circ)$ and current local likelihood $\log b_{s_{best}}^{\lambda}(t)$ respectively.

The contribution of the current frame x_t to the accumulated likelihood is therefore determined by the following logic.

$$\delta_t^\lambda(s) = \begin{cases} \max_i \delta^\lambda_{t^\circ}(i)\, b^\lambda_s(x_t) & if\ d(Z_a, Z_b) > T^\lambda(s) \\ \delta^\lambda_{t^\circ}(s) & else \end{cases}$$

where $s = 1, 2, \ldots N$ represent state index of a HMM, and $T^\lambda(s)$ denotes a model-state-dependent threshold (MSDT) estimated from HMM training stage.

Thus when the process is of self-loop transition and the current frame is sufficiently close to the previous frames in a sense of likelihood, one can be sure that the information it carries is insignificant and therefore can be compressed. The recognition process within the Viterbi algorithm needs to be modified. If a frame of speech is decided not to used in the computation of the accumulated likelihood, all the partial likelihood associated with each state is unchanged, i.e. all the values remains the same as evaluated at frame t°. Otherwise, the conventional likelihood update is applied, and the active frame index t° is set to be the current frame index. It is not difficult to see that the actual number of signal frames used in the calculation of final likelihood score is generally different for different model. To determine the rank in the recognition, the global likelihood resulted from each HMM should be normalized with its effective length. This task can be easily accomplished by counting effective frames in the matching process for each model.

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 the acoustic feature vectors of each training token to the best state sequence. The average value of the Tanimoto ratio calculated on local likelihood distributed on each state is then weighted empirically as $T^\lambda(i)$.

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.

Telephone bandwidth: Sampling frequency of 8 kHz.
Front End parameters:
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.
Training Method: Baum-Welch algorithm
Additional parameters:
Model-dependent state-dependent threshold $T^{\lambda}(i)$ for each model and each state.

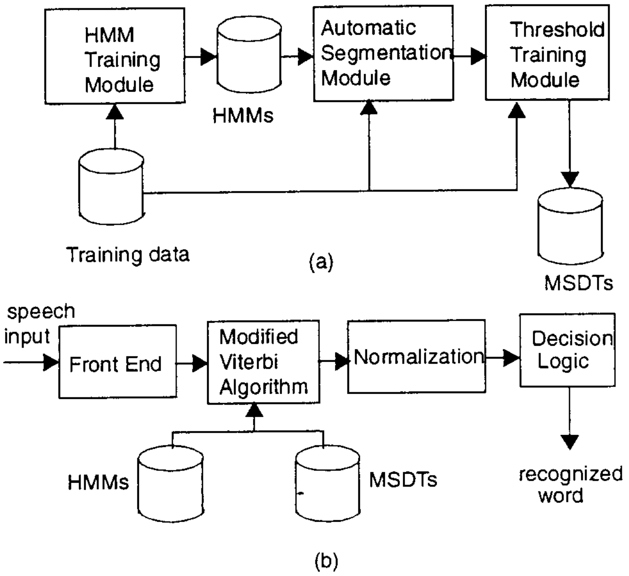

FIGURE 2. Block diagram for (a) Training phase (b) Recognition phase

A conventional HMM recognition system is based on the one described in [3], 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 the degree of stationarity of each signal frame by the HMM transition pattern and the value of Tanimoto ratio. Whether this frame will contribute to the final likelihood score will be decided by the state transition pattern 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.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] L. Deng, K. Hassanein, M. Elmasry, (1994) "*Analysis Of The Correlation Structure For A Neural Predictive Model With Application To Speech Recognition*", Neural Networks, Vol. 7, No. 2, pp. 331-339.

[7] 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.