SPEECH RECOGNITION USING A STRONG CORRELATION ASSUMPTION FOR THE INSTANTANEOUS SPECTRA

J. Ming, P. O'Boyle, J. McMahon, and F. J. Smith

School of Electrical Engineering and Computer Science
The Queen's University of Belfast
Belfast BT7 1NN, UK

ABSTRACT

The conventional independence assumption made for the evolving speech spectra is replaced by a strong correlation assumption, which then leads to a new stochastic model. This model implements a nonlinear interpolation between the lower and upper bounds of the joint probability distributions. The advantage of the new model over other correlation-based modelling approaches is that it has a low parameter complexity, the same as that in models based on the independence-assumption. Experiments on a speaker-independent E-set database show the effectiveness of this new modelling approach.

1. INTRODUCTION

Given a time series of instantaneous speech spectra $X=(x_1, \cdots, x_T)$, the problem of speech recognition can be formulated as an estimation of the joint probability $P_\Lambda(X)$, based on some probabilistic model of X with parameter set $\Lambda=(\Lambda_1, \cdots, \Lambda_T)$. For the estimation of this joint probability, it is known that

$$\prod_{t=1}^{T} P_{\Lambda_t}(x_t) \le P_\Lambda(X) \le \min\left\{P_{\Lambda_1}(x_1), \cdots, P_{\Lambda_T}(x_T)\right\} \quad (1)$$

Therefore, there are at least two ways to obtaining an approximate of $P_\Lambda(X)$. The first is to use the lower bound, where $x_1, \cdots, x_T$ are assumed to be independent or weakly correlated; and the second is to use the upper bound, which is valid if there is a strong correlation between x_t's. It would be ideal to calculate $P_\Lambda(X)$ directly. However, the large parameter sizes, and the lack of an adequate joint probabilistic model, usually make this calculation impractical, if not impossible [1-2].

Most traditional approaches, e.g. HMM, use the lower-bound approximation. This has the advantage of low model complexity but the disadvantage of losing the temporal correlation (i.e. dynamic spectral) information. Extensive studies have been carried out toward improving this disadvantage. Typical approaches include the bigram model [3, 4], the vector-valued linear-predictive model [5], and a family of segment-based models (see, for example, [1, 6-8]). In spite of the differences, most of these approaches perform the modelling of a segment of speech frames by directly calculating some type of conditional or joint probability distributions, and therefore are inevitably challenged by the increasing parameter sizes with the increase in the length of the segments being modelled. In this paper we study the modelling of the correlation in a speech spectral sequence from a different point of view, from the upper bound of the joint probability distributions (i.e. the right-hand side of inequality (1)). This study is built upon the well-known fact that there is strong correlation between the evolving speech spectra. An obvious advantage of this new approach, in comparison to all other existing approaches, is the promising low parameter complexity: the same as that in the independence-assumption based models.

2. A CORRELATION MODEL

For the estimation of the parameter set, a continuous and differentiable probability function of Λ, which has as its limits both the lower and upper bounds of inequality (1), is defined as

$$P_\Lambda(X) = \exp\left(-\left[\sum_{t=1}^{T}\left(-\ln P_{\Lambda_t}(x_t)\right)^\beta\right]^{\gamma/\beta}\right) \quad (2)$$

where it is assumed that $0 \le P_{\Lambda_t}(x_t) \le 1$, i.e. $P_{\Lambda_t}(x_t)$ is a probability measure. It can be shown that (2) reduces to the lower bound of inequality (1) when $\beta=1$ and to the upper bound when $\beta \to \infty$. It can also be shown that different values of the parameter β between unity and infinity approximate to various degrees of assumed correlation within the spectral sequence, from the zero correlation when $\beta=1$ to the maximum possible correlation when $\beta \to \infty$. In the following we discuss some variants of (2) in dealing with the actual speech spectral sequences.

2.1. Accommodation of Sequences of Variable Lengths

To accommodate the variable length of the observed spectral sequences, following the traditional HMM method, it can be assumed that each sequence is segmented into M segments, where M is a prechosen index, and that the instantaneous spectra in each segment are subject to an identical distribution, of the segment-based parameter set. Thus, (2) can be rewritten as

$$P_\Lambda(X) = \exp\left\{ -\left[\sum_{i=1}^{M} \sum_{t \in T_i} \left(-\ln P_{\Lambda_i}(x_t) \right)^\beta \right]^{1/\beta} \right\} \qquad (3)$$

where Λ_i is the parameter set of the i'th segment, $T_i = \{t : x_t \in i\text{'th segment}\}$, $i = 1, \cdots, M$, and $\Lambda = (\Lambda_1, \cdots, \Lambda_M)$. The segmentation of the sequences can be performed based on the conventional Viterbi algorithm, or based on the linear sampling/segmentation principle, as is used in stochastic segment models on phoneme levels [1-2].

2.2. Representation of the Independence between Spectral Vector Components

For certain type of spectral parameter vectors such as the cepstral coefficients, it can be reasonably assumed that the individual components within each vector are independent. This results in the diagonal-covariance observation structure in the conventional HMM's. The same assumption can be made for the model defined in (3). However, instead of using (3) with each $P_{\Lambda_t}(x_t)$ of a diagonal type covariance matrix, we decompose the vector-valued time series, $x_1, \cdots, x_T$, into N component time series, $\{x_1^n, \cdots, x_T^n\}_{n=1, N}$, where N is the size of each vector, and then model each of the one-dimensional time series by a joint probability function of the form as defined in (3). The individual component time series are assumed to be independent of each other. Therefore we have the model

$$P_\Lambda(X) = \prod_{n=1}^{N} P_{\Lambda(n)}(x_1^n, \cdots, x_T^n) \qquad (4)$$

where each $P_{\Lambda(n)}(x_1^n, \cdots, x_T^n)$ is represented by

$$P_{\Lambda(n)}(x_1^n, \cdots, x_T^n) = \exp\left\{ -\left[\sum_{i=1}^{M} \sum_{t \in T_i} \left(-\ln P_{\Lambda_i(n)}(x_t^n) \right)^\beta \right]^{1/\beta} \right\} \qquad (5)$$

and $\Lambda(n) = (\Lambda_1(n), \cdots, \Lambda_M(n))$ is the parameter set corresponding to the n'th spectral component sequence. Using (4) and (5) rather than (3) with diagonal covariance matrices has been found to be beneficial in the training process. Since in (4) and (5) we work with the scalar space, it becomes easier to obtain convergent estimates of the model parameters.

3. TRAINING OF THE MODEL

In the training stage, the model parameter set $\Lambda = \{\Lambda_1(n), \cdots, \Lambda_M(n)\}_{n=1, N}$ is estimated based on (4) and (5) by using the classic gradient descent algorithm, given the segmentation of the training sequences and the parameter β. Assuming multiple independent training sequences $\{X^k\}$, it can be shown that the derivative of

$$J(\Lambda) = -\sum_{n=1}^{N} \sum_{k} \ln P_{\Lambda(n)}(x_1^{n,k}, \cdots, x_T^{n,k}) \qquad (6)$$

with respect to each $\Lambda_i(n)$, $1 \le i \le M$, $1 \le n \le N$, is given by

$$\frac{\partial J(\Lambda)}{\partial \Lambda_i(n)} = -\sum_{k} w_{n,k} \sum_{t \in T_i^k} h_{i,n,k,t} \frac{dP_{\Lambda_i(n)}(x_t^{n,k}) / d\Lambda_i(n)}{P_{\Lambda_i(n)}(x_t^{n,k})} \qquad (7)$$

where

$$w_{n,k} = \left[\sum_{j=1}^{M} \sum_{t \in T_j^k} (-\ln P_{\Lambda_j(n)}(x_t^{n,k}))^\beta \right]^{1/\beta - 1} \qquad (8)$$

and

$$h_{i,n,k,t} = (-\ln P_{\Lambda_i(n)}(x_t^{n,k}))^{\beta - 1} \qquad (9)$$

Therefore each $\Lambda_i(n)$ is updated, at the r'th iteration, according to $\Lambda_i^r(n) = \Lambda_i^{r-1}(n) - \mu \partial J(\Lambda) / \partial \Lambda_i(n) \big| \Lambda = \Lambda^{r-1}$, where μ is the updating stepsize.

In our experiments we assume that each individual $P_{\Lambda_i(n)}(x_t^n)$ is a scalar Gaussian function, therefore it is easy to obtain the last part of (7), i.e. the derivatives of the Gaussian function with respect to the mean and variance components, respectively. To ensure that $P_{\Lambda_i(n)}(x_t)$ is a probability, a minimum variance floor can be set such that $0 \le P_{\Lambda_i(n)}(x_t^n) \le 1$ holds for all x_t^n. Another alternative is to define the individual probabilities as $P_{\Lambda_i(n)}(x_t^n) \Delta x^n$, where Δx^n is a small positive constant, such that $P_{\Lambda_i(n)}(x_t^n) \Delta x^n$ approximates to the probability measure $\text{Prob}(-1/2\, \Delta x^n \le x_t^n \le 1/2\, \Delta x^n)$. Both methods were tested and they were shown to give similar results. Particularly, with cepstral lifting, which results in approximately normalized, yet increased, variances over all the cepstral coefficients, a constant

$\Delta x^n = 1$ can be chosen without numerical difficulties in the calculation.

In the experiments, the initial values of the parameters are set to those of the corresponding models assuming observation independence. It is observed in the experiments that the above training algorithm achieves convergence within 30-50 iterations.

4. EXPERIMENTS AND DISCUSSION

Experiments are performed based on a speaker-independent E-set (b, c, d, e, g, p, t, v) database, extracted from the Connex Alphabet Database provided by British Telecom Research Laboratories. This database contains 104 speakers, each speaker contributing 3 utterances to each word. Among the 104 speakers, 52 have been designated for training and the remaining 52 for testing. About 155 utterances are available for training a model for each word and, for the E-set part, a total of 1219 utterances are available for testing. The speech is divided into frames, each of a span 25.6 ms, on which the instantaneous spectra are calculated. We chose the 12th-order mel-frequency cepstral coefficients (MFCC's) as the spectral parameters for each frame.

4.1. Experiments Using Static Spectra

Firstly, we examine the new model using the static spectral parameters, MFCC's, alone. In this experiment, the parameters of the correlation model are initialised by a left-to-right HMM, and accordingly, the segmentation of the observation sequences in both training and recognition is performed with the HMM using the Viterbi algorithm. The state-transition probabilities of the HMM are found to be insignificant for the HMM (about $\pm 1\%$ differences in the performance) and therefore are ignored, which thus makes the comparison of the two models be focused on their respective observation structures, independence or correlation. Figure 1 shows the recognition results of the correlation model (CM) compared to that of the HMM for some typical values of the parameter β. Different numbers of segments (states) are examined.

4.2. Incorporation of Delta Spectra

We then examine the correlation model by incorporating the delta spectral parameters, ΔMFCC's, calculated over ± 2 frames. Denoting by ΔX the delta spectral sequence, it is assumed in the correlation model that $P(X, \Delta X) = P(X)P(\Delta X)$, where $P(\Delta X)$ is the joint probability distribution of ΔX, which is defined to be of a similar form to that in (4) and (5). The parameters of $P(\Delta X)$ then can be estimated separately using the same algorithms as described in Section 3. The experimental results for the 15 segment/state case are shown in Figure 2, where $\beta(X)$ and $\beta(\Delta X)$ represent the parameters β's used in $P(X)$ and $P(\Delta X)$, respectively.

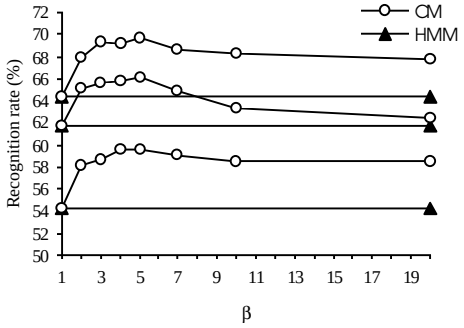

Figure 1: Comparison between the correlation model (CM) and HMM using MFCC's alone. The pairs of curves from top to bottom correspond to 15 segments/states, 10 segments/states, and 5 segments/states, respectively.

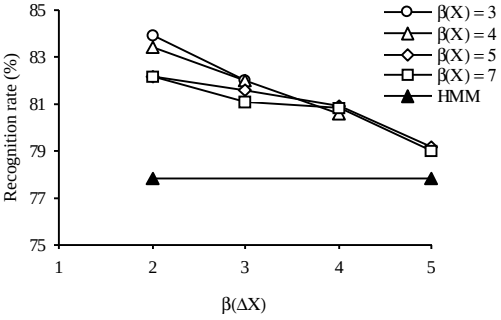

Figure 2: Comparison between the correlation model and HMM incorporating ΔMFCC's for the 15 segment/state case.

4.3. Discussions

Both Figure 1 and Figure 2 have shown improvements of the new model over the independence-assumption based models. As indicated in Section 1, the strong correlation assumption leads to the association of the joint probability of a time series with its individual observation probabilities of smaller values. The recognition based on the strong correlation assumption thus depends mainly on the small probability events occurring in a sequence of observations for a given model. Since the small probability events always occur between an observation and an

assumed model in their most different parts, the system implements an automatic discriminative selection and weighting of the parts of the signal during recognition. This discriminative analysis is made to be effective by the training process, which essentially performs a mini-max process, i.e. maximizing the probabilities of occurrence of those acoustic events with smaller likelihoods for the same class of signals. As an example, we plot in Figure 3 the error rate against the number of gradient descent iterations used for training the models. The result is obtained for the 15 segment case, using MFCC's alone.

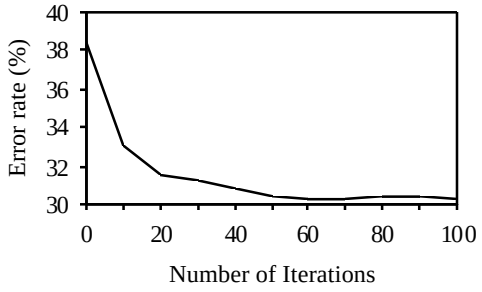

Figure 3: A typical example of the recognition error rate against the number of gradient descent iterations in the training stage.

5. CONCLUSIONS

This paper described a new stochastic modelling approach for capturing the statistical correlation in a time series. The new model is drawn from the upper bound of the joint probability distributions, corresponding to a strong correlation assumption, and is adjusted to represent various degrees of correlation by using a nonlinear interpolation between the lower and upper bounds of the joint probability distributions, where the lower bound corresponds to the statistical independence assumption.

When there exists a correlation in a time series, a joint probabilistic model starting from the upper bound of the joint probability distributions may reach a closer solution to the correct answer than assuming independence over time. Recognition then may benefit from a more accurate characterization of the dynamic information. This has been confirmed, preliminarily, by the experiments above based on a specific application of this principle to isolated word speech recognition. The advantage of this new model is that it has a low parameter complexity, and a flexible model structure, similar to that of a standard HMM. Further modifications of this model may include incorporation of certain fine-tuning and discriminative techniques proved to be successful in traditional HMM's, and investigation of more correlation units, e.g. phoneme or state.

ACKNOWLEDGEMENT

The authors thank British Telecom Research Laboratories for providing the test database.

REFERENCES

1. Ostendorf, M., and Roucos, S. "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. ASSP-37*, pp. 1857-1869, 1989.

2. Ghitza, O., and Sondhi, M. M. "Hidden Markov models with templates as non-stationary states: an application to speech recognition," *Computer Speech and Language*, Vol. 2, pp. 101-119, 1993.

3. Paliwal, K. K. "Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer," *ICASSP-93*, pp. 215-218, 1993.

4. Smith, F. J., Ming, J., O'Boyle, P., and Irvine, A. "A hidden Markov model with optimized inter-frame dependence," *ICASSP-95*, pp. 209-212, 1995.

5. Kenny, P., Lennig, M., and Mermelstein, P. "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. ASSP-38*, pp. 220-225, 1990.

6. Digalakist, V., Rohlicek, J. R., and Ostendorf, M. "A dynamical system approach to continuous speech recognition," *ICASSP-92*, pp. 289-292, 1992.

7. Russell, M. J. "A segmental HMM for speech pattern modelling," *ICASSP-93*, pp. 499-502, 1993.

8. Zavaliagkos, G., Zhao, Y., Schwartz, R., and Makhoul, J. "A hybrid segmental neural net/hidden Markov system for continuous speech recognition," *IEEE Trans. SAP-2*, pp. 151-159, 1994.