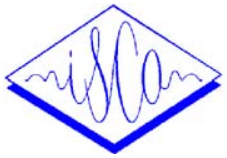

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Modelling inter-frame dependence with preceeding and succeeding frames[†]

P.Hanna, J.Ming, P. O'Boyle & F.J. Smith

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

E-mail: P.Hanna@qub.ac.uk, J.Ming@qub.ac.uk, P.OBoyle@qub.ac.uk, FJ.Smith@qub.ac.uk
Tel: (+44 1232) 245133 x 4538 Fax: (+44 1232) 666520

Abstract

This paper explores the modelling of inter-frame dependence as a means of improving the performance of HMMs. More specifically, a model based on the IFD-HMM (Ming & Smith, 1996) that assumes a dependency upon both succeeding and preceeding frames is proposed. The means by which a dependency upon succeeding frames might be integrated into a HMM framework are explored, and a mathematical outline of the proposed extension given. The results of various tests aimed towards exploring the consequences of introducing succeeding frame dependencies are included. It was found that a dependency upon succeeding frames enabled dynamic spectral information, not found in the preceeding frames, to be usefully employed; resulting in a significant increase in the recognition accuracy. Additionally, it was shown that modelling of the dynamic spectral information (using time-lag sequences) was at least as important as improved modelling of the instantaneous spectra (using multiple mixtures).

10.21437/Eurospeech.1997-15

1. Introduction

A number of different approaches have been adopted towards weakening the independence assumption of frame vectors within the framework of an HMM. Choice examples include the linear-predictive HMM (Kenny, *et al.*, 1990; Woodland, 1992), the bi-gram constrained HMM (Paliwal, 1993) and more recently the IFD (Interframe dependence)-HMM (Ming & Smith, 1995, 1996). Generally speaking, these approaches assume each observed frame is dependent upon one or more previous frames (using some form of conditional observation density).

iven the non-stationary nature of spoken speech, it is reasonable to assume that for a particular frame, the succeeding frames contain useful dynamic information which may not be encapsulated in the preceeding frames. In this paper, we investigate this idea by extending the IFD-HMM so that a dependency upon both succeeding and precceding frames can be explored.

In effect, the inclusion of frame dependencies results in an improved modelling of the dynamic signal features. Conversely, the use of multiple mixtures can be viewed as an improved modelling of the instantaneous spectra. Ideally both approaches would be employed. However, sometimes this is not possible; either due to constrains on computational complexity, or a limited amount of training data. In this paper, we compare the relative advantages of each approach.

This paper is organised as follows. In section 2, the original formulation of the IFD-HMM is outlined. Additionally, the means by which a dependence upon preceeding frames might be introduced, and the model changes arising from such an extension are detailed. Section 3 outlines the experimental conditions and presents a selection of the collected experimental results. Finally, in section 4 various conclusions are drawn.

2. Mathematical Formulation

2.1. Original IFD-HMM formulation

The original formulation of the IFD-HMM (Ming & Smith, 1996) assumes a dependence upon multiple *previous* frames; the observation probability can be expressed as:

$$p_\lambda(x \mid s,\tau) = \prod_{t=1}^{T} b_{s_t}(x_t \mid x_{t-\tau_t(1)} \ldots x_{t-\tau_t(N)}) \quad (1)$$

where x is a frame observation sequence, s a state sequence and $\tau = \{\tau_t(n):\ 1 \le t \le T; 1 \le n \le N\}$ defines a series of time-lag sequences of frame dependencies relative to the current frame. In this paper, we assume that the time-lag sequence does not vary with time, hence we define $\tau = (\tau(1) \ldots \tau(n))$.

The observation density function, $b_i(\cdot)$, is approximated as a weighted mixture of a set of first-order conditional densities, each mixture component accounting for a particular conditional variable, i.e.

$$b_i(x_t \mid x_{t-\tau(1)} \ldots x_{t-\tau(N)}) \approx \sum_{n=1}^{N} w_{in} f_{in}(x_t \mid x_{t-\tau(n)}) \quad (2)$$

[†] This work is supported by an EPSRC grant, number GR/K8205.

Where f_{in} are the conditional densities and w_{in} the corresponding weights. By assuming x_t and $x_{t-\tau(n)}$ are jointly Gaussian distributed, the density function $f_{in}(\cdot)$ can be defined as a multivariate conditional Gaussian density of the standard form (Ming & Smith, 1996), i.e.

$$f(z \mid z_n) \propto \exp\{-1/2h(z, z_n)\}$$

Where $h(z, z_n) = (z - Hz_n - \mu)^T U_{11}(z - Hz_n) + \log |U_{11}^{-1}|$ given that μ, H and U_{11} are the density parameters.

Given a set of model parameters and a time-lag sequence, the joint likelihood of a given observation and state sequence can be expressed as:

$$p(x, s \mid \tau, \lambda) = p_\lambda(x \mid s, \tau) p(s \mid \lambda)$$

$$= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \left[\sum_{n=1}^{N} w_{s_t n} f_{s_t n}(x_t \mid x_{t-\tau(n)}) \right] \quad (3)$$

Training is based on an EM process. During the expectation step the parameter set λ is estimated based on the given time-lag sequence, τ. The estimation of the parameters is accomplished using a forward-backward re-estimation algorithm. During the maximisation step the time-lag sequence is optimised against the previously estimated model. The EM process is iterated until a sufficient level of convergence has been achieved.

Ming & Smith (1996) show that the parameter estimation algorithm for the IFD-HMM is similar to that of the traditional multiple-mixture Gaussian HMM. In additional, both approaches have comparable complexity.

2.2. Extension of the IFD-HMM

We define a time-lag sequence consisting solely of succeeding frames as a *succeeding only* time-lag sequence. Likewise, a *preceeding only* time-lag sequence entirely specifies preceeding frames. A *bi-directional* time-lag sequence consists of both succeeding and preceeding frames.

Assume two time-lag sequences; τ^- and τ^+. Where τ^- defines a time-lag sequence of preceeding frames, and τ^+ defines a time-lag sequence of succeeding frames. Hence $p_\lambda(x \mid s, \tau^-)$ and $p_\lambda(x \mid s, \tau^+)$ are the respective probabilities of an observation sequence, dependent on the defined time-lag sequences.

A joint model, which takes into account both the *preceeding only* and *succeeding only* time-lag sequences, can be formed from various combinations of the above two IFD-HMMs. Two of the most obvious formulations are as follows:

$$\mu \times p_\lambda(x \mid s, \tau^+) + (1 - \mu) \times p_\lambda(x \mid s, \tau^-) \quad (4a)$$

(Linear Interpolation)

$$p_\lambda(x \mid s, \tau^+) \times p_\lambda(\mathbf{x} \mid \mathbf{s}, \ ^-) \quad (4b)$$

(Geometric combination)

Adoption of a linear interpolation of the two IFD-HMMs is more complicated than that of the geometric combination; i.e. we need to estimate μ, using some form of deleted interpolation. Conversely, if a geometric combination is used, then the modifications to the IFD-HMM formulation are relatively modest, as will be shown. Hence, for practical reasons, a geometric combination was employed, resulting in the definition of a new joint likelihood, shown below:

$$q_\lambda(x \mid s, \tau^-, \tau^+) = p_\lambda(x \mid s, \tau^-) \times p_\lambda(x \mid s, \tau^+) \quad (5)$$

The likelihood of a given observation sequence, given a state sequence and two time-lag sequences can now be expressed as:

$$q_\lambda(x \mid s, \tau^-, \tau^+) = \prod_{t=1}^{T} \left(\begin{array}{l} \sum_{n=1}^{N} w^-_{s_t n} f^-_{s_t n}(x_t \mid x_{t-\tau^-(n)}) \\ \times \sum_{n=1}^{N} w^+_{s_t n} f^+_{s_t n}(x_t \mid x_{t-\tau^+(n)}) \end{array} \right) \quad (6)$$

A Baum-Welsh re-estimation procedure can be readily applied to the above model. The parameter estimation procedure mirrors that of the estimation of the joint densities of normal and delta spectral parameters, e.g. $p_\lambda(x \mid s, \tau^-)$ corresponds to the density based on the normal spectral components, and $p_\lambda(x \mid s, \tau^+)$ corresponding to the density based on the delta spectral components. Hence, using a Baum-Welsh re-estimation algorithm both densities can easily be jointly optimised, effectively enabling *bi-directional* time-lag sequences to be employed.

For a given number of frame dependencies, the complexity of the modified IFD-HMM is the same as that of the original formulation (i.e. the estimation of a model using 4 frame *preceeding only* time-lag sequence requires the same amount of computation as a joint model using a 2 frame *preceeding only* and a 2 frame *succeeding only* time-lag sequence).

The above reformulation does not extend to the maximisation of the time-lag sequence (the M-step). The computationally intensive nature of this procedure precluding it from this initial investigation. Ming & Smith (1996) reported up to 11% reduction in recognition error rate when the time-lag sequence is maximised.

3. Experimental results

3.1. Experimental set-up

The same experimental conditions are previously used by Woodland (1992) and Ming & Smith (1996) were adopted in order to provide ease of comparison with other results. Hence, all experiments are based on the Connex speaker-independent alphabetic database (provided by British Telecom Research Laboratories). The database contains three repetitions of each word by a total of 104 speakers; the database is roughly balanced with respect to age and gender. Experiments were based on the highly-confusable E-set (b, c, d, e, g, p, t and v),

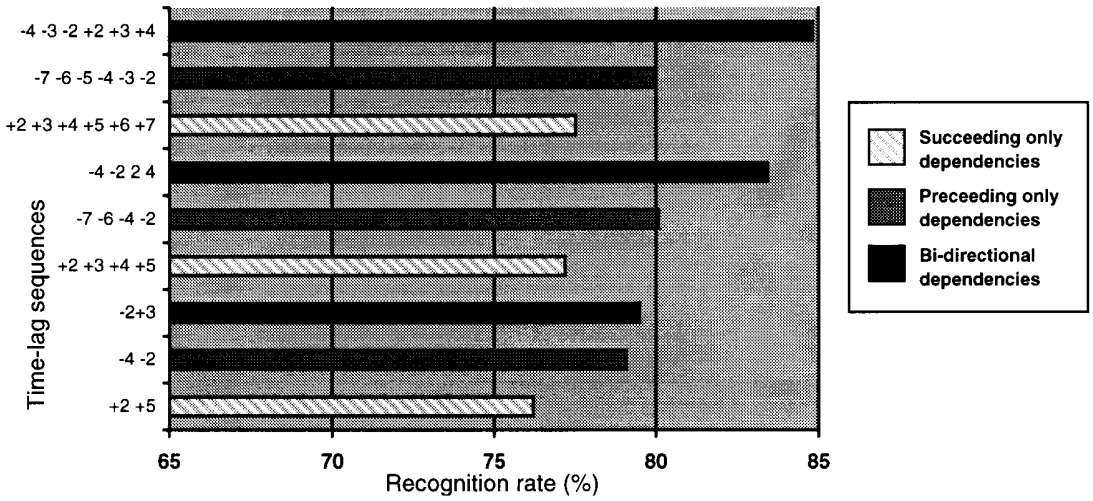

Figure 1 - Recognition results for various time-lag sequences

with 52 speakers designated for training (a total of 1219 utterances) and 52 speakers for testing. The speech, which was sampled at 20kHz, was divided into 25.6ms frames with a consecutive frame overlap of 15.6ms. Each frame was passed through a 27 band-pass mel-frequency filter from which 12 MFCCs were extracted.

3.2. Recognition accuracy of bi-directional time-lag sequences vs. uni-directional time-lag sequences.

The results presented in this section test the hypothesis that a dependency upon succeeding frames provides useful discriminative dynamic information. The results were obtained for a 5 state left-to-right topology.

In order to permit an unbiased comparison, only time-lag sequences with the same number of frame dependencies will be compared with one another. This ensures that the compared results are for a fixed model parameter size. Additionally, for a given number of frame dependencies, the best results obtained using *succeeding only* and *preceeding only* frame-lag sequences are compared against the best result obtained using a *bi-directional* time-lag sequences. The results are drawn from approximately 150 tests cases which were balanced in terms of the type of time-lags (*succeeding only,* etc.) and the number of frame dependencies.

A representational sample of the collected results is shown in Figure 1. Evidently, use of a *bi-directional* time-lag sequence results in the highest recognition accuracy. Conversely, the *succeeding only* time-lag sequences return the lowest performance. Table 1 shows the average percentage reduction in error rates averaged over all the collected results.

Number of frame dependencies	*Succeeding only* → *bi-directional* time-lags	*Preceeding only* →*bi-directional* time-lags	*Succeeding only* → *preceeding only* time-lags
2	13.8	2.0	12.1
4	27.4	16.9	12.6
6	32.5	24.2	10.9

Table 1 – Percentage reduction in error rates for changes in time-lag sequences

The collected results provide good evidence that, not only do succeeding frames provide useful dynamic information, but that it is profitable (in terms of increased recognition accuracy) to use this information in addition to that obtained from preceeding frames.

The higher recognition rate of *preceeding only* time-lag sequences over *succeeding only* time-lag sequences is an artefact of the tests being based on the E-set, where most of the useful discriminative information is found at the start of the utterances. Hence, the *preceeding only* time-lag sequence will result in better use of the discriminative information. Confirmation of this was obtained from tests based on a non E-set subset of the Connex database; where similar performance for both types of time-lag sequences was found.

3.3. Comparison of modelling the instantaneous spectra with modelling the dynamic features.

Use of multiple mixtures can be viewed as increasing the representational accuracy of the instantaneous spectra. In contrast to this, dependency upon either preceeding or succeeding frames can be viewed as a more accurate modelling of the dynamic features.

Employment of either of these approaches increases the model's parameter size; leading to lengthened training and recognition times, and the need for a greater minimal amount of training data to ensure good parameter estimation. In this section we compare the two approaches in terms of performance accuracy for a given model size.

Table 2 shows the results obtained using a 5 state IFD-HMM. The figures in brackets show the parameter size relative to a 5 state IFD-HMM without multiple mixtures and using a two frame time-lag. Generally speaking, an analysis of the results obtained for 5 state IFD-HMMs shows that, for a given parameter size, emphasis on an improved modelling of the dynamic information normally results in improved performance over that obtained from an emphasis on improved modelling of the instantaneous spectra. However, this finding was not so strongly supported when a 15 state IFD-HMM, with

the last 9 states tied, was used as the base model. The results can be seen in Table 3 (the figures in brackets denote the model's parameter size relative to a 15 state IFD-HMM without multiple mixtures and using a four frame time-lag sequence).

An analysis of Table 3 fails to statistically distinguish one approach as offering improved results over the other. Unfortunately, a complete comparison could not be made due to the fact that the 150 utterances available for training each word model proved to be insufficient when multiple mixtures were combined with long time-lag sequences (a relative model size of 3 was the largest that could be adequately trained).

The highest recognition accuracy shown in Table 3 is 92.4%, obtained for a IFD-HMM using a long time-lag sequence, but without multiple mixtures. It is reasonable to expect that the additional introduction of multiple mixtures, coupled with a greater amount of training data, would lead to an increased recognition accuracy.

Time-lag sequence	1 mixture	2 mixtures	3 mixtures	4 mixtures
-2 +2	79.1 (1)	80.2 (2)	84.7 (3)	84.3 (4)
-3 -2 +2 +3	82.5 (2)	85.4 (4)	82.6 (6)	83.7 (8)
-4–3–2+2+3 +4	84.8 (3)	85.4 (6)	83.0 (9)	83.9 (12)

Table 2 – Combinations of time-lag sequence and multiple mixture for a 5 state IFD-HMM (bracketed figures show the model's parameter size relative to a 5 state IFD-HMM without multiple mixtures and using a two frame time-lag).

Time-lag sequence	1 mixture	2 mixtures	3 mixtures	4 mixtures
-4–2+2+4	91.2 (1)	90.7 (2)	91.9 (3)	91.5 (4)
-4–3–2+2+3+4	90.9(1.5)	91.8 (3)	90.8(4.5)	90.7 (6)
-7-6-5-4-3-2 +2+3+4+5+6+7	92.4 (3)	91.7 (6)	90.5 (9)	89.5 (12)

Table 3 – Combinations of time-lag sequence and multiple mixture for a 15 state IFD-HMM with the last 9 states tied (bracketed figures show the model's parameter size relative to a 15 state IFD-HMM without multiple mixtures and using a four frame time-lag sequence).

3.4. Comparisons

Table 4 shows a range of recognition accuracies obtained for a 15 state topology with the last nine states of all word models tied.

Model	Recognition accuracy
Standard HMM with delta spectral features[†]	85.7
Preceeding only IFD-HMM	89.7
Preceeding only IFD-HMM with time-lag optimisation	90.7
Bi-directional IFD-HMM	92.4

Table 4 – Recognition accuracies obtained for 15 state (last 9 tied) topologies

[†] Woodland & Cole, 1991

The IFD-HMM using a *bi-directional* time-lag sequence offers a 47% reduction in the error rate over a standard HMM using delta spectral features.

4. Conclusion

The extension of the IFD-HMM to include a dependency upon succeeding frames permits an improved modelling of the dynamic spectral information, resulting in a significant decrease in error recognition rate compared to either comparable *succeeding only* or *preceeding only* time-lag sequences.

The highest recognition accuracy obtained using a *bi-directional* time-lag sequence was 92.4% (obtained for the E-set of the Connex alphabetic database). This result was attained without either the use of multiple mixtures or the optimisation of the time-lag sequence (the M-step of the IFD-HMM training procedure). Should sufficient training data be available then it is reasonable to expect the introduction of these approaches to result in an increased recognition accuracy.

The introduction of *bi-directional* time-lag sequences to the IFD-HMM was based on the assumption that the non-stationary nature of speech entails that succeeding frames contain discriminative dynamic information not found in the preceeding frames. The collected results show that this assumption is indeed correct. Additionally, the introduction of *bi-directional* time-lag sequences does not entail an increased model parameter size or increased training complexity for a given number of dependencies.

The collected results also show that emphasis on modelling the dynamic information is at least as profitable, if not more so, than emphasis on modelling the instantaneous spectra.

References

Kenny, P. Lennig, M. & Mermelstein, P. (1990). *A linear predictive HMM for vector-valued observations with applications to speech recognition.* IEEE Transactions on Acoustic., Speech and Signal Proceesing, 38, 220-225

Woodland, P.C. & Cole, D.R. (1991) *Optimizing hidden Markov models using discriminative output distributions.* Proceedings of ICASSP'91 pp545-548

Woodland, P.C. (1992) *Hidden Markov models using vector linear prediction and discriminative output distributions.* Proceedings of ICASSP'92, 509-512

Paliwal, K.K. (1993) *Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer.* Proceedings of ICASSP'93, 209-212

Ming, J. and Smith, F.J. (1996) *Modelling of the interframe dependence in an HMM using conditional Gaussian mixtures.* Computer Speech and Language (1996) 10, 229-247

Smith, F.J., Ming, J., O'Boyle, P., Irvine, A.D. (1995) *A hidden Markov model with optimized inter-frame dependence.* Proceedings of ICASSP'95 pp209-212