© IEE 1998 3 March 1998
Electronics Letters Online No: 19980618

F.J. González-Serrano (*DTC - ETSI Telecomunicación, Universidad de Vigo, Campus Universitario, s/n. 35200 Vigo, Spain*)

F. Pérez-Cruz and A. Artés-Rodríguez (*DSSR - ETSI Telecomunicación, UPM, Ciudad Universitaria, s/n. 28040 Madrid, Spain*)

References

1 ALBUS, J.: 'A new approach to manipulator control: the cerebellar model articulation controller', *Trans. ASME, J. Dyn. Syst., Meas. Control*, 1975, **63**, pp. 220–227

2 GONZÁLEZ-SERRANO, F.J., ARTÉS-RODRÍGUEZ, A., and FIGUEIRAS-VIDAL, A.R.: 'The generalized CMAC'. Proc. ISCAS, IEEE, Atlanta, GA, USA, May 1996

3 RUMELHART, D.E., HINTON, G.E., and WILLIAMS, R.J.: 'Learning internal representations by error propagation, Vol. 1' (MIT Press, Cambridge, MA, USA, 1986)

Variation of features of interframe dependent HMM for speech recognition

P. Hanna, N. Harte, J. Ming, S. Vaseghi and F.J. Smith

The effects are explored of using different dynamic features in conjunction with an HMM that permits a dependency on both preceeding and succeeding frames. In particular, features which capture dynamic information are varied in the interframe dependent HMM. A recognition accuracy of 93.9% for the Connex speaker-independent E-set was obtained using tied states.

Introduction: The interframe dependent (IFD)-HMM was originally proposed [1] as a means of overcoming the independence assumption of speech frames within the framework of an HMM. In particular, the observed frame was assumed to be dependent on a number of preceeding frames. Later work [2] extended the model, forming the bi-directional IFD-HMM, which permitted both succeeding and preceeding frame dependencies to be employed; the non-stationary nature of speech entails that the succeeding frames include dynamic spectral information not contained in the preceeding frames.

We explore the union of the IFD-HMM and features that themselves capture dynamic spectral information (for example differential MFCC features, etc.). Both approaches aim to capture all the useful discriminative spectral information contained in the speech utterance. However, should both approaches only capture a subset of the available information, then it is possible that the union of the two approaches will result in a model that contains more information than either of the two approaches considered in isolation. To test this hypothesis, we explore the effects of introducing two different types of dynamic features to the IFD-HMM.

Model definition: The probability density function (PDF) of the bi-directional IFD-HMM observation sequence, given a state sequence and set of frame dependencies, is expressed as the geometric combination of two uni-directional PDFs (one conditional on succeeding frames, and the other on preceeding frames):

$$p(x|s,\tau^-,\tau^+,\lambda) = \prod_{t=1}^{T} b_{s_t}\left(x_t|x_{t-\tau^-(1)}\cdots x_{t-\tau^-(N)}\right) \cdot b_{s_t}\left(x_t|x_{t+\tau^+(1)}\cdots x_{t+\tau^+(N)}\right) \quad (1)$$

where $\tau^{+,-} = \{\tau^{+,-}(1) \ldots \tau^{+,-}(N)\}$ defines the dependency relationship between the observed frame and the preceeding and succeeding frames, respectively, with each $\tau^{+,-}(n) > 0$ and $\tau^{+,-}(n-1) < \tau^{+,-}(n)$; $b_s(z|z_1 \ldots z_N)$ is the conditional observation density for state s, which is approximated as a weighted mixture of a set of first-order Gaussian conditional densities [1]:

$$b_s(z|z_1 \ldots z_N) \simeq \sum_{n=1}^{N} w_{sn} f_{sn}(z|z_n) \quad (2)$$

where $f_{sn}(z|z_n)$ represents the nth component Gaussian conditional density in state s. The IFD-HMM can be estimated through the application of a Baum-Welsh re-estimation procedure [1, 2]. The estimation procedure is similar to that of the traditional multiple mixture Gaussian HMM. However, in the multiple-mixture Gaussian procedure, the mixture is formed over the instantaneous observation space, whereas for the IFD-HMM it is formed over the observation history.

Experiments and discussion: All experiments are based on the highly-confusable E-set (b, c, d, e, g p, t and v) of the Connex speaker-independent alphabetic database (provided by British Telecom Research Laboratories). Precise details of the experimental setup can be found elsewhere [1, 2]. Three types of feature vector were used in conjunction with the IFD-HMM, namely: 12 standard MFCC coefficients (MFCC); 12 MFCC coefficients augmented with first-order differential coefficients (MFCC+ΔMFCC); 24 segmental cepstral-time matrix based features (S-CTM). The S-CTM features are detailed in [3], whereby a second-order derivative cepstral-time matrix is used to capture the short-term temporal dynamics and an additional DCT is applied to the calculated sub-segmental cepstral-time matrix in order to determine the longer term transitional dynamics. Entropic's Hidden Markov Model Toolkit, HTK, was used to construct and train the standard HMMs.

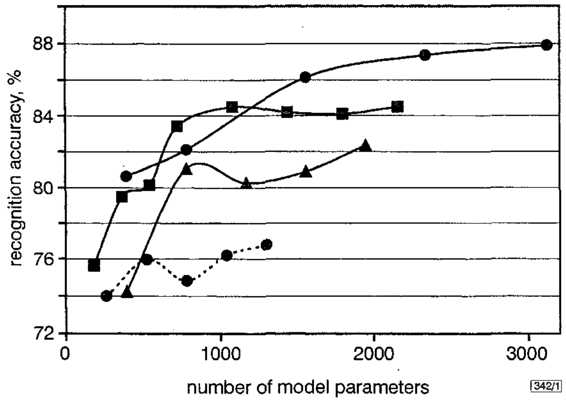

Fig. 1 *Five state HMM and IFD-HMM recognition accuracies against model parameter sizes using differential MFCC features*

- -●- - Standard HMM with MFCCs+ΔMFCCs
—▲— Standard HMM with MFCCS+ΔMFCCs+ΔΔMFCCs
—■— Bi-directional IFD-HMM with MFCCs
—●— Bi-directional IFD-HMM with MFCCs+ΔMFCCs

We first compare the IFD-HMM to standard HMMs employing differential MFCC based features, as shown in Fig. 1 where recognition accuracy is expressed as a function of the model parameter size, which is increased through the use of additional mixtures (for the standard HMMs) or frame dependencies (for the IFD-HMMs). We conclude that when MFCC features are used, the IFD-HMM offers a significantly higher recognition accuracy for a given model parameter size, than a corresponding HMM using differential MFCC features. Furthermore, the IFD-HMM using MFCC+ΔMFCC features permits the highest levels of recognition accuracy, and provides a means of overcoming the almost static level of recognition accuracy reached when the addition of new frame dependencies does not significantly provide any new information.

Although the addition of MFCC+ΔMFCC to the bi-directional IFD-HMM resulted in increased recognition accuracies, when S-CTM based features were used, lower levels of recognition accuracy resulted. Fig. 2 shows the recognition accuracies for the IFD-HMM using a number of two-frame dependencies (one succeeding and one preceeding). When S-CTM based features are used, the resultant recognition accuracy is somewhat insensitive to the choice of frame dependencies, whereas MFCC based features are highly sensitive to the choice of frame dependencies. Based on the collected findings, we conclude that the dependent frames should be selected so that there is little temporal overlap between the features. Maximal performance should be expected to arise from that set of frame dependencies which minimises the temporal overlap between features whilst still ensuring that a maximum amount of useful dynamic information is captured. However, when S-CTM based features are employed, due to the much longer temporal coverage of this feature, any dependencies close to the observed frame will have a considerable degree of overlap. However, if the overlap is minimised then the frame dependencies are too far removed from the observed frame to capture much useful information.

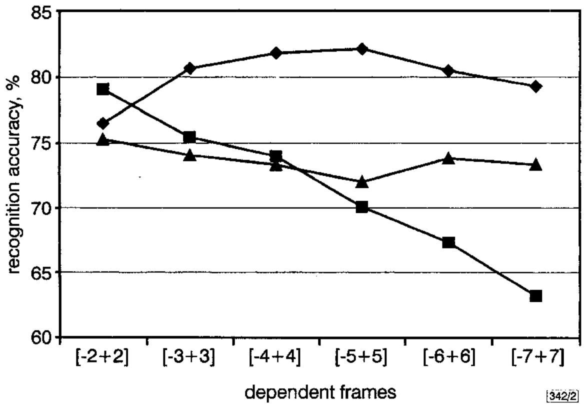

Fig. 2 *Recognition accuracy for bi-directional IFD-HMM using two-frame dependency (one succeeding, one preceeding) combined with different features*

—◆— MFCCs+ΔMFCCs
—■— MFCCs
—▲— S-CTM

Table 1: Recognition accuracies for various 15 state (last nine tied) topologies

Model	Feature	Recognition accuracy
Standard HMM	MFCC+ΔMFCC	% 85.7
Preceeding frame IFD-HMM	MFCC	89.7
S-CTM (full covariance)	24 S-CTM+LDA	92.2
Bi-directional IFD-HMM	MFCC	92.4
Bi-directional IFD-HMM	MFCC+ΔMFCC	93.9

The use of MFCCs augmented with delta spectral components has enabled the previous bi-directional IFD-HMM's highest recognition accuracy of 92.4% [2] to be extended. A fifteen state, of which the last nine states were tied [4], IFD-HMM using an four preceeding and four succeeding dependent frames combined with MFCC+ΔMFCC features obtained an E-set recognition accuracy of 93.9%. This result was obtained without the use of multiple mixtures for the component densities, and to the authors knowledge is the highest result yet published for this database. A comparison of fifteen state models (the last nine states tied) is shown in Table 1. The standard HMM result is taken from [4] and the S-CTM result from [3].

In conclusion, the use of MFCC+ΔMFCC in conjunction with the IFD-HMM provides a mechanism through which more useful dynamic information can be captured, resulting in an increased recognition accuracy over that obtained by either the standard HMM using dynamic features or the IFD-HMM using static features alone.

Acknowledgments: The authors would like to thank BT for financial support. Additionally this work is supported by an EPSRC grant, number GR/K82505.

© IEE 1998 *9 March 1998*
Electronics Letters Online No: 19980640

P. Hanna, N. Harte, J. Ming, S. Vaseghi and F.J. Smith (*School of Electrical Engineering and Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, United Kingdom*)

References

1 MING, J., and SMITH, F.J.: 'Modelling of the inteframe dependence in an HMM using conditional Gaussian mixtures', *Comput. Speech and Lang.*, 1996, **10**, pp. 229–247

2 HANNA, P., MING, J., O'BOYLE, P.O., and SMITH, F.J.: 'Modelling interframe dependence with preceeding and succeeding frames'. Eurospeech 97, Vol. 3, pp. 167–1170

3 HARTE, H., VASEGHI, S., and MILNER, B.: 'Dynamic features for segmental speech recognition'. ICSLP 96, pp. 933–936

4 WOODLAND, P.C., and COLE, D.R.: 'Optimizing hidden Markov models using discriminative output distributions'. Proc. ICASSP '91, pp. 545–548

Very low bit rate speech coding using a diphone-based recognition and synthesis approach

M. Felici, M. Borgatti and R. Guerrieri

High compression rates of speech signals may be achieved by coding schemes based on relevant linguistic segments. A system is described that relies on a diphone recogniser as the coder and on a speech synthesiser reproducing speech starting from a diphone codebook as the decoder. The spoken message is encoded in textual (phoneme labels) plus prosody representation. This speech coding technique may be used for voice mail or phone communication over low bit rate channels.

Introduction: From a linguistic point of view, speech may be considered as a concatenation of a limited number of basic units such as phonemes. An adequate set of these segments may be gathered in a codebook used to synthesise speech. Hence, a high compression rate can be achieved by transmitting only the phoneme indexes [1, 2].

Naturalness in a text-to-speech synthesiser can be improved by using diphones (the part of speech between the centres of two subsequent phonemes) instead of phonemes, since diphones capture transitions between sounds [3]. This suggests that using diphones as the basic speech unit may result in better quality compared to previous phoneme-based vocoders.

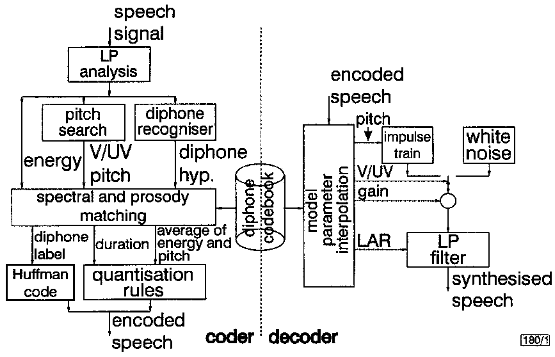

Fig. 1 *Diphone speaker dependent coder-decoder*

This Letter presents a speech compression system based on diphones, outlined in Fig. 1. The coder is composed of a diphone recogniser which determines diphone labels and boundaries, and a prosody analyser. At the decoder stage, speech is obtained by concatenation of diphone templates coming from a speaker dependent codebook. The synthesiser is based on a source-filter vocal model, allowing easy modification of prosody parameters. A different codebook is required for each user, unless a standard voice is used.

The Italian language has been used to test the performance of the system. Our phonetic transcription is based on 30 different symbols (sounds) plus a symbol representing silence. Of the 961 combinations of two symbols, only 517 are encountered in the Italian language.

Diphone segmentation and prosody analysis: A neural network (NN) speaker independent diphone recogniser has been used to determine diphone probabilities [4]. The recognition and synthesis processes rely on linear predictive analysis: a 14th order filter is computed for each frame. The incoming speech signal (sampled at 8kHz) is preemphasised, following which Hamming windowing is