4 RUMYANTSEV, S., LEVINSHTEIN, M.E., GASKA, R., SHUR, M.S., YANG, J.W., and KHAN, M.A.: 'Low-frequency noise in AlGaN/GaN HFETs on SiC and sapphire substrates', submitted to *J. Appl. Phys.*, 1999

5 ASIF KHAN, M., HU, X., SIMIN, G., LUNEV, A., YANG, J., GASKA, R., and SHUR, M.S.: 'AlGaN/GaN metal-oxide-semiconductor field effect transistor', to be published in *IEEE Electron Device Lett.*, 2000

6 HOOGE, F.N.: '1/f noise sources', *IEEE Trans.*, 1994, ED-41, (11), pp. 1926–1935

7 RUMYANTSEV, S., LEVINSHTEIN, M.E., GASKA, R., SHUR, M.S., KHAN, A., YANG, J.W., SIMIN, G., PING, A., and ADESIDA, T.: 'Low 1/*f* noise in AlGaN/GaN HFETs on SiC substrates'. Abstracts of 3rd Int. Conf. on Nitride Semiconductors (ICNS3), Montpellier, France, 4–9 July 1999, pp. 125–126

8 KUKSENKOV, D.V., TEMKIN, H., GASKA, R., and YANG, J.W.: 'Low-frequency noise in AlGaN/GaN heterostructure field effect transistors', *IEEE Electron Device Lett.*, 1998, 19, (7), pp. 222–224

9 GARRIDO, J.A., CALLE, F., MUNOZ, E., IZPURA, I., SANCHEZ-ROJAS, J.L., LI, R., and WANG, K.L.: 'Low frequency noise and screening effects in AlGaN/GaN HEMTs', *Electron. Lett.*, 1998, 34, (24), pp. 2357–2359

10 BALANDIN, A., MOROZOV, S.V., CAI, S., LI, R., WANG, K.L., WIJERTANE, G., and VISWANATHAN, C.R.: 'Low flicker-noise GaN/AlGaN heterostructure field effect transistors for microwave communications', *IEEE Trans.*, 1999, MTT-47, (8), pp. 1413–1417

Discriminative multi-resolution sub-band and segmental phonetic model combination

P. McCourt, N. Harte and S. Vaseghi

A joint discriminative framework for multi-resolution sub-band HMMs and a hybrid segmental phonetic model is presented which combines independent likelihood scores using class dependent weightings trained to a minimum classification error criteria. This successfully extends the performance on a phonetic classification task to 76.3% compared to a full-band HMM score of 69.2%.

Introduction: While the use of hidden Markov models (HMMs) of phonetically defined sub-word units remains dominant in current large vocabulary automatic speech recognition systems, there has been much recent exploration of alternative modelling strategies, particularly 'segmental' modelling [1]. Segment based models aim to model sequences of feature vectors in order to account for the constraints of feature dynamics inherent in the speech production process. The segment-based features introduced in [2] for example give similar phonetic classification performance to HMM monophones trained on mel-frequency cepstral coefficient (MFCC) features with first- and second-order time derivatives. Despite equal classification scores, differences exist in the discriminative capabilities of each model set. In the example quoted, correct classification by one model type and not the other occurs for 12% of the total number of test tokens. This suggests that exploiting the complementary discriminative properties of alternative acoustic models bears the potential for increased performance. Direct likelihood combination from independent acoustic models representing the same phonetic class is explored here as a solution to combined decoding decisions. Multi-resolution sub-band modelling and a novel segmental model are described which independently outperform MFCC trained HMMs. Linear log-likelihood score combination based on an independence assumption is then demonstrated to increase the phonetic classification performance over either model set in isolation. Finally, discriminative training according to the minimum classification error (MCE) criteria [3] of a class-dependent weight set for linear log-likelihood combination is demonstrated to extend this performance advantage yet further.

Multi-resolution sub-band features and models: For standard MFCC features, cepstral analysis is performed on the mel-spaced filterbank log energy vector $\mathbf{E}$ of each short-time analysis frame, as expressed by the linear transformation $\mathbf{X} = \mathbf{A}\mathbf{E}$ where $\mathbf{A}$ represents the DCT. The log energy vector $\mathbf{E}$ can be split into N sub-vectors $\mathbf{E} = [\mathbf{E}_1^T \cdots \mathbf{E}_b^T \cdots \mathbf{E}_N^T]^T$ (where T indicates matrix transpose) such that each sub-vector $\mathbf{E}_b$ effectively represents a grouped bandwidth of log energies. Separate cepstral analysis using appropriately dimensioned DCT transforms $\mathbf{A}_b$ yields new sub-band cepstral vectors $\mathbf{X}_b$. Thus

$$[\mathbf{X}_1^T \cdots \mathbf{X}_N^T]^T = [(\mathbf{A}_1\mathbf{E}_1)^T \cdots (\mathbf{A}_N\mathbf{E}_N)^T]^T \quad (1)$$

The usefulness of these features is based on the conjecture that important cues for discrimination exist in the local spectral correlates not captured by full band cepstra. Unlike much recent sub-band modelling work, e.g. in [4], the sub-band models trained on these features are used to supplement rather than replace full band models. Multi-resolution analysis implies feature extraction from alternative sub-band decompositions.

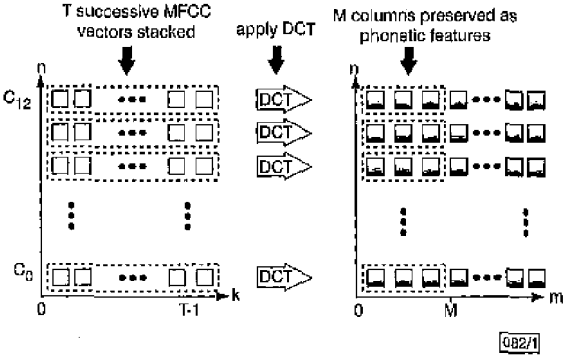

Fig. 1 *Transformation across segment of MFCC vectors to yield phonetic features*

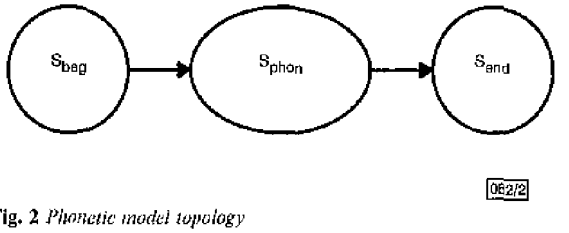

Fig. 2 *Phonetic model topology*

Phonetic segmental model: For a given unit of speech of length T vectors, identified as a phonetic unit, the phonetic features for that segment can be calculated as

$$\mathbf{Y} = \mathbf{A}_T\mathbf{X} \quad (2)$$

where $\mathbf{X} = [\mathbf{x}_t, ..., \mathbf{x}_{t+T-1}]$ is the segment and $\mathbf{A}_T$ is a transformation dependent on the segment length T. Here $\mathbf{A}_T$ is the T length DCT used to encode the transitional dynamics across the duration of the phonetic event. $\mathbf{Y}$ hence denotes the phonetic features for that segment derived via a DCT on the stacked cepstral vectors $\mathbf{X}$. This is illustrated in Fig. 1. The first M columns of the matrix are preserved as phonetic features for the complete segment such that a fixed length representation is yielded from variable length sequences.

The phonetic model in Fig. 2 looks similar to the standard monophone HMM. Transition probabilities between states are omitted however. The beginning and end state model the conventional cepstral feature frames bounding the segment. The middle or phonetic state is dedicated to modelling the segmental phonetic features. Given a segment $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T]$ with known boundaries, the likelihood for the segment using the full phonetic model $\lambda^{(seg)}$ is expressed as

$$P(\mathbf{X}|\lambda^{(seg)}) = P(\mathbf{x}_1|s_b) \cdot P(Y(2, T-1)|s_{ph}) \cdot P(\mathbf{x}_T|s_e) \cdot P(T|\gamma^{(dur)}) \quad (3)$$

where $Y(2, T-1)$ are the transformed phonetic features with the individual scores for the beginning frame, the phonetic feature columns and the end frame conditioned on the relevant state. $P(T|\gamma^{(dur)})$ is a probability measure of a segment having a duration of T frames calculated from a gamma distribution derived from duration statistics.

Model combination: Let $\lambda_j^{(rb)}$ denote independent phoneme models for each band b of resolution decomposition r, and $\lambda_j^{(seg)}$ an independent hybrid segmental model, for a phoneme j. The log-likelihood score of the segment belonging to class j is

$$g_j(\mathbf{X}) = \left[\sum_{r=0}^{R}\sum_{b=1}^{B_r}\omega_j^{(rb)}\log\left(\mathbf{X}_j^{(rb)}|\lambda_j^{(rb)}\right)\right] + \omega_j^{(seg)}\log\left(\mathbf{X}_j^{(seg)}|\lambda_j^{(seg)}\right) \quad (4)$$

Here B_r identifies the number of bands in sub-band decomposition level r ($B_0 = 1$ represents the full band case). It is assumed that $\mathbf{X}$ represents the complete sequence of observation vectors for the segment with the model-specific subsets appropriately identified as shown. Minimum classification error (MCE) training [3] of the class-dependent weights $\omega_j^{(rb)}$ and $\omega_j^{(seg)}$ is defined as follows. The misclassification measure $d_k(\mathbf{X})$ for an observation known to belong to class k is given by

$$d_k(\mathbf{X}) = -g_k(\mathbf{X}) + g_\eta(\mathbf{X}) \quad (5)$$

where η represents the model with the nearest score, i.e. the most confusable class. A smoothed continuous loss function is defined as a sigmoidal function of $d_k(\mathbf{X})$ (γ controls the slope of the sigmoid function), i.e.

$$\Gamma_k(\mathbf{X}) = \frac{1}{1+e^{-\gamma d_k(\mathbf{X})}} \quad (6)$$

Generalised probabilistic descent (GPD) token-by-token training implies that the gradient of the class-specific loss function drives the parameter updates [3]. Thus the weight update equation is

$$\omega_k^{n+1} = \omega_k^n - \varepsilon\frac{\partial\Gamma_k(\mathbf{X})}{\partial\omega_k^n} \quad (7)$$

where ω_k^n is a model weight after the nth iteration, $\partial\Gamma_k(\mathbf{X})/\omega_k$ is the gradient of the local loss function and ε is a small positive learning constant. The gradient function expanded according to the chain rule of calculus gives the update equations for the $(n+1)$th iteration below, where $m \in \{rb, seg\}$ identifies the model type.

$$\omega_k^{(m)^{n+1}} = \omega_k^{(m)^n} - \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X})-1])\log\left(\mathbf{X}_k^{(m)}|\lambda_k^{(m)}\right) \quad (8a)$$

$$\omega_\eta^{(m)^{n+1}} = \omega_\eta^{(m)^n} + \varepsilon(\Gamma_\eta(\mathbf{X})[\Gamma_\eta(\mathbf{X})-1])\log\left(\mathbf{X}_\eta^{(m)}|\lambda_\eta^{(m)}\right) \quad (8b)$$

Table 1: TIMIT classification results for model combination

Model combination	Unity weights	Trained weights
FB, SB_{21}, SB_{22}	70.4	72.8
FB, SB_{31}, SB_{32}, SB_{33}	71.5	73.7
FB, SB_{21}, SB_{22}, seg	73.3	75.3
FB, SB_{21}, SB_{22}, SB_{23}, seg	74.5	76.3

Baseline scores: FB = 69.2%, seg = 70.6%

Experimental results: Results are reported for phoneme classification on the core test set of the TIMIT database. This task remains important in assessing new acoustic modelling strategies with many recent results available for comparison. For the phonetic segmental model, the segmental features are calculated from MFCC vectors extracted with a window length of 15ms at frame rate of 1.5ms. The first four columns of the trajectory matrix (Fig. 1) were retained in training 48 mixture continuous density Gaussian distribution models. Sub-band MFCCs for the multi-resolution models were extracted at a 10ms rate for 25ms frames and supplemented by delta and delta-delta coefficients. For full band MFCCs, 13 coefficients are retained, with seven from each sub-band (regardless of the number of bands). Continuous density Gaussian HMMs with 20 mixtures were trained in each case. Equal width sub-band decomposition in the mel-frequency scale was applied, i.e. for two bands the band edges are (0, 2kHz, 7.9kHz) and for three bands (0, 0.9kHz, 2.7kHz, 7.9kHz). The baseline score for standard full-band HMMs is 69.2%. The phonetic segmental model itself improves on this giving a performance of 70.6. Table 1 gives the classification scores for various model combinations for unity and discriminative trained weightings. The multi-resolution models for full-band (FB) combined with two sub-bands (SB_{21}, SB_{22}) or three sub-bands or (SB_{31}, SB_{32}, SB_{33}) are all seen to improve the performance compared to full-band HMMs. The combination of the multi-resolution models with the segmental (seg) models yields yet a further significant increase in performance. This result indicates that likelihood combination based on an independence assumption is effective in creating a joint discriminative function that extends on the performance of any model type in isolation. This improvement is greatest for the segmental model combined with the full-band model and three sub-band models. The use of discriminatively trained weights extends the performance advantage in all combination variations. Changes to the weights are of the order of or ±5% from unity. As weight updates are proportional to the independent model score (eqn. 8), the effect of training the weights is to slightly modify the relative dynamic ranges of the model scores. The best classification rate of 76.3% achieved for a full-band model combined with three sub-band models and the segmental model compares among the best reported for this task.

Conclusions: A multi-resolution sub-band model and a novel segmental phonetic model have been described which independently improve the phonetic classification performance compared to that of full-band HMMs. It has been clearly demonstrated that a linear combination of the log-likehood scores from these different acoustic modelling strategies is effective in facilitating a joint discriminative framework. The use of class-dependent model weights trained according to a minimum classification error objective successfully extends this performance advantage. The highest phonetic classification rate reported of 76.3% improves significantly on a full-band HMM score of 69.2% and compares among the best recently reported for this task.

Acknowledgment: This work is supported by EPSRC grant GR/L60463.

18 November 1999
Electronics Letters Online No: 20000049
DOI: 10.1049/el:20000049

P. McCourt, N. Harte and S. Vaseghi (*School of Electrical and Electronic Engineering, Queens University Belfast, United Kingdom*)

References

1 OSTENDORF, M., DIGILAKIS, V., and KIMBALL, O.: 'From HMMs to segment models: A unified view of stochastic modelling for speech recognition', *IEEE Trans. Speech and Audio Processing*, 1996, **4**, (5), pp. 360–378
2 HARTE, N., VASEGHI, S., and MCCOURT, P.: 'A novel model for phoneme recognition using phonetically derived features'. Proc. EUSIPCO, 1998, pp. 1485–1488
3 JUANG, B., CHOU, W., and LEE, C.: 'Minimum classification error rate methods for speech recognition', *IEEE Trans. Speech and Audio Process.*, 1997, **5**, (3), pp. 257–265
4 BOURLARD, H., DUPONT, S., HERMANSKY, H., and MORGAN, N.: 'Towards sub-band based speech recognition'. Proc. EUSIPCO, 1996, pp. 1579–1582

Q-factor measurement of nonlinear superconducting resonators

X.S. Rao, C.K. Ong and Y.P. Feng

A novel method, in which a multi-bandwidth measurement technique and an extrapolation procedure are combined, is proposed for extracting the loaded Q-factor (Q_L) with improved accuracy from the non-Lorentzian resonances of nonlinear superconducting resonators.

Introduction: The nonlinear microwave surface impedance ($Z_S = R_S + jX_S$) of high temperature superconductors, i.e. its power dependence $Z_S(P)$, is of interest both for practical applications