USE OF GENERALIZED DYNAMIC FEATURE PARAMETERS FOR SPEECH RECOGNITION: MAXIMUM LIKELIHOOD AND MINIMUM CLASSIFICATION ERROR APPROACHES

C. Rathinavelu and L. Deng

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

In this study we implemented a speech recogniser based on the integrated view, proposed first in [2], on the speech pre-processing and speech modeling problems in the recogniser design. The integrated model we developed generalizes the conventional, currently widely used delta-parameter technique, which has been confined strictly to the pre-processing domain only, in two significant ways. First, the new model contains state-dependent weighting functions responsible for transforming static speech features into the dynamic ones in a slowly time-varying manner. Second, novel maximum-likelihood and minimum-classification-error based learning algorithms are developed for the model that allows joint optimization of the state-dependent weighting functions and the remaining conventional HMM parameters. The experimental results obtained from a standard TIMIT phonetic classification task provide preliminary evidence for the effectiveness of our new, general approaches to the use of the dynamic characteristics of speech spectra.

1. INTRODUCTION

During the past decade, use of the dynamic feature parameters associated with speech spectra has resulted in demonstrable success in enhancing the performance of speech recognition systems. In practically all these systems, however, the way in which the speech spectral dynamics is represented has been as naive as simply taking the differences of or taking other experimentally chosen combinations of the "static" feature parameters. This representation has been confined strictly within the speech preprocessing domain in the speech recogniser design.

The objective of the research reported in this paper is to generalise the already successful, despite its empirical nature, delta-cepstrum technique such that the design of the dynamic features of speech is gracefully integrated into the overall speech recogniser design including optimization of the speech model parameters. Although the basic principle guiding our research is sufficiently general and can be applied to all types of speech recognizers, we restrict our presentation to only the recognizer based on hidden Markov model (HMM) representation of the speech spectra/cepstra.

2. A STATISTICAL MODEL OF SPEECH INCORPORATING GENERALIZED DYNAMIC FEATURE PARAMETERS

The statistical model, called the *integrated HMM*, which incorporates generalised dynamic speech features described in this paper is an extension of the model from the earlier *unimodal* Gaussian version of the integrated HMM [2] to the current Gaussian *mixture* version. This statistical model integrates the dynamic features that belong traditionally to the preprocessing domain into the speech modeling process. The integration is accomplished by defining a set of HMM-state-dependent weighting functions, which serve the role of converting the static features to the dynamic ones in a time-varying manner, as a set of intrinsic parameters of the model that can be learned from the speech data.

Let $\mathcal{X} = \{\mathcal{X}^1, \mathcal{X}^2, \cdots, \mathcal{X}^L\}$ denote a set of L static-feature (vector) sequences (i.e., L variable-length tokens), and let $\mathcal{X}^l = \{\mathcal{X}_1^l, \mathcal{X}_2^l, \cdots, \mathcal{X}_{T^l}^l\}$ denote the l-th sequence having the length of T^l frames. The dynamic feature vector $\mathcal{Y}_t^l$ at time frame t is defined as a linear combination of the static features stretching over the interval f frames forward and b frames backward according to

$$\mathcal{Y}_t^l = \sum_{k=-b}^{f} w_{k,i,m} \mathcal{X}_{t+k}^l, \qquad 1 \le l \le L,\ 1 \le t \le T^l \quad (1)$$

where $w_{k,i,m}$ is the kth weighting coefficient associated with the mth mixture residing in the Markov state i.

A finite mixture Gaussian density associated with each integrated HMM state i (a total of N states) assumes the form

$$b_i(\mathcal{O}_t^l) = b_i(\mathcal{X}_t^l, \mathcal{Y}_t^l) = \sum_{m=1}^{M} c_{i,m} b_{i,m}(\mathcal{X}_t^l) b_{i,m}(\mathcal{Y}_t^l), \qquad (2)$$

where M is the number of mixture components, and $c_{i,m}$ is the mixture weight for the mth mixture in state i. In eqn.(2), $b_{i,m}(\mathcal{X}_t)$ and $b_{i,m}(\mathcal{Y}_t)$ are d-dimensional unimodal Gaussian densities.

3. THE TRAINING ALGORITHMS

In this paper, we report two distinct approaches to the training of the parameters of the integrated HMM. The first approach, based on the maximum-likelihood (ML) principle, has been described in [2] in detail and will not be given

0-7803-2431-5/95 $4.00 © 1995 IEEE

here. This section is thus concentrated on the training algorithm developed using the minimum-classification-error (MCE) approach.

Consider L variable-length training tokens, with each token consisting of a vector-valued augmented data sequence (l-th token):

$$\mathcal{O}^l = \{\mathcal{X}^l, \mathcal{Y}^l\} = \{(\mathcal{X}_1^l, \mathcal{Y}_1^l), (\mathcal{X}_2^l, \mathcal{Y}_2^l), \cdots, (\mathcal{X}_{T^l}^l, \mathcal{Y}_{T^l}^l)\},$$

where the dynamic portion ($\mathcal{Y}^l$) of the data sequence is related to the static one ($\mathcal{X}^l$) in a state-dependent manner according to Eqn.1.

In the supervised training mode which we assume, each training token $(\mathcal{X}_{T^l}^l, \mathcal{Y}_{T^l}^l)$ is known to belong to one of $\mathcal{K}$ classes $\{C^j\}_{j=1}^{\mathcal{K}}$. The goal of the MCE training is to find the classifier parameter set, denoted by $\Phi = \{\Phi^j\}_{j=1}^{\mathcal{K}}$, such that the probability of misclassifying any $\mathcal{O}^l$ is minimized and the resulting Φ gives the optimal solution of the classifier. In the integrated HMM, the classifier parameter set consists of all the state-dependent, mixture-dependent weighting functions $w_{k,i,m}$, together with the conventional HMM parameters (including Markov transition probabilities $a_{i,j}$, mixture weights $c_{i,m}$, mixture Gaussian mean vectors $(\mu_{x,i,m}, \mu_{y,i,m})$, and mixture Gaussian covariance matrices $(\Sigma_{x,i,m}, \Sigma_{y,i,m})$), for all the models each representing a distinctive class of the speech sounds to be classified.

3.1. The MCE Optimization Criterion

The first step in the formulation of the objective function is to choose an appropriate discriminant function $g_\kappa(\mathcal{O}^l, \Phi)$ according to the following decision rule for classification:

$$C(\mathcal{O}^l) = C^\kappa, \; if \; g_\kappa(\mathcal{O}^l, \Phi) = \max_j \; g_j(\mathcal{O}^l, \Phi) \qquad (3)$$

where $C(.)$ is the class associated with the test data $\mathcal{O}^l$ as determined by the classifier.

In our implementation of the integrated HMM, we choose the most likely (optimal) state path traversing the Markov model as the basis for defining the discriminant function. The log-likelihood score of the input utterance $\mathcal{O}^l$ along the optimal state sequence $\Theta^\kappa = \{\theta_1^\kappa, \theta_2^\kappa \cdots, \theta_{T^l}^\kappa\}$ for the model associated with the κth class Φ^κ can be written as

$$\begin{aligned} g_\kappa(\mathcal{O}^l, \Phi) &= \log P(\mathcal{O}^l, \Theta^\kappa | \Phi^\kappa) \\ &= \log P(\mathcal{O}^l | \Theta^\kappa, \Phi^\kappa) + \log P(\Theta^\kappa | \Phi^\kappa) \\ &= \sum_{t=1}^{T^l} \log b_{\theta_t^\kappa}(\mathcal{O}_t^l) + \sum_{t=1}^{T^l-1} \log a_{\theta_t^\kappa \theta_{t+1}^\kappa} \qquad (4) \end{aligned}$$

where $b_{\theta_t^\kappa}(\mathcal{O}_t^l)$ is the probability of generating the feature vector $\mathcal{O}_t^l$ at time t in state θ_t^κ by the model for class κth, $a_{\theta_t^\kappa \theta_{t+1}^\kappa}$ is the transition probability of the κth model, and T^l is the number of frames of the lth observation sequence.

Given a discriminant function, a misclassification measure for an input training utterance $\mathcal{O}^l$ from class κ can be defined as follows to quantify the classification behavior:

$$d_\kappa(\mathcal{O}^l, \Phi) = -g_\kappa(\mathcal{O}^l, \Phi) + \log \left[\frac{1}{\mathcal{K}-1} \sum_{j \neq \kappa} e^{g_j(\mathcal{O}^l, \Phi)\eta} \right]^{\frac{1}{\eta}}$$

where η is a positive number and $\mathcal{K}$ is the total number of classes. $d_\kappa(\mathcal{O}^l, \Phi)$ above is a quantity that indicates the degree of confusion between the correct class and the other competing classes for a given input utterance $\mathcal{O}^l$. When η approaches ∞, the misclassification measure becomes

$$\begin{aligned} d_\kappa(\mathcal{O}^l, \Phi) &\Rightarrow -g_\kappa(\mathcal{O}^l, \Phi) + \max_{j \neq \kappa} g_j(\mathcal{O}^l, \Phi) \\ &= -g_\kappa(\mathcal{O}^l, \Phi) + g_i(\mathcal{O}^l, \Phi), \qquad (5) \end{aligned}$$

with C^i being the most confusable class. Clearly, a positive value of $d_\kappa(\mathcal{O}^l, \Phi)$ indicates a misclassification and a negative value of $d_\kappa(\mathcal{O}^l, \Phi)$ implies a correct decision.

Given a misclassification measure, we further define a smoothed loss function for each class κ:

$$\Upsilon_\kappa(\mathcal{O}^l, \Phi) = \frac{1}{1 + e^{-\rho d_\kappa(\mathcal{O}^l, \Phi)}}, \quad \rho > 0 \qquad (6)$$

which approximates the classification error count. That is, the loss function assigns near-zero penalty when an input is correctly classified and assigns a near-unity penalty when an input is misclassified. The parameter ρ controls the slope of the above smoothed zero-one function.

Finally, given a loss function defined for each class, we define the overall loss function for the entire classifier as

$$\Upsilon(\mathcal{O}^l, \Phi) = \sum_{\kappa=1}^{\mathcal{K}} \Upsilon_\kappa(\mathcal{O}^l, \Phi) \delta[\mathcal{O}^l \in C^\kappa] \qquad (7)$$

where $\delta[\xi]$ is the Kronecker indicator function of a logic expression ξ that gives value 1 if the value of ξ is true and value 0 otherwise.

3.2. Gradient Computation

In the MCE discriminative training, the integrated HMM parameters are adaptively adjusted to reduce the overall loss function along a gradient descent direction. The following gradient equations are obtained by computing the partial derivatives of $\Upsilon(\mathcal{O}^l, \Phi)$ with respect to each integrated HMM parameter for a given training token $\mathcal{O}^l$ belonging to class κ:

$$\begin{aligned} \frac{\partial \Upsilon(\mathcal{O}^l, \Phi)}{\partial w_{k,i,m}(j)} &= -\psi_j \sum_{t \in T_i^l(j)} \gamma_{i,m,t}(j) \Lambda_{y,i,m}^{Tr}(j) \Sigma_{y,i,m}^{-1}(j) \mathcal{X}_{t+k}^l \\ \frac{\partial \Upsilon(\mathcal{O}^l, \Phi)}{\partial \mu_{x,i,m}(j)} &= \psi_j \sum_{t \in T_i^l(j)} \gamma_{i,m,t}(j) \Sigma_{x,i,m}^{-1}(j) \Lambda_{x,i,m}(j) \\ \frac{\partial \Upsilon(\mathcal{O}^l, \Phi)}{\partial \mu_{y,i,m}(j)} &= \psi_j \sum_{t \in T_i^l(j)} \gamma_{i,m,t}(j) \Sigma_{y,i,m}^{-1}(j) \Lambda_{y,i,m}(j) \\ \frac{\partial \Upsilon(\mathcal{O}^l, \Phi)}{\partial \tilde{c}_{i,m}(j)} &= \psi_j \sum_{t \in T_i^l(j)} \gamma_{i,m,t}(j)[1 - c_{i,m}(j)], \end{aligned}$$

where the set $T_i^l(j)$ includes all the time indices such that the state index of the state sequence at time t belongs to state ith in the Markov chain, i.e.

$$T_i^l(j) = \{t | \theta_t^j = i\}, \;\; 1 \le i \le N, \;\; 1 \le t \le T^l$$

Other quantities in the above equations are

$$\psi_j = \begin{cases} -\rho \Upsilon_\kappa(\mathcal{O}^l, \Phi)[1 - \Upsilon_\kappa(\mathcal{O}^l, \Phi)] & \text{if } j = \kappa \\ \rho \Upsilon_\kappa(\mathcal{O}^l, \Phi)[1 - \Upsilon_\kappa(\mathcal{O}^l, \Phi)] \frac{e^{\eta g_j(\mathcal{O}^l, \Phi)}}{\sum_{\iota \neq \kappa} e^{\eta g_\iota(\mathcal{O}^l, \Phi)}} & \text{if } j \neq \kappa \end{cases}$$

$$\begin{aligned} \Lambda_{x,i,m}(j) &= \mathcal{X}_t^l - \mu_{x,i,m}(j) \qquad (8) \\ \Lambda_{y,i,m}(j) &= \mathcal{Y}_t^l - \mu_{y,i,m}(j) \; and \\ \gamma_{i,m,t}(j) &= \frac{c_{i,m}(j) b_{i,m}^j(\mathcal{X}_t^l) b_{i,m}^j(\mathcal{Y}_t^l)}{b_i^j(\mathcal{O}_t^l)} \end{aligned}$$

Note in the above that ψ_j of eqn. (8) serves as adaptive step size of parameter adjustment. It can be seen from eqn. (8) that a substantial parameter adjustment is made when the absolute value of $d_\kappa(\mathcal{O}^l, \Phi)$ is small — that is, when the training token is likely to be misclassified. On the other hand, when the absolute value of d_κ is large, that is, when the input token is either unlikely to cause confusion or obviously an extreme outlier, then the amount of adjustment is accordingly reduced.

4. EXPERIMENTS

4.1. Task and corpus

The experiments described in this section are aimed at classifying the 61 quasi-phonemic labels defined in the TIMIT database. In keeping with the convention adopted by the speech recognition community, we folded 22 phone labels into the remaining 39 ones in determining classification accuracy.

The training set consists of 442 speakers, of both male and female, resulting in 3536 sentences from a training subset of the TIMIT database. The test set consists of 160 sentences (a total of 5775 phone tokens), spoken from 20 speakers completely disjoint from the training set.

4.2. Computation of the static speech features

The raw speech data in TIMIT was in the form of waveforms. The following is the analysis condition under which the static speech features are computed:

Frame size : 10 msec (160 samples)
Window type : Hanning (modified Hamming)
Window length : 32 msec (512 samples)
Features : Mel-frequency cepstrum coefficients (MFCC)

For the computation of MFCC, 25 triangular band pass filters are simulated, spaced linearly from 0 to 1 kHz and exponentially from 1 kHz to 8.86 kHz, with the adjacent filters overlapped in the frequency range by 50%. The FFT power spectrum points are combined using a weighted sum to simulate the output of the triangular filter. The MFCC (static features) are then computed according to

$$MFCC(p) = \sum_{r=1}^{25} S_r \cos\left(p \times [r - 0.5] \times \frac{\pi}{25}\right), \; 0 \leq p \leq 7$$

where S_r is the log-energy output of the rth mel-filter. An eight-component static feature vector is extracted every 10 msec throughout the signal. For the integrated HMM, only the static feature vectors are used as the raw data to the recogniser, which constructs the dynamic feature parameters internally within the recogniser.

4.3. Experimental Setup and Results

The main goal of the experiments designed in this study is to investigate the relative effectiveness of the generalized dynamic-parameter technique in comparison with the conventional one. Therefore, we have attempted to keep all other aspects of the speech models associated with both the conventional and the generalised techniques as much in common as possible, and to keep the recogniser structure as simple as possible. For both the integrated HMM and its benchmark counterpart (i.e. the conventional HMM using the pre-processor that appends the delta feature vectors into the static ones), each phone is represented by a three-state, left-to-right HMM with no skips. The covariance matrices in all the states of all the models are diagonal and are not tied.

For the ML approach, we have implemented three types of the integrated HMM according to the different constraints imposed on the state-dependent weights that define the generalised dynamic parameters. These three constraints are: "Relaxed" Constraint (RC), Linear Constraint (LC), and Nonlinear Constraint (NC) (see [2] for detail). For the MCE approach, the three types of the integrated HMM we have implemented differ from each other according to the three different ways of using initial model parameters before the MCE training takes place. These three initial model parameters were directly taken from the integrated HMMs trained by the ML criterion with RC, LC, and NC constraints.

Further, for all versions of the integrated HMM and the benchmark HMM, we have explored both context independent and the context dependent versions of the phonetic model. For the context independent version, a total of 39 models (39 ×3 = 117 states) were constructed, one for each of the 39 classes intended for the classification task. For the context dependent version, a total of 1209 states were constructed, with each three-state combination out of these states representing one allophone conditioned on predefined merged phonetic classes as left and right contexts (i.e. generalised triphone). These pre-defined merged phonetic classes (15 in total) were modified from the merged classes published in [4] and in [3].

The phonetic classification results are obtained to show dependence of the classification rate on a variety of factors including the nature of the dynamic-parameter constraints used to construct (in the case of ML training) or to initialise (in the case of MCE training) the integrated HMM, the number of Gaussian mixtures in the HMM state, and the context dependence/independence in the classification task. The results shown in Table 1 are obtained using the ML training. First, use of five mixtures in the HMM states produces significantly better results, uniformly across all four types of the speech models (both context dependent and context independent ones), than use of the unimodal Gaussian HMM. Second, compared with context independent models, use of context dependent models typically reduces the classification errors by about 40%, independent of other factors. Third, among all four types of the model evaluated, the NC version of the integrated HMM performs bet-

Type of Model	Context-Ind Rate		Context-Dept Rate	
	1 Mix	5 Mixs	1 Mix	5 Mixs
Benchmark	58.28%	62.41%	74.18%	76.98%
RC-IHMM	58.24%	62.84%	74.79%	77.14 %
LC-IHMM	50.48%	53.58%	71.73%	73.20%
NC-IHMM	58.79%	63.36%	75.79%	78.58%

Table 1. TIMIT 39-phone context independent (left) and context dependent (right) classification rate using ML training.

Type of Model	Context-Ind Rate		Context-Dep Rate	
	1 Mix	5 Mixs	1 Mix	5 Mixs
Benchmark	63.19%	67.19%	78.94%	80.04%
RC-Initial	64.25%	67.46%	79.64%	79.88%
LC-Initial	56.95%	58.74%	74.25%	76.45%
NC-Initial	64.59%	68.23%	79.81%	81.45%

Table 2. TIMIT 39-phone context independent and context dependent classification rate using MCE training.

ter than any of the remaining. Fourth, the superior performance achieved by the NC version of the integrated HMM is more significant for the context-dependent experiment than for the context-independent one, and more significant when five mixtures are used. The error rate reduction with use of the NC version of the integrated HMM is 7.0% relative to the benchmark HMM. Finally, the LC version of the integrated HMM produces more classification errors than any other model types evaluated, including the benchmark HMM.

The phonetic classification results shown in Table 2 are obtained using the MCE training. We observe from Table 2 that for both the context dependent and context independent classification tasks, the integrated HMM initialised by the ML-trained model with nonlinear constraint (last row, NC-Initial) is superior to the integrated HMM initialised otherwise and to the benchmark HMM. The best classification rate, 81.45%, achieved with use of the context-dependent integrated HMM trained by the MCE criterion starting from the nonlinear-constraint ML version of the integrated HMM, represents a 7.1% error rate reduction compared with the benchmark HMM (80.04%). It also represents a 13.4% error rate reduction compared with the same version of the integrated HMM except with only the ML training (78.58%, see Table 1). In general, moving from the ML training to the MCE training, we are able to achieve a classification error rate reduction ranging from 10% to 25% for the integrated HMM as well as for the benchmark HMM. To the best of our knowledge, even for the benchmark HMM (i.e. the same HMM evaluated in [1]), the results we report in the present study are the first demonstrating the effectiveness of the MCE training for the standard TIMIT 39-phone classification task. We also note from Table 2 the quantified superiority in performance of the context-dependent models over the context-independent ones, and of the five-mixture models over the unimodal Gaussian models, for both the benchmark models and the integrated models.

5. SUMMARY AND CONCLUSION

In comparison with the conventional technique exploring the dynamic features, our new, generalised dynamic-feature technique is based on a solid theoretical ground. Within the theoretical framework described in this paper, use of dynamic features of speech is automatically integrated as a sub-component of the overall speech modeling strategy, rather than being treated as just a narrow signal processing problem. Specifically, the new integrated HMM generalises the currently widely used dynamic-parameter (delta-cepstrum) technique in two ways. First, the model contains *state-dependent* weighting functions for transforming static speech features into the dynamic ones, instead of having the weights be pre-fixed by the pre-processor. Second, the theoretically motivated EM-like algorithm and the MCE procedure are developed for the integrated HMM that allows joint optimisation of the state-dependent weighting functions and the remaining conventional HMM parameters.

Starting from the original proposal for the integrated HMM [2], we find that moving from the ML training to the MCE training is particularly desirable. With the conventional HMM, the sole motivation for the use of the MCE training in place of the ML one is from the general consideration of minimizing error rate due to poor approximation of the HMM as a source model to true statistical characteristics of the speech process. While the same motivation applies to the integrated HMM, the MCE approach automatically eliminates the need for use of unrealistic and artificial constraints that are essential for the formulation of the integrated HMM based on the ML design philosophy. The constraints have been on the state-dependent weighting functions in the definition of the generalised dynamic parameters. Elimination of these constraints by moving away from the ML approach appears to be a significant contributing factor for the improvement of the classifier performance from the best classification rate of 78.58% obtained by the ML approach to that of 81.45% by the MCE approach.

REFERENCES

[1] W. Chow, C. H. Lee, B. H. Juang, and F. K. Soong, "A minimum error rate pattern recognition approach to speech recognition," *International J. Pattern Recognition and Artificial Intelligence*, Vol. 8, No. 1, 1994, pp. 5–31.

[2] L. Deng, "Integrated optimisation of dynamic feature parameters for hidden Markov modeling of speech", *IEEE Signal Processing Letters*, Vol.1, No. 4, 1994, pp. 66-69.

[3] L. Deng, M. Lennig, F. Seitz and P. Mermelstein. "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," *Computer Speech and Language*, Vol. 4, 1990, pp. 345-357.

[4] A. Ljolje, "High accuracy phone recognition using context clustering and quasi-triphonic models", *Computer Speech and Language*, Vol.8, pp. 129-151, 1994.