USE OF VECTOR-VALUED DYNAMIC WEIGHTING COEFFICIENTS FOR SPEECH RECOGNITION: MAXIMUM LIKELIHOOD APPROACH

Rathinavelu Chengalvarayan

Currently at: Speech Processing Group, Bell Labs
Lucent Technologies, Naperville, IL 60566, USA
Tel: (630) 224 6398, Fax: (630) 979 5915
Email: rathi@lucent.com

ABSTRACT

In this paper, an integrated approach to vector dynamic feature extraction is proposed in the design of a hidden Markov model (HMM) based speech recognizer. The integrated model we developed in this study generalizes the conventional, currently widely used dynamic-parameter technique, which has been confined strictly to the pre-processing domain only, in two significant ways. First, the new model contains state-dependent, vector-valued weighting functions responsible for transforming static speech features into the dynamic ones in a slowly time-varying manner. Second, a novel maximum-likelihood based training algorithm is developed for the model that allows joint optimization of the state-dependent, vector-valued weighting functions and the remaining conventional HMM parameters. The experimental results on alphabet classification demonstrate the effectiveness of the new model relative to standard HMM using dynamic features that have not been subject to optimization during training.

1. INTRODUCTION

In speech recognition, it is desirable to extract features that are focused on discriminating between classes. The spectral dynamic characteristics are shown to play a crucially important role in speech perception, and consonants are mainly perceived on the basis of the spectral transition into the following vowels [5]. In the past few years, use of the coefficients that measure dynamic changes in the spectra has resulted in demonstrated success in enhancing the performance of both speech recognition and speech parameter generation systems [1], [6], [10], [16]. 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 over an empirically determined fixed time span.

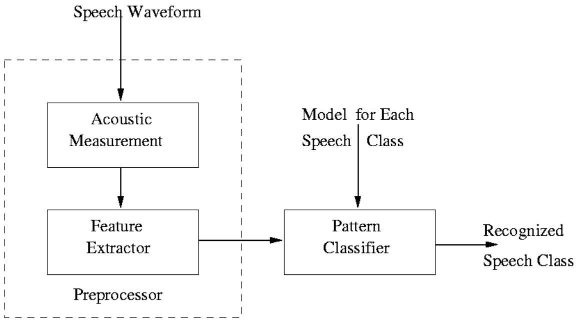

Figure 1. A block diagram of a typical speech recognizer.

The structure of many successful speech recognition systems typically consists of a feature analysis-extraction procedure followed by a statistical pattern classifier as shown in Figure 1. Usually, the back-end classifier is designed independently of the front-end preprocessor. However, there is no evidence that such a design stategy is the best for speech recognition. The recent advent of discriminative feature extraction (DFE) showed that improved recognition results can be obtained by using an integrated optimization of both the preprocessing and classification stages [2], [4], [9], [11], [13], [15]. In the conventional recognizer, features are extracted and then the classifier performs a mapping from feature space to discrimination space. The new integrated recognizer maps from the original acoustic measurement space to the optimized feature space and then maps from the optimized feature space to the discriminative space.

More recently, the cepstral time matrix [7], matrix coefficient filter [8] and time-varying linear filter coefficients [14] have shown to provide an optimal construction of dynamic parameters from existing static ones. We present in this paper an integrated model that 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, vector-valued weighting functions responsible for transforming static speech features into the dynamic ones in a slowly time-varying manner. Second, a novel maximum-likelihood (ML) based learning algorithm is developed for the model that allows joint optimization of the state-dependent weighting functions and the remaining conventional hidden Markov model (HMM) parameters. Only the static feature vectors are used as the raw data to the recognizer, which constructs the dynamic feature parameters internally within the recognizer.

2. CONSTRUCTION OF STATE-DEPENDENT VECTOR-VALUED DYNAMIC FEATURE PARAMETERS

The statistical model, called the *vector-valed dynamic integrated HMM* (VVD-IHMM), which incorporates generalized dynamic speech features described in this paper is an extension of the *scalar-valed dynamic integrated HMM* (SVD-IHMM) [14]. The state-dependent weights to tranform static speech features into dynamic ones as explained in [14] were considered as scalar valued. The more general case of vector-valued weighting coefficients is developed and implemented in this work. We show that our approach based on this technique is appropriate to model the dynamics of cepstra since regression is done independently for each dimension of the transformed cepstral space. 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 vector-valued 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 $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_T\}$ denote the the vector sequence of static feature parameters having the length of T frames. The dynamic feature vector $\mathcal{Y}_t$ at time frame t is defined as a simple combination of the static features stretching over the interval f frames forward and b frames backward according to

$$\mathcal{Y}_t \quad = \quad \mathcal{W}_{f,i}\mathcal{X}_{t+f} + \mathcal{W}_{b,i}\mathcal{X}_{t-b}, \qquad 1 \leq t \leq T, \qquad (1)$$

where $\mathcal{W}_{f,i}$ and $\mathcal{W}_{b,i}$ are the vector-valued weighting coefficients associated with the Markov state i. To simplify the discussion, we assume that the static and dynamic features are statistically independent. A Gaussian density associated with each VVD-IHMM state i (a total of N states) assumes the form

$$b_i(\mathcal{O}_t) \;=\; b_i(\mathcal{X}_t, \mathcal{Y}_t) \;=\; b_i(\mathcal{X}_t)b_i(\mathcal{Y}_t), \qquad (2)$$

where $\mathcal{O}_t$ is the augmented feature parameters at frame t consists of both static and the dynamic feature vectors. In the above equation $b_i(\mathcal{X}_t)$ and $b_i(\mathcal{Y}_t)$ are d-dimensional unimodal Gaussian densities for static and dynamic features respectively, as

$$b_i(\mathcal{X}_t) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{x,i}|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}[\mathcal{X}_t - \mu_{x,i}]^{Tr}\Sigma_{x,i}^{-1}[\mathcal{X}_t - \mu_{x,i}]\right)$$

$$b_i(\mathcal{Y}_t) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{y,i}|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}[\mathcal{Y}_t - \mu_{y,i}]^{Tr}\Sigma_{y,i}^{-1}[\mathcal{Y}_t - \mu_{y,i}]\right)$$

where variables $\mathcal{X}$ and $\mathcal{Y}$ indicate the static and the dynamic features, respectively. The parameters $\mu_{x,i}$, $\mu_{y,i}$ are the state-dependent Gaussian mean vectors and $\Sigma_{x,i}$, $\Sigma_{y,i}$ are the state-dependent diagonal covariance matrices. Superscripts -1 and Tr denote matrix inversion and vector transposition with d being the dimension of the static and dynamic feature vectors.

3. THE MAXIMUM LIKELIHOOD TRAINING PROCEEDURE

In this section, we describe closed-form solutions for jointly training all the integrated HMM parameters using the celebrated EM algorithm according to the maximum likelihood criterion. The algorithm consists of iterative E (expectation)-step and M (maximization)-step. The desirable objective function, which becomes suitable for maximization in the M-step, is established through a set of simplified E-step procedures [3]:

$$Q(\Phi|\Phi_0) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{t,i}\log\left(b_i(\mathcal{X}_t)b_i(\mathcal{Y}_t)\right). \qquad (3)$$

where the weight $\gamma_{t,i}$ is the probability of being in state i at time t which accounts for the observation data $\mathcal{X}_t, \mathcal{Y}_t$ conditioned on the previous model Φ_0 and this weight can be computed by using the standard forward and backward algorithms [12]. Re-estimates for the model parameters are obtained in the M-step via maximization of the objective function with respect to all model parameters. The re-estimation formulae for joint optimization of the state-dependent vector-valued weighting functions defining the dynamic features and of the Gaussian means for the dynamic features are derived as follows. The re-estimation formulae for the remaining parameters are similar to those for the conventional HMM [12].

After dropping several optimization-independent terms, the objective function equivalent to that in eqn. (3) is

$$Q_0(\mu_{y,i}, \mathcal{W}_{f,i}, \mathcal{W}_{b,i}) = \sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{t,i}[\mathcal{Y}_t - \mu_{y,i}]^{Tr}\Sigma_{y,i}^{-1}[\mathcal{Y}_t - \mu_{y,i}]$$

In order for the maximum-likelihood approach to the problem of estimating dynamic feature parameters to be sensible, constraints on the parameters $\mathcal{W}_{f,i}$, and $\mathcal{W}_{b,i}$ must be provided. This is so because infinitely high likelihood would be achieved by uniformly setting $\mathcal{W}_{f,i} = 0$ and $\mathcal{W}_{b,i} = 0$ without discriminability among different speech classes. In this study, we explore non-linear type of constraint that is imposed on the solution of the problem of jointly optimizing $\mathcal{W}_{f,i}$, $\mathcal{W}_{b,i}$ and $\mu_{y,i}$ in the M-step of the EM algorithm: $\mathcal{W}_{f,i}^2 + \mathcal{W}_{b,i}^2 = \mathcal{C}$, where $\mathcal{C} \neq 0$ is a model-specific constant, serving the role of eliminating the possibility that all $\mathcal{W}_{f,i}$ and $\mathcal{W}_{b,i}$ are set to zero (singularity) thereby giving infinitely large but senseless likelihood. Note that the use of discriminative training of parameter learning could eliminate the need for the non-linear constraint [14]. The simple dynamic features become a degenerative instance of the above model when $\mathcal{C} = 2$, $\mathcal{W}_{f,i}^2 = 1$ and $\mathcal{W}_{b,i}^2 = -1$.

The Lagrangian of $Q_0(\mu_{y,i}, \mathcal{W}_{f,i}, \mathcal{W}_{b,i})$ with respect to the non-linear constraint can be written as

$$Q_0^L = Q_0(\mu_{y,i}, \mathcal{W}_{f,i}, \mathcal{W}_{b,i}) + \sum_{i=1}^{N} \lambda_i \left(\mathcal{W}_{f,i}^2 + \mathcal{W}_{b,i}^2 - \mathcal{C} \right),$$

where λ_i's are Lagrange multipliers. Setting the partial derivatives of Q_0^L with respect to $\mathcal{W}_{f,i}$, $\mathcal{W}_{b,i}$ λ_i and $\mu_{y,i}$ to zero, we establish the following set of non-linear system of equations:

$$\begin{aligned}
\sum_{t=1}^{T} \gamma_{t,i} [\mathcal{Y}_t - \hat{\mu}_{y,i}]^{Tr} \Sigma_{y,i}^{-1} \mathcal{X}_{t+f} + \hat{\lambda}_i \hat{\mathcal{W}}_{f,i} &= 0, \\
\sum_{t=1}^{T} \gamma_{t,i} [\mathcal{Y}_t - \hat{\mu}_{y,i}]^{Tr} \Sigma_{y,i}^{-1} \mathcal{X}_{t-b} + \hat{\lambda}_i \hat{\mathcal{W}}_{b,i} &= 0, \\
\hat{\mathcal{W}}_{f,i}^2 + \hat{\mathcal{W}}_{b,i}^2 - \mathcal{C} &= 0, \\
\sum_{t=1}^{T} \gamma_{t,i} [\mathcal{Y}_t - \hat{\mu}_{y,i}] &= 0.
\end{aligned}$$

There are no general methods for solving systems involving more than one non-linear equations in a closed form. The re-estimation formulae is then established by solving a system of the above four non-linear equations, we applied the Newton-Raphson method to obtain an iterative solution [3], with respect to each of the unknown model paramerters.

4. ENGLISH ALPHABET CLASSIFICATION EXPERIMENTS

The experiments conducted to evaluate the various integrated IHMMS are aimed at recognizing the 26 letters in the English alphabet, contained in the TI46 speaker dependent isolated word corpus. The speaker-independent training set consists of 10 tokens per word from two male and two female speakers (m1, m2, f1 and f2). The remaining 16 tokens per word for each of the above four speakers is used as test data. The preprocessor produces a vector of 13 Mel-frequency cepstral coefficients (MFCCs) for every 10 msec throughout the signal. The augmented feature vectors used for the benchmark HMM consist of 26-elements, with 13 cepstrum coefficients and 13 delta cepstra. The delta MFCCs are constructed by taking the difference between two frame forward and two frame backward of the MFCCs. This window length of 50ms is found to be optimal in capturing the slope of the spectral envelope, i.e. the transitional information [14]. For the integrated HMM, only the static feature vectors are used as the raw data to the recognizer, which constructs the dynamic feature parameters internally within the recognizer according to (1). To be consistent with the conventional delta parameter techniques, the window variables f, b and the model-specific constant C are set to 2.

Type of Model	Classification Rate
Conventional HMM	80.11%
SVD-IHMM	81.55%
VVD-IHMM	82.57%

Table 1. TI 26-alphabet speaker-independent classification rate as a function of the model type.

The main goal of the experiments designed in this study is to investigate the relative effectiveness of the vector-valued dynamic-parameter technique in comparison with the conventional and scalar-valued dynamic-parameter techniques. Each word is represented by a single left-to-right, three-state HMM (no skips), with single Gaussian state observation densities. The covariance matrices in all the states of all the models are diagonal and are not tied. All transition probabilities are uniformly set to 0.5 (all transitions from a state are considered equally likely) and are not learned during the training process. The conventional HMM models are trained from training data using five-iterations of the ML training with single mixture for each state in the HMMs [12]. The scalar-valued dynamic integrated HMM (SVD-IHMM) are trained using five-iterations of the ML algorithm with non-linear type constraint [14]. The vector-valued dynamic integrated HMM (VVD-IHMM) are trained according to the training procedure outlined in the previous section.

The experimental results are summarized in Table 1. We observe from Table 1 that both the integrated HMM trained by nonlinear constraint is superior to the conventional HMM. The SVD-IHMM based classifier produces 81.55% accuracy with an error rate reduction of 7.2% compared with the convention HMM classifier's performance. This error rate reduction is consistent with our previous experiments using TIMIT database reported in [3]. From the final classifier based on VVD-IHMM, which incorporated vector-valued dynamic weighting functions, the best classification results have been obtained, shown as VVD-IHMM in Table 1. The recognition rate using the VVD-IHMM improved from 80.11% (conventional ML-trained HMM) to 82.57% which translates to 12.4% error rate reduction. It also represents a 5.5% error rate reduction compared with SVD-IHMM. Among all three types of the model evaluated, the VVD-IHMM performs better than the any of the remaining.

5. CONCLUSIONS

We have proposed in this paper, an integreted view on speech preprocessing and speech modeling in the design of HMM-based speech recognizers. The new integrated HMM generalizes the currently widely used dynamic-parameter technique in two ways. First, the model contains *state-dependent, vector-valued* weighting functions for transforming static speech features into the dynamic ones. Second, the EM algorithm is developed for the integrated HMM that allows joint optimization of the state-dependent weighting functions and the remaining conventional HMM parameters.

The state-dependent weights to tranform static speech features into dynamic ones as explained in [14] were considered as scalar valued. The more general case of vector-valued weighting coefficients is developed and implemented in this work. We found that our approach based on this technique is appropriate to model the dynamics of cepstra since regression is done independently for each dimension of the transformed cepstral space. The best error rate reduction of 12.4% is obtained using the new model, tested on a TI alphabet classification task, relative to conventional HMM. Compared across all three classifiers, VVD-IHMM produced the lowest error rate and is the new efficient way of describing the dynamic characteristics of speech cepstra. Although we restrict our presentation to only the recognizer based on HMM, the basic principle guiding our research is sufficiently general and can be applied to all types of speech recognizers.

REFERENCES

[1] T. Beppu and K. Aikawa, "Spontaneous Speech Recognition Using Dynamic Cepstra Incorporating Forward and Backward Masking Effect", *Proc. EUROSPEECH*, Vol. 1, pp. 511-514, Madrid, September, 1995.

[2] A. Biem and S. Katagiri, "Cepstrum-Based Filter-Design Using Discriminative Feature Extraction Training at Various Levels", *Proc. ICASSP*, Vol. 2, pp. 1503-1506, Munich, April, 1997.

[3] L. Deng and C. Rathinavelu, "Constructon of State-Dependent Dynamic Parameters Using the Maximum Likelihood Approach: Applications to Speech Recognition", *Signal Processing*, Vol. 55, pp. 149-165, December, 1996.

[4] S. Euler, "Integrated Optimization of Feature Transformation for Speech Recognition", *Proc. EUROSPEECH*, Vol. 1, pp. 109-112, Madrid, September, 1995.

[5] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", *IEEE Transactions on ASSP*, Vol. 34, No. 1, February, 1986.

[6] V. N. Gupta, M. Lennig and P. Mermelstein, "Integration of Acoustic Information in a Large Vocabulary Word Recognizer", *Proc. ICASSP*, Vol. 2, pp. 697-700, Dallas, 1987.

[7] N. Harte, S. Vaseghi and B. Milner, "Dynamic Features for Segmental Speech Recognition", *Proc. ICSLP*, Vol. 2, pp. 933-936, Philadelphia, October, 1996.

[8] K. Katagishi, H. Singer, K. Aikawa and S. Sagayama, "Feature Extraction Using a Matrix Coefficient Filter for Speech Recognition", *Speech Communication*, Vol. 13, pp. 297-306, December, 1993.

[9] C. S. Liu, "A General Framework of Feature Extraction: Application to Speaker Recognition", *Proc. ICASSP*, Vol. 2, pp. 669-672, Atlanta, May 1996.

[10] J. Nouza, "On the Speech Feature Selection Problem: Are Dynamic Features More Important Than the Static Ones?", *Proc. EUROSPEECH*, Vol. 2, pp. 919-922, Madrid, September, 1995.

[11] K. K. Paliwal, M. Bacchiani and Y. Sagisaka, "Minimum Classification Error Training Algorithm for Feature Extractor and Pattern Classifier in Speech Recognition", *Proc. EUROSPEECH*, Vol. 1, pp. 541-544, Madrid, September, 1995.

[12] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *The IEEE proceedings*, Vol.77, No. 2, pp. 257-285, February, 1989.

[13] M. Rahim and C. H. Lee, "Simultaneous ANN Feature and HMM Recognizer Design Using String-Based Minimum Classification Error (MCE) Training", *Proc. ICSLP*, Vol. 3, pp. 1824-1827, Philadelphia, October, 1996.

[14] C. Rathinavelu and L. Deng, "Use of Generalized Dynamic Feature Parameters for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, pp. 232-242, May, 1997.

[15] E. G. Schukat-Talamazzini, J. Hornegger and H. Niemann, "Optimal Linear Feature Transformations for Semi-Continuous Hidden Markov Models", *Proc. ICASSP*, Vol. 1, pp. 369-372, Detroit, May 1995.

[16] K. Tokuda, T. Kobayashi and S. Imai, "Speech Parameter Generation form HMM Using Dynamic Features", *Proc. ICASSP*, Vol. 1, pp. 660-663, Detroit, 1995.