HMM-BASED SPEECH RECOGNITION USING STATE-DEPENDENT, LINEAR TRANSFORMS ON MEL-WARPED DFT FEATURES

C. Rathinavelu and L. Deng

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

ABSTRACT

In this paper, we investigate the interactions of front-end feature extraction and back-end classification techniques in HMM based speech recognizer. This work concentrates on finding the optimal linear transformation of Mel-warped short-time DFT information according to the minimum classification error criterion. These transformations, along with the HMM parameters, are automatically trained using the gradient descent method to minimize a measure of overall empirical error count. The discriminatively derived state-dependent transformations on the DFT data are then combined with their first time derivatives to produce a basic feature set. Experimental results show that Mel-warped DFT features, subject to appropriate transformation in a state-dependent manner, are more effective than the Mel-frequency cepstral coefficients that have dominated current speech recognition technology. The best error rate reduction of 9% is obtained using the new model, tested on a TIMIT phone classification task, relative to conventional HMM.

1. INTRODUCTION

The recent advent of discriminative feature extraction showed that improved recognition results can be obtained by using an integrated optimization of both the preprocessing and classification stages [4]. Various modeling techniques such as filter bank, lifter and generalized dynamic feature design have been proposed for combining the preprocessing stage with the classification stage [1, 4, 5]. This problem is important because as the modeling technique is drastically improving over the recent past, further advances in speech recognition will likely to come from better feature extraction. 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.

Discrete cosine transform (DCT) is a linear operation that can be used for mapping Mel-warped Discrete Fourier Transform (DFT) (in the form of Mel filter bank (MFB) log-channel energies) into a lower dimensional feature space, giving rise to Mel-frequency cepstral coefficients (MFCC) widely in use for speech recognition [2]. Despite the empirical superiority of MFCC over other types of signal processing techniques, there are no theoretical reasons why the linear transformation associated with DCT, which is fixed a-priori and independent of HMM states and of speech classes, on MFB log-channel energies is an optimal one as far as the speech recognition performance is concerned. This work concentrates on finding the optimal linear transformation of Mel-warped short-time DFT information according to the minimum classification error (MCE) criterion. These transformations, along with the HMM parameters, are automatically trained using the gradient descent method to minimize a measure of overall empirical error count. These discriminatively derived state-dependent transformations on the DFT data are then combined with their first time derivatives to produce a basic feature set. The new model, which we call optimum-Transformed HMM (THMM), uses only the MFB log-channel energies derived from Mel-warped short-time DFT as the raw data to the recognizer, both static and dynamic features are automatically constructed within the recognizer.

2. CONSTRUCTION OF STATE-DEPENDENT LINEAR TRANSFORMS

The THMM described in this paper integrates the input features into the modeling process using the transformation matrices as a set of trainable parameters of the model. Let $\mathcal{F} = \{\mathcal{F}^1, \mathcal{F}^2, \cdots, \mathcal{F}^L\}$ denote a set of L MFB log-channel energy vector sequences (vector is of n dimension and L is the total number of tokens), and let $\mathcal{F}^l = \{\mathcal{F}^l_1, \mathcal{F}^l_2, \cdots, \mathcal{F}^l_{T^l}\}$ denote the l-th sequence having the length of T^l frames. The static feature vector $\mathcal{X}^l_t$ at time frame t of l-th token is defined as a linear combination of each row of transformation matrix with each element of MFB log-channel energy vector at time t, according to

$$\mathcal{X}^l_{p,t} = \sum_{q=1}^{n} \mathcal{B}_{p,q,i,m} \mathcal{F}^l_{q,t} \quad p = 1, 2, \cdots, n \quad t = 1, 2, \cdots, T^l$$

$$\mathcal{X}^l_t = \mathcal{B}_{i,m} \mathcal{F}^l_t$$

In the matrix form, the above equation can be written as

$$\begin{pmatrix} \mathcal{X}^l_{1,t} \\ \mathcal{X}^l_{2,t} \\ \vdots \\ \mathcal{X}^l_{d,t} \end{pmatrix} = \begin{pmatrix} \mathcal{B}_{1,1,i,m} & \mathcal{B}_{1,2,i,m} & \cdots & \mathcal{B}_{1,n,i,m} \\ \mathcal{B}_{2,1,i,m} & \mathcal{B}_{2,2,i,m} & \cdots & \mathcal{B}_{2,n,i,m} \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{B}_{d,1,i,m} & \mathcal{B}_{d,2,i,m} & \cdots & \mathcal{B}_{d,n,i,m} \end{pmatrix} \begin{pmatrix} \mathcal{F}^l_{1,t} \\ \mathcal{F}^l_{2,t} \\ \vdots \\ \mathcal{F}^l_{n,t} \end{pmatrix},$$

where $\mathcal{B}_{p,q,i,m}$ is the pq-th element of the transformation matrix $\mathcal{B}_{i,m}$ associated with the m-th mixture residing in

0-7803-3192-3/96 $5.00©1996 IEEE

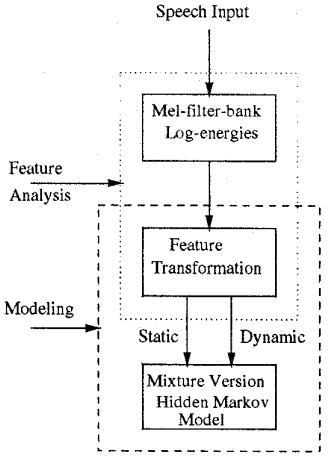

Figure 1. A block diagram of the optimum-Transformed HMM.

the Markov state i, n is the number of MFB log-channel energies for each frame and d is the number of static feature elements used in the modeling process. The dynamic feature vectors at time t are constructed by taking the difference between 2 frame forward and 2 frame backward of the static features according to

$$\begin{aligned} \mathcal{Y}_t^l &= \mathcal{X}_{t+2}^l - \mathcal{X}_{t-2}^l \\ &= \mathcal{B}_{i,m}\mathcal{F}_{t+2}^l - \mathcal{B}_{i,m}\mathcal{F}_{t-2}^l \\ &= \mathcal{B}_{i,m}[\mathcal{F}_{t+2}^l - \mathcal{F}_{t-2}^l] \end{aligned}$$

This window length of 40ms is found to be optimal in capturing the slope of the spectral envelope i.e. the transitional information [5]. The augmented static and dynamic features are provided as the data input for every frame of speech into the modeling stage as shown in Figure 1. A finite mixture Gaussian density associated with each state i (a total of N states) assumes the form

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

where $\mathcal{O}_t^l$ is the augmented feature vector of the l-th token at frame t, M is the number of mixture components, and $c_{i,m}$ is the mixture weight for the mth mixture in state i. In the above equation, $b_{i,m}(\mathcal{X}_t)$ and $b_{i,m}(\mathcal{Y}_t)$ are d-dimensional the unimodal Gaussian densities, variables $\mathcal{X}$ and $\mathcal{Y}$ indicate the static and the dynamic features.

3. MCE CRITERION FOR TRAINING MODEL PARAMETERS

Discriminative training algorithm has been successfully used by several researchers in speech recognition tasks to improve the ML criterion [3]. In the supervised training mode, each training token $\mathcal{O}^l$ is known to belong to one of $\mathcal{K}$ classes $\{C^j\}_{j=1}^{\mathcal{K}}$. The recognizer is represented as a set of parameters $\Phi = \{\Phi^j\}_{j=1}^{\mathcal{K}}$, which includes the feature extraction parameters as well as the classification parameters. The goal is to reduce the number of misclassifications occuring over this set through a minimization of the overall loss function $\Upsilon(\mathcal{O}^l, \Phi)$, which is a reflection of the classification errors. In THMM, the classifier parameter set consists of all the state-dependent, mixture-dependent transformation matrices $\mathcal{B}_{i,m}$ together with the conventional HMM parameters (including 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. The overall loss function is constructed and minimized through the following steps:

1. *Discriminant function*: The log-likelihood score of the input utterance $\mathcal{O}^l$ along the optimal state sequence $\Theta = \{\theta_1, \theta_2 \cdots, \theta_{T^l}\}$ for the model associated with the κth class Φ^κ can be written as

$$g_\kappa(\mathcal{O}^l, \Phi) = \sum_{t=1}^{T^l} \log b_{\theta_t}^\kappa(\mathcal{O}_t^l)$$

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 by the model for κth class. The implied decision rule for classification is defined as

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

2. *Misclassification measure*: Given a discriminant function, the misclassification measure for an input training utterance $\mathcal{O}^l$ from class κ becomes

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

where C^χ is the most confusable class. $d_\kappa(\mathcal{O}^l, \Phi) > 0$ implies misclassification and $d_\kappa(\mathcal{O}^l, \Phi) \leq 0$ means correct classification.

3. *Loss function*: The loss function is defined as a sigmoid, non-decreasing function of d_κ:

$$\Upsilon_\kappa(\mathcal{O}^l, \Phi) = \frac{1}{1 + e^{-d_\kappa(\mathcal{O}^l, \Phi)}}$$

which approximates the classification error count.

4. *Overall loss function:* The overall loss function for the entire classifier is defined for each class as

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

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.

5. *Minimization*: The loss function $\Upsilon(\mathcal{O}^l, \Phi)$ is minimized, each time a training token $\mathcal{O}^l$ is presented, by adaptively adjusting the parameter set Φ according to

$$\Phi_{l+1} = \Phi_l - \epsilon \nabla \Upsilon(\mathcal{O}^l, \Phi_l),$$

where Φ_l is the parameter set at the lth iteration, $\nabla\Upsilon(\mathcal{O}^l, \Phi_l)$ is the gradient of the loss function for training sample $\mathcal{O}^l$ and ϵ is a small positive learning constant.

4. GRADIENT CALCULATION

The THMM parameters are adaptively adjusted to reduce the overall loss function along a gradient descent direction. The gradient equations are obtained by computing the partial derivatives of $\Upsilon(\mathcal{O}^l, \Phi)$ with respect to each THMM parameter for a given training token $\mathcal{O}^l$ belonging to class κ. For the sake of keeping the discussion simple, we present the gradient derivations for the newly introduced feature parameters. Let $\mathcal{B}^j_{i,m}$ denote a feature extraction parameter associated with model j, then in the case of token-by-token training, we can write the gradient as

$$\begin{aligned}\frac{\partial \Upsilon(\mathcal{O}^l,\Phi)}{\partial \mathcal{B}^j_{i,m}} &= \frac{\partial}{\partial \mathcal{B}^j_{i,m}}\left(\sum_{\kappa'=1}^{\mathcal{K}} \Upsilon_{\kappa'}(\mathcal{O}^l,\Phi)\delta[\mathcal{O}^l \in C^{\kappa'}]\right) \\ &= \frac{\partial}{\partial \mathcal{B}^j_{i,m}}\Upsilon_\kappa(\mathcal{O}^l,\Phi) \\ &= \frac{\partial \Upsilon_\kappa(\mathcal{O}^l,\Phi)}{\partial d_\kappa(\mathcal{O}^l,\Phi)}\frac{\partial d_\kappa(\mathcal{O}^l,\Phi)}{\partial g_j(\mathcal{O}^l,\Phi)}\frac{\partial g_j(\mathcal{O}^l,\Phi)}{\partial \mathcal{B}^j_{i,m}} \quad (1)\end{aligned}$$

The first two factors in the right-hand-side of eqn. (1) can be simplified to

$$\psi_j = \begin{cases} \Upsilon_\kappa(\mathcal{O}^l,\Phi)[\Upsilon_\kappa(\mathcal{O}^l,\Phi)-1] & \text{if } j=\kappa \\ \Upsilon_\kappa(\mathcal{O}^l,\Phi)[1-\Upsilon_\kappa(\mathcal{O}^l,\Phi)] & \text{if } j=\chi \end{cases} \quad (2)$$

The third factor of the right-hand-side of eqn. (1) can be modified to

$$\begin{aligned}\frac{\partial g_j(\mathcal{O}^l,\Phi)}{\partial \mathcal{B}^j_{i,m}} &= \frac{\partial}{\partial \mathcal{B}^j_{i,m}}\sum_{t=1}^{T^l}\log b^j_{\theta_t}(\mathcal{O}^l_t) \quad (3) \\ &= \sum_{t\in T^l_i}\frac{1}{b^j_i(\mathcal{O}^l_t)}\frac{\partial}{\partial \mathcal{B}^j_{i,m}}\sum_{k=1}^{M} c^j_{i,k} b^j_{i,k}(\mathcal{X}^l_t) b^j_{i,k}(\mathcal{Y}^l_t) \\ &= \sum_{t\in T^l_i}\frac{c^j_{i,m} b^j_{i,m}(\mathcal{X}^l_t) b^j_{i,m}(\mathcal{Y}^l_t)}{b^j_i(\mathcal{O}^l_t)}\frac{\partial}{\partial \mathcal{B}^j_{i,m}}\frac{-1}{2} \\ &\quad \left([\mathcal{X}^l_t-\mu^j_{x,i,m}]^{Tr}\Sigma^{-1}_{x,i,m}(j)[\mathcal{X}^l_t-\mu^j_{x,i,m}]\right. \\ &\quad \left.+[\mathcal{Y}^l_t-\mu^j_{y,i,m}]^{Tr}\Sigma^{-1}_{y,i,m}(j)[\mathcal{Y}^l_t-\mu^j_{y,i,m}]\right)\end{aligned}$$

where the set T^l_i 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.

$$T^l_i = \{t|\theta_t = i\}, \quad 1\le i\le N, \quad 1\le t\le T^l$$

the *a posteriori* probabilities are defined as:

$$\gamma^j_{i,m}(t) = \frac{c_{i,m} b_{i,m}(\mathcal{X}^l_t) b_{i,m}(\mathcal{Y}^l_t)}{b_i(\mathcal{O}^l_t)}$$

In the remaining of this section, class index j will be omitted for clarity of presentation. Using eqns. (2), (3) and applying the chain rule results in eqn. (1) the gradient calculation becomes:

$$\begin{aligned}\frac{\partial \Upsilon(\mathcal{O}^l,\Phi)}{\partial \mathcal{B}_{i,m}} &= -\psi\sum_{t\in T^l_i}\gamma_{i,m}(t)\left(\Sigma^{-1}_{x,i,m}[\mathcal{X}^l_t-\mu_{x,i,m}][\mathcal{F}^l_t]^{Tr}\right. \\ &\quad \left.+\Sigma^{-1}_{y,i,m}[\mathcal{Y}^l_t-\mu_{y,i,m}][\mathcal{F}^l_{t+2}-\mathcal{F}^l_{t-2}]^{Tr}\right)\end{aligned}$$

To reduce the computational complexity as well as the model complexity, we tied all the mixtures for feature transformation matrices $\mathcal{B}_{i,m}$ to a single state parameter $\mathcal{B}_i$. For this special case, the gradient can be given by:

$$\begin{aligned}\frac{\partial \Upsilon(\mathcal{O}^l,\Phi)}{\partial \mathcal{B}_i} &= -\psi\sum_{t\in T^l_i}\sum_{m=1}^{M}\gamma_{i,m}(t)\left(\Sigma^{-1}_{x,i,m}[\mathcal{X}^l_t-\mu_{x,i,m}]\right. \\ &\quad \left.[\mathcal{F}^l_t]^{Tr}+\Sigma^{-1}_{y,i,m}[\mathcal{Y}^l_t-\mu_{y,i,m}][\mathcal{F}^l_{t+2}-\mathcal{F}^l_{t-2}]^{Tr}\right) \quad (4)\end{aligned}$$

The gradient formulae for the remaining parameters are similar to those for the conventional HMM.

5. EXPERIMENTAL EVALUATION

The THMM described above is evaluated on a standard TIMIT speaker independent database, aiming at classifying 61 quasi-phonemic TIMIT labels folded into 39 classes. Classification is performed, instead of recognition, to focus on the front-end processing and aimed at observing the accuracy of THMM on the speech representation and speech modeling. The training-set consists of 3536 sentences from 442 speakers and the test-set consists of 160 sentences from 20 speakers. MFB log-channel energies, are computed by simulating 21 triangular filters spacing linearly, from 0 to 500Hz, and exponentially, from 500Hz to 8500Hz, and overlapped by 50% for every 10ms of speech. Each phone is represented by a simple left-to-right, 3-state HMM with mixture Gaussian state observation densities. We perform a total of 5 epochs of training and only the best-incorrect-class is used in the misclassification measure. For context-independent (CI) model, a total of 39 models ($39 \times 3 = 117$ states) were constructed, one for each of the 39 classes intended for the classification task. The procedure outlined in paper [5] has been adopted to create context-dependent (CD) models, which results in a total of 1209 HMM states. The ML trained benchmark HMM with state-dependent DCT matrices is provided as the initial model for MCE training of THMM.

First, preliminary experiments are conducted using a subset of training-set, which consists of 320 sentences from each of 40 speakers, and test-set with single mixture CI phone models. The results are plotted in Figure 2 showing classification rate as a function of the number of rows in the feature transformation matrix, with ML trained HMM as dash-dash line, MCE trained HMM as dash-dot line and MCE trained THMM as dot-dot line. From these results, we conclude that the performance remains fairly constant after 12 dimensions. As one might expect, increasing the number of dimensions does help, but only upto a point (performance starts degrading after 18 dimension). In our following experimental evaluation we choose the dimension of the linear transformation matrix to be 12×21 as optimal.

Given the 12 dimensions determined from above, then a series of comparative experiments are carried out using full sets of training and test data in TIMIT, to examine the effectiveness of MCE training on the proposed THMM. The classification rates for various experimental conditions are summarized in Figure 3. For performance comparison, a conventional HMM was first implemented. The conventional ML-HMM is trained using 5-iterations of

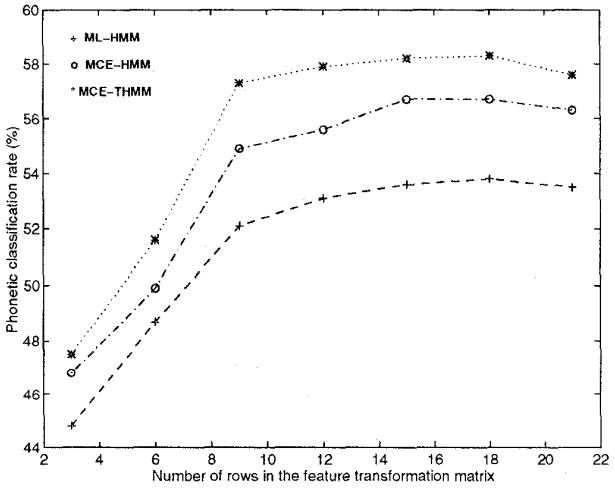

Figure 2. Dimensionality selection for the feature transformation matrix.

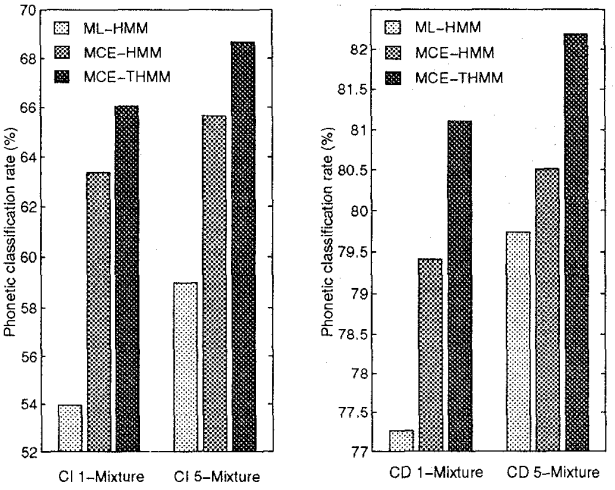

Figure 3. TIMIT 39-phone context independent (the six representations on the left) and context dependent (the six representations on the right) classification rate as a function of the model type (all using MCE training except the ML trained initial HMM) of the number of Gaussian mixtures in the HMM state.

Baum-Welch re-estimation and MCE-HMM is obtained by discriminative training. As can be seen from Figure 3, the performance is significantly improved by the MCE training method. For the THMM, the initial state-dependent DCT matrices are discriminatively trained according to eqn. (4). The results corresponding to 5-mixtures CD model (82.19%) indicate a significant reduction in error rate (9%) compared to the MCE-HMM result. From the results shown in Figure 3, the THMM outperforms the MCE trained HMM by about 7% in error rate on average for all cases. It is interesting to observe that, the single mixture THMM performs better than the 5-mixtures MCE-based HMM in case of both CI and CD models, indicating a clear superiority of THMM with the comparable number of state parameters. (Number of state parameters for 1-mixture THMM is $21 \times 12 + 26 + 26 = 304$ and similarly for 5-mixtures HMM it is $5 \times (26 + 26 + 1) = 265$). The results clearly demonstrated the effectiveness of new approach.

6. CONCLUSIONS

We have proposed an integrated technique, based on discriminative feature extraction for feature reduction of the MFB log-channel energy space. The entire HMM recognizer, consisting part of the preprocessing as well as the classifier, was trained with the MCE training algorithm. We presented experimental results for the optimally designing generalized feature (cepstrum) representation for phone classification. The best classification rate (an error rate reduction of 9%) of 82.19% was obtained using 5-mixtures context-dependent THMM, tested on a TIMIT phone classification task, compared to 80.52% with the conventional MCE trained HMM. Compared across all three classifiers, THMM produced the lowest error rate and is the new efficient way of utilizing the input data. We first showed that Mel-warped DFT features, subject to appropriate transformation in a state-dependent manner, are more effective than the MFCCs that have dominated current speech recognition technology. Further improvement of the performance can be expected by incorporating both the state-dependent generalized dynamic feature parameters [5] and the state-dependent linear transforms to obtain the combined advantages of individual parameters. The proposed integrated technique for feature design, based on discriminative feature reduction, is sufficiently general and can be applied to all types of pattern classifiers.

REFERENCES

[1] A. Biem E. McDermott and S. Katagiri, "A discriminative filter bank model for speech recognition," *European Conference on Speech Communication and Technology*, 1995, pp. 545-548.

[2] C. R. Jankowski, H. Vo and R. P. Lippman, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech and Audio Processing*, Vol.3, No.4, 1995, pp. 286-293.

[3] W. Chou, B. H. Juang and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," *IEEE Proc. ICASSP*, 1992, pp. 473-476.

[4] S. Katagiri, B. H. Juang and A. Biem, "Discriminative feature extraction", in *Artificial neural networks for speech and vision*, Chapman & Hall, London, 1993, pp. 278-293.

[5] C. Rathinavelu and L. Deng, "Use of generalized dynamic feature parameters for speech recognition: maximum likelihood and minimum classification error approaches", *IEEE Proc. ICASSP*, 1995, pp. 373-376.