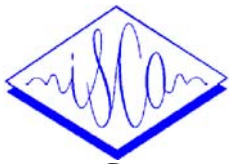

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

OPTIMAL STATE DEPENDENT SPECTRAL REPRESENTATION FOR HMM MODELING : A NEW THEORETICAL FRAMEWORK

C. Mokbel*, G. Gravier** and G. Chollet**
*France Télécom - CNET - DIH/RCP, 2 av. Pierre Marzin, 22307 Lannion, France
** ENST, dept Signal, 46 rue Barrault, 75634 Paris cedex 13, France

*Tel. +33 2 96 05 39 28, FAX: +33 2 96 05 35 30, E-mail : mokbel@lannion.cnet.fr

ABSTRACT

In this paper we propose a theoretical framework to extend classical continuous density HMM in order to consider different spectral representations depending on the state. We stress the need for a reference space and for spectral transformations between the model spectral representation spaces and the reference space. We show that this framework permits to obtain more precise pdfs in the reference space. Preliminary speech recognition experiments for two spectral representations MFCC and linear frequency scale cepstral coefficients show no improvements ; however they identify that the choice of the spectral representations is crucial and the determination of the spaces transformations is a complex problem

1. INTRODUCTION

Acoustic speech signal modeling systems are generally formed of two stages. In the first one, an analysis module extracts from the speech signal a sequence of feature vectors that describes the speech in a time-frequency space. ''Mel Frequency based Cepstral Coefficients'' (MFCC) are a popular feature set. In the second stage, stochastic modeling of the feature sequences is performed, generally using ''Hidden Markov Models'' (HMM) [8].

In order to compute the MFCC coefficients a spectral analysis with a filterbank defined on a MEL scale is first performed, then the logarithm operator is applied on the filterbank energies followed by a cosine transform. MEL frequency scale, a psycho-acoustic scale, is characterized with a higher resolution in the low frequency bands with respect to the high frequency bands. Besides the psycho-acoustic characteristics, increasing the frequency resolution in the low frequency bands is adequate with the nature of the speech signal where a maximum of information is concentrated in the low frequencies. Other spectral representations are used for speech recognition[4] ; SMC[3], PLP[1], LSP[7]. Bilinear transformation is also used to approximate the MEL scale and to increase the spectral resolution in the low frequency bands [6]. Even if MFCC is suitable globally for speech representation, it may not be the most appropriate to represent local information. For example, increasing the spectral resolution in the high or middle frequency bands may be suitable to represent the acoustic information relative to some phonemes. Work on speaker normalization by frequency warping [9] prove that warping the frequency scale differently by speaker does improve the recognition performances.

In this paper we suggest to define a specific spectral representation for each part of the speech signal model in order to increase the modeling precision. This produces a generalization of the classical CDHMM modeling where each state, for which corresponds a sub-process, is characterized by a given spectral representation in addition to the classical parameters of the sub-process density function. The spectral representation for a given state in the HMM is automatically determined during the training process. Thus, we called the approach : ''Optimal State Dependent Spectral Representation for HMM''. In this paper we define the theoretical general framework and discuss the implementation issues for a particular case where different spectral resolution cepstral coefficients are considered as the possible spectral representations.

2. THEORETICAL FRAMEWORK

2.1 Model Definition

Classically an HMM λ, with Q states and Gaussian sub-processes pdfs, is defined using the parameters :

- $\pi = \{\pi_i\}_{i=1,...,Q}$: the probabilities of occupying the state i at the first instant.
- $A = \{a_{ij}\}_{i,j=1,...,Q}$: the transition probabilities from state i to state j.
- $B = \{b_i(\underline{X}_\tau)\}_{i=1,...,Q}$: the Gaussian sub-processes associated with the different states of the model. Each Gaussian distribution has two sets of parameters : its mean $\underline{\mu}_i$ and its covariance matrix $\underline{\Gamma}_i$.

Here, we propose to associate a given spectral representation identified by its index α with each state in the model. We assume that the different feature vectors belongs to $\Re^p$. This assumption is necessary in order to perform ''Maximum Likelihood'' training and classification. In such case the sub-processes pdfs are defined :

- $B = \{b_i(\underline{X}_{\tau,\alpha_i})\}_{i = 1,...,Q}$ and $\alpha_i \in \{1, ..., M\}$: where each distribution is identified by the spectral representation index in addition to the classical mean vector and covariance matrix.

However, as defined, the new HMM cannot be trained nor used for classification since the observed outputs do not belong to a predefined and unique space, even if the different representation spaces have the same dimension p. Actually, the observation spectral representation depends on the state of the HMM which is hidden and not observed. This means that the spectral representation of frame may vary for several word hypotheses and within a single hypothesis during the training, making difficult to tie the measured likelihood to the probability. A solution to this problem consists in defining a reference space with a spectral representation $\hat{\alpha}$ (MFCC for example). Considering that a function $\mathbf{T}_{\alpha/\hat{\alpha}}$ permits to match the space $\boldsymbol{X}_\alpha$ on to the space $\boldsymbol{X}_{\hat{\alpha}}$:

$$\underline{X}_{\tau,\hat{\alpha}} = \mathbf{T}_{\alpha/\hat{\alpha}}(\underline{X}_{\tau,\alpha}) \qquad (1)$$

then the density for a state i can be written :

$$b_i(\underline{X}_{\tau,\hat{\alpha}}) = b_i(\underline{X}_{\tau,\alpha_i}) / \|\boldsymbol{J}(\underline{X}_{\tau,\alpha_i})\| \qquad (2)$$

where $\boldsymbol{J}(\underline{X}_{\tau,\alpha_i})$ is the Jacobian matrix whose $(k,l)^{-th}$ element is :

$$\boldsymbol{J}_{k,l}(\underline{X}_{\tau,\alpha}) = \frac{\partial \mathbf{T}_{\alpha/\hat{\alpha}}(\underline{X}_{\tau,\alpha})_k}{\partial \underline{X}_{\tau,\alpha\,l}} \qquad (3)$$

Using the Eq. (2) and (3) the observations of the new defined HMM belong to a unique space $\boldsymbol{X}_{\hat{\alpha}}$. A very important question appears at this stage :

- ***In what the new model differs from an HMM completely defined in the space $\boldsymbol{X}_{\hat{\alpha}}$?***

Actually, the new model differs from the $\boldsymbol{X}_{\hat{\alpha}}$ model in the form of its sub-processes distributions. It may be more reliable to approximate with a Gaussian the distribution of the data in a space $\boldsymbol{X}_\alpha$ for a given state, than to approximate with a Gaussian the distribution of the same data in the reference space $\boldsymbol{X}_{\hat{\alpha}}$. Thus, the data in $\boldsymbol{X}_{\hat{\alpha}}$ may be distributed following a more complex and precise probability density function depending on the space transformation. Hence, through the Jacobian matrices, the corresponding distribution in the reference space would be more precise. Moreover, more precise distributions lead generally to a more precise and robust model. Nevertheless, the estimation of these Jacobian matrices remains a major problem as we will see in the following.

2.2 Computation of the Space Transformation function

The determination of the transformation $\mathbf{T}_{\alpha/\hat{\alpha}}$ that allows to match a given space on to the reference space is not always obvious. Thus, one can approximate this transformation by a simpler function chosen for some mathematical attractiveness such as a regression matrix, a LMR, a linear quadratic function, a Volterra function, etc. Once defined the parameters of these functions must be estimated. This can be done using classical criteria such as ''Minimum

Mean Square Error'' (MMSE). We can notice that the transformation parameters estimation may be easier here since there is no alignment problem since the different analysis techniques are derived for the same frames of signal.

2.3 Training Algorithm

The parameters of the new model may be trained using the classical EM algorithm. In the ''Estimate'' step the auxiliary function relative to a given state is computed for the M possible spectral representations. During the ''Maximize'' step, in order to maximize the auxiliary function, the mean and covariance matrices would be determined for all the spectral representations, and for a given distribution the optimal spectral representation is chosen following:

$$\alpha_{opt}=\arg\max_{\alpha}\sum_{\tau}[-1/2\log||\underline{\Gamma}_{i,\alpha}|| - \log||\boldsymbol{J}(\underline{X}_{\tau,\alpha})||] \quad (4)$$

Looking to this equation it appears clearly that the Jacobian impact is crucial to compensate for decreasing the variations in the state.

3. PARTICULAR APPLICATION : VARIABLE RESOLUTION CEPSTRAL COEFFICIENTS

Different spectral representations may be obtained by varying the spectral resolution of the filter-bank used to compute the MFCC coefficients. In [6], a computational method using FFT is proposed for spectra with non-uniform resolution on the frequency scale. A procedure for implementing a frequency warping all-pass transformation, known as bilinear transformation, estimates the AR parameters (of order p) using the outputs of p successive filters. These dephasing filters are preceded by a corrective filter in order to obtain an all-pass filter. If first order dephasing filters are considered then, the frequency transformation is expressed by the following relation in the normalized frequency domain:

$$\omega' = 2\pi f' = \theta(2\pi f) = tg^{-1}\left|\frac{(1-\alpha^2)\sin(2\pi f)}{(1-\alpha^2)\cos(2\pi f) - 2\alpha}\right| \quad (5)$$

where $-1 \le \alpha \le 1$ (the filter parameter).

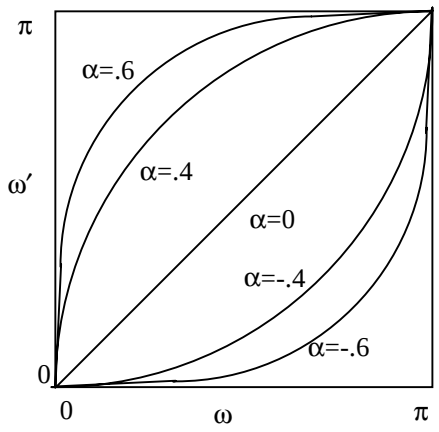

Figure 1 : The frequency transformation obtained with bilinear transformation.

As this warping transformation corresponds to a good approximation of the Mel scale it is of great interest for speech recognition, and has been integrated with SMC analysis, as described in [4]. Using different values of α, several spectral resolutions can be defined. The frequency transformation can be computed for each α which allows to define a corresponding filter-bank. Supposing that the filterbank is applied on the spectral density defined on the logarithmic scale, then the cepstral coefficients can be derived from the filter-bank outputs by applying the cosine transform matrix. Let us call $\underline{S}_w$ the vector representing the spectral density in the logarithmic scale. The Cepstral coefficients relative to a spectral resolution α can be written :

$$\underline{X}_\alpha = \underline{A}_\alpha \cdot \underline{S}_w \quad (6)$$

If we consider now that the space transformation between $\boldsymbol{X}_\alpha$ and $\boldsymbol{X}_{\hat{\alpha}}$ is a simple regression matrix $\underline{R}_\alpha$:

$$\underline{X}_{\tau,\hat{\alpha}} = \mathbf{T}_{\alpha/\hat{\alpha}}(\underline{X}_{\tau,\alpha}) = \underline{R}_\alpha \cdot \underline{X}_{\tau,\alpha} \quad (7)$$

Replacing Eq. 5 into Eq. 6 , leads to :

$$\underline{A}_{\hat{\alpha}} \cdot \underline{S}_{\tau,w} = \underline{R}_\alpha \cdot \underline{A}_\alpha \cdot \underline{S}_{\tau,w} \quad (8)$$

$\underline{R}_\alpha$ must be chosen in order to satisfy Eq. 8 whatever the spectral density $\underline{S}_{\tau,\omega}$. A MMSE solution of Eq. 8 can be derived :

$$\tilde{\underline{R}}_\alpha = \underline{A}_{\hat{\alpha}} \cdot \underline{A}_\alpha^{\dagger} \quad (9)$$

where $\underline{A}_\alpha^\dagger$ is the pseudo-inverse of the matrix $\underline{A}_\alpha$ defined by:

$$\underline{A}_\alpha^\dagger = \underline{A}_\alpha{}^T \cdot \left[\underline{A}_\alpha \cdot \underline{A}_\alpha{}^T\right]^{-1} \qquad (10)$$

4. SPEECH RECOGNITION EXPERIMENTS

This algorithm is implemented within the CNET speech recognition system [2]. Two spectral representations were used : MFCC and linear frequency scale cepstral coefficients LFCC. The MFCC was considered as the base spectral representation. The training and testing programs are modified to integrate this new algorithm. Eq. 10 is first used to compute the Jacobian matrix. No significant results are obtained on a digit PSN/GSM database [5]. This can be explained by the fact that a unique regression matrix do not change the distribution form in the reference space that remains a Gaussian. It is thus more suitable to have a Jacobian that depends on the feature vector than a constant on the whole space. Thus we tried to use LMRs associated to a codebook of size 32 constructed in the MFCC space using LBG algorithm. Here the obtained determinant of the regression matrices are found to be too small (order of 10^{-3}) which makes the training algorithm not select any of the LFCC space. Afterwards, we have tried to consider the transformation matrix $\underline{R}_{q,LFCC}$ for each class q in the codebook to be :

$$\underline{R}_{q,LFCC} = \underline{\Gamma}_{q,MFCC}^{1/2} \cdot \underline{\Gamma}_{q,LFCC}^{-1/2} \qquad (11)$$

Using this last approach, the training algorithm considers LFCC space in a large part of the model. The obtained HMM produces good results on the training set. However, no improvements was obtained on the test part of the database.

Based on these results it appears that the choice of the spectral representation spaces is very crucial and the determination of the space transformations is a very complex problem. Large work must be done in order to define the transformations between different spaces. We do believe that such a work help to build robust speech models.

5. CONCLUSIONS

We have presented a new theoretical framework where the HMM outputs are derived with different spectral representations depending on the state. The association between the spectral representations and the HMM states is data-driven. In order to develop such models we have stressed the need of transformations matching any of the available spectral representations on to a reference acoustical space. We have described a particular case of application of this approach where the spectral representations are cepstral coefficients with filter-banks defined using different spectral resolution. This new modeling approach of the acoustical signal may be used in both speech and speaker recognition systems. On the basis of few speech recognition experiments we have shown that large work remains to be done in order to define suitable speech representation spaces and reliable transformations between these spaces.

REFERENCES

[1] Hermansky H., Morgan N. and Hirsch H.G., ''Compensation for the Effect of Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP),'' *Proc. EuroSpeech*, pp. 1367-1370, 1991.

[2] Jouvet D., Bartkova K. and Monné J., "On the Modelization of Allophones in a HMM Based Speech Recognition System," *Proc. EuroSpeech*, pp. 923-926, 1991.

[3] Mansour D. and Juang B.H., ''The Short-time Modified Coherence Representation an Noisy Speech Recognition,'' IEEE Trans. on ASSP, Vol. 37, n° 6, pp. 795-804, June 1989.

[4] Mokbel C. and Chollet G., ''Automatic Word Recognition in Cars,'' IEEE Trans. on SAP, Vol. 3, n° 5, pp. 346-356, September 1995.

[5] Mokbel C. and Jouvet D., ''Recognition of Digits over PSN & GSM Networks,'' *Proc. IEEE Workshop on ASR*, pp. 167-168, December 1995.

[6] Oppenheim A.V., Johnson D.H. and Steiglitz S., ''Computation of Spectra with Unequal Resolution using Fast Fourier Transform,'' Proc. of the IEEE, Vol. 59, pp.299-301, February 1971.

[7] Paliwal K.K., ''On the Use of Line Spectral Frequency Parameters for Speech Recognition,'' Digital Signal Processing, Vol. 2, pp. 80-87, 1992.

[8] Rabiner L. and Juang B.-H., ''Fundamentals of Speech Recognition,'' Prentice Hall, Englewood Cliffs, 1993.

[9] Rose R.C. and Potamianos A., ''Improving Robustness in HMM Based Speech Recognition Through Simultaneous Frequency Warping and Spectral Shaping,'' Proc. of ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, April 1997.