Joint Optimization of the Frequency-domain and Time-domain Transformations in Deriving Generalized Static and Dynamic MFCCs

Yiu-Pong Lai, *Student Member, IEEE,* Manhung Siu, *Senior Member, IEEE,* and Brian Mak, *Member, IEEE*

***Abstract*— Traditionally, static mel-frequency cepstral coefficients (MFCCs) are derived by discrete cosine transformation (DCT), and dynamic MFCCs are derived by linear regression. Their derivation may be generalized as a frequency-domain transformation of the log filter-bank energies (FBEs) followed by a time-domain transformation. In the past, these two transformations are usually estimated or optimized separately. In this paper, we consider sequences of log FBEs as a set of spectrogram images, and investigate an image compression technique to jointly optimize the two transformations so that the reconstruction error of the spectrogram images is minimized; there is an efficient algorithm that solves the optimization problem. The framework allows extension to other optimization costs as well.**

***Index Terms*— low-rank approximation of matrices, time-frequency representation, mel-frequency cepstral coefficients, discrete cosine transform**

I. Introduction

Mel-frequency cepstral coefficients (MFCCs) are the most commonly used acoustic features in automatic speech recognition. Traditionally, static MFCCs are derived by discrete cosine transformation (DCT), and dynamic MFCCs are derived by linear regression. The use of DCT is motivated by the need to de-correlate the cepstral elements so that Gaussians with diagonal covariances may be used in hidden Markov model states, while the use of linear regression gives a robust estimate of the temporal derivative. The MFCC extraction procedure may be generalized as two transformations or filtering on the time-frequency (TF) representation of speech: a frequency-domain transformation or filtering of the log filter-bank energies (FBEs) followed by a time-domain transformation or filtering. Various well-known statistical techniques like PCA, LDA, ICA, and NLDA have been investigated and compared for frequency transformation [1], [2]. For time-domain transformation, similar transformation-based techniques are used in cepstral-time matrices [3]; on the other hand, [4], [5] perform filtering across successive cepstral vectors. The time-domain transformation introduces a smoothing across time and can improve the robustness property of the features but may degrade performance when the data is clean.

In the past, the two TF transformations or filtering processes are usually estimated or optimized separately in a sequential manner. In this paper, we propose a joint optimization of the two TF transformations (JOTFT) so that the reconstruction error of the log FBEs is minimized. That is, if a speech signal is represented by its 2-dimensional spectrogram (using log FBEs), we would like to find an optimal acoustic representation of reduced dimension such that the spectrogram can be reconstructed with minimum squared error.

Our work is motivated by a recent work in low-rank approximation of matrices [6]. There is an efficient algorithm that solves the optimization problem, and it runs directly on the TF representation without building any acoustic models.

II. Generalization of Time-Frequency Transformations

The feature extraction of static and dynamic mel-frequency cepstral coefficients (MFCCs) can be generalized as two time-frequency (TF) transformation on the log filter-bank energies (FBEs). Let $\mathbf{S}\in \mathbb{R}^{r\times c}$ be a sequence of log FBEs, $\mathbf{s}_t\in \mathbb{R}^r$ around time t, and $\mathbf{S} = [\mathbf{s}_{t-d}, \ldots, \mathbf{s}_{t+d}]$ where $c = 2d + 1$. The acoustic feature vector that may be called the "generalized static and dynamic MFCCs" for automatic speech recognition, $\mathbf{X}\in \mathbb{R}^{l_1\times l_2}$ can be computed by one matrix multiplication to its left and another matrix multiplication to its right as follows:

$$\mathbf{X} = \mathbf{L}'\mathbf{S}\mathbf{R}, \tag{1}$$

where $\mathbf{L}\in \mathbb{R}^{r\times l_1}$ represents a frequency-domain transformation that computes the static cepstra, and $\mathbf{R}\in \mathbb{R}^{c\times l_2}$ represents a time-domain transformation that computes the dynamic cepstra[1].

For automatic speech recognition, it is generally desirable that the frequency transformation $\mathbf{L}$ can separate the phonetic information from the speaker information, or the more discriminative information from the less discriminative information. Thus, the frequency-domain transformation may also be considered as an information selection, and some coefficients are usually removed. Meanwhile, the time transformation $\mathbf{R}$ usually has to derive useful dynamic features over a long period of time. As a result, there is usually a dimensionality reduction and the dimension of $\mathbf{X}$ is smaller than that of $\mathbf{S}$. i.e. $l_1 < r$ and $l_2 < c$.

Dr. Manhung Siu and Yiu-Pong Lai are with the Department of Electrical and Electronic Engineering, the Hong Kong University of Science and Technology (HKUST), Clear Water Bay, Hong Kong. Email: eemsiu@ee.ust.hk, harry@ust.hk.

Dr. Brian Mak is with the Department of Computer Science, HKUST, Hong Kong. Tel: +852 2358-7012; Fax: +852 2358-1477; E-mail: mak@cs.ust.hk.

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST6201/02E, CA02/03.EG04,and CA02/03.EG05.

[1] In this paper, vector or matrix quantities are bold-faced, and $'$ represents their transpose.

Let us look at two special cases of the generalized TF transformations.

A. Mel-frequency Cepstral Coefficients (MFCCs)

The most common acoustic features for speech recognition are the static MFCCs with their first and the second order time derivatives. In the computation of standard MFCCs,

- the frequency-domain transformation matrix $\mathbf{L}$ is the DCT matrix
$$L'(i,j) = \sqrt{\frac{2}{N}} \cos\left(\frac{\pi i}{N}(j-0.5)\right) \tag{2}$$
where N is the size of the filter-bank.
- the time-domain transformation matrix $\mathbf{R}$ represents linear regression over 9 frames of speech:
$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0.04 \\ 0 & 0 & 0.04 \\ 0 & -0.2 & 0.01 \\ 0 & -0.1 & -0.04 \\ 1 & 0 & -0.10 \\ 0 & 0.1 & -0.04 \\ 0 & 0.2 & 0.01 \\ 0 & 0 & 0.04 \\ 0 & 0 & 0.04 \end{bmatrix} \tag{3}$$

B. Two-dimensional DCT

Two-dimensional DCT (2D-DCT) has also been proposed for the computation of static and dynamic MFCCs and is equivalent to a 1D-DCT on the frequency domain followed by another 1D-DCT on the time domain. The main difference from the standard MFCC extraction is that the linear regression over consecutive time frames is replaced by another DCT. Thus,

- the frequency-domain transformation matrix $\mathbf{L}$ is the same DCT matrix as given by Eqn. (2).
- the time-domain transformation matrix $\mathbf{R}$ is also similar to the L given by Eqn. (2) except that N is now replaced by M which is the number of consecutive speech frames. That is,
$$R'(i,j) = \sqrt{\frac{2}{M}} \cos\left(\frac{\pi i}{M}(j-0.5)\right) \ . \tag{4}$$

C. Log Frame Energy

As the log frame energy is usually used together with MFCCs, we augment the $\mathbf{S}$ matrix with a sequence of log frame energies $\mathbf{E}$ as follows:

$$\hat{\mathbf{S}} = \begin{bmatrix} \mathbf{S} \\ \mathbf{E} \end{bmatrix} , \tag{5}$$

and the frequency-domain transformation matrix becomes

$$\hat{\mathbf{L}} = \begin{bmatrix} \mathbf{L} & 0 \\ 0 & 1 \end{bmatrix} . \tag{6}$$

As a results, the last row of the acoustic feature matrix $\mathbf{X}$ represents the static and dynamic log frame energies.

III. Joint Optimization of the Two Time-Frequency Transformations

A sequence of log FBEs, $\mathbf{S} = [\mathbf{s}_{t-d}, \ldots, \mathbf{s}_{t+d}]$, as defined in Section II may be considered as a 2D spectrogram image. It is well-known that an image may be compressed (and smoothed or de-noised as a side effect) by treating the image as a matrix and applying low-rank approximation. In this paper, we investigate the use of a recent work in low-rank approximation of matrices proposed by Ye [6] to extract static and dynamic acoustic features through smoothing speech spectrograms. The method allows us to cast the feature extraction problem as an optimization problem with a well-defined cost function which is the error in reconstructing the original spectrograms from the compressed spectrograms.

Formally, if we have a set of n original spectrograms represented by the matrices $\{\mathbf{S}_i \in \mathbb{R}^{r \times c}, i = 1, \ldots, n\}$, we would like to compress them into n matrices of reduced dimensions, $\{\mathbf{X}_i \in \mathbb{R}^{l_1 \times l_2}, i = 1, \ldots, n\}$. Each original spectrogram $\mathbf{S}_i$ can be reconstructed with some error from $\mathbf{X}_i$ with the help of two transformation matrices, $\mathbf{L} \in \mathbb{R}^{r \times l_1}$ and $\mathbf{R} \in \mathbb{R}^{c \times l_2}$, both having orthonormal column vectors as $\mathbf{L}\mathbf{X}_i\mathbf{R}'$. Our goal is to find the optimal $\mathbf{L}$ and $\mathbf{R}$ so that the squared reconstruction error (SRE) over the set of n spectrogram images (i.e. log FBEs sequences) is minimized. The squared reconstruction error is defined as

$$\text{SRE} = \sum_{i=1}^{n} ||\mathbf{S}_i - \mathbf{L}'\mathbf{X}_i\mathbf{R}||_F^2 \tag{7}$$

where $||\mathbf{A}||_F^2$ is the Frobenius norm of matrix $\mathbf{A}$.

It was proved in [6] that

- minimizing the SRE of Eqn. (7) is equivalent to maximizing
$$\sum_{i=1}^{n} ||\mathbf{L}\mathbf{S}_i\mathbf{R}'||_F^2 \ , \tag{8}$$
subject to the required dimensions of $\mathbf{L}, \mathbf{R}$, and $\mathbf{S}_i, i = 1, \ldots, n$.
- for the optimal solution, we will have
$$\forall i, \ \ \mathbf{X}_i = \mathbf{L}'\mathbf{S}_i\mathbf{R} \ . \tag{9}$$

Comparing the solution of minimizing SRE in Eqn. (9) with the generalized MFCCs in Eqn. (1), we notice that the compressed spectrograms may be interpreted as the generalized static and dynamic MFCCs, and the $\mathbf{L}$ and $\mathbf{R}$ matrices as the frequency- and time-domain transformation respectively.

The optimization problem has no closed form solution, but an iterative algorithm was proposed by Ye [6] which is given in Fig. 1. The solution is based on the following key observation:

- For a given $\mathbf{R}$, $\mathbf{L}$ consists of the l_1 eigenvectors of the matrix $\mathbf{A}_L = \sum_{i=1}^{n} \mathbf{S}_i\mathbf{R}\mathbf{R}'\mathbf{S}_i'$ corresponding to the largest l_1 eigenvalues.
- For a given $\mathbf{L}$, $\mathbf{R}$ consists of the l_2 eigenvectors of the matrix $\mathbf{A}_R = \sum_{i=1}^{n} \mathbf{S}_i'\mathbf{L}\mathbf{L}'\mathbf{S}_i$ corresponding to the largest l_2 eigenvalues.

Computationally, the learning of JOTFT matrices is approximately linearly dependent on the data size and proportional to the square of r and c. Once the R and L matrix are known, there is no difference in computation between the three extraction methods.

```
Algorithm

Input: matrices $\mathbf{S}_i$, $1 \le i \le n$.
Output: matrices $\mathbf{L}$, $\mathbf{R}$, and $\mathbf{X}_i$, $1 \le i \le n$.
  1) Obtain an initial $\mathbf{L}_0$ and set $i \leftarrow 1$
  2) While convergence is not reached
       a) form the matrix $\mathbf{A}_R = \sum_{j=1}^{n} \mathbf{S}'_j \mathbf{L}_{i-1} \mathbf{L}'_{i-1} \mathbf{S}_j$
       b) compute the $l_2$ eigenvectors $\phi_j^R, 1 \le j \le l_2$
          of $\mathbf{A}_R$ corresponding to the largest $l_2$ eigenvalues
       c) $\mathbf{R}_i \leftarrow [\phi_1^R, \ldots, \phi_{l_2}^R]$
       d) form the matrix $A_L = \sum_{j=1}^{n} \mathbf{S}_j \mathbf{R}_{i-1} \mathbf{R}'_{i-1} \mathbf{S}'_j$
       e) compute the $l_1$ eigenvectors $\phi_j^L, 1 \le j \le l_1$
          of $\mathbf{A}_L$ corresponding to the largest $l_1$ eigenvalues
       f) $\mathbf{L}_i \leftarrow [\phi_1^L, \ldots, \phi_{l_1}^L]$
       g) $i \leftarrow i + 1$
  3) EndWhile
  4) $\mathbf{L} \leftarrow \mathbf{L}_{i-1}$
  5) $\mathbf{R} \leftarrow \mathbf{R}_{i-1}$
  6) $\forall j = 1$ to $n$, $\mathbf{X}_j = \mathbf{L}' \mathbf{S}_j \mathbf{R}$
```

Fig. 1. The optimization algorithm of finding the two TF transformation matrices in JOTFT.

IV. Experimental Evaluation

Generalized MFCCs computed by the following 3 different time-frequency (TF) transformation methods:

- **Standard**: TF transformations for the standard MFCCs.
- **2D-DCT**
- **JOTFT**

were compared in two speech tasks: TIMIT phoneme recognition, and Resource Management (RM) speech recognition.

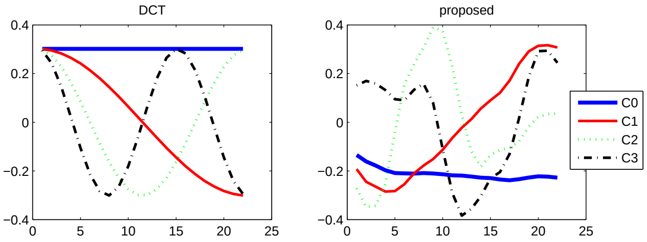

Fig. 2. The left frequency-domain transformation matrix.

A. Task I: TIMIT Phoneme Recognition

1) Feature Extraction: Through FFT-based filter-bank analysis, log FBEs were extracted over each 25ms-window of speech at a frame rate of 100Hz. There were 23 triangular filters uniformly lying along the mel-scaled frequency spectrum of 0–8kHz with 50% overlap. Following the standard procedure for generating the standard 39-dimensional MFCC acoustic vectors (which contains 12 static MFCCs plus the log frame energy, and their first- and second-order temporal derivatives), all the three TF transformations were performed on successive blocks of 9 frames of log FBEs. That is, the dimension of the signal blocks $\mathbf{S}_i$ is 23×9, and the dimension of the extracted feature matrices $\mathbf{X}_i$ is 13×3. The feature matrices are vectorized to obtain the final 39-dimensional MFCC acoustic vectors. After feature extraction, cepstral mean normalization (CMN) was carried out to remove the channel effect.

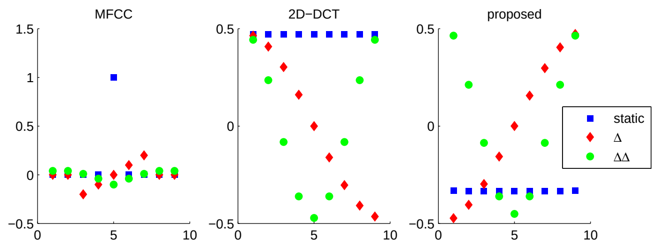

Fig. 3. The right time-domain transformation matrix.

2) JOTFT: The jointly optimized TF transformations were estimated by the JOTFT algorithm described in Fig. 1 using 100 utterances randomly selected from the TIMIT training set. We find that using more utterances does not make much difference to the transforms.

The JOTFT algorithm requires an initial frequency-domain transformation $\mathbf{L}_0$ to start. Here, we simply employ the DCT which is used in the generation of the standard MFCC. The first four basis vectors of the frequency-domain transformation $\mathbf{L}$ found by our JOTFT method are plotted in Fig. 2 together with those of DCT. Except for a change in sign, they are quite similar, but their leading basis vectors are more similar than their trailing basis vectors. For example, it is found that the first basis vector of JOTFT's $\mathbf{L}$ matrix is fairly constant for high frequency bands; its role is believed to be similar to that of the standard $c0$ or log frame energy. And the second basis vector of JOTFT's $\mathbf{L}$ matrix is almost as linear as that of DCT.

Fig. 3 shows the time-domain transformation $\mathbf{R}$ found by JOTFT as well as that of 2D-DCT, and the linear regression matrix for the generation of standard dynamic MFCCs. Again except for some sign changes, the $\mathbf{R}$ matrix found by JOTFT is very similar to that of 2D-DCT. This may not be surprising as in the current algorithm, both $\mathbf{L}$ and $\mathbf{R}$ matrices are composed of orthonormal eigenvectors as in 2D-DCT and we start with $\mathbf{L} = \text{DCT}$.

3) Phoneme recognition: The standard TIMIT training and testing data without the "sa" utterances were used. The training set consists of 3696 utterances from 462 speakers, and the test set consists of 1344 utterances from 168 speakers. Each of the 61 phones in the standard TIMIT phone sets was modeled as a strictly left-to-right hidden Markov model (HMM) with 3 emitting states, and each HMM state was a mixture of Gaussians. Context-independent HMMs were employed. Finally decoding was carried out with a bigram language model that was trained only from the training set.

The recognition results are folded into the standard 39

TABLE I

TIMIT PHONEME ACCURACY.

TF Transformation	1-mixture	4-mixture	8-mixture
Standard	53.69%	61.38%	64.14%
2D-DCT	53.91%	61.60%	64.55%
JOTFT	53.82%	62.48%	65.28%

TABLE II

RM WORD ERROR RATE.

TF Transformation	2-mixture	7-mixture
Standard	5.22%	3.86%
JOTFT (RM on RM)	5.06%	3.98%
JOTFT (TIMIT on RM)	5.50%	4.84%

phoneme classes to compute the phoneme recognition accuracies. The phoneme recognition accuracies for HMMs with various numbers of Gaussian mixtures are summarized in Table I. It is obvious that while MFCCs generated by 2D-DCT is only marginally better the standard MFCCs, MFCCs generated by our newly proposed JOTFT gives additional improvement over 2D-DCT.

B. Task II: RM Speech Recognition

The generalized MFCCs computed by JOTFT algorithm was also compared with the standard MFCCs on Resource Management (RM1) speech recognition. All the 3990 SI training utterances from 109 speakers in RM1 were used for training acoustic models, and evaluation was performed on 4 common RM test sets: feb89, oct89, feb91, and sep92 test set, which consist of totally 1200 testing utterances from 40 speakers. The experimental procedure is similar to that of the TIMIT experiment except for the following:

- the spectrogram frame size for JOTFT was reduced from 9 to 5 as we found that 5 frames gave a slightly better performance than 7 or 9 frames.
- 47 phoneme HMMs were estimated.
- the standard RM word-pair grammar was used for decoding the test utterances.

The results are shown in Table II. In this task, the relative performance of the standard MFCCs and the generalized MFCCs computed by JOTFT varies with the complexity of the acoustic models and the difference is small. Furthermore, the use of cross-domain estimation for JOTFT degrades performance significantly.

C. Reconstruction Error

For JOTFT, the extracted features are selected by optimizing the reconstruction error as measured by Eqn. (7) which can also be used as a criterion for feature comparison. In Fig. 4, the rate-distortion curves (ratio of the dimensions before and after extraction vs. SNR as measured by the distortion over the signal energy) of RM are plotted by changing both l_1 and l_2. The top two curves are from JOTFT: one learned from RM and one from TIMIT and the mis-match gives a slightly worse SNR. The third and fourth curves, labeled "2D-DCT" and "MFCC" are results of reconstruction from the 2D-DCT and MFCC features [2]. Because both approaches are not optimized for reconstruction errors, their distortions are higher.

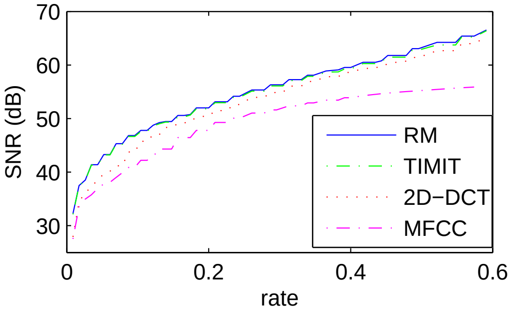

Fig. 4. The rate-distortion curves of various feature extraction methods using RM utterances.

V. CONCLUSIONS

In this paper, we propose a data-driven approach to jointly optimize the two time-frequency (TF) transformations used in the generation of MFCCs. It is observed that both the frequency- and time-domain transformations found by our JOTFT method are similar to those of 2D-DCT. This may not be too surprising given the fact that we use DCT as the initial frequency-domain transformation to run our iterative JOTFT algorithm, and both JOTFT and DCT employ orthonormal eigenvectors. In the TIMIT phoneme recognition task, the MFCCs generated by our TF transformations outperform those generated by 2D-DCT and the standard MFCC generation method, but the result is not conclusive in the RM speech recognition task.

In the current JOTFT method, the generation of generalized MFCCs is cast as an optimization problem that tries to minimize the reconstruction error of spectrograms. We believe that the framework may be generalized with other optimization criteria, e.g. minimum classification errors, as well. Moreover, besides generation of MFCC features for speech recognition, the proposed joint optimization method may also be used for speech coding to represents the spectrum with reduced number of parameters.

REFERENCES

[1] P. Somervuo, "Experiments with linear and nonlinear feature transformations in HMM based phone recognition," in *Proc. of ICASSP*, 2003, vol. 1, pp. 52–55.

[2] I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Spectral and cepstral projection bases constructed by independent component analysis," in *Proc. of ICSLP*, 2000, vol. 3, pp. 63–66.

[3] B. Milner, "Cepstral-time matrices and LDA for improved connected digit and sub-word recognition accuracy," in *Proc. of Eurospeech*, 1997, vol. 1, pp. 405–408.

[4] D. Macho et al., "Comparison of time and frequency filtering and cepstral-time matrix approaches in ASR," in *Proc. of Eurospeech*, 1999.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on SAP*, vol. 2, pp. 578–589, 1984.

[6] J. Ye, "Generalized low rank approximations of matrices," in *ICML*, 2004.

[2] In the reconstruction using MFCC, the first column of R is replaced by an all one vector to make R orthogonal.