A Mathematical Relationship between Full-band and Multi-band Mel-frequency Cepstral Coefficients

Brian Mak, *Associate Member, IEEE*

Abstract

Recently, it has been shown that robustness of automatic speech recognition (ASR) against band-limited additive noises may be improved by multi-band ASR (MBASR) approaches. In an M-subband MBASR system, the channels in the full-band filter-bank are divided into M subbands, usually of equal partitions, and subband *mel-frequency cepstral coefficients* (MFCCs) are computed from each filter-bank partition using the Discrete Cosine Transform. However, there is not as yet any analysis on the relationship between full-band and multi-band MFCCs. In this paper, we show that the (Mj)-th full-band MFCC is the sum of or difference between the M j-th multi-band MFCCs multiplied by $\sqrt{M}$.

Keywords

multi-band automatic speech recognition, mel-frequency cepstral coefficients, discrete cosine transform

Brian Mak is with the Department of Computer Science, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. Tel: +852 2358-7012; Fax: +852 2358-1477; E-mail: mak@cs.ust.hk.

I. Introduction

One of the major challenges to automatic speech recognition (ASR) is to maintain good performance or to degrade gracefully under the adverse effect of noise. Many attempts have been tried: for instance, signal-based approaches such as cepstral mean subtraction [1], spectral subtraction [2], signal bias removal [3], and RASTA processing [4]; model-based approaches such as parallel model combination [5], and noise adaptation. In recent years, multi-band ASR (MBASR) [6], [7], [8], [9], [10] has been proposed as another model-based method to improve robustness of ASR against band-limited additive noises. In an M-subband MBASR system, the channels in the full-band filter-bank are divided into M subbands, usually of equal partitions, and subband *mel-frequency cepstral coefficients* (MFCCs) are computed from each filter-bank partition using the Discrete Cosine Transform (DCT). While the exact method of modeling varies from system to system, the subband MFCCs are usually employed in one of the following two common ways:

(1) in *parallel* mode [6], [11]: M independent recognizers are developed, one for each subband, and they are recombined at some temporal unit during recognition.

(2) in *concatenative* mode [12], the subband MFCCs are concatenated into one feature vector to develop one single "quasi-full-band" recognizer.

The feature extraction process for the two MBASR approaches is depicted in Figure 1 for a 2-subband system with $2N$ channels and $L/2$ MFCCs per subband.

Both approaches demonstrate some successes. Table I shows the performance[1] of the two MBASR approaches and compares them with their corresponding full-band system on the connected TIDIGITS task [14]. The full-band system employs continuous-density hidden Markov models (HMM) with feature vectors consisting of 12 MFCCs plus normalized energy and their first- and second-order derivatives (a total of 39 parameters). The MFCCs are computed from the outputs of a filter-bank of 26 channels covering the bandwidth of 0–4000Hz. The multi-band system have two subbands covering the bandwidths, 0 – 1080Hz and 1000 – 4000Hz. Six MFCCs are computed from the outputs of the first 13 channels and the last 13

[1]Notice that our recognition accuracies on clean connected TIDIGITS are lower than other reported results by about 1%. However, during our signal analysis, speech is band-passed to a bandwidth of 0–4000Hz while most reported results employ the full 10kHz bandwidth of TIDIGITS. Our system performance should better be compared with those based on Aurora [13] since the setup is very similar except for the down-sampling and filtering processes. Aurora's benchmark word accuracy on clean speech is 99.02% [13].

channels of the full-band filter-bank respectively for the two subband systems. The final 21-dimensional subband feature vectors also contain the subband normalized energy as well as the first- and second-order derivatives of their 6 MFCCs and energy. The parallel MBASR system recombines subband log-likelihoods at state level with equal subband weightings. That is,

$$b_j(x_t) = b_j^{(1)}(x_t^{(1)})\, b_j^{(2)}(x_t^{(2)})$$

where, $b_j(x_t)$ is the probability of observation x_t at state j of an HMM, and $b_j^{(k)}(x_t^{(k)})$, $k = 1, 2$, is the corresponding observation probability of each of the two subbands. All corresponding subband and full-band HMMs are whole digit models with the same topology (6 states and 4 Gaussian mixtures per state).

From Table I, we see that all the three systems have comparable word error rates in clean and noisy speech. However, we should note that by exploiting asynchrony and optimal weighting, MBASR may outperform full-band ASR as shown in [11], [15].

In any case, we are not aware of any detailed analysis of the relationship between full-band MFCCs and multi-band MFCCs. Intuitively, since they are all derived from the same spectrum, they should exhibit some relationship. In the next section, we will show mathematically that there is a linear relationship between the j-th subband MFCCs and the (Mj)-th full-band MFCCs.

II. Relationship Between Full-band and Multi-band MFCCs

MFCCs of a frame of speech is commonly computed by performing the DCT on the logarithm of the channel energy amplitudes of a filter-bank. That is, if there are N channels in the filter-bank, and x_i is the logarithm of the i-th channel energy amplitude, then the j-th MFCC c_j is given by

$$c_j = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} x_i \cos\left[(i - 0.5)\frac{j\pi}{N}\right] \quad , \quad 1 \le j \le L \le N \ . \tag{1}$$

Let us denote each cosine transformation term as

$$W_{N,j,i} = \sqrt{\frac{2}{N}} \cos\left[(i - 0.5)\frac{j\pi}{N}\right] \quad , \quad 1 \le i \le N \ , \ 1 \le j \le L \tag{2}$$

and the whole cosine transformation vector and log channel energy vector as

$$\boldsymbol{W_{N,j}} = [W_{N,j,i} \quad , \quad i = 1, 2, \ldots, N]^T \tag{3}$$

and

$$\boldsymbol{x_N} = [x_i \quad , \quad i = 1, 2, \ldots, N]^T \tag{4}$$

respectively. (As usual, vector or matrix quantities are boldfaced.) Then, the j-th MFCC can be rewritten as

$$c_j \ = \sum_{i=1}^{N} x_i W_{N,j,i} = \boldsymbol{W_{N,j}^T x_N} \ . \tag{5}$$

A. Multi-band MFCCs

Let us consider a MBASR system with M subbands, each with a filter-bank of N channels. According to Eqn.(5), the j-th MFCC of the k-th subband of a frame of speech is

$$c_j^{(k)} \quad = \quad \boldsymbol{W_{N,j}^T x_N^{(k)}} \ . \tag{6}$$

B. Full-band MFCCs

The full-bank counterpart of the M-subband MBASR system in the previous section has a filter-bank of $M \times N$ channels. Therefore, from Eqn.(5), the j-th MFCC of a frame of full-band speech is

$$c_j \quad = \quad \boldsymbol{W_{MN,j}^T x_{MN}} \ . \tag{7}$$

Now, let us break up the full-band log energy vector $\boldsymbol{x_{MN}}$ into the M subband log energy vectors

$$\boldsymbol{x_{MN}^T} \ = [\boldsymbol{x_N^{(1)^T}} : \boldsymbol{x_N^{(2)^T}} : \ldots : \boldsymbol{x_N^{(M)^T}}] \tag{8}$$

where the symbol ":" is the concatenation operator. On the other hand, we may express the N log energies of the k-th subband as an (MN)-dimensional full-band log energy vector $\boldsymbol{\tilde{x}_{MN}^{(k)}}$ which contains $\boldsymbol{x_N^{(k)}}$ and is padded with zeros in all the remaining positions corresponding to subbands other than the k-th subband. That is,

$$\boldsymbol{x_{MN}} \ = \boldsymbol{\tilde{x}_{MN}^{(1)}} + \boldsymbol{\tilde{x}_{MN}^{(2)}} + \ldots + \boldsymbol{\tilde{x}_{MN}^{(M)}} \ . \tag{9}$$

Substituting Eqn.(9) into Eqn.(7),

$$c_j \ = \sum_{k=1}^{M} \boldsymbol{W}_{\boldsymbol{MN,j}}^{\boldsymbol{T}} \tilde{\boldsymbol{x}}_{\boldsymbol{MN}}^{(\boldsymbol{k})} \ . \tag{10}$$

Eqn.(10) shows that a full-band MFCC may be considered as the sum of M subband MFCCs, $\tilde{c}_j^{(k)}$, $k = 1, 2, \ldots, M$, each of which is computed from a set of subband filters *under the full-band settings*. Hence, the contribution of the k-th subband toward a full-band MFCC is

$$\tilde{c}_j^{(k)} \ = \boldsymbol{W}_{\boldsymbol{MN,j}}^{\boldsymbol{T}} \tilde{\boldsymbol{x}}_{\boldsymbol{MN}}^{(\boldsymbol{k})} = \boldsymbol{W}_{\boldsymbol{MN,j}}^{(\boldsymbol{k})^{\boldsymbol{T}}} \boldsymbol{x}_{\boldsymbol{N}}^{(\boldsymbol{k})} \tag{11}$$

and

$$c_j \ = \sum_{k=1}^{M} \tilde{c}_j^{(k)} \tag{12}$$

where, $\boldsymbol{W}_{\boldsymbol{MN,j}}^{(\boldsymbol{k})^{\boldsymbol{T}}}$ is the portion of $\boldsymbol{W}_{\boldsymbol{MN,j}}^{\boldsymbol{T}}$ that multiplies with outputs of the k-th subband channel filters. That is,

$$\boldsymbol{W}_{\boldsymbol{MN,j}}^{(\boldsymbol{k})} = [W_{MN,j,i}^{(k)} \quad , \quad i = 1, 2, \ldots, N]^T$$

and

$$\begin{aligned} W_{MN,j,i}^{(k)} \ &= \ \sqrt{\frac{2}{MN}} \cos\left[((k-1)N + i - 0.5)\frac{j\pi}{MN}\right] \\ &= \ \sqrt{\frac{2}{MN}} \cos\left[(k-1)\frac{j\pi}{M} + (i-0.5)\frac{j\pi}{MN}\right] \ . \end{aligned} \tag{13}$$

C. Comparing full-band and subband MFCCs

Comparing Eqn.(11) and Eqn.(6), the contribution of the k-th subband to a full-band MFCC, $\tilde{c}_j^{(k)}$, is very similar to the corresponding MFCC, $c_j^{(k)}$, of the k-th subband in a multi-band system, and the difference is only in their DCT weightings (c.f. Eqn.(13) against Eqn.(2)). Specifically, the values of $W_{MN,j,i}^{(k)}$ for $k = 1, 2, 3, \ldots$ are

$$\begin{aligned} k = 1, \quad W_{MN,j,i}^{(1)} \ &= \ \sqrt{\frac{2}{MN}} \cos\left[(i-0.5)\frac{j\pi}{MN}\right] \\ k = 2, \quad W_{MN,j,i}^{(2)} \ &= \ \sqrt{\frac{2}{MN}} \cos\left[\frac{j\pi}{M} + (i-0.5)\frac{j\pi}{MN}\right] \end{aligned}$$

$$k=3, \quad W^{(3)}_{MN,j,i} = \sqrt{\frac{2}{MN}}\cos\left[\frac{2j\pi}{M}+(i-0.5)\frac{j\pi}{MN}\right]$$
$$\vdots$$

With the above values of $W^{(k)}_{MN,j,i}$ in mind, and comparing Eqn.(11) with Eqn.(6), we get

$$\begin{aligned} c_j^{(1)} &= \sqrt{M}\,\tilde{c}^{(1)}_{Mj} \\ c_j^{(2)} &= (-1)^j\sqrt{M}\,\tilde{c}^{(2)}_{Mj} \\ &\vdots \\ c_j^{(k)} &= \begin{cases} \sqrt{M}\,\tilde{c}^{(k)}_{Mj} & \text{if } k=\text{odd} \\ (-1)^j\sqrt{M}\,\tilde{c}^{(k)}_{Mj} & \text{if } k=\text{even} \end{cases} \end{aligned} \tag{14}$$

Thus, it means that except for a factor of $\sqrt{M}$ the j-th MFCC of the k-th subband in a multi-band system is equivalent to the contribution of the k-subband filters to the (Mj)-th full-band MFCC when k is odd; and, when k is even, the former is similar to the latter except that its sign toggles depending on the value of j. For example, when the MFCCs of a 2-subband multi-band system are compared with those of the full-band system:

Subband #1: $c_1^{(1)}=\sqrt{2}\,\tilde{c}_2^{(1)}$, $c_2^{(1)}=\sqrt{2}\,\tilde{c}_4^{(1)}$, $c_3^{(1)}=\sqrt{2}\,\tilde{c}_6^{(1)}$, $c_4^{(1)}=\sqrt{2}\,\tilde{c}_8^{(1)}$, ...
Subband #2: $c_1^{(2)}=-\sqrt{2}\,\tilde{c}_2^{(2)}$, $c_2^{(2)}=\sqrt{2}\,\tilde{c}_4^{(2)}$, $c_3^{(2)}=-\sqrt{2}\,\tilde{c}_6^{(2)}$, $c_4^{(2)}=\sqrt{2}\,\tilde{c}_8^{(2)}$, ...
Full-band: $c_2=\tilde{c}_2^{(1)}+\tilde{c}_2^{(2)}=\frac{1}{\sqrt{2}}(c_1^{(1)}-c_1^{(2)})$, $c_4=\tilde{c}_4^{(1)}+\tilde{c}_4^{(2)}=\frac{1}{\sqrt{2}}(c_2^{(1)}+c_2^{(2)})$, ...

III. Discussion

The following points are noteworthy.

Point #1 : Without any truncation of the cepstrum, full-band and multi-band MFCCs are equivalent and are different representations of the same spectral information. However, in practice, to capture only the slow-varying vocal tract information, only the low-order MFCCs are used in speaker-independent speech recognition. It is still not clear if truncating NM MFCCs to L MFCCs in the full-band is equivalent to truncating N MFCCs to L/M MFCCs in each of the M subbands.

Point #2 : MFCCs from an M-subband multi-band system appear in the (Mj)-th MFCCs of a full-band system as some linear combination of the former. From Eqn.(14), it is now clear how a corrupted subband will affect some of the full-band MFCCs.

Point #3 : More accurately speaking, the derived relationship is a property of the DCT.

IV. Conclusion

In this paper, we derive a mathematical relationship between every (Mj)-th full-band MFCCs with the corresponding j-th subband MFCCs if the full spectrum is partitioned into M equal subbands. It is still an open question if there is a simple relation between other full-band MFCCs and subband MFCCs.

V. Acknowledgements

I would like to thank Olivier Siohan of the Bell Labs, and my colleague Manhung Siu for valuable comments on the paper. This work is supported by the Hong Kong RGC under the grant number DAG00/01.EG09.

References

[1] F.H. Liu, R. Stern, A. Acero, and P.J. Moreno, "Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparison," in *Proceedings of ICASSP*, 1994, vol. II, pp. 61–64.

[2] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 113–120, 1979.

[3] M.G. Rahim and B.H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 19–30, Jan 1996.

[4] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[5] A.P. Varga and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," in *Proceedings of ICASSP*, 1990, pp. 845–848.

[6] H. Bourlard and S. Dupont, "A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," in *Proceedings of ICSLP*, October 1996.

[7] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on Partially Corrupted Speech," in *Proceedings of ICSLP*, October 1996.

[8] C. Cerisara, J.-F. Mari J.-P. Haton, and D. Fohr, "A Recombination Model for Multi-band Speech Recognition," in *Proceedings of ICASSP*, 1998, vol. II, pp. 717–720.

[9] N. Mirghafori and N. Morgan, "Sooner or Later: Exploring Asynchrony in Multi-Band Speech Recognition," in *Proceedings of the 6th Eurospeech*, 1999.

[10] K. Daoudi, D. Fohr, and C. Antoine, "A New Approach for Multi-band Speech Recognition Based on Probabilistic Graphical Models," in *Proceedings of ICSLP*, 2000, vol. 1, pp. 329–332.

[11] B. Mak and Y. C. Tam, "Asynchrony with Re-Trained Transition Probabilities Improves Performance in Multi-Band Speech Recognition," in *Proceedings of ICSLP*, 2000, vol. IV, pp. 149–152.

[12] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band Speech Recognition in Noisy Environments," in *Proceedings of ICASSP*, May 1998, vol. II, pp. 641–644.

[13] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium,"* September 2000.

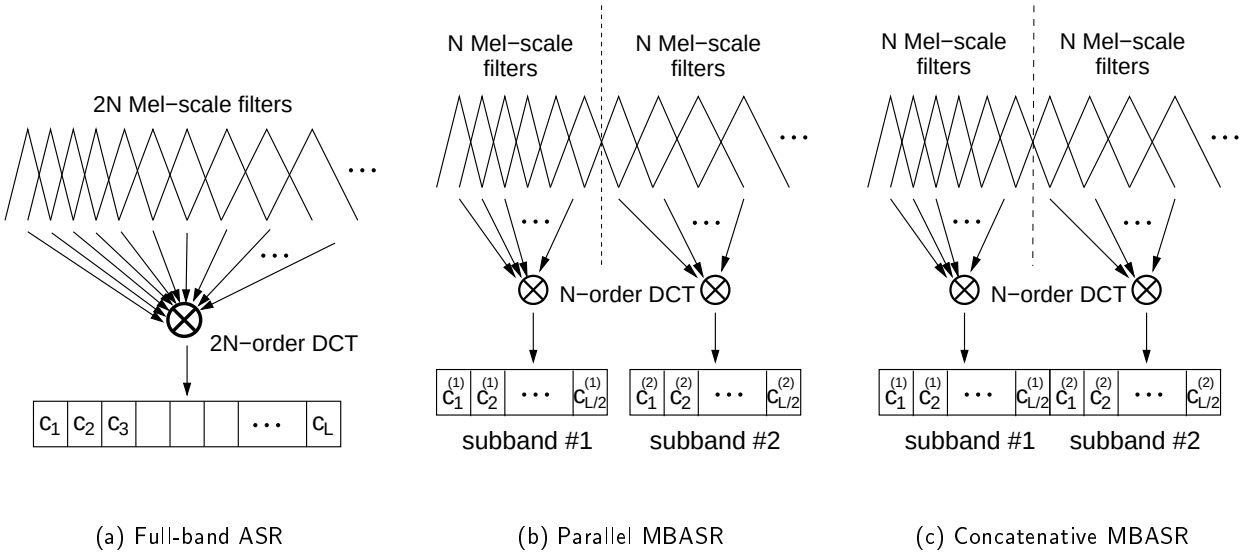
[14] R.G. Leonard, "A Database for Speaker-Independent Digit Recognition," in *Proceedings of ICASSP*, 1984.

[15] Y. C. Tam and B. Mak, "Optimization of Sub-Band Weights Using Simulated Noisy Speech in Multi-Band Speech Recognition," in *Proceedings of ICSLP*, 2000, vol. I, pp. 313–316.

TABLE I

THE PERFORMANCE OF A FULL-BAND ASR (FB), A PARALLEL MBASR (PMB), AND A CONCATENATIVE MBASR (CMB) ON CLEAN AND NOISY CONNECTED TIDIGITS. (NOISES ARE ARTIFICIALLY ADDED.)

	String Accuracy (%)			Word Accuracy (%)		
Noise	FB	CMB	PMB	FB	CMB	PMB
clean	94.00	94.31	93.60	98.80	98.86	98.78
white 0db	22.06	21.38	16.72	70.90	71.80	71.40
babble 0db	14.64	10.62	10.16	62.84	62.60	61.34



(a) Full-band ASR (b) Parallel MBASR (c) Concatenative MBASR

Fig. 1. Feature extraction for full-band and 2-subband multi-band ASR