TWO-DIMENSIONAL MULTI-RESOLUTION ANALYSIS OF SPEECH SIGNALS AND ITS APPLICATION TO SPEECH RECOGNITION

C.P. Chan, Y.W. Wong, Tan. Lee and P.C. Ching
Department of Electronic Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
{cpchan, ywwong, tlee1, pcching}@ee.cuhk.edu.hk

ABSTRACT

This paper describes a novel approach of using multi-resolution analysis (MRA) for automatic speech recognition. Two-dimensional MRA is applied to the short-time log spectrum of speech signal to extract the slowly varying spectral envelope that contains the most important articulatory and phonetic information. After passing through a standard cepstral analysis process, the MRA features are used for speech recognition in the same way as conventional short-time features like MFCCs, PLPs, etc. Preliminary experiments on both clean connected speech and noisy telephone conversation speech show that the use of MRA cepstra results in a significant reduction in insertion error when compared with MFCCs.

1. INTRODUCTION

In general, the speech short-time spectrum *S(f)* can be expressed as,

$$S(f) = G(f)V(f) \qquad (1)$$

$$\log S(f) = \log G(f) + \log V(f) \qquad (2)$$

where *G(f)* is the excitation and *V(f)* is the vocal tract spectral envelope. The logarithm operation turns the spectrum into a summation of two parts, namely the fine spectral details *log G(f)* and the spectral envelope *log V(f)*. For speech recognition purpose, only *V(f)* needs to be extracted since it carries most of the phonetic information.

Previous research shows that human articulators move at rates between 2-12Hz [7] while others demonstrate that modulations at rates above 16Hz are not required for speech intelligibility [8]. In [5,6], PLP and LPC cepstra are band-pass filtered respectively and results showed that most of the linguistic information lie in the range between 1 and 16Hz of the cepstral coefficients. These provide evidences that too many temporal details might not be necessary for recognition task. This is especially true for connected or continuous speech which require continuous movement of articulators. Therefore, the dynamic aspect of articulatory information lies with the temporal variation of the spectral envelope.

Multi-resolution analysis (MRA) is a signal decomposition technique that decomposes a signal into different frequency bands [1]. In this study, we use two-step MRA in both frequency and time domain, to: 1) separate the spectral envelop from fine spectral details; and 2) extract smooth temporal trajectory of the spectral envelopes. Conventionally, time trajectory of spectral features is represented by the first and second derivatives of the cepstral coefficients. Our proposed approach finds its novelty in that it applies a unified spectro-temporal analysis directly to the time-varying spectral envelope.

In the following sections, the feature extraction procedure and the preliminary experimental results will be described. Section 2 describes the 2-D MRA process and the method to obtain warped spectrum MRA cepstra. Section 3 introduces the experiments with clean connected speech and noisy telephone conversation speech to compare the proposed MRA cepstra and the conventional Mel-Frequency Cepstral Coefficients (MFCCs).

2. FEATURE EXTRACTION

2.1 2-D multi-resolution analysis of spectrograms

Multi-resolution analysis is a signal decomposition technique to separate the signal into certain number of frequency bands. MRA decomposes a signal *s(n)* into the detail components d_{mn} , m=1,2,…,L and the coarsest approximation c_{Ln} as shown in Figure 1. This is done by using multiple stages of identical low-pass filters *g(p)* and high-pass filters *h(p)*. The output of each stage is sub-sampled by two. The MRA filters are used to divide the spectrum successively by two.

The 2-D MRA extends the 1-D MRA to 2-D case. The idea is to first form a 1-D column sequence from the 2-D image, perform 1-D MRA, then restore the MRA output to the 2-D format and do another 1-D MRA to the row sequence. Figure 2 illustrates the 2-D MRA decomposition. Separable filters [10] are used in this study rather than non-separable one [9] because it is simple to implement and the 1-D filters can be re-used. Since we are not interested in the spectro-temporal details of the speech waveform, we perform low-pass filtering in the 2-D MRA and obtain the approximation coefficients while ignoring all detail coefficients.

Typical Daubechies-2 wavelet filter [11], with extremal phase and highest number of vanishing moments, is used in this study. Moreover, short filter has the advantage of reducing the time delay introduced by filtering in practical automatic speech recognition (ASR).

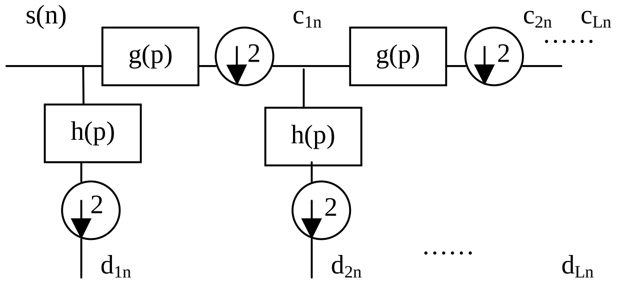

Figure 1: 1-D MRA decomposition

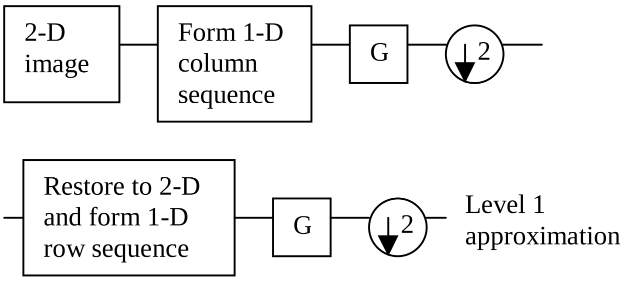

Figure 2: Lowpass filtering of 2-D MRA decomposition

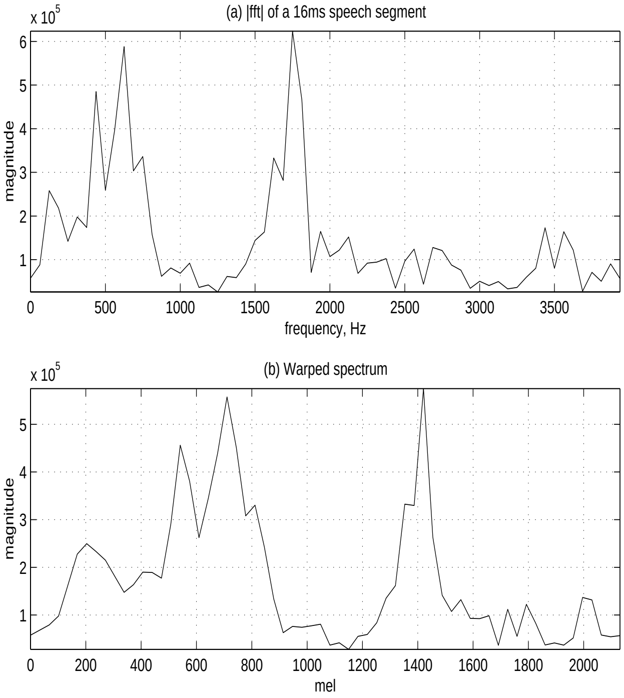

Figure 3: (a) STFT of a 16ms speech segment. (b) Warped spectrum in mel-scale

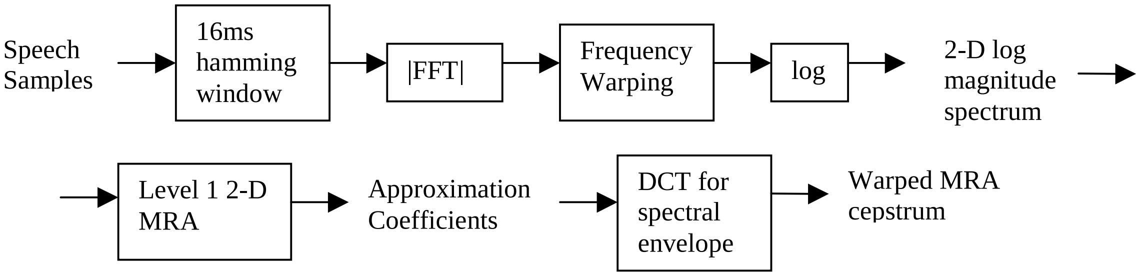

Figure 4: Procedure for computation of warped MRA cepstrum

2.2 Warped frequency scale

Psychophysical studies have shown that human perception of the frequency content of sounds does not follow a linear scale. Each pure tone with an actual frequency is perceived as a subjective pitch [3]. In order to simulate the perceptual properties of human beings, we warp the frequency spectrum based on mel-scale before performing the MRA**.** The mel-scale is defined by

$$Mel(f) = 2595*log_{10}(1+f/700) \quad (3)$$

We transform the constant frequency spacing FFT point to constant mel frequency interval by eqn. (3) and linear interpolation. Figure 3 gives an example of a pair of unwarped and warped short-time frequency spectrums.

2.3 The warped MRA cepstrum

The cepstrum of the spectral envelope (column sequence) extracted by the 2-D MRA is calculated by applying discrete cosine transform (DCT) to obtain 12 coefficients. These 12 MRA cepstra are used for comparison with the MFCCs commonly used in ASR. The MRA cepstrum is calculated as in Figure 4. Figure 5 shows the difference in time trajectory of one of the cepstrum (c1) of MFCC and warped MRA cepstra. The time trajectory of the warped MRA cepstrum contains less fluctuation than that of MFCC.

In the following experiments, 16ms window width is used and level 1 approximation is good enough to extract the spectral envelope. Only 0-4kHz spectrum is used for MRA.

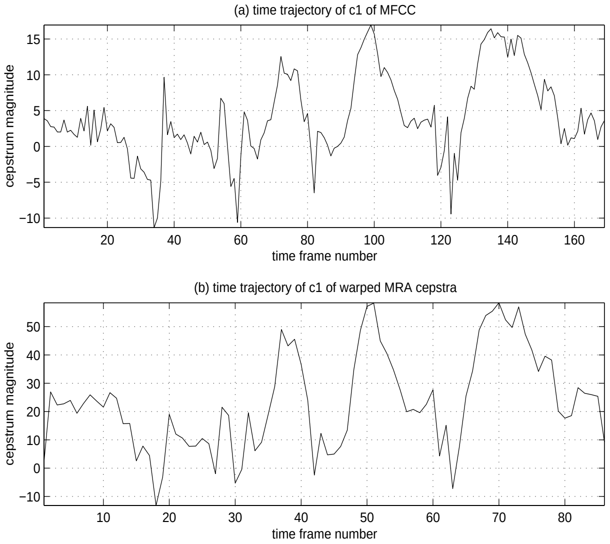

Figure 5: Time trajectory of cepstrum-1 for a) MFCC and b) warped MRA cepstra for Cantonese utterance "saa1-tin4-daai6-wui6-tong4".

3. EXPERIMENTAL RESULTS

3.1 Experiment 1 – Cantonese CUWORD Corpus

The Cantonese CUWORD database [13] was designed to provide Cantonese speech data for acoustic modeling at syllable or sub-syllable level. The basic corpus consists of 2,527 polysyllabic words, each of which is 1 - 7 syllables in length. 13 male and 15 female speakers were recorded. Each of them read the entire corpus once in a moderately quiet room. As a result, about 70,000 utterances were obtained and manually transcribed. The speech data from 22 speakers (10 male and 12 female) are designated for the training of speech recognition systems while the remaining data are for performance evaluation purpose.

Syllable-level HMMs have been trained with 2-D MRA cepstra and MFCCs. In each case, a total of 573 syllable models with 3 Gaussian mixtures per state were trained using the HMM toolkit (HTK) [12]. For MFCCs, the analysis window length was 25 msec and frame period was 10 msec. Each feature vector had 26 components, including 12 MFCCs, energy and their delta.

We compared the MFCC baseline system with the following MRA parameter sets:

CU-MRA-D :	12 unwarped spectrum MRA cepstra, 12 delta cepstra
CU-WMRA-D :	12 warped spectrum MRA cepstra, 12 delta cepstra

Table 1 and 2 show the utterance-level recognition rate and number of syllable-level recognition errors respectively. Without using any lexical and grammatical constraints, 25.49% of the test utterances can be correctly recognized by the baseline system. The use of MRA cepstra results in a significant improvement to 30.91% utterance recognition rate. As show in Table 2, the major improvement is attained by reducing 81.13% of the insertion errors (from 7886 to 1488) while the number of deletions and substitutions only increase slightly. If the warped MRA cepstra are used instead, the number of all types of errors are reduced and the resulted utterance recognition rate is 33.1%

	CU-Baseline	CU-MRA-D	CU-WMRA-D
Utterance correctness	25.49%	30.91%	**33.10%**

Table 1: Utterance-level recognition rate with MRA cepstra and MFCC for CUWORD corpus (Total utterance number: 14896)

	Deletion	Substitution	Insertion
CU-Baseline	28	13302	**7886**
CU-MRA-D	85	14675	**1488**
CU-WMRA-D	76	14004	**1307**

Table 2: Syllable recognition error with MRA cepstra and MFCC for CUWORD corpus (Total number of syllables in test data: 41170)

3.2 Experiment 2 - Mandarin Call Home Corpus

The second release (Apr95) of Mandarin Call Home corpus [4] produced by the Linguistic Data Consortium (LDC) was used. It is a Mandarin corpus of telephone conversations. There are 80 conversations in the training set and 20 in development test set. Channel noise and distortions occur in Call Home because of signal transmission over international connections. The purpose of the experiment on Call Home is to investigate the effectiveness of MRA cepstra on noisy, distorted telephone channel.

12 MFCC and their delta were used in the baseline system, forming a 24-component vector for each frame. The frame size was 16 msec and frame period was 10 msec. Syllable back-off bigram [12] built on the Apr95 training set was also added to the baseline system. A total of 388 syllable HMMs, with 8 Gaussian mixtures per state, were trained using HTK.

We compared the baseline system with the following MRA parameter set:

CH-WMRA-D:	12 warped spectrum MRA cepstra, 12 delta cepstra.

And the effect of MRA features was also investigated with the bigram added.

Table 3 shows the syllable-level recognition errors for Mandarin Call Home corpus. The performance of the warped MRA cepstra without language model (LM) is more or less the same as to the baseline system (with LM). Each type of the error is very close in both cases. When both MRA cepstra and language model are used together, the number of insertion errors can be further reduced by 58.3% (from 5981 to 2493) and the utterance recognition rate becomes 11.32% (Table 4). This may suggest that the functionality of MRA and the language model are complementary to each other.

The above recognition results are consistent with the experiment using clean speech in CUWORD corpus. This indicates that the MRA method also works well for noisy and distorted telephone channel with conversation speech.

	Deletion	Substitution	Insertion
CH-Baseline	1576	19179	**5948**
CH-WMRA-D	1323	19692	**5981**
CH-WMRA-D w/ bigram	2720	17277	**2493**

Table 3: Syllable recognition error using MRA cepstra and MFCC for Mandarin Call Home corpus (total number of syllables in test data: 27436)

	CH-Baseline	CH-WMRA-D	CH-WMRA-D w/ bigram
Utterance correctness	7.08%	8.75%	**11.32%**

Table 4: Utterance correctness using MRA cepstra and MFCC for Mandarin Call Home corpus (total utterance number: 3660)

4. DISCUSSION

From both experiments described above, we consistently observe that the MRA cepstra can greatly improve the utterance-level recognition accuracy by reducing insertion errors. This may be due to that, as shown in Figure 5, the proposed MRA features have a smoother time trajectory which would cause fewer undesirable state transitions in the decoding process of HMM. Such reduction of insertion errors finds its importance in continuous speech recognition, especially for Chinese language because of its phonetic structure.

Mandarin Call Home is a very difficult task since it deals with noisy spontaneous speech. In [4], a character recognition accuracy of about 35% was attained by making use of a lot of linguistic constraints including a 44,000-word lexicon, a word trigram, etc… In our experiment, we have used a simple syllable bigram which covers a small portion of lexical and grammatical information. Therefore, the use of MRA can only attain a low utterance recognition rate of 11.32% for the Mandarin Call Home experiment. Nevertheless, it still outperforms MFCCs by having much less insertion errors. Indeed, the utterance-level recognition accuracy of 33.10% for the CUWORD corpus is highly impressive if we consider the fact that no linguistic information has been incorporated yet.

In addition to the great improvement in insertion error, the recognition speed is faster when using MRA cepstra because the number of frames is smaller due to the sub-sampling by 2 in MRA.

5. CONCLUSION

2-D MRA has been proposed as a unified analysis technique to extract spectro-temporal features for speech recognition. It is aimed at preserving the most important spectral and temporal features while discarding the irrelevant ones. The preliminary experimental results show that the 2-D MRA cepstra reduces a significant amount of insertion error when compared with standard MFCC for both clean connected speech and noisy, distorted telephone conversation speech. Although it is too early to conclude that the proposed MRA features would generally outperform conventional acoustic features in speech recognition applications, our work presents a promising direction of research along which much more work need to be done.

ACKNOWLEDGEMENT

This work is supported by research grants from Hong Kong Research Grant Council and Industry Department of Hong Kong SAR Government.

REFERENCES

[1] S.G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Trans. Pattern Analysis, Machine Intelligence*, Vol. 11, No. 7, pp. 674-693, July 1989.

[2] S. Greenberg and Brian E.D. Kingsbury, "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech", *Proc. ICASSP-97*, pp. 1647-1650, 1997.

[3] S.S. Stevens and J. Volkmann, "The relation of pitch of frequency: A revised scale". *Am. J. Psychol.*, 53: 329-353, 1940.

[4] F.H. Liu, M. Picheny, P. Srinivasa, M. Monkowski, and J. Chen, "Speech Recognition on Mandarin Call Home: A Large-Vocabulary, Conversational and Telephone Speech Corpus". *Proc. ICSLP*-96, pp. 157-160, 1996.

[5] N. Kanedera, H. Hermansky and T. Arai, "On Properties of Modulation Spectrum for Robust Automatic Speech Recognition", Proc. *ICASSP-98*, pp. 1448-1451, 1998.

[6] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelopes", *Proc. ICSLP*-96, pp.2490-2493, 1996

[7] Caroline L. Smith, Catherine P. Browman, Richard S. McGowan, and Bruce Kay, "Extracting dynamic parameters from speech movement data", *J. Acoust. Soc. Am.*, 93(3) : 1580-1588, March, 1993.

[8] Rob Drullman, Joost M. Festen, and Reinier Plomp. "Effect of temporal envelope smearing on speech reception", *J. Acoust. Soc. Am.*, 95(2):1053-1064, February 1994.

[9] G. Karlsson, M. Vetterli, "Theory of Two-Dimensional Multirate Filter Banks", *IEEE Trans. Acoust. Speech and Signal Processing*, Vol. 38, No. 6 pp. 925-937, June 1990.

[10] Y. T. Chan, *Wavelet Basics*, Kluwer Academic Publishers, 1995

[11] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992

[12] Steve Young, *The HTK Book*, Cambridge University, 1995.

[13] W.K. Lo, Tan Lee and P.C. Ching, "Development of Cantonese Spoken Language Corpora for Speech Applications", to appear in *Proc. ISCSLP-98*, Singapore, Dec.1998.