International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

A COMPARATIVE STUDY ON WAVELET PACKET BASED FRONT-END IN CONNECTED MANDARIN DIGIT RECOGNITION

Xiu-Ping Wang   Chuan-Qi Zhu   Zong-Ge Li

Department of Computer Science, Fudan University, Shanghai
xpwang_fd@hotmail.com

ABSTRACT

This paper investigates the wavelet packet based front-ends for the connected mandarin digit recognition task. Firstly an ERB-like wavelet packet basis is proposed. Then two kinds of wavelets are selected for comparison. One is the Vaidyanathan wavelet, which has good frequency selectivity but big shift variance. The other is the reverse biorthogonal spline wavelet with excellent shift invariant property. Thirdly, the Teager-Kaiser energy operator (TEO) based subband cepstral (TC) feature parameters are extracted from the wavelet packet derived multi-frequency channels. The recognition results of the new front-ends are tested and compared with the popular MFCC parameter on the 8K 16-bit speaker-independent mandarin connected digit corpora. Apart from clean data condition, the performances of the new front-ends are further compared in various noisy conditions.

1. INTRODUCTION

Wavelet analysis is suitable for non-stationary signal representation because it has the ability to capture localized features of signals. Recently, some interests were cast on the wavelet packet based speech recognition front-end for its flexibility in tiling of the time-frequency plane and computational efficiency. Erzin [1] proposed the subband based cepstral coefficients that outperform the MFCC by 10% in Turkish digits recognition with the presence of car noise. Jabloun [2] introduced TEO to calculate the subband energy, from which a more robust version of subband cepstral coefficient was obtained. Sarikaya [3][4] achieved 24% and 17.2% monophone recognition improvement for loud and angry stressed speech over MFCC parameters by wavelet packet based parameters. Kryze [5] compared the robustness of wavelet packet derived feature vectors with PLP. A 20% error rate relative improvement compared to PLP was obtained in context-dependent modeling under the condition of 5 dB SNR car noise.

This paper studies the effect of wavelet packet front-ends on the 8kHz-sampled speaker-independent connected mandarin digit recognition task. Firstly an ERB-like wavelet packet basis is proposed. Secondly, two kinds of wavelet filters are selected for comparison. One is the orthogonal Vaidyanathan filter and the other is the reverse biorthogonal spline wavelet. The vaidyanathan filter [6][7] is chosen because it was optimized for speech coding. The biorthogonal spline wavelet [8][9] is chosen because of its translation invariance. Thirdly, the Teager-Kaiser energy operator (TEO) based subband cepstral (TC) feature parameters are extracted from the wavelet packet derived multi-frequency channels. Finally, the performances of the wavelet packet front-ends are compared with the popular MFCC parameter in four kinds of conditions, i.e., clear data condition, car noise at 5 dB SNR, car noise at 10 dB SNR, and white noise condition. Simulation results show that in clean data condition and car noise condition, the proposed Vaidyanathan wavelet packet based front-end has equivalent performance with the popular MFCC. However, in the case of white noise, the performances of the both wavelet packet based front-ends degrade considerably.

2. WAVELET PACKET BASED FEATURE EXTRACTION

2.1 Wavelet packet based filter banks

Wavelet analysis uses high pass and low pass filter iteratively to decompose speech signals into a dyadic tree of approximations and details. In discrete time wavelet transform, only the low pass branches (approximations) are split and iterated on. Wavelet packets offer more flexibility in frequency-time tiling because in wavelet packet analysis, the details as well as the approximations are split. Various bases can be selected from the wavelet packets tree so as to achieve different frequency-time tiling. Supposing there are 2^d nodes in level d of the basis tree, given the Nyquist frequency f_N, the center frequency f_c of the wavelet filter bank at node (d, b) is approximately:

$$f_c = \frac{2b+1}{2^{d+1}} f_N \quad , \qquad (b = 0, \dots , 2^d) \tag{1}$$

The Equivalent Rectangular Bandwidth (ERB) is given by:

$$ERB(f) = 24.7(0.00437 f + 1) \tag{2}$$

26 ERB spaced center frequencies from 100 Hz to 3800 Hz are calculated by [10]:

$$f_c = \frac{676170.4}{47.06538 - \exp(0.08950404 \cdot r)} - 14678.49 \tag{3}$$

where r takes the value of 26 linearly spaced points from 100 to 3800. Comparing the center frequencies derived by equation (1) and (3), an ERB-like wavelet packet basis is selected and shown in Figure 1. The first two filter banks in this complete basis are ignored because their center frequencies are below 100Hz. Thus we have 26 filter banks in total.

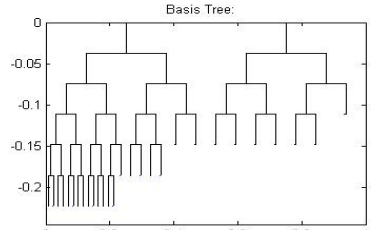
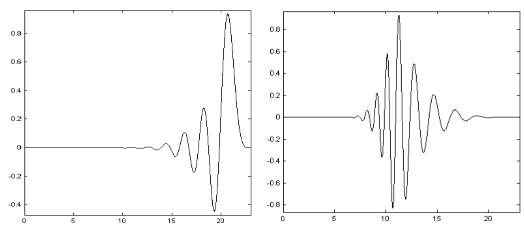


Figure 1: wavelet packet basis

For comparison, two kinds of wavelet bases are considered in this paper. One is the Vaidyanathan and the other is the reverse biorthogonal spline wavelet family. The Vaidyanathan filter (Vaid) is a 24-parameter QMF optimized for speech coding. Figure 2 illustrates its scale function and wavelet function. The Vaid wavelet has the advantage of good frequency selectivity, but its shift variance is large. Shift variance is caused by decimation. The discrete wavelet coefficients of signal $s(k)$ at octave i is calculated by

$$w(2^i, 2^i n) = \frac{1}{\sqrt{2^i}} \sum_k \overline{\Psi}(\frac{k}{2^i} - n) s(k) \tag{4}$$

where Ψ is the mother wavelet, the bar represents complex conjugate, n stands for time translation. The shift invariance [11] can be easily seen by substituting $s(k-r)$ for $s(k)$ which produces $w(2^i, 2^i(n - r/2^i))$, an integer translation of $w(2^i, 2^i n)$ only if r is a multiple of 2^i. Large shift variance makes the coefficients of different subbands inconsistent. Compensations have to be made to get reasonable phase plane of the signal. To investigate the effect of wavelets with good shift invariant property, we selected another wavelet named the reverse biorthogonal spline wavelet (rbio) family. The rbio6.8 filter pair from this wavelet family is chosen. For further information about rbio6.8 filters, refer to [9].



(a) (b)

Figure 2: Vaidyanathan wavelet
(a) Scale function (b) Wavelet function

2.2 The Teager-Kaiser energy operator (TEO) based subband cepstral (TC)

The Teager-Kaiser energy operator (TEO) [2] tracks the energy of the source producing an oscillation signal $x(t)$, the discrete-time version of the TEO is given by:

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{5}$$

The TEO based frame energy of subband d is

$$E_d = \frac{1}{N_d} \sum_i^{N_d} |\Psi[x_d(i)]| \quad ; \quad d = 1, \ldots, D \tag{6}$$

N_d is the window length of subband d.

Taking the discrete cosine transform of the log subband TEO energies derives the TC parameters

$$TC(k) = \sum_{d=1}^{D} \log(E_d) \cos(\frac{k(d-0.5)}{D} \pi), \ \mathrm{k} = 1, 2, \ldots, 13 \tag{7}$$

The first 13 $TC(k)$ coefficients, together with their deltas and acceleration coefficients are used in the feature vector. A final feature vector with dimension 39 is obtained and is used for training and recognition.

3. CORPUS

The corpus was recorded in laboratory at 16 KHz 16 bits and down sampled to 8 KHz 16 bits. It contains 2051 utterances (1741 4-digit-strings and 310 8-digit-strings) of connected mandarin digits spoken by 138 adult speakers (68 females and 70 males). The speakers were instructed to read from the list in a naturally speaking way. The corpus adds up to 50 minutes excluding silence. The vocabulary consists of 11 Chinese digits 0~9 and a 'yao1', which is an option for '1/yi1/'. Refer to Table 1 for a summary of the corpus.

DATA SET	TRAIN	TEST
No. of 4-digit-strings	1603	138
No. of 8-digit-strings	201	29
No. of female speakers	64	4
No. of male speakers	66	4

Table 1: Summary of the corpus

The car noise [8] was recorded at 19.98 KHz by the Institute for perception-TNO, The Netherlands, and was also down sampled to 8 KHz. The white noise is generated using the Matlab random number generator. The noisy speech was obtained by adding the clean speech and noise data proportionally according to a preset SNR.

4. IMPLEMENTATION ISSUES

In implementation, DC mean is removed and pre-emphasis ($1 - 0.97z^{-1}$) is processed before feeding speech signal into the wavelet packet filter banks.

Since the wavelet packet filters process the whole utterance continuously, the conventional step of beforehand windowing is replaced by averaging afterwards in subband based front ends. This gives more flexibility in choosing frame sizes and frame rates. In the Vaid case, we chose the 32 ms frame size and 8 ms frame rate. As to the rbio6.8 version of feature parameters, the frame size is 25 ms and the frame rate is 10ms.

There are different numbers of coefficients in different levels of wavelet packet decomposition. With good shift invariant property, the rbio6.8 wavelet packet coefficients can be easily shaped into the form of frame-by-frame feature vectors by using windows of different lengths at various levels. This is not the case with the Vaid wavelet base, for which we have to compensate the shift variance effect by calculating the wavelet packet atoms' contribution on the frames.

Figure 3 is the spectrograms of the mandarin digit string '8275'. Figure 3 (c) and Figure 3 (d) are subband log TEO energy plots using the two proposed wavelet packets respectively. For comparison, the spectrogram based on short time Fourier transform (SFT) is plotted in Figure 3 (a) and the correspondent 26 Mel-filter banks log energy is plotted in Figure 3 (b). As shown in Figure 3, dyadic sampling makes the subband energy plots not so continuous as that of the Mel-filter bank.

Chinese is a tonal language with one syllable per character. The difficulty of connected mandarin digits recognition mainly lies in three aspects. Firstly there are three single vowel digits pronounced /yi1/, /er4/ and /wu3/. Segmentation of vowel-vowel sequence such as /yi1 yi1/ is difficult. Secondly, the sequences /qi1 yi1/, /jiu3 wu3/ and /liu4 wu3/ are confusing with the single word /qi1/, /jiu3/ and /liu4/. These result in high deletions and insertions. Thirdly, there exist several confusing pairs, e.g. {/er4/ /ba1/} {/liu4/ /jiu3/} {/liu4/ /ling2/} etc.

To tackle these difficulties, firstly, crossword triphone HMMs are build to handle the coarticulation among the digits. The baseline recognizer is a speaker independent, cross-syllable-triphone, decision-tree-based tied-state system and is trained incrementally with the HTK3.0. The final HMMs have a 3-state left to right structure with 6 Gaussian mixtures per state. The 39 dimensional feature vectors of the baseline system consists of 12 MFCC's, a C0 and their first, second order derivates.

5. SIMULATION RESULTS

5.1 Results in clean data condition

Table 2 shows the recognitions scores of clean data. Both the context independent (CI) recognition result of monophone model and the context dependent (CD) final result of the tied-triphone model are compared. Percent correct (PC) stands for word correct percentage and WA stands for word accuracy. From this table we can see that the performance of MFCC and TC (Vaid) are very close in clean data condition and the performance of the TC (rbio6.8) is relatively low.

Analysis type	CI		CD	
	PC%	WA%	PC%	WA%
MFCC	80.43	74.42	96.43	94.52
TC (Vaid)	80.13	73.65	95.92	94.01
TC (rbio6.8)	77.58	72.01	92.60	91.07

Table 2: comparative recognition results of clean speech

5.2 Results in noisy condition

The following simulation results are obtained in matched conditions, i.e., the HMMs are trained and tested in the same noise conditions. Car noise is band limited, low frequency noise. Table 3 is the recognition result with presence of a 10dB car noise. It's interesting that most of the recognition results improve a little in this noisy condition. It seems that a little of low frequency noise helps in improving recognition accuracy. Table 4 is the recognition results with presence of a 5dB car noise. Table 5 is the results of white noise corrupted data at 10 dB SNR. The wavelet packet based front-ends degrades significantly with the presence of white noise.

Analysis type	CI		CD	
	PC%	WA%	PC%	WA%
MFCC	82.93	79.78	96.30	94.77
TC (Vaid)	78.39	74.14	95.41	94.26
TC (rbio6.8)	78.32	72.97	92.35	91.45

Table 3: comparative recognition results for noisy speech data with car noise at 10dB SNR

Analysis type	CI		CD	
	PC%	WA%	PC%	WA%
MFCC	84.98	80.00	95.66	94.52
TC (Vaid)	80.44	73.11	92.60	92.47
TC (rbio6.8)	76.56	72.82	92.60	91.96

Table 4: comparative recognition results for noisy speech data with car noise at 5dB SNR

Analysis	CI		CD	
	PC%	WA%	PC%	WA%
MFCC	67.4	63.4	92.09	90.69
TC (Vaid)	64.03	57.00	84.18	82.40
TC (rbio6.8)	64.10	55.38	80.99	79.46

Table 5: comparative recognition results of white noise corrupted speech at 10dB SNR

6. CONCLUSION

In this paper, the Vaidyanathan wavelet packet and the reverse biorthogonal spline wavelet are used for feature extraction in the speaker-independent connected mandarin digit string recognition task. Simulation results show that in clean data condition and car noise condition, the proposed Vaidyanathan wavelet packet based front-ends have equivalent performance with the popular MFCC. Though recognition result of the rbio6.8 wavelet is the lowest, it is the most computationally efficient method. In the white noise corrupted data condition, the performances of wavelet packet based front-ends degrade considerably.

7. DISCUSSION

Wavelet analysis provides an alternative for speech feature extraction. The main advantages of the wavelet packet based front-ends lie in its computational efficiency and diversity. However, the dyadic sampling and the resultant shift variance make it difficult to shape the wavelet packet coefficients into the frame-by-frame form required by most popular recognition tools such as htk. There are two ways to tackle this problem. One is to find a wavelet function with better frequency selectivity and small shift variance. Since these filters are designed specially for speech recognition, the perfect reconstruction condition can be relaxed. The other way is to construct a special HMM structure for wavelet packets coefficients, as is done in [12].

8. REFERENCES

[1] Erzin E., Cetin A.E., and Yardimci Y. "Subband analysis for robust speech recognition in the presence of car noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, May 1995.

[2] Jabloun F., Cetin A. E. "The Teager energy based feature parameters for robust speech recognition in car noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

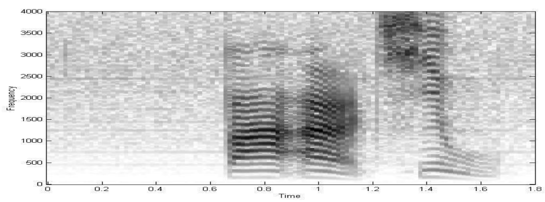
[3] Sarikaya R. and Hansen J. H. L. "High resolution speech feature parameterization for monophone-based stressed speech recognition," *IEEE Signal Processing Letters*, 7(4): 182-185, 2000.

[4] Sarikaya R. and Gowdy J. N. "Subband based classification of speech under stress," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Seattle, May 1998.

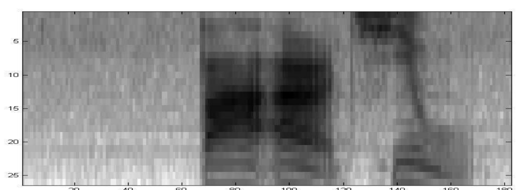
[5] Kryze D., Rigazio L., Applebaum T., and Junqua J-C. "A new noise-robust subband front-end and its comparison to PLP," *Automatic Speech and Understanding Workshop*, I-55, December, Keystone, Colorado, 1999.

[6] Vaidyanathan P. P. and Huong P.-Q. "Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks," *IEEE Trans. on*

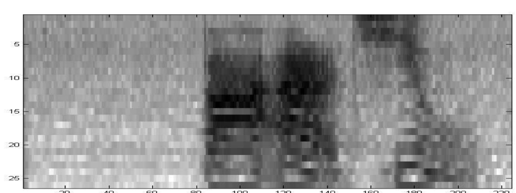
Acoustics, Speech, and Signal Processing, 36(1): 81-94, 1988.
[7] WaveLab 802 (http://www-stat.stanford.edu/~wavelab)
[8] Daubechies I., *Ten lectures on Wavelets*, SIAM, 1992, pp. 271-280.
[9] Matlab Wavelet Toolbox 2.1
[10] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[11] Shensa M. J., "The discrete wavelet transform: wedding the a trous and Mallat algorithms," *IEEE Trans. On Signal Processing,* 40(10): 2464-2482, 1992.
[12] Crouse M. S., Nowak R. D., Barahuik R.G., "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. On Signal Processing*, 46(4): 886-902, 1998.



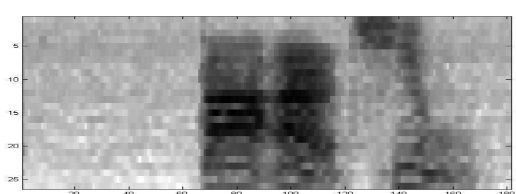
(a)



(b)



(c)



(d)

Figure 3: Spectrograms of mandarin digit string '8275'
(a) Spectrogram based on SFT
(b) Mel-filter bank log energy, 25ms/frame, shift=10ms
(c) Vaid subband log TEO energy, 32ms/frame, shift=8ms
(d) rbio6.8 subband log TEO energy, 25 ms/frame, shift=10ms