A STUDY ON CEPSTRAL SUB-BAND NORMALIZATION FOR ROBUST ASR

Syu-Siang Wang[1], *Jeih-weih Hung*[2], *Yu Tsao*[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]Dept. of Electrical Engineering, National Chi Nan University, Nantou, Taiwan

ABSTRACT

In this paper, we propose a cepstral subband normalization (CSN) approach for robust speech recognition. The CSN approach first applies the discrete wavelet transform (DWT) to decompose the original cepstral feature sequence into low and high frequency band (LFB and HFB) parts. Then, CSN normalizes the LFB components and zeros out the HFB components. Finally, an inverse DWT is applied on LFB and HFB components to form the normalized cepstral features. When using the Haar functions as the DWT bases, the calculation of CSN can be processed efficiently with a 50% reduction on the amount of feature components. In addition, our experimental results on the Aurora-2 task show that CSN outperforms the conventional cepstral mean subtraction (CMS), cepstral mean and variance normalization (CMVN), and histogram equalization (HEQ). We also integrate CSN with advanced front-end (AFE) for feature extraction. Experimental results indicate that the integrated AFE+CSN achieves notable improvements over the original AFE. The simple calculation, compact in form, and effective noise robustness properties enable CSN to perform suitably for mobile applications.

Index Terms— discrete wavelet transform, CMS, CMVN, RASTA, noise robust, speech recognition.

1. INTRODUCTION

Degradation on automatic speech recognition (ASR) performance under noisy conditions is a crucial drawback. To fix this issue, many approaches have been proposed to reduce the effect of noise components from speech data by means of normalizing speech features. Cepstral mean subtraction (or normalization, CMS, CMN) [1] [2] is a successful method to normalize cepstral features by subtracting the means from speech frames. Cepstral mean and variance normalization (CMVN) [3] and higher order cepstral moment normalization (HOCMN) [4] use second and higher order cepstral moment normalization to adjust the distribution of noisy speech features closer to that of the clean ones. In addition, histogram equalization (HEQ) [5] applies a mapping function to convert the noisy speech features to another predefined (or referenced) distribution to alleviate the mismatch cased by noise.

Other than normalizing speech features to improve ASR performance, filter design is another method to suppress noise effect in speech features. All the approaches are usually applied assuming that major speech components are located around the low modulation frequency parts (except for the DC component). A notable example is the relative spectral (RASTA) bandpass filter [6], which preserves the informative speech components around 4 Hz while suppresses components at other frequencies in the modulation frequency domain. Another successful approach filters out less important speech components based on the decorrelation property of discrete cosine transform (DCT) by deriving a band-pass filter using DCT techniques [7]. A DC-removed DCT-based filter is proposed to achieve further improvements [8].

Recently, a novel subband feature statistics normalization technique has been proposed [9]. This technique first applies the discrete wavelet transform (DWT) [10] to decompose full-band speech features into several subbands. Speech components in each subband are normalized separately by CMVN or HEQ [9] processes. This subband normalization technique provides further improvements over the conventional full-band-based normalization techniques because each subband carries distinct speech and noise information.

In this paper, we propose a cepstral subband normalization (CSN) approach. By applying the Haar function [10] as DWT bases, the CSN procedure can be processed easily with a 50% reduction on the amount of feature components. In addition, our experimental results indicate that CSN approach outperforms the conventional CMS, CMVN, and HEQ techniques on the Aurora-2 [11] speech recognition tasks. Furthermore, we integrate CSN with the advanced front-end (AFE) [12] for feature extraction. Our experimental results show that the integrated AFE+CSN provides better recognition performance than the AFE alone.

The remainder of this paper is organized as follows: section 2 briefly introduces the DWT theory. Section 3 presents the proposed CSN approach. Section 4 shows the experimental setup and discusses the experimental results. Finally, section 5 concludes this study.

2. WAVELET TRANSFORM

Fig. 1 shows the flowchart of wavelet transform (WT) and inverse wavelet transform (IWT). For a signal, $f(t)$, we apply

WT to decompose it into two parts, $a(k)$ and $b(k)$ (equation (1)) carrying information of the lower and higher-frequency components of $f(t)$, respectively.

$$f(t) = \sum_k a(k)\phi_k(t) + \sum_k b(k)\psi_k(t), \tag{1}$$

where,

$$\phi_k(t) = \sqrt{2}\phi(2t + k),$$
$$\psi_k(t) = \sqrt{2}\psi(2t + k).$$

Parameters k and t are the time indices in Eq. (1). $\phi_k(t)$ and $\psi_k(t)$, called scale and wavelet functions, are designed as low-pass and high-pass filters and orthogonal to each other:

$$\langle \phi_k(t), \psi_k(t) \rangle = 0; k \in \mathcal{Z}, t \in \mathcal{R}. \tag{2}$$

Meanwhile, the scale and wavelet functions satisfy

$$\begin{aligned} \langle \phi_k(t), \phi_l(t) \rangle = \int \phi_k(t)\phi_l(t)dt = \delta(l, k), \\ \langle \psi_k(t), \psi_l(t) \rangle = \int \psi_k(t)\psi_l(t)dt = \delta(l, k), \end{aligned} \tag{3}$$

where $\delta(l, k)$ is the Kronecker delta function.

To perform WT, we calculate $a(k)$ and $b(k)$ in Eq. (1) by

$$\begin{aligned} a(k) = \langle f(t), \phi_k(t) \rangle = \int f(t)\sqrt{2}\phi(2t - k)dt, \\ b(k) = \langle f(t), \psi_k(t) \rangle = \int f(t)\sqrt{2}\psi(2t - k)dt, \end{aligned} \tag{4}$$

where the constant $\sqrt{2}$ is used to preserve the norm of time-scaling functions. On the other hand for IWT, we reconstruct a signal, $\tilde{f}(t)$, with $a(k)$ and $b(k)$ by

$$\tilde{f}(t) = \sum_k a(k)\tilde{\phi}_k(t) + \sum_k b(k)\tilde{\psi}_k(t), \tag{5}$$

where $\tilde{\phi}_k(t)$ and $\tilde{\psi}_k(t)$ have the same properties as $\phi_k(t)$ and $\psi_k(t)$ in Eqs. (2) and (3). With a careful design of $\phi_k(t)$, $\psi_k(t)$, $\tilde{\phi}_k(t)$ and $\tilde{\psi}_k(t)$, we can apply IWT to perfectly recover the original signal ($\tilde{f}(t) = f(t)$).

Based on the WT theory, the discrete WT (DWT) theory has been derived to process discrete-time signals. The DWT theory uses the same concepts as WT that performs decomposition and reconstruction on signals with designed scale and wavelet functions. In this study, we propose a filtering process based on DWT to normalize speech features to enhance speech recognition performance under noisy conditions.

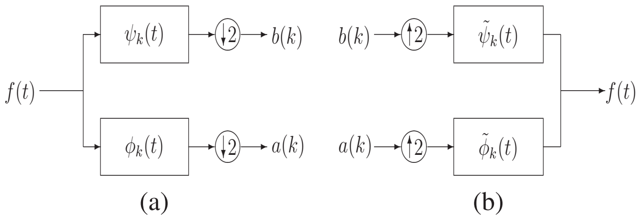

Fig. 1. Flowcharts for (a) decomposition and (b) reconstruction process, where downarrow- and upperarrow- 2 represent 2-order down-sampling and up-sampling processes.

3. CEPSTRAL SUBBAND NORMALIZATION (CSN)

The cepstral subband normalization (CSN) algorithm is derived, considering that the noise-affected cepstral features are located in higher modulation frequency bands. By applying DWT, CSN decomposes the original cepstral feature sequence into low- and high- frequency band parts (LFB and HFB). Then, CSN normalizes the LFB and zeros out the HFB components. Finally, we apply inverse DWT (IDWT) on the LFB and HFB components to form normalized cepstral features. The procedure of CSN is demonstrated in Fig. 2.

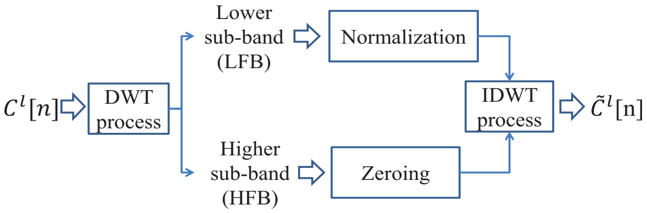

Fig. 2. The flowchart for CSN procedure

Many functions can be used as scale and wavelet functions for DWT bases. In this study, the Haar functions are applied, which design $\phi[n]$ and $\psi[n]$, $\tilde{\phi}[n]$, and $\tilde{\psi}[n]$ by

$$\left\{ \begin{array}{l} \phi_0[n] = \{\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\}, \\ \psi_0[n] = \{\frac{-\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\}, \end{array} \right. \quad \left\{ \begin{array}{l} \tilde{\phi}_0[n] = \{\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\}, \\ \tilde{\psi}_0[n] = \{\frac{\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2}\}, \end{array} \right. \tag{6}$$

where n is the time index. With the designed DWT bases in Eq. (6), speech cerpstral features can be decomposed into LFB and HFB components. Next, CSN applies a normalization algorithm on LFB and zeros out HFB components.

$$\begin{array}{l} a[n] = H_L\{\mathbf{C}^l[n]\} \\ b[n] = \mathbf{0} \end{array} , n \in \text{integer}, \tag{7}$$

where H_L is an operator that extracts LFB components from $\mathbf{C}^l[n]$ and performs normalization; "**0**" represents the zeroing process to HFB components; $a[n]$ and $b[n]$ in Eq. (7) represent the processed LFB and HFB components, respectively. Meanwhile, note that the lengths of both $a[n]$ and $b[n]$ are half of the original cepstral feature stream, $\mathbf{C}^l[n]$, because the down-sampling process is conducted in the DWT procedure.

With the calculated $a[n]$ and $b[n]$ from Eq. (7), IDWT is performed using the designed $\tilde{\phi}[n]$ and $\tilde{\psi}[n]$ from Eq. (6) to obtain the final cepstral feature vectors, $\tilde{\mathbf{C}}^l[n]$:

$$\tilde{\mathbf{C}}^l[n] = \sum_k a[k]\tilde{\phi}_k[n] + \sum_k b[k]\tilde{\psi}_k[t] \tag{8}$$

The CSN process can be considered as a filter-based algorithm because the zeroing process removes components in the high frequency subband, as shown in Eq. (7). Fig. (3) compares the frequency response of CMS, CSN, and RASTA filtering processes. In the CSN procedure, CMS is applied to perform normalization process and thus is denoted by CSN(M) in Fig. (3). From Fig. (3), we can see that CSN(M), conventional CMS and RASTA algorithms remove the DC components while the RASTA and CSN(M) techniques further suppress higher frequency components. The difference between CSN and RASTA lies in that the frequency response of CSN(M) is relatively smooth while the frequency response of RASTA has a zero in the high-half frequency band.

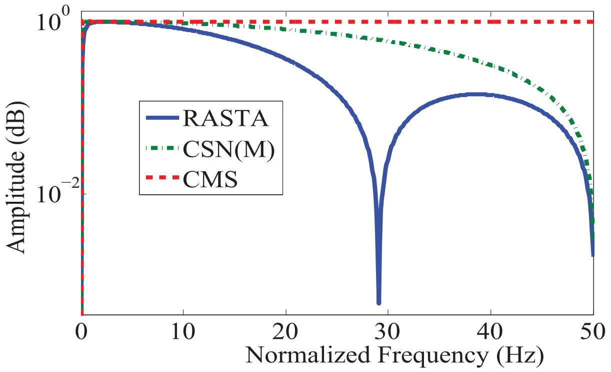

Fig. 3. The frequency response for three robustness techniques, provided that the frame rate is 100 Hz.

4. EXPERIMENT RESULTS AND ANALYSES

In this section, we provide experimental setup, recognition results and discussions.

4.1. Experimental Setup

We conducted the speech recognition experiment on the Aurora-2 task [11], which is a standardized database widely used for evaluating robustness algorithms. Aurora-2 includes three test sets: Test Sets A , B and C. Speech signals in Test Sets A and B are distorted by additive noise (in Set A, the noise types are subway, babble, car, and exhibition; in Set B, the noise types are restaurant, street, airport, and train station), and speech signals in Test Set C are distorted by additive noise and channel effects (subway and street noises together with an MIRS channel mismatch). Each noise instance is added to the clean speech at six SNR levels (ranging from 20 dB to -5 dB). Aurora-2 has two training sets: clean and multi-condition. The clean-condition training set includes 8440 speech utterances, all recorded from a clean condition. The multi-condition training set includes the same 8440 utterances with artificially affected by the same four types of additive noise as those in Test Set A, at different SNRs: 5 dB, 10 dB, 15 dB, 20 dB, and clean condition.

Each utterance in the training or testing sets was first converted into a sequence of Mel-frequency cepstral coefficients, including 13 static components plus their first- and second-order time derivatives. The frame length and frame shift are set to 32 ms and 10 ms, respectively. In addition to MFCC, we test performance with using the AFE technique for a further comparison. All the following experiments are applied on the MFCC or AFE speech features. Besides, hidden Markov model kit (HTK) [13] was adopted for the training and recognition processes. Acoustic models include 11 digit models (zero, one, two, three, four, five, six, seven, eight, nine and oh) with silence and short pause models. Each digit model contains 16 states and 20 Gaussian mixtures per state. Silence and short pause models include three and one states, respectively, both with 36 Gaussian mixtures per state [14].

4.2. Recognition Results

In this study, the recognition performance is evaluated based on word error rate (WER). Results for three different test sets, performed between 0- to 20-dB SNR condition, are reported in the following experiments. An additional "Average" column indicates the average performance over the three sets. Experimental results are presented in two parts. First, we compare CSN with several well-known normalization-based and filter-based robustness algorithms. Next, we investigate the performance of integrating CSN with AFE. Based on Eq. (7), we implement CSN-based CMS and CMVN, denoted as CSN(M) and CSN(M+V), respectively. Note that to compensate the scalars (of DWT bases) normalized by variance normalization, CSN(M+V) conducts an additional scaling process on $a[n]$ before performing IDWT in Eq. (8).

4.2.1. Comparing with Normalization Techniques

Table 1 shows the results of CMS, CMVN, CSN(M), and CSN(M+V) using the clean-condition trained HMM set. The baseline is also listed in the first row.

Table 1. Averaged recognition accuracy and word error rate (%) based on the clean-condition training set.

Set	Set A	Set B	Set C	Average	WER
MFCC baseline	60.70	54.36	72.38	60.50	39.50
CMS	68.29	73.43	69.10	70.51	29.49
CMVN	79.41	80.12	80.71	79.96	20.04
CSN(M)	69.15	74.10	69.99	71.30	28.70
CSN(M+V)	81.09	81.81	82.21	**81.61**	**18.39**

From Table 1, both CSN(M) and CSN(M+V) outperform their conventional counterparts, namely CMS and CMVN, respectively, for the three test sets and the average results. CSN(M+V) achieves the best performance among these four approaches with a significant 53.44% WER reduction (from 39.50% to 18.39%) in average over the baseline result.

In Table 2, recognition results of CSN(M), CSN(M+V), CMS and CMVN are presented using the HMM set prepared by the multi-condition training data. From this table, CSN(M) and CSN(M+V) again outperform CMS and CMVN, respectively. In addition, CSN(M+V) gives the best performance among the four approaches with an average of 25.08% WER reduction (from 9.41% to 7.05%) over the baseline result.

Table 2. Averaged recognition accuracy and word error rate (%) based on the multi-condition training set.

Set	Set A	Set B	Set C	Average	WER
MFCC baseline	91.71	90.14	89.26	90.59	9.41
CMS	92.71	92.55	93.13	92.73	7.27
CMVN	93.13	92.50	92.69	92.79	7.21
CSN(M)	92.93	92.71	93.28	92.91	7.09
CSN(M+V)	93.12	92.67	93.17	**92.95**	**7.05**

4.2.2. *Comparing with Filter-based Techniques*

Next, the proposed CSN approach is compared with filter-based methods, including RASTA and subband feature statistics compensation technique. Here, we conduct the subband CMVN (SB-CMVN) [9] as a representative, because it is confirmed to provide very good performance among the subband feature statistics compensation techniques. Briefly speaking, SB-CMVN first uses a 2-level DWT to split the full-band temporal sequence into four several sub-band sequences; then mean and variance normalization is performed on some or all sub-band sequences; finally IDWT is applied to construct the new full-band sequence. The results of RASTA, SB-CMVN$_{(1,2)}$ (in which the subscript (1,2) indicates that only the first and second lower sub-band sequences, roughly within the ranges $[0, 6.25\text{Hz}]$ and $[6.25\text{Hz}, 12.5\text{Hz}]$, respectively, are processed by MVN), and CSN(M+V) are showed in Table 3. According to the report in [9], SB-CMVN$_{(1,2)}$ gives nearly optimal accuracy compared with the other forms of SB-CMVN. The results for HEQ is also included in this table for comparison.

Table 3. Average recognition results for filter-based techniques on the multi-condition training set.

Set	Set A	Set B	Set C	Average	WER
HEQ	93.04	92.64	92.95	92.86	7.14
RASTA	90.83	90.65	90.97	90.79	9.21
SB-CMVN$_{(1,2)}$	92.30	92.40	92.33	92.35	7.65
CSN(M+V)	93.12	92.67	93.17	**92.95**	**7.05**

From Table 3, CSN(M+V) outperforms HEQ, RASTA, and SB-CMVN$_{(1,2)}$. The results first confirm that CSN(M+V) achieves better performance on noise robustness than HEQ, which serves as a better normalization technique than CMS and CMVN. Next, since one difference between CSN(M+V) and SB-CMVN$_{(1,2)}$ is that CSN(M+V) zeros out the HFB components (roughly corresponding to the sub-band $[25\text{Hz}, 50\text{Hz}]$) while SB-CMVN$_{(1,2)}$ still keeps the sub-band sequences (approximately within the range $[12.5\text{Hz}, 50\text{Hz}]$) unchanged, the better performance achieved implies that the zeroing process in HFB is effective to alleviate noise components. Finally, the results suggest that CSN(M+V) has better noise-suppressed ability than the RASTA filter.

4.2.3. *Integrating with AFE*

Finally, CSN is performed on the AFE features. Table 4 shows the results of AFE and the integrated AFE+CSN.

Table 4. AFE-based averaged recognition accuracy and word error rate (%) on the multi-condition training set.

Set	Set A	Set B	Set C	Average	WER
AFE	94.14	93.35	92.94	93.58	6.42
AFE+CSN(M)	93.98	93.50	93.62	**93.72**	**6.28**
AFE+CSN(M+V)	93.73	93.18	92.98	93.36	6.64

From Table 4, CSN(M) can further improve the recognition performance of AFE, especially for Set C. The overall improvement achieved by CSN is a 2.18% WER reduction over the AFE (from 6.42% to 6.28%). However, CSN(M+V) does not enhance the AFE-preprocessed features to achieve better results. One possible explanation is that AFE has done the noise reduction very well. Further normalizing the variance of features very probably lessens the components corresponding to the difference among various acoustic units, thereby to result in worse recognition accuracy.

In addition to the performance improvements, note that the CSN procedure is simple in computation by following Eq. (7). Moreover, because a down-sampling process is applied in the DWT procedure, CSN provides a 50% reduction on the amount of feature components. These advantages make CSN particularly suitable for mobile applications.

5. CONCLUSION

This paper proposes a novel CSN approach for noise robust speech recognition. CSN combines DWT and normalization processes to suppress noise components in noisy speech signals. The CSN procedure can also be processed easily with reducing 50% amount of the original speech features. The evaluations were conducted on the Aurora-2 task. For the MFCC tests, experimental results show that CSN(M) and CSN(M+V) outperform the conventional CMS and CMVN, respectively. In addition, CSN(M+V) achieves better performance than HEQ, RASTA, and SB-CMVN. For the AFE tests, the recognition results reveals that the integrated AFE+CSN(M) outperforms the original AFE.

6. REFERENCES

[1] O. Viikki, K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," *Acoustics, Speech and Signal Processing*, vol. 2, pp. 733-736, 1998.

[2] H. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," *IEEE Trans. Speech Audio Proc.*, vol. 11, pp. 435-446, 2003.

[3] S. Tibrewala and H. Hermansky, "Multiband and adaptation approaches to robust speech recognition," *in Proc. Eurospeech*, pp. 2619-2622, 1997.

[4] C. W. Hsu and L. S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition," *in Proc. ICASSP*, pp. 197-200, 2004.

[5] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 845-854, 2006.

[6] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578-589, 1994.

[7] J. Yeh and C. Chen, "Noise-robust speech features based on cepstral time coefficients," *Conference on Computational Linguistics and Speech Processing (ROCLING 2009)*, pp. 31-38, 2009.

[8] W. C. Lin, H. T. Fan, and J. W. Hung, "DCT-based processing of dynamic features for robust speech recognition," *in Proc. ISCSLP*, pp. 12-17, 2010

[9] H. T. Fan and J. W. Hung, "Sub-band feature statistics normaliztion techniques based on discrete wavelet transform for robust speech recognition," *in Proc. ICME*, pp. 586-589, 2009.

[10] M. Vetterli and J. Kovaevi, "Wavelets and subband coding," *Prentice-Hall PTR*, 1995.

[11] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ICSA ITRW ASR2000*, 1999.

[12] ETSI,Speech processing, transmission and quality aspects (STQ) "Distributed speech recognition; Advanced front-end feature extraction algorithm," *ETSI standard document ES 202 050*, 2002.

[13] http://htk.eng.cam.ac.uk/

[14] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," *in Proc. ICSLP*, pp. 17-20, 2002.