Temporal Structure Normalization of Speech Feature for Robust Speech Recognition

Xiong Xiao, *Student Member, IEEE*, Eng Siong Chng, *Senior Member, IEEE*, and Haizhou Li, *Senior Member, IEEE*

***Abstract*—This letter presents a new feature normalization technique to normalize the temporal structure of speech features. The temporal structure of the features is partially represented by its power spectral density (PSD). We observed that the PSD of the features varies with the corrupting noise and signal-to-noise ratio. To reduce the PSD variation due to noise, we propose to normalize the PSD of features to a reference function by filtering the features. Experimental results on the AURORA-2 task show that the proposed approach when combined with the mean and variance normalization improves the speech recognition accuracy significantly; the system achieves 69.11% relative error rate reduction over the baseline.**

***Index Terms*—Feature normalization, robust speech recognition, temporal structure, temporal filter.**

I. Introduction

SPEECH recognition performance degrades dramatically when speech is corrupted by background noise and channel distortion. To address the problem, several normalization techniques have been proposed to counter the adverse environmental effects, e.g. cepstral mean normalization (CMN) [1], cepstral variance normalization (CVN) [2], and histogram equalization (HEQ) [3] techniques. The CMN technique removes the mean of speech features in an utterance to reduce the effect of convolutional interference. The CVN technique normalizes the variance of the feature to a fixed range and is usually used together with the CMN to counter noise and channel effects. The HEQ technique equalizes the histogram of the speech feature to a reference function, e.g. Gaussian function, through nonlinear transformations. From a statistical viewpoint, all these methods attempt to normalize the distribution of speech features. The CMN and CVN normalize the first and second order moments of the speech features respectively, while the HEQ normalizes the probability distribution function.

A sequence of speech features can be viewed as a realization of a random process, and can be described in two aspects: one is the statistical distribution of the individual samples of the process, the other is the temporal structure of the process, partially represented by the autocorrelation function or the power spectral density (PSD) function. When speech is corrupted by noise, both its statistical distribution and temporal structure are distorted. Hence, it is desirable to normalize the temporal structure of the features as well.

Recently, several filters have been proposed to smooth the speech feature along the temporal axis. In the MVA [4] technique, the mean subtraction (CMN), variance normalization (CVN) and low--pass autoregressive moving average (ARMA) filtering operations are cascaded for post-processing of features. The MVA technique is motivated by the observation that the PSD of the noisy speech feature usually has a larger high-frequency component than its clean counterpart. Despite its simplicity, the MVA achieves significant performance improvement in the AURORA-2 task. In another approach, a filter is designed for each feature based on certain criteria, such as linear discriminant analysis (LDA), principal component analysis (PCA), and minimum classification error (MCE), to project the speech feature into a subspace for enhanced discriminative ability. These filters are reported to be effective in a connected Chinese digital string task [5].

The above filters are typically designed offline. We believe that better performance can be achieved if we are able to dynamically adapt the filter to the acoustic environment utterance by utterance. To this end, we propose two methods to estimate the reference functions from the clean speech utterances and carry out utterance-based temporal structure normalization (TSN) with respect to the reference functions. The proposed technique is evaluated on the AURORA-2 framework [9].

II. Normalization of Temporal Structure

A. Temporal Structure Varies With Noise

Consider using the Mel-scaled filterbank cepstral coefficients (MFCC) for speech recognition. Let $x_k(n)$ be the cepstral coefficient of the n^{th} frame and k^{th} MFCC channel of an utterance. Hence, we have K time series, $x_1(n)$ to $x_K(n)$, where K is the number of channels. In our experiments, the raw features are the $c0-c12$ cepstral coefficients, delta and acceleration coefficients, thus $K = 39$. The time series of these raw features are first processed by CMN and CVN, referred to as mean and variance normalization (MVN) hereafter, then normalized by the proposed temporal structure normalization (see Fig. 1). The MVN preprocessing removes the DC offset and normalizes the power so that the features are scaled roughly to the same range under different noise conditions. We are interested to study the temporal structure of these time series from the viewpoint of their PSD functions.

Without loss of generality, we examine a speech utterance that is corrupted by additive car noise. Fig. 2(a)–(c) shows the time series of the first MFCC feature $c1$ after MVN processing for three SNR levels. Fig. 2(d) illustrates their corresponding

Manuscript received September 3, 2006; revised November 17, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dominic K. C. Ho.

X. Xiao and E. S. Chng are with the School of Computer Engineering, Nanyang Technological University, Block N4, Singapore 639798 (e-mail: xiao0007@ntu.edu.sg; aseschng@ntu.edu.sg).

H. Li is with the Institute for Infocomm Research, Singapore 119613 and also with the School of Computer Engineering, Nanyang Technological University, Block N4, Singapore 639798 (e-mail: hli@i2r.a-star.edu.sg).

Digital Object Identifier 10.1109/LSP.2006.891341

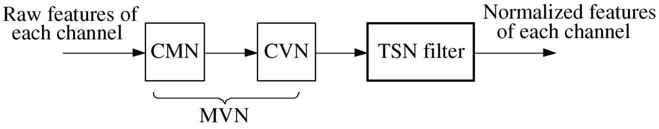
Raw features of each channel
CMN
CVN
TSN filter
Normalized features of each channel
MVN

Fig. 1. Block diagram of the proposed framework.

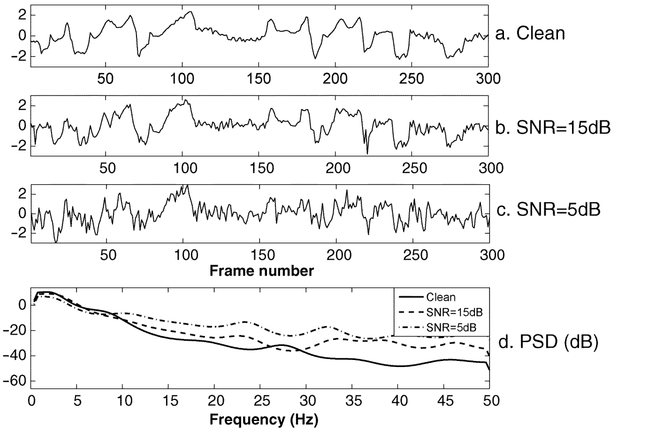

Fig. 2. Time series of MFCC feature $c1$ corrupted by additive car noise after MVN operation in three SNR levels: (a) clean, (b) 15 dB, (c) 5 dB, and (d) their corresponding PSD which is estimated using the Yule-Walker method with model of order 15.

PSD functions. We observe that the PSD functions of the feature for the same utterance under different SNR levels differ due to noise. An interesting observation is that the noisy PSD functions are close to the clean PSD function in the low frequency range (less than 10Hz) but have higher power in the high-frequency range. This explains why the noisy features are more spiky than the clean one as illustrated in Fig. 2(a)–(c). Motivated by the success of other normalization techniques [1]–[3], we propose to normalize the feature PSD to reduce noise effect. It is however not desirable to match the feature PSD functions of every utterance to a single group of reference functions exactly, as that will reduce the discriminative power of the features. We propose to only normalize the trend of the feature PSD functions for each utterance. The effect of normalizing the feature PSD is the same as that of normalizing the temporal structure of the feature time series. The task of normalizing the PSD can be divided into two subtasks, one subtask is to find a reference PSD function, and the other subtask is to filter the speech feature to equalize the feature PSD to the reference function. We call the filter for this purpose the temporal structure normalization (TSN) filter (Fig. 1).

B. Normalizing Temporal Structure Using TSN Filters

We first introduce the proposed filtering technique in this section, and then discuss the training of the reference functions in the next section. Given a reference PSD function for each channel, we can find the required filter transfer functions of the TSN filters in the frequency domain for each utterance, and then estimate the FIR filters' weights using the windowed method [8]. Let $X_k(\omega)$ and $Y_k(\omega)$ denote the two-sided utterance-dependent PSD and reference PSD for the k^{th} channel respectively, where ω is the normalized angular frequency. To equalize the noisy PSD $X_k(\omega)$ to the reference PSD $Y_k(\omega)$, the required magnitude response of the filter is

$$|H_k(\omega)| = \sqrt{Y_k(\omega)/X_k(\omega)}, \quad \omega \in [-\pi, \pi]. \tag{1}$$

Using the above magnitude response, the PSD function after filtering will be $\hat{X}_k(\omega) = |H_k(\omega)|^2 X_k(\omega) = Y_k(\omega)$. The filter's weights can be found by

$$w_k(l) = \text{IDFT}\left(|H_k(\omega)|\right), \quad l = 1, \ldots, L \tag{2}$$

where $\text{IDFT}(\cdot)$ represents the inverse discrete Fourier transform and L is the number of weights. The weights $w_k(l)$ are symmetric w.r.t. the center tap and the most significant weights are concentrated around this tap. Only the central weights are extracted to form the filter of a desired length. By not using all the weights, the realized filter's transfer function is smoothed. The number of weights to use is chosen such that the filter will mainly normalize the trend of the PSD function and leaves the detailed speech information less affected. A Hanning window is applied to the truncated weights to reduce truncation effect. Finally, the sum of the windowed weights are normalized to one to ensure that features are properly scaled. The filters for all channels thus designed have identical linear phase response.

C. Finding the Reference PSD Functions

As each MFCC channel possesses an unique temporal characteristic, we design a reference PSD function for each channel. Each reference PSD function is trained using the same clean speech data as that for the acoustic models training. As the feature PSD of each utterance is partially dependent on both speech content and speaker, we propose to average the PSD over a collection of utterances to produce for each channel a general reference function. Similar PSD averaging scheme has been used by Nadeu *et al.* [6] to analyze the properties of speech features. The averaging process reduces speaker effect that is detrimental for speaker-independent speech recognition. The averaging process also smoothes out the speech detail but retains the common trend of the clean feature PSD. The obtained reference functions are found to be suitable for the design of the TSN filter in Section II-B.

In addition, we also consider to perform a low-pass filtering operation on the features before training the reference functions. This is motivated by the observation that smoothing the speech features improves recognition performance [4], [5]. Our experimental results show that TSN using reference functions with additional low-pass filtering produces slightly better performance for the AURORA-2 task.

III. Experiments and Results

A. Experiment Settings

The sampling rate of the feature time series is the same as the frame rate which is 100 Hz. The autoregressive model used in the Yule-Walker PSD estimation method has an order of 15 to obtain the desired smoothness for the PSD. The number of bins for the two-sided PSD is 256 to ensure sufficient samples in the

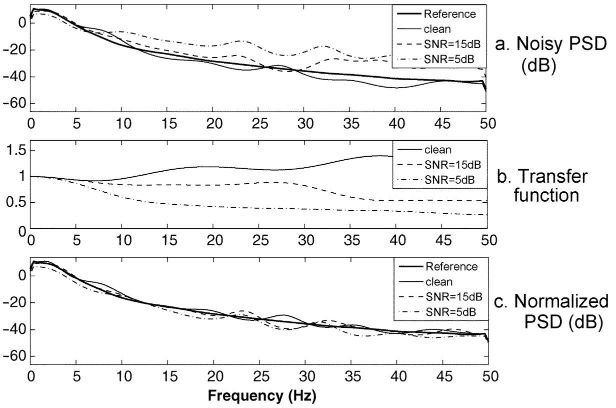

Fig. 3. Effect of Scheme A on the feature PSD of Fig. 2(d).

frequency domain. As described in Section II-C, we propose two reference function training schemes, Scheme A and B. They are summarized as follows.

- Scheme A: $features \rightarrow \mathrm{MVN} \rightarrow \mathrm{PSD\ estimation} \rightarrow \mathrm{Averaging} \rightarrow Scheme\ A's\ ref.\ functions$
- Scheme B: $features \rightarrow \mathrm{MVN} \rightarrow \mathrm{ARMA}\ filter \rightarrow \mathrm{PSD}\ estimation \rightarrow \mathrm{Averaging} \rightarrow Scheme\ B's\ ref.\ functions$

The ARMA filter in Scheme B is the same as that in MVA with its order M empirically set to be 3. The TSN filter's length is empirically set to 21. We use the same 1000 utterances to train the reference functions of the two schemes. The utterances are extracted from the clean training set of AURORA-2 [9] with selection based on gender and speaker balance.

B. Normalization Effect

1) TSN With Scheme A Reference Functions:

Fig. 3(a) shows the same PSD as that in Fig. 2(d) along with the reference PSD function. Fig. 3(b) illustrates the magnitude response of the TSN filters for the different SNR levels. For the clean case, the magnitude response is slightly high-pass while for the $\mathrm{SNR} = 15$ dB and 5 dB cases, the filters are both low-pass. Fig. 3(c) shows that the normalized PSD functions are very similar to the reference PSD function.

Fig. 4(b)–(d) compare the normalized version of the time series of Fig. 2(a)–(c) with their original clean version in Fig. 4(a). It is observed that the smoothness of the filtered features in different SNR levels are similar. In particular, the normalized clean features are almost identicial to the original clean features. This indicates that the filtering process does not alter the clean features adversely. On the other hand, the normalized $\mathrm{SNR} = 15\ \mathrm{dB}$ and 5 dB features are much smoother than their original features. This shows that the TSN filters the features of different SNRs differently.

From the magnitude responses in Fig. 3(b), under noisy cases the filters attenuate high-frequency components that are mainly due to noises while preserving low frequency speech information. This agrees with the findings by Kanedera *et al.* [7] where it is reported that the most important modulation frequencies of the features for recognition are around 4 Hz.

2) TSN With Scheme B Reference Functions: Fig. 5, 6 show the normalization effect using the same configurations as those of Figs. 3 and 4 but using the Scheme B's reference functions. As a result of the additional low-pass ARMA filtering, the reference PSD function of Scheme B Fig. 5(a) has a lower power density in high frequency than that of Scheme A Fig. 3(a). The filters in Fig. 5(b) for all three SNR levels are low-pass filters, with the gain for high-frequency decreasing with SNR level. Like the previous example, the normalized PSD for all three SNR levels in Fig. 5(c) are very similar to the reference PSD. Fig. 6(b)–(d) shows the features after normalization. As compared to Fig. 4(b)–(d), all the features are smoother, including the clean one. This is attributed to the smoothing of MVA in

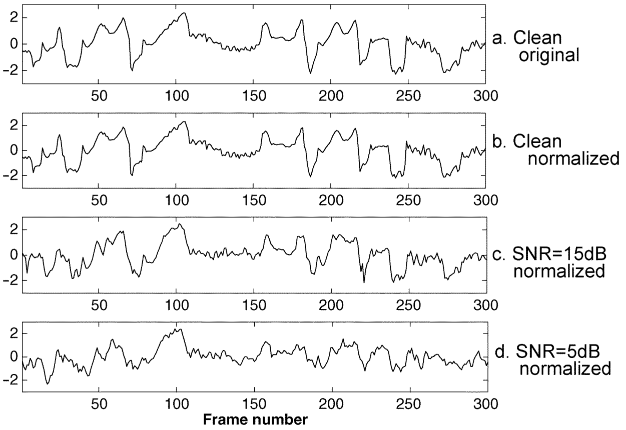

Fig. 4. Effect of Scheme A on the speech features of Fig. 2(a)–(c).

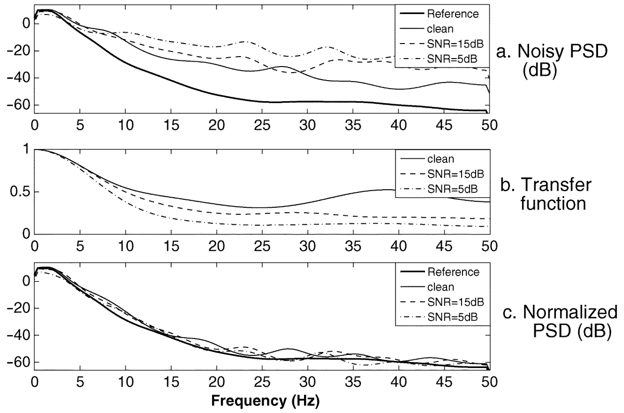

Fig. 5. Effect of scheme B on the feature PSD of Fig. 2(d).

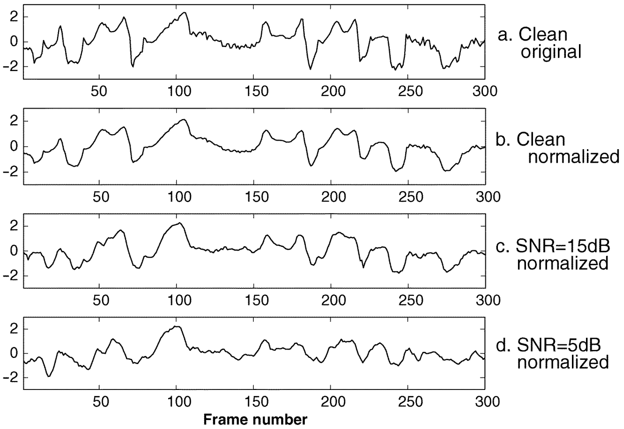

Fig. 6. Effect of scheme B on the speech features of Fig. 2(a)–(c).

TABLE I
ACCURACY (%) FOR AURORA-2 TASK AVERAGED ACROSS THE SNR BETWEEN 0 AND 20 dB. AR (%) AND RR (%) ARE THE ABSOLUTE AND RELATIVE ERROR RATE REDUCTIONS OVER THE BASELINE

Method	**Set A**	**Set B**	**Set C**	**Avg.**	**AR**	**RR**
Baseline	53.17	47.89	63.05	53.03	-	-
MVN	77.91	79.48	77.70	78.49	25.46	54.20
MVA	84.18	85.16	84.28	84.59	31.56	67.19
TSN+A	84.27	85.87	83.62	**84.78**	31.75	67.60
TSN+B	84.72	86.59	84.80	**85.49**	32.46	69.11

TABLE II
ACCURACY (%) FOR AURORA-2 TASK ACHIEVED BY MVA AND $\mathrm{TSN}+\mathrm{B}$ FOR EACH SNR LEVEL AVERAGED ACROSS TEN NOISE CASES. AR (%) AND RR (%) ARE THE ABSOLUTE AND RELATIVE ERROR RATE REDUCTIONS OVER MVA

Method	**Clean**	**20dB**	**15dB**	**10dB**	**5dB**	**0dB**	**-5dB**
MVA	99.10	97.81	95.95	91.38	80.43	57.39	27.09
TSN+B	99.26	97.93	96.13	92.06	81.76	59.56	28.13
AR	0.16	0.12	0.18	0.68	1.33	2.17	1.04
RR	17.59	5.52	4.30	7.88	6.81	5.11	1.44

Scheme B training of the reference functions. After normalization, the smoothness of the speech features in different SNR levels become more similar.

C. Recognition Results

Now we proceed with the performance evaluation in the AURORA-2 framework [9]. We follow the scripts provided by the AURORA-2 framework for training and testing of the recognition engine, except that the cepstral energy $c0$ is used, rather than the log energy. In all the experiments, the 13 MFCC features, $c0 - c12$, together with their delta and acceleration features are generated prior to any post-processing. After these 39 MFCC features are generated for each utterance, a post-processing step is applied on each of the 39 MFCC channels. We evaluate four different post-processing schemes.

a) MVN: $\mathrm{CMN} + \mathrm{CVN}$.
b) MVA: $\mathrm{MVN} + \mathrm{ARMA}$ filtering.
c) $\mathrm{TSN} + \mathrm{A}$: $\mathrm{MVN} + \mathrm{TSN}$ with Scheme A ref. functions.
d) $\mathrm{TSN} + B$: $\mathrm{MVN} + \mathrm{TSN}$ with Scheme B ref. functions;

The experimental results are summarized in Tables I and II. In Table I, the MVN result shows that the normalization of the first and second order moments of the feature improves the accuracy significantly over the baseline with absolute error rate reduction of 25.46%. The MVA further improves the performance by filtering out some feature variations between the clean and noisy features. The $\mathrm{TSN} + \mathrm{A}$ and $\mathrm{TSN} + \mathrm{B}$ further improve the performance.

In Table II, we compare the performance of MVA to $\mathrm{TSN}+\mathrm{B}$. It is observed that $\mathrm{TSN}+\mathrm{B}$ outperforms MVA in all SNR levels, especially in low SNR cases with absolute error rate reduction of 2.17% at 0 dB in particular. The improvement is also observed in clean case which can be attributed to the speaker normalization effect of reference functions. We also find that $\mathrm{TSN} + \mathrm{B}$ outperforms $\mathrm{TSN} + \mathrm{A}$ in general. This suggests that, besides temporal structure normalization, proper smoothing to both clean and noisy features are beneficial for speech recognition, which agrees with the findings in other recent reports [4]–[7].

D. Discussion

The MVA method [4] and the data-driven filters [5] smooth the features without explicitly normalizing its temporal structure. The utterance-based TSN introduces the normalization of the feature PSD as the objective for temporal filtering under different noise and SNR situations. Apparently a fixed filter would not be optimal for varying acoustic environment. In this letter, we propose adapting filters utterance by utterance. Experimental results show that our method results in different degree of smoothing for different noise and SNR situations and improves the accuracy across all SNR levels over reported results in [4].

We note that TSN introduces extra computational cost by involving the IDFT and Yule-Walker PSD estimation operations. Both operation however can be implemented by efficient algorithms, such as the fast Fourier transform (FFT) and the Levinson-Durbin recursion [8], respectively.

IV. CONCLUSION

In this letter, we studied the TSN filter to normalize the speech feature's temporal structure explicitly in the form of PSD utterance by utterance. The TSN method filters the speech features aiming to equalize the feature PSD functions of an utterance to the desired reference functions. Experimental results show that TSN improves the recognition accuracy in both clean and noisy cases for the AURORA-2 task and its improvement is greater than that of MVA [4] due to the adaptive filtering mechanism.

REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[2] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in *Proc. ICASSP*, 1998, vol. 11, pp. 733–736.

[3] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005.

[4] C.-P. Chen, J. Blimes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on AURORA 2.0," in *Proc. ICSLP*, 2002, pp. 2445–2448.

[5] J.-W. Hung and L.-S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 808–832, May 2006.

[6] C. Nadeu, P. Paches-Leal, and B.-H. Juang, "Filtering the time sequences of spectral parameters for speech recognition," *Speech Commun.*, vol. 22, pp. 315–332, Sep. 1977.

[7] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Commun.*, vol. 28, pp. 43–55, 1999.

[8] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing-Principles, Algorithms, and Applications*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[9] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recogntion systems under noisy conditions," in *Proc. ICSLP*, 2000, pp. 29–32.