Evaluating the Temporal Structure Normalisation Technique on the Aurora-4 Task

Xiong Xiao[1,2], *Eng Siong Chng*[1], *Haizhou Li*[1,2]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Institute for Infocomm Research, Singapore
xiao0007@ntu.edu.sg, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

10.21437/Interspeech.2007-106

Abstract

We evaluate the temporal structure normalisation (TSN), a feature normalisation technique for robust speech recognition, on the large vocabulary Aurora-4 task. The TSN technique operates by normalising the trend of the feature's power spectral density (PSD) function to a reference function using finite impulse response (FIR) filters. The features are the cepstral coefficients and the normalisation procedure is performed on every cepstral channel of each utterance. Experimental results show that the TSN reduces the average word error rate (WER) by 7.20% and 8.16% relatively over the mean-variance normalisation (MVN) and the histogram equalisation (HEQ) baselines respectively. We further evaluate two other state-of-the-art temporal filters. Experimental results show that among the three evaluated temporal filters, the TSN filter performs the best. Lastly, our results also demonstrates that fixed smoothing filters are less effective on Aurora-4 task than on Aurora-2 task.

Index Terms: robust automatic speech recognition, feature normalisation, temporal filter, Aurora task

1. Introduction

The current state-of-the-art speech recognition system is a statistical pattern classification machine. If there is a mismatch between the statistics of the training and test speech features, the system performance can be seriously affected. For example, if the microphone used to record the training and test utterances are different, the resultant mean of the cepstral features may differ. In this case, mean normalisation can be used to minimise the feature mismatch. Many methods have been proposed to normalise other global statistics of the features. If we view the time series of each cepstral feature, also known as the cepstral channel, as realisations of a random process, we can normalise the process's first-order, second-order, and higher-order statistics [1] to improve the robustness of feature. Existing normalisation techniques can be classified into two categories: the first-order statistics normalisation (FOSN) and the second-order statistics normalisation (SOSN). To the authors' best knowledge, there is still no reported work that normalises the higher order statistics for robust speech recognition.

Techniques that normalise either the moments or the whole probability density function (PDF) of the feature belong to the FOSN category. Examples of these techniques include, the cepstral mean normalisation (CMN) [2], the cepstral variance normalisation (CVN) [3], the third-order normalisation (CTN) [4], the higher order cepstral moment normalisation (HOCMN) [5, 6] and the histogram equalisation (HEQ) technique [7, 8]. The HEQ technique which normalises the histogram of the feature, i.e. the PDF of the feature, in theory normalises all the moments of the feature. In general, by normalising more moments, better recognition performance is achieved.

The SOSN techniques on the other hand aims to normalise the inter-frame relationship (i.e., temporal structure) of the feature. Usually filters are used to realise the normalisation. The inter-frame relationship can be examined through the feature's autocorrelation function (ACF) or PSD function. In our previous work, we proposed a TSN technique [9, 10] which normalises the feature's PSD function to a reference function by filtering the feature time series. We reported significant improvement to the recognition accuracy over the mean and variance normalisation (MVN) baseline on the Aurora-2 task.

Another SOSN filter is the temporal smoothing filter (TES) [11]. Although the TES and TSN share a similar motivation, their implementations are quite different. The TES is a group of autoregressive moving average (ARMA) filters designed from the feature's ACF, while the TSN is a group of FIR filters designed from the feature's PSD functions.

As the FOSN and SOSN techniques normalise different statistics of the features, they are hence complementary and can be applied jointly. Our experimental results suggest that the combination of the two usually yields better performance.

Other researchers have also reported on the use of temporal filters to filter the feature time series [12, 14, 15]. However, as compared to SOSN techniques, these filters are not designed with the objective to normalise the feature's temporal structure. Instead they are designed based on either empirical observations [12, 15] or some data driven approaches to optimise the discriminative power of the features [14]. For example, the RASTA filter [12] is empirically designed to filter out the very low modulation frequency (0-1Hz) and the high modulation frequency (16-50Hz, assume the frame rate is 100Hz), as these two portions of the modulation spectrum are considered as less relevant to the speech recognition problem [13]. Another empirically designed temporal filter is an ARMA filter that operates after the MVN operation and is known as the MVA [15]. The MVA improves the recognition accuracy significantly on the Aurora-2 task. In [14], several temporal filters are designed based on criteria such as linear discriminative analysis (LDA), the principle component analysis (PCA) and the minimum classification error analysis (MCE). Filters designed using these criteria are usually band-pass or low-pass and experimental results on a Chinese digital string task showed good performance.

In this paper, we evaluated the TSN technique on the Aurora-4 large vocabulary task and compared the performance of the TSN, MVA and RASTA filters in combination with the MVN or HEQ preprocessing. In section 2, we briefly introduce the TSN filter and the HEQ technique. In section 3, we describe the experimental setup and present the results and discussion.

Finally, we conclude in section 4.

2. Normalising the Statistics of the Features

This section discusses the TSN and HEQ techniques.

2.1. Temporal Structure Normalisation

Our analysis showed that the PSD function of the feature time series varies with the acoustic environment conditions such as changes in signal-to-noise ratio (SNR) [9]. To reduce the effect of acoustic distortion, the TSN technique normalises the feature's PSD function to a reference PSD function. Note that the PSD function is not normalised exactly to the reference function, rather only the trend of the PSD function is normalised. By doing this, only the variation due to the environmental distortion will be reduced [9] while the speech related information will be retained.

FIR filters are used to normalise the feature's PSD functions. For each feature channel of an incoming utterance, the transfer function of the FIR filter is obtained by

$$|H(\omega)| = \sqrt{P_{ref}/P_{test}} \quad (1)$$

where P_{ref} is the reference PSD function and P_{test} is the PSD function of the feature to be processed, both PSD functions are two-sided. The reference PSD functions are found by averaging the feature's PSD function across clean utterances. Equation (1) is shown to be the transfer function of the square-root Wiener filter under certain assumptions [10]. The filter's coefficients can be found by taking the inverse fast Fourier transform (IFFT) of $|H(\omega)|$. The FIR filter so designed will normalise the P_{test} to P_{ref} exactly, as the PSD of the filtered feature will be $P_{out} = |H(\omega)|^2 P_{test} = P_{ref}$. If we wish only to normalise the trend of the P_{test}, a smoothed version of the $|H(\omega)|$ should be applied instead. This can be achieved by extracting only the middle part of the filter's coefficients which have the most significant weights. This truncation of filter's coefficients results in smoothed transfer function for the realised filter. Hence, only the overall trend of the feature's PSD function is normalised. After the truncation, the Hanning window is applied on the coefficients and finally, the sum of the coefficients are scaled to one to ensure unit gain at zero frequency. For a more complete description of the filter, please refer to [9, 10].

2.2. Histogram Equalisation

The HEQ [7, 8] is a general technique to transform the histogram of the input data to any desired histogram. As the environmental distortion is one important factor that contributes to the variations of the feature's histogram, the HEQ is used to normalise the feature's histogram to reduce the effect of the distortion. When the histogram, i.e. the PDF of the feature, is normalised, all the moments of the feature are normalised theoretically. Therefore, the HEQ can be seen as a generalisation of the MVN that only normalises the first two moments of the feature. Experimental results have also shown that the HEQ provides better performance than the MVN [7].

For completeness, the HEQ technique is briefly described. Given both the histogram of the input data and the reference histogram, the transformation that normalises the input histogram to the reference histogram is:

$$\hat{x} = C_{ref}^{-1}(C_{test}(x)) \quad (2)$$

where the $\hat{x}$ and x are the output and input features, respectively. The $C_{ref}()$ and $C_{test}()$ are the reference and input cumulative density functions (CDF) respectively, which can be obtained from their corresponding histograms.

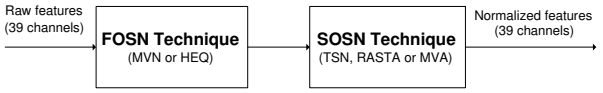

Figure 1: The block diagram of the proposed feature normalisation framework

To implement the HEQ, we need to have both the input histogram and the reference histogram. The input histogram is estimated from the incoming utterance, either from a segment of several seconds [8] or from the whole utterance [7]. The reference histogram can be either sampled from a standard PDF, e.g., the Gaussian PDF, or estimated by averaging the histogram over some clean training utterances.

3. Experiments

The general framework of our experiment for feature normalisation is to cascade the FOSN techniques and the SOSN techniques as shown in Fig. 1. We conducted experiments using two FOSN techniques, i.e., the MVN and the HEQ, and three SOSN techniques, i.e., the RASTA, the MVA and the TSN.

3.1. Speech Recognition System

In the Aurora-4 database [16], speech is sampled in both 8kHz and 16kHz, of which only 8kHz speech is used in this paper. We use the HTK speech recognition toolkit (HTK) [17] to build a speech recognition system with cross-word triphone hidden Markov models. In every triphone model, there are three emitting states, each with 4 mixtures. The recognition system is trained from 7,138 clean utterances (Training Set 1 described in [16]) and the training process follows that of the ISIP baseline system (page 9 of [16]). For all experiments, there are around 3,150 tied states, which is similar to the number used in the ISIP baseline system. The standard bigram language model from the Wall Street Journal (WSJ01) database is modified to include several missing words in the training and testing data using HTK. For all recognition tests, the weight of the language model in final likelihood, the word insertion penalty and the pruning threshold are set to 16, -10 and 250, respectively.

There are 14 test cases in the Aurora-4 task. The test case 01 is the clean test and the test cases 02-07 are noisy tests, each case is corrupted by one type of noise. The utterances in each noisy test case are corrupted in different SNR levels from 15dB to 5dB with an average SNR of 10dB. The test cases 08-14 are the same as the test cases 01-07 except that the microphone mismatch is introduced. For all the 14 test cases, there are a big test (330 utterances) and a small test (166 utterances). Although the small test is a subset of the big test, its word count distribution and utterance duration distribution match closely to the big test. To save time, our experiments are all carried out on the small test. The total length of the speech in one small test case is 20 mins. The computation time required for the small test is about 1x-2x real time for clean test case (test 01) and 3x-8x real time for noisy test cases (test 02-14) on a Pentium IV 2.8 GHz PC.

3.2. Feature Normalisation Techniques

The raw features are generated from the WI007 feature extraction program that is distributed with the Aurora-2 database [18]. The 13 static features and their delta and acceleration features

Table 1: WER (%) for AURORA-2 task averaged across the SNR between 0 and 20 dB. RR (%) is the relative error rate reduction over the MVN baseline

Method	Set A	Set B	Set C	Avg.	RR
MVN baseline	22.09	20.52	22.3	21.51	-
MVN+RASTA	18.94	17.31	18.29	18.16	15.57
MVN+MVA	15.82	14.84	15.72	15.41	28.36
MVN+TSN	15.73	14.13	16.38	**15.22**	29.24

form the 39 dimension feature vector. The first MFCC feature c0 is used instead of the log energy. After the 39 features are generated, they are preprocessed by either MVN or HEQ, then further processed by the temporal filters: the RASTA, the MVA and the TSN filters. All processing is performed on each feature individually. The following test setups are examined:

- MVN baseline: utterance-based MVN
- HEQ baseline: utterance-based HEQ
- MVN+RASTA: MVN followed by RASTA filter
- MVN+MVA: MVN followed by ARMA filter
- MVN+TSN: MVN followed by TSN filter
- HEQ+RASTA: HEQ followed by RASTA filter
- HEQ+MVA: HEQ followed by ARMA filter
- HEQ+TSN: HEQ followed by TSN filter
- AFE [19], advanced feature extraction program of ESTI

The MVN and HEQ are implemented utterance by utterance. The number of bins for the histogram used in the HEQ is 100 and the reference histogram is sampled from the standard Gaussian PDF. The RASTA filter is implemented using the equation (1) in [12], with the pole value set to 0.94 for better performance. We find that ARMA filters in MVN+MVA and HEQ+MVA work the best when their order are two and one respectively. The ESTI's advanced feature extraction program (AFE) [19] is also evaluated for comparison.

In the TSN, the feature's PSD function is estimated using the Yule-Walker method with the autoregressive model of order 15 to obtain the desired smoothness of PSD. The number of bins for the two sided PSD is 256 to ensure sufficient samples in the frequency domain. The reference PSD functions, one for each feature channel, are averaged over the whole 7,138 clean training utterances of the Aurora-4. The raw features of each utterance are processed by either the MVN or the HEQ before the calculation of their PSD functions and the filtering. The non-causal FIR filter's order is chosen to be 21, which is the same as that used for the Aurora-2 task [9][10]

3.3. Experiment results

For ease of comparison, we extract our previously published results [10] in Table 1 with MVN preprocessing on the Aurora-2 task. The results (see the Avg. column) show that MVN+TSN and MVN+MVA resulted in very similar improvements over the baseline. The improvements were 28.36% and 29.24% respectively.

The average WERs with MVN and HEQ preprocessing on the Aurora-4 task are shown in Fig. 2 and the detailed results are shown in Table 2. The RASTA and MVA filters improve the performance over the baselines slightly. Only the TSN filter provides significant improvements over their corresponding

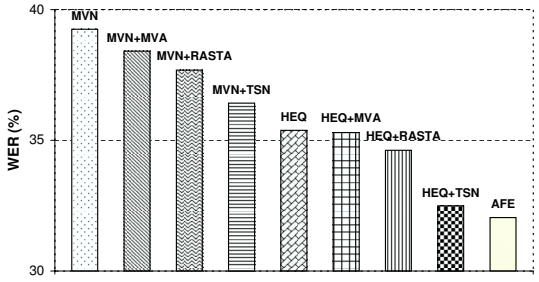

Figure 2: The average WER of different feature normalisation schemes on the Aurora-4 task.

baselines, specifically the MVN+TSN and the HEQ+TSN reduce the average WER by 7.20% and 8.16% respectively. As shown in Table 2, the use of the RASTA and MVA filters produce worse results than the baselines for the clean test 01. However, the TSN filter delivers comparable or even better results than the baselines for test 01. The results show that the TSN is complementary to MVN or HEQ. The best combination was achieved using the HEQ+TSN scheme. Its average result of 32.49% is comparable to the AFE results of 32.05%.

3.4. Discussion

We note that the effect of the environmental distortion can be reduced by low-pass filtering the features. However, some speech information that exists in the high modulation frequency ranges will also be lost. For a small vocabulary task such as the Aurora-2 (11 word models), the low-pass filtered features remain sufficiently discriminative even though the high modulation frequencies are totally smoothed out. However, for large vocabulary task such as the Aurora-4, when the number of models become significantly larger (>3000 tied states), the features must not be overly smoothed so as to retain their discriminative ability to achieve good recognition performance. The basic consideration in smoothing the features is to have a balance between reducing feature mismatch due to environmental distortion and the feature's discriminative power.

The TSN achieves a good balance by adapting its level of smoothing to different distortion situations. For example, under high SNR level, mild smoothing is used to retain more speech details. While for low SNR level utterances, more aggressive smoothing is applied. Readers are referred to [9][10] for an illustration of the adaptive transfer functions in different SNR levels. In summary, we find that the TSN is effective in feature normalisation. It benefits from the adaptive ability in different SNR levels.

Unlike the adaptive filtering used in the TSN, the RASTA and MVA use fixed low-pass filters (RASTA also suppresses the very low modulation frequency). Hence, these two techniques cannot achieve the same level of performance as TSN on a large vocabulary task (Table 2).

4. Conclusions

In this paper, we compared several start-of-the-art temporal filters for robust speech recognition. Our experimental results show while these filters can all improve the recognition performance on small vocabulary Aurora-2 task, only the TSN remains effective on the large vocabulary Aurora-4 task. The results reaffirm that temporal feature normalisation - such as the

Table 2. The WERs of different feature normalisation schemes on the Aurora-4 Task. RR (%) is the relative error rate reduction of the temporal filters over their corresponding baselines. The result of the HEQ*+TES is taken from [8] directly as we didn't implement the TES filter. This result cannot be compared directly to that of the HEQ+TSN as different types of HEQ are used in HEQ+TSN and HEQ*+TES and settings are also different in the recognition systems.

Method	01	02	03	04	05	06	07	08	09	10	11	12	13	14	Avg.	RR
MVN baseline	13.59	25.30	38.31	39.34	40.04	37.38	42.73	22.32	35.17	49.02	52.30	55.25	48.51	50.24	39.25	-
MVN+MVA	15.14	27.11	37.68	40.74	38.27	39.89	39.78	22.50	33.04	45.38	50.67	51.23	48.91	47.48	38.41	2.14%
MVN+RASTA	15.76	24.16	36.80	39.89	37.90	36.80	40.63	22.80	33.26	45.89	48.66	50.83	45.38	48.84	37.68	3.99%
MVN+TSN	14.03	25.60	33.81	36.21	36.76	34.14	38.82	22.36	34.73	43.35	47.66	50.87	43.65	47.96	36.42	7.20%
HEQ baseline	13.41	22.39	34.29	35.84	35.29	34.18	37.42	20.52	30.83	43.43	47.26	49.65	43.61	47.22	35.38	-
HEQ+MVA	14.44	21.99	33.96	37.38	35.40	35.40	36.32	21.80	29.57	41.62	48.95	47.78	44.64	45.01	35.29	0.24%
HEQ+RASTA	14.92	22.69	31.79	37.13	33.59	34.07	36.46	21.47	30.46	41.47	45.89	47.26	42.14	45.30	34.62	2.16%
HEQ+TSN	12.85	20.85	30.76	32.45	32.38	32.41	35.14	20.96	29.36	38.71	42.14	45.41	39.63	41.84	32.49	8.16%
AFE	14.51	19.59	30.09	36.17	31.23	34.73	32.97	23.35	26.08	37.83	43.39	41.51	40.41	36.80	32.05	-
HEQ*+TES	11.42	21.40	35.25	37.53	34.59	36.17	38.56	20.15	28.80	43.87	47.66	49.83	44.75	46.77	35.48	-

TSN - is effective for robust speech recognition, especially on large vocabulary tasks. In the future, we would like to investigate SNR-dependent temporal filtering scheme, in which temporal filters are designed and applied according to SNR levels.

5. References

[1] A. Papoulis and S. U. Pillai, "Probability, Random Variables and Stochastic Processes", 4th Ed., McGraw-Hill Higher Education, 2002

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

[3] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise", *In Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1998, vol. 11, pp. 733-736

[4] Y. H. Suk, S. H. Choi and H. S. Lee, "Cepstrum third-order normalisation method for noisy speech recognition", *IEE Electronics letter*, vol. 35, no. 7, pp. 527-528, 1999

[5] C.-W. Hsu and L.-S. Lee, "Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition", *In Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. I, pp. 197-200, Montreal, Canada, May 17-21, 2004

[6] C.-W. Hsu and L.-S. Lee, "Extension and Further Analysis of Higher Order Cepstral Moment Normalization (HOCMN) for Robust Features in Speech Recognition", *In Proceedings of InterSpeech 2006*, pp. 369-372, Sep. 17-21, Pittsburgh, USA.

[7] 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, 2005

[8] F. Hilger and H. Ney, "Quantile Based Histogram Equalization for noise Robust Large Vocabulary Speech Recognition", *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no 3, pp. 845-854, 2006

[9] X. Xiao, E. S. Chng, and H. Li, "Temporal Structure Normalization of Speech Feature for Robust Speech Recognition", to appear in *IEEE Signal Processing Letters*.

[10] X. Xiao, E. S. Chng, and H. Li, "Normalizing the Speech Modulation spectrum for robust speech recognition", *to appear in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Honolulu, US, 2007

[11] L. García , J.C. Segura , C. Benítez, J. Ramírez, A. de la Torre, "Normalization of the inter-frame information using smoothing filtering", *In Proceedings of InterSpeech 2006*, pp. 369-372, Sep. 17-21, Pittsburgh, USA.

[12] H. Hermansky and N. Morgan, "RASTA processing of speech", *IEEE Trans. Speech, Audio Proceess.*, vol. 2, no. 4, pp. 578-589, 1994

[13] 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 Commnunication*, vol. 28, No. 1, pp. 43-55, 1999

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

[15] C.-P. Chen and J. A. Bilmes, "MVA Processing of Speech Features", *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no 1, pp. 257-270, Jan 2007

[16] N. Parihar and J. Picone, "Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02", Institute for Signal and Infomation Processing, Mississippi State Univ.

[17] Cambridge Univ. Engineering Dept., "HTK speech recognition toolkit", available online: http://htk.eng.cam.ac.uk/

[18] 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

[19] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms", ETSI ES 202 050 Recommendation, 2002.