A RECURSIVE FEATURE VECTOR NORMALIZATION APPROACH FOR ROBUST SPEECH RECOGNITION IN NOISE

Olli Viikki[1], David Bye[2], Kari Laurila[1]

[1]Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
[2]Nokia Mobile Phones R&D, Camberley, Surrey, UK
Email: {olli.viikki, kari.laurila}@research.nokia.fi, david.bye@nmp.nokia.com

ABSTRACT

The acoustic mismatch between testing and training conditions is known to severely degrade the performance of speech recognition systems. Segmental feature vector normalization [8] was found to improve the noise robustness of MFCC feature vectors and to outperform other state-of-the-art noise compensation techniques in speaker-dependent recognition. The objective of feature vector normalization is to provide environment-independent parameter statistics in all noise conditions. In this paper, we propose a more efficient implementation approach for feature vector normalization where the normalization coefficients are computed in a recursive way. Speaker-dependent recognition experiments show that the recursive normalization approach obtains over 60%, the segmental method approx. 50%, and Parallel Model Combination 14% overall error rate reduction, respectively. Moreover, in the recursive case, this performance gain is obtained with the smallest implementation costs. Also in speaker-independent connected digit recognition, over 16% error rate reduction is obtained with the proposed feature vector normalization approach.

1. INTRODUCTION

The development of noise robust speech recognition algorithms is becoming increasingly important as speech technology is currently widely applied to real world applications. Although noise robustness has recently gained a great deal of interest in speech recognition research, practical speech recognition systems still provide a moderate performance even in simple recognition tasks if the testing and training environments do not match each other acoustically.

Several techniques have been proposed for reducing the mismatch between the testing and training environments. Many of these methods operate either in spectral [5], or in cepstral domain [6]. In addition to various normalization approaches, noise robust feature extraction techniques, such as the RASTA method [3], have also been developed. The mismatch effects can also be compensated in the recognition unit, as done in Parallel Model Combination (PMC) [2]. Most noise compensation methods require a good noise estimate in order to work properly. To compute a reliable noise estimate, an accurate Voice Activity Detector (VAD) is needed. Since the accuracy of VADs is poor in noisy environments, many compensation techniques tend to fail in adverse conditions.

In [8], we presented a segmental mean and variance compensation approach for feature vectors. All feature vectors were normalized to have the same segmental parameters statistics in all noise conditions. This joint mean and variance compensation was found to improve significantly the robustness of the Mel-Frequency Cepstral Coefficients (MFCC) against various additive noise types and microphone mismatch. Speaker-dependent recognition experiments showed that in terms of recognition rate segmental feature vector normlization was superior to other state-of-the-art noise compensation methods. Here, we describe a recursive implementation approach for the previously presented segmental normalization method. Using the recursive normalization approach we can implement feature vector normalization more efficiently without compromising in recognition performance.

The viability of the proposed recursive normalization method is tested in speaker-dependent isolated name recognition and speaker-independent connected digit recognition tasks which are the two key speech recognition applications for hands-free voice dialling systems.

2. NORMALIZATION ALGORITHM

Earlier, we proposed a segmental feature vector normalization technique [8] to deal with the performance degradation due to the acoustic mismatch between training and testing environments. The objective was to normalize the feature vectors to have a zero mean and unity variance within a segment of interest as follows

$$\hat{\mathbf{x}} = \sqrt{\Lambda^{-1}} \cdot (\mathbf{x} - \mathbf{m}) \qquad (1)$$

where $\mathbf{x}$ is the original feature vector, and $\hat{\mathbf{x}}$ is its normalized version, respectively. This kind of normalization generally requires some information on the statistics of feature vectors over the whole utterance. Therefore, the normalization coefficients, the cepstral mean vector $\mathbf{m}$, and the inverse of the diagonal covariance matrix $\Lambda^{-1} = \lceil 1/\sigma^2_{11},\ldots, 1/\sigma^2_{LL} \rfloor$ were computed over a sliding finite length normalization window (1.0 sec.). The feature vector to be normalized was located at the center of the window. The advantages of this type of normalization are the environment-independent parameter statistics, fast adaptation to changing noise conditions, and independence of VAD.

Previously, a similar normalization approach has widely been applied to neural network based speech recognizers to speed up the parameter estimation process. However, in those systems the normalization coefficients are typically computed over the whole utterance which is not a feasible solution in real-time applications due to the unnecessary long processing delay. To avoid this delay, it was proposed in [7] that mean and standard deviation should be computed as the long-term estimates over the several past utterances. Furthermore, a recursive technique for computing the normalization coefficients was suggested to avoid the problems associated with buffering feature vectors of the past utterances. It is nevertheless obvious that this kind of normalization approach is not well applicable for practical speech recognizers. At first, it cannot cope with rapidly changing noise conditions, and secondly, it assumes

0-7803-4428-6/98 $10.00 © 1998 IEEE

that the past utterances are spoken in similar noise conditions, as the current utterance.

2.1 Recursive Feature Vector Normalization

The length of the normalization window is the major drawback associated with the segmental feature vector normalization approach. It was shown in [8] that approximately a 1.0 sec. speech segment is long enough to guarantee robust normalization coefficients. From implementation point of view, the window length should be as short as possible, since memory consumption and the processing delay are directly proportional to the normalization window length.

The normalization coefficients can be estimated recursively as shown in [7]. By combining the recursive normalization coefficient computation and the previous segmental feature vector normalization, we can significantly shorten the normalization window length without decreasing the recognition performance. During the first N feature vectors, we only compute the values for sample sum and sample square sum. Thus, the initial mean and standard deviation estimates can be determined for each feature vector component i as

$$m_t(i) = \frac{1}{N}\sum_{t=1}^{N} o_t(i) \quad \text{and}$$
$$\sigma_t(i) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left[o_t^2(i)\right] - \left[m_t(i)\right]^2} = \sqrt{\overline{s_t^2(i)} - \left[m_t(i)\right]^2} \qquad (2)$$

where $o_t(i)$ denotes the ith feature vector component at time t. Once the initial estimates are known, the first feature vector in the window can be normalized as

$$\hat{x}_{t-N}(i) = \frac{x_{t-N}(i) - m_t(i)}{\sigma_t(i)}. \qquad (3)$$

All feature vectors are thus decoded with the delay of N frames. For each new incoming feature vector, we slide the normalization window forwards and iteratively update the cepstral mean and sample square estimates as

$$m_t(i) = \lambda \cdot m_{t-1}(i) + (1-\lambda)\cdot o_t(i) \qquad (4)$$

and

$$\overline{s_t^2(i)} = \lambda \cdot \overline{s_{t-1}^2(i)} + (1-\lambda)\cdot o_t^2(i) \qquad (5)$$

where λ is the step-size of the update. Experimental results show that we can use significantly shorter (approximately 50-80%) window for computing the normalization coefficients, and still slightly improve the recognition performance compared to the segmental normalization method.

3. ANALYSIS OF NORMALIZED FEATURE VECTORS

The feature vector normalization transform given in Eq. (1) can be divided into mean removal and Automatic Gain Control (AGC) parts. Mean removal can be regarded as linear high-pass filtering and division by standard deviation takes care of the AGC function so that the parameter statistics are always the same irrespective of noise conditions.

3.1 Time Trajectories of Normalized Features

If the normalization coefficients are computed over the whole utterance, the shape of feature vector time trajectories is not altered at all. However, since the normalization window does not expand over the whole utterance, the time-domain trajectories of feature vectors are distorted. This phenomenon is illustrated in Figs. 1 and 2 where the time-domain trajectory of the first cepstral coefficient (C1) is drawn with and without recursive normalization in clean and noisy environments (N=30, $\lambda = 0.955$). The vertical lines denote the start- and endpoints of the utterance, respectively.

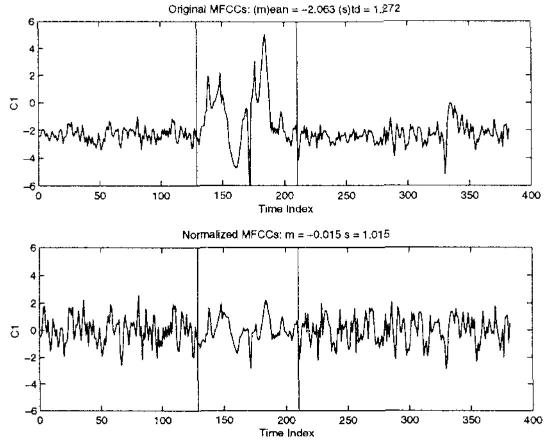

Fig. 1: Time domain trajectory of the C1 in a clean environment with and without feature vector normalization.

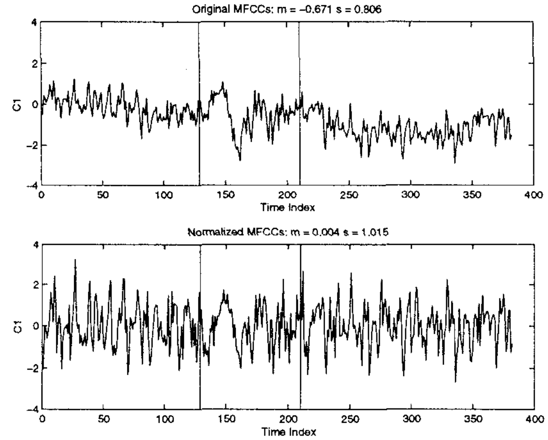

Fig. 2: Time domain trajectory of the C1 in the presence of car noise (SNR = -5 dB) with and without feature vector normalization.

The lower displays of Figs. 1 and 2 show that both in the clean and noisy case, the parameter statistics appear very similar. Since the variance of background noise is very small, particularly in a clean environment (Fig. 1), one has to emphasize the noise portions of the utterance in order to meet the unit variance requirement within a segment of interest. Therefore, the speech and background noise portions of the utterance can visually be difficult to distinguish in the case of normalized features. This event is also visible in spectral domain as shown in the next section.

3.2 Spectral Representation of Normalized MFCCs

We used mel-log power spectrogram displays for investigating the spectral representation of normalized feature vectors. The spectrograms were plotted both in the case of original and normalized MFCCs. Fig. 3 illustrates the spectrograms of the digit "three" for original and normalized MFCCs in a clean environment. In Fig. 4, the same displays are plotted for the same digit spoken by the same speaker in a noisy environment (car environment, driving approx. 120 km/h).

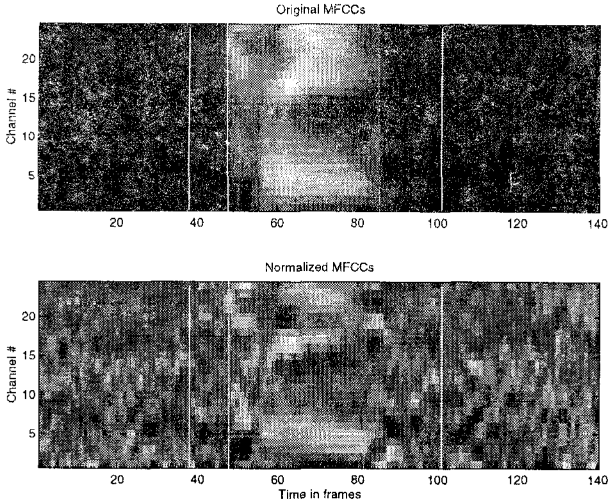

Fig. 3: Mel-log power spectrograms for original and normalized MFCCs in a clean environment.

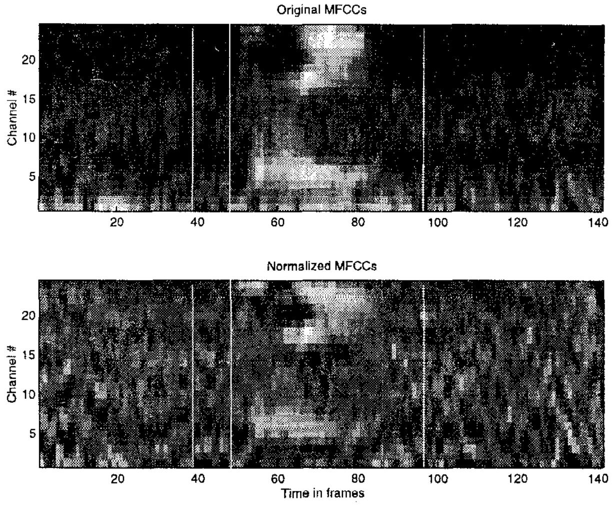

Fig. 4: Mel-log power spectrograms for original and normalized MFCCs in the presence of noise.

By studying the upper displays of Figs. 3 and 4, it can be seen that ambient background noise effectively masks speech resulting in very different feature representation in clean and noisy conditions. However, the lower displays of Figs. 3 and 4 show that in the case of normalized features, noise has less effected the spectrograms. As shown in Fig. 3, the use of feature vector normalization in a clean environment makes the spectrograms to look more "noisy", and hence, it is more difficult to detect speech events. Recognition experiments nevertheless indicate that a fuzzy boundary region between speech and noise does not decrease the recognition performance in clean conditions. Moreover, a comparison of the lower displays of Figs. 3 and 4 show that non-speech regions appear similar. It is difficult to give a precise explanation for the improvement in recognition accuracy that we see when applying this normalization approach to mismatch conditions. The most obvious explanation is that the feature vector statistics are made similar in different noise conditions.

4. EXPERIMENTAL RESULTS

The recursive feature vector normalization approach was evaluated in speaker-dependent name recognition and speaker-independent connected digit recognition tasks. In all the tests, 13 MFCCs (including the zeroth cepstral coefficient), their first- and second-order time derivatives were extracted from the incoming signal.

In name recognition, a state duration constrained HMM [4] for each vocabulary word was estimated from a single noise-free training utterance. Each training utterance was automatically endpointed based on frame powers and zero-crossing rates. Due to the lack of training data, all HMMs shared the same diagonal covariance matrix (unity matrix in the case of normalized MFCCs). In testing, clean waveforms were artificially corrupted by adding noise at various Signal-to-Noise Ratios (SNR). All given recognition percentages are average rates over all test speakers.

The TIDIGITS speech database was used in speaker-independent connected digit experiments. Again, to obtain noise corrupted utterances, car noise was added to clean waveforms. Whole word digit HMMs were estimated using an equal amount of training data from selected noise conditions according to the Maximum Likelihood (ML) principle.

4.1 Results with Recursive Normalization

At first, the effect of normalization window length on the recognition accuracy was studied in the case of recursive feature vector normalization. The window length N (in terms of frames) and the update step-size λ were coupled as

$$1-\lambda^N = \frac{1}{\sqrt{2}}. \qquad (6)$$

Figure 5 shows the results obtained in speaker-dependent name recognition with various values of N at different SNRs. Clearly, one has to buffer at least 20 feature vectors (0.2 secs.) in order to achieve a good recognition performance at low SNRs as well. Based on these results, the window length of 30 was chosen for recursive normalization in the following experiments conducted in this paper.

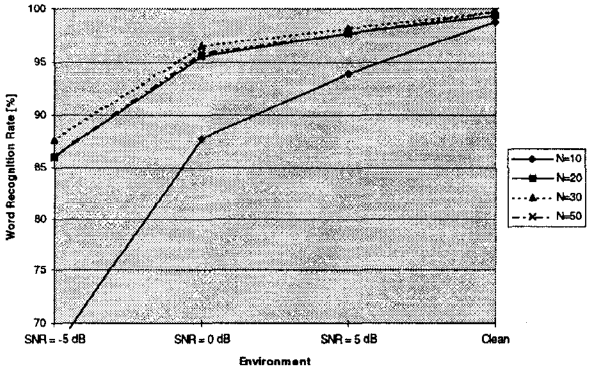

Fig. 5: Effect of normalization window length on the recognition accuracy.

4.2 Comparison of Noise Robust Techniques

The viability of the proposed recursive feature vector normalization approach was studied both in speaker-dependent and speaker-independent recognition tasks. The performance of the recursive normalization approach was compared in speaker-dependent name dialling to original MFCCs, PMC, and segmental normalized MFCCs [8]. Our PMC implementation [9-10] relies on VAD [1] whose decisions are used to control the noise estimate computation. In connected digit recognition, the performance of the recursively normalized MFCCs was compared to original MFCCs when using multi-environment HMMs.

Figure 6 illustrates the recognition accuracy in name dialling at various noise conditions. In the matched case (clean), a high recognition rate was obtained in all cases. If no noise compensation algorithms were applied, the recognition performance begins to decrease drastically. PMC slightly improved the performance, but the recognition rates were still fairly low. By means of feature vector normalization, a high performance could also be achieved in adverse conditions. According to Figure 6, it can be seen that the recursive approach slightly outperforms the segmental normalization approach. The average error rate reduction in the recursive case was 62.0%, and using the earlier segmental method a 49.5% overall error rate reduction was obtained. Moreover, regarding the implementation aspects the recursive approach was superior to the segmental method, as only 30 feature vectors were needed for normalization, whereas the segmental method required 100 frames [8] in order to achieve approximately the same recognition performance. Both feature normalization methods were superior to PMC which obtained a 14.0% overall error rate reduction.

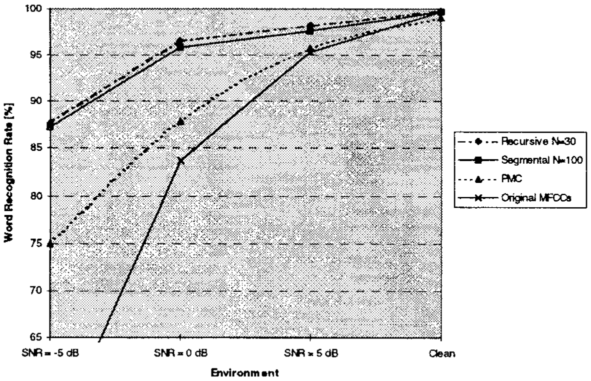

Fig. 6: The performance of various noise robust techniques in name recognition.

Speaker-independent connected digit tests were carried out in three different environments using three different sets of HMMs. Each model set had a constant number of mixtures (1-3) in each HMM state. As Fig. 7 shows, the use of normalization always improved the recognition rates with respect to original MFCCs (baseline).

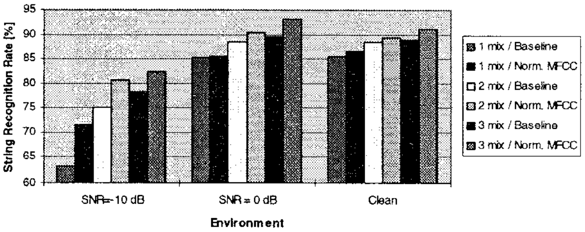

Fig. 7: Recognition results in speaker-independent connected digit tests with and without normalization.

Due to normalization, the average error rate reduction over all the model sets and environments was 16.3%. The greatest absolute performance improvement was achieved with single mixture HMMs in the most noisy conditions. Even though feature vector normalization was found to improve the recognition accuracy, the performance gains were not as great as expected based on the speaker-dependent experiments. Large speaker variability and various co-articulation effects possibly reduced the performance in these speaker-independent experiments.

5. CONCLUSIONS

In this paper, we proposed and investigated a recursive implementation approach for feature vector normalization. The main advantage of the recursive normalization approach is its more efficient implementation compared to the previously described segmental method. Recognition experiments indicate that the recursive approach also provides marginally better recognition performance in speaker-dependent name dialling than obtained with segmental normalization. In the speaker-independent tests, recursive feature vector normalization was found to improve the performance in a multi-environment connected digit recognition task. However, we are not fully satisfied with the obtained performance gains. The focus of the future work is thus on improving the performance of normalization in the speaker-independent case as well.

REFERENCES

[1] D. K. Freeman, G. Cosier, C. B. Southcott, I. Boyd, "The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service", *Proc. ICASSP*, Glasgow, Scotland, pp. 369-372, 1989.

[2] M. Gales, S. Young, "Cepstral Parameter Compensation for HMM Recognition", *Speech Communication*, Vol. 12, No. 3, pp. 231-239, 1993.

[3] H. Hermansky, N. Morgan, H.-G. Hirsch,"Recognition of Speech in Additive and Convolutive Noise Based on RASTA Spectral Processing", *Proc. ICASSP*, Minneapolis, USA, pp. II-83 - II-86, 1995.

[4] K. Laurila, "Noise Robust Speech Recognition with State Duration Constraints", *Proc. ICASSP*, Munich, Germany, pp. 871-874, 1997.

[5] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", *Speech Communication*, Vol. 11, No. 2-3, pp. 215-228, 1992.

[6] A. Rosenberg, C.-H. Lee, F. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", *Proc. ICSLP*, Yokohama, Japan, pp. 1835-1838, 1994.

[7] S. Tiberwala, H. Hermansky, "Multi-band and Adaptation Approaches to Robust Speech Recognition", *Proc. EUROSPEECH'97*, Rhodes, Greece, pp. 2619-2622, 1997.

[8] O. Viikki, K. Laurila, "Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature Vector Normaliztion", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-à-Mousson, France, pp. 107-110, 1997.

[9] R. Yang, M. Majaniemi, P. Haavisto, "Dynamic Parameter Compensation for Speech Recognition in Noise", *Proc. EUROSPEECH'95*, Madrid, Spain, pp. 469-472, 1995.

[10] R. Yang, P. Haavisto, "An Improved Noise Compensation Algorithm for Speech Recognition in Noise", *Prc. ICASSP*, Atlanta, USA, pp. 49-52, 1996.