

# NOISE ROBUST HMM-BASED SPEECH RECOGNITION USING SEGMENTAL CEPSTRAL FEATURE VECTOR NORMALIZATION

Olli Viikki and Kari Laurila

email: {olli.viikki,kari.laurila}@research.nokia.fi

Nokia Research Center, Speech and Audio Systems Laboratory

P. O. Box 100, FIN-33721 Tampere

FINLAND

## ABSTRACT

To date, speech recognition systems have been applied in real world applications in which they must be able to provide a satisfactory recognition performance under various noise conditions. However, a mismatch between the training and testing conditions often causes a drastic decrease in the performance of the systems. In this paper, we propose a segmental feature vector normalization technique which makes an automatic speech recognition system more robust to environmental changes by normalizing the output of the signal-processing front-end to have similar segmental statistics in all noise conditions. The viability of the suggested technique was verified in various experiments using different background noises and microphones. In an isolated-word recognition task, the proposed normalization technique reduced the error rates over 70% in noisy conditions with respect to the baseline tests, and in a microphone mismatch case, over 75% error rate reduction was achieved.

## 1. INTRODUCTION

It is well known that a mismatch between the training and testing environments produces degradation in the recognition performance. This mismatch can be caused for example due to different microphones, noise conditions, or communication channels. In order to make the recognition performance more independent of the usage environment, a number of different noise compensation techniques have been suggested. At the feature representation level, various normalization methods and noise robust feature extraction techniques have been developed [1-3]. Noise compensation can also be carried out in the recognition unit. In the technique called Parallel Model Combination (PMC) [4], the HMMs estimated in a clean environment are modified to characterize the current noise conditions. Regardless of the used normalization or compensation technique, a noise estimate is needed to characterize the noise conditions of the present usage environment. To perform a fast noise estimate adaptation, a reliable Voice Activity Detector (VAD) is required in practical systems to make the decision whether the input frame is speech or noise. As the speech frame classification accuracy of VAD depends heavily on the noise level, the performance of many normalization or compensation techniques decreases in very noisy conditions.

In the current state-of-the-art speech recognition systems, the Mel-Frequency Cepstral Coefficients (MFCC) are widely

used to characterize the speech input. Since statistics of the MFCCs vary depending on noise conditions, we propose, in this paper, a normalization technique which converts the output of the Feature Extraction (FE) unit to have equal segmental statistics in all noise conditions in order to reduce the mismatch between the training and testing conditions. The proposed normalization approach operates independently of VAD, and thus, a good performance is obtained in very noisy conditions as well. Due to the pure segmental nature of the proposed normalization technique, a fast adaptation to new background noise conditions is obtained. Experiments indicate that the proposed normalization technique is robust to different noise conditions and microphones. Compared to other noise compensation techniques, the proposed approach clearly outperforms them in terms of recognition accuracy.

## 2. SEGMENTAL FEATURE VECTOR NORMALIZATION

The fundamental idea behind the segmental feature vector normalization technique is that irrespective of noise conditions, the FE output is forced to the same numerical range. In this paper, the MFCCs are normalized to zero mean and unit variance within a segment of interest. Previously, a similar type of normalization approach has been applied in the context of neural network classifiers to speed up the network parameter estimation, when the normalization coefficients, i.e., mean and standard deviation, were calculated over the whole utterance. However, it is obvious that this type of implementation is not useful for real-time applications. In our normalization approach, the normalization coefficients are calculated over a relatively short sliding window. Therefore, our approach can be utilized in a real-time speech recognizer.

### 2.1 Segmental Normalization Algorithm

Before performing the Viterbi decoding with feature vector, it is normalized as follows

$$\hat{x}_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

where  $x_i$  is the  $i$ th component of the original feature vector, and  $\hat{x}_i$  is its normalized version, respectively. One can also perform simplified normalization, when division with standard deviation is not done. Cepstral mean estimate calculation is the major difference of this Segmental Cepstral Mean Normalization (SCMN) compared to the conventional Cep-

stral Mean Normalization (CMN) approach. The normalization coefficients, mean  $\mu_i$  and standard deviation  $\sigma_i$ , for each feature vector component  $i$ , are calculated over the sliding normalization window as

$$\mu_i = \frac{1}{N} \sum_{t=1}^N x_{i,t} \quad (2)$$

and

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{t=1}^N (x_{i,t} - \mu_i)^2} \quad (3)$$

where the  $N$  denotes the normalization segment length. The feature vector to be normalized and recognized is located at the center of the window. At the beginning of recognition, a shorter length normalization segment is used to move the feature vector to be processed to the segment center point, and to minimize the delay at the end of recognition. The first feature vectors are not normalized until there are  $N/2$  vectors stored to the normalization segment. Once there are enough vectors stored in the buffer, the first vector is normalized and decoded, and a new feature vector is inserted to the buffer. This procedure is continued until the normalization segment is full. At this point, the first  $N/2$  vectors have already been decoded, and the next vector to be processed is located at the center of normalization segment. Now, the normalization segment can be slid over feature vectors until the end of utterance is detected, when the last  $N/2$  feature vectors in the buffer are normalized and recognized using the same normalization coefficients.

Using the proposed approach, the delay due to feature vector buffering at the end of recognition corresponds to the half length of the normalization segment. The major advantage of the segmental normalization technique is that no VAD is needed, hence, the VAD inaccuracy in very noisy conditions does not have any effect to performance. Consequently, there is no delay associated with normalization coefficient computation, as mean and standard deviation calculations require no iterative formulas, but the coefficients are computed over the feature vector buffer. Due to this, adaptation to new noise conditions is very fast. Since feature vectors are normalized to unit variance within a segment, all components of diagonal covariance matrix can be set to unity. This reduces the HMM memory consumption and enables a faster observation probability calculation.

## 2.2 Feature Vector Trajectories

The shape of the feature vector trajectories does not change if the normalization coefficients are calculated over the entire utterance. In segmental normalization, mean and standard deviation are nevertheless determined over a finite length window, when it is not possible to maintain the shape of the original feature vector trajectories. Clearly, one cannot use too short segment length since the trajectory would be too severely destroyed. Our goal is to find the minimum segment length which provides reliable enough normalization coefficients and which does not change the trajectory too much. Figures 1 and 2 show the trajectory of the first cepstral component (C1) for the clean and noisy (SNR=0 dB) utterances without any normalization, and with the entire utterance (file) based normalization, and segmental normalization approach using the segment lengths of 50 and 100 (0.5 and 1.0 seconds, respectively).



Figure 1: C1 trajectory of the noise-free utterance without and with (utterance based and segmental) normalization.



Figure 2: C1 trajectory of the noisy utterance (SNR = 0 dB) without and with normalization.

## 2.3 Normalization Segment Length

From the implementation point of view, the length of normalization segment is a critical parameter which should be as short as possible. Both computational complexity and memory requirements of the proposed algorithm are directly proportional to the segment length. The normalization segment length has also a major effect to the recognition accuracy. Therefore, special attention must be paid to the decision of the segment length. In Figure 3, the recognition accuracy is given as a function of the normalization window length in an isolated word recognition task with different noise levels (car noise). The window length of 0 (marked with the circles) corresponds to the experiments where original feature vectors were used, and correspondingly, the segment length of 500 is the case where coefficients were computed over the entire utterance.

As Figure 3 shows, the normalization segment must extend over 30 frames in noisy conditions, and over 50 frames in a clean environment in order to achieve performance gain. It can be seen that the recognition performance saturates around the segment length of 100. Compared to the original feature vectors, the performance improvement due to normalization increases at the lower SNRs.



Figure 3: Dependency of the recognition accuracy on the normalization window length.

### 3. TEST DATABASES AND EXPERIMENTS

The viability of the proposed normalization method was tested in various experiments with different noise types and levels. In all experiments, the feature vectors consisted of the FFT-based MFCCs, log-energy, and their first and second order time-derivatives. In the first set of experiments, the performance of the segmental normalization scheme was studied in the environmental mismatch case against two other well-known noise compensation methods, namely Cepstral Mean Normalization (CMN) [7] and PMC. Both mean and variance compensations for static and dynamic coefficients were carried out in the used real-time PMC implementation [8-9], and an adaptive single mixture noise model was adjusted according to the noisy feature vectors. Also in our CMN implementation, a VAD [10] was required, as the mean of each cepstral value was estimated only over the noise portions of the utterance. In the second experiment, the recognition performance of the proposed algorithm was evaluated in a microphone mismatch case against original MFCCs.

#### 3.1 Tests on Various Noise Types

The test database consisted of isolated words spoken by five different speakers in a clean office environment. For each speaker, speaker-dependent, single mixture, state duration constrained [5] HMMs were estimated using a *single* training token spoken in a *clean* environment. An automatic endpointing algorithm based on the frame power and zero crossings was used to determine the starting and ending points of the training utterances. In the experiments with original feature vectors, all HMMs shared the same variance vector (grand variance) due to the lack of training data, whereas in the case of segmental normalized feature vectors, a unit variance vector was used.

The vocabulary consisted of 30 words (Finnish first names). In the testing phase, different background noise types were subsequently added to the clean speech waveforms at various SNRs. Background noise was modelled with a garbage modelling technique presented in [6]. The following four different background noise types were added to the noise-free utterances at various SNRs:

- Stationary, narrow-band car noise
- Multi-talker, wide-band babble noise
- Classical music (instrumental)
- Non-stationary, impulsive machine gun noise

The normalization segment length was set to 100 when using normalized feature vectors. In the presence of car noise, the performance of SCMN was also tested. Because segmental normalization with standard deviation division provided much better recognition performance, SCMN was not used with other noise types. No normalization or compensation schemes were used in the baseline tests. In the CMN tests, the VAD decisions controlled the update of cepstral mean estimates. Estimates were initialized with the mean of the first 20 feature vectors of each test utterance which were assumed to be non-speech. Thereafter, the estimates were iteratively updated every frame when VAD did not detect speech. In the case of PMC, a noise model was also adapted every frame when VAD did not detect speech. Only in the presence of machine gun noise, possibly due to poor accuracy of VAD, PMC could not achieve performance gain. Otherwise PMC performed much better than CMN. However, the performance of both CMN and PMC decreased quite drastically at lower SNRs. Results are summarized in Tables 1-4. In Table 1, the row “clean” corresponds to the experiment in which both the training and testing were carried out in the same noise-free environment.

SNR	Ba-	CMN	PMC	SCMN	SG_NOR
<b>Clean</b>	96.9	96.8	96.9	97.0	97.5
<b>5</b>	95.3	95.6	95.5	95.8	96.3
<b>0</b>	94.0	93.5	94.8	94.4	96.1
<b>-5</b>	89.3	87.7	91.6	89.6	94.6
<b>-10</b>	70.1	74.5	83.3	74.9	91.2

Table 1: Recognition performance in car noise.

SNR	Baseline	CMN	PMC	SG_NOR
<b>25</b>	96.3	95.8	96.2	96.6
<b>20</b>	95.6	94.7	95.8	96.1
<b>15</b>	93.9	92.8	94.5	95.7
<b>10</b>	87.8	86.0	91.2	93.0
<b>5</b>	69.7	66.8	79.4	84.5

Table 2: Recognition performance in babble noise.

SNR	Baseline	CMN	PMC	SG_NOR
<b>20</b>	93.6	93.0	94.9	95.8
<b>15</b>	87.9	87.6	92.3	94.8
<b>10</b>	75.4	75.9	87.3	92.7
<b>5</b>	54.2	55.5	73.9	86.3
<b>2</b>	39.3	40.1	60.6	79.1

Table 3: Recognition performance in music.

SNR	Baseline	CMN	PMC	SG_NOR
<b>10</b>	95.2	95.3	95.2	96.2
<b>5</b>	93.6	92.4	93.9	95.8
<b>0</b>	88.9	88.1	89.4	94.6
<b>-5</b>	83.7	83.9	83.6	90.9
<b>-10</b>	77.9	78.4	76.9	84.7

Table 4: Recognition performance in machine gun noise.

Irrespective of noise type or level, the proposed feature vector normalization approach clearly outperformed PMC, CMN, and baseline results. For example, in -10 dB SNR car noise and in +5 dB SNR music background, the proposed normalization technique reduced the error rates over 70% with respect to the baseline tests, and over 47% compared to the PMC tests. Not surprisingly, the highest recognition accuracy was achieved in highly stationary car noise. In general, it can be concluded that the lower the baseline recognition rate, the higher the gained performance improvement due to feature vector normalization.

### 3.2 Tests with Different Microphones

In the last experiment, the effect of different microphones in training and testing was studied. Test settings were the same as in the experiments described in Section 3.1. Test utterances (isolated names) were spoken by six different speakers, and each of them had a different vocabulary. All speech was recorded using the following four different microphones:

- AKG C410/B - head-mounted, close talking microphone (CT)
- Primo EMU 4705 - hands-free microphone (HF)
- WM-62A - omnidirectional, close-talking, electret condenser microphone (OC)
- CMP-202 "Fico" - PC microphone (PC)

Normalized feature vectors (segment length 100) were compared to the original MFCCs. All different microphones were used both in training and testing. Experiments were done only in noise-free conditions. Tables 5 and 6 show the recognition results with standard and segmental normalized MFCCs. The recognition accuracies in the matched cases are located on the diagonal of Table in the shaded cells.

By studying Table 5, one can notice that the average matched microphone recognition rate was 98.7% whereas the average mismatch microphone recognition rate was 98.0%, which corresponds to an increase of 50% in the error rate. Table 6 shows that the average matched microphone recognition rate was 99.4% and the average mismatch microphone recognition rate was 99.5%, which actually means decrease of 15% in the error rate. When the average microphone mismatch figures for the original MFCCs and segmental normalized MFCCs were compared, one can notice 75% reduction of the error rate due to normalization.

Train/Test	CT	HF	OC	PC
CT	99.2	99.5	98.8	99.5
HF	96.5	97.9	96.1	96.6
OC	97.9	97.9	98.1	98.3
PC	97.6	99.0	98.4	99.5

Table 5: Recognition accuracy with different microphones using the standard MFCCs.

Train/Test	CT	HF	OC	PC
CT	99.5	99.6	99.5	99.6
HF	99.5	99.4	99.5	99.7
OC	99.0	99.2	99.2	99.7
PC	99.6	99.7	99.6	99.6

Table 6: Recognition accuracy with different microphones using the segmental normalized MFCCs.

## 4. CONCLUSIONS

This paper has reported a segmental feature vector normalization technique for noise robust HMM-based speech recognition. The feature vectors to be recognized are normalized to zero mean and unit variance. The normalization coefficients are calculated over a sliding finite length normalization segment. In the performed experiments very encouraging results were obtained with a mismatch between the training and testing conditions. With all noise types and levels, the proposed normalization technique provided better recognition accuracy than baseline, CMN, and PMC techniques. The proposed normalization scheme also decreased the performance degradation due to microphone mismatch.

## REFERENCES

- [1] D. van Compernelle, T. Claes, "SNR-Normalisation for Robust Speech Recognition", *Proc. of ICASSP'96*, pp. 331-334, Atlanta, USA, 1996.
- [2] A. Acero, R. M. Stern, "Cepstral Normalization for Robust Speech Recognition", *Proc. of Speech Processing in Adverse Conditions*, pp. 89-92, Cannes-Mandelieu, France, 1992.
- [3] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *Journal Acoust. Soc. Am.*, Vol. 87, No. 4, pp. 1738-1752, 1990.
- [4] M. Gales, S. Young, "Cepstral Parameter Compensation for HMM Recognition", *Speech Communication*, Vol. 12, No. 3, pp. 231-239, 1993.
- [5] K. Laurila, "Noise Robust Speech Recognition with State Duration Constraints", To be appeared in ICASSP'97, Munich, 1997.
- [6] J. Iso-Sipilä, K. Laurila, P. Haavisto, "Optimal Adaptive Garbage Modeling in Speech Recognition", *Proc. of IEEE Nordic Signal Processing Symposium*, pp. 107-110, Espoo, Finland, 1996.
- [7] A. Rosenberg, C.-H. Lee, F. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", *Proc. of ICSLP'94*, pp. 1835-1838, Yokohama, Japan, 1994.
- [8] R. Yang, M. Majaniemi, P. Haavisto, "Dynamic Parameter Compensation for Speech Recognition in Noise", *Proc. of EUROSPEECH'95*, pp. 469-472, Madrid, Spain, 1995.
- [9] R. Yang, P. Haavisto, "An Improved Noise Compensation Algorithm for Speech Recognition in Noise", *Proc. of ICASSP'96*, pp. 49-52, Atlanta, USA, 1996.
- [10] D. K. Freeman, G. Cosier, C. B. Southcott, I. Boyd, "The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service", *Proc. of ICASSP'89*, pp. 369-372, Glasgow, Scotland, 1989.