ELSEVIER

Speech Communication 25 (1998) 133–147

SPEECH COMMUNICATION

Cepstral domain segmental feature vector normalization for noise robust speech recognition

Olli Viikki *, Kari Laurila

Nokia Research Center, Speech and Audio Systems Laboratory, P.O. Box 100, FIN-33721 Tampere, Finland

Received 1 September 1997; accepted 1 February 1998

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 parameter 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 by over 70% in noisy conditions with respect to the baseline tests, and in a microphone mismatch case, over 75% error rate reduction was achieved. In a multi-environment speaker-independent connected digit recognition task, the proposed method reduced the error rates by over 16%.

Zusammenfassung

Heutzutage eingesetzte Spracherkennungssysteme müssen in der Lage sein, unter verschiedenartigen Störeinfluessen zuverlässig zu arbeiten. Unterschiedliche Trainings- und spätere Anwendungsbedingungen schränken dies jedoch oft erheblich ein. In diesem Bericht erläutern wir ein neues Verfahren zur Normierung der Ergebnisse der Vorverarbeitung. Dadurch besitzen die entsprechenden Signale auch bei unterschiedlichen Störigeräuschen ähnliche statistische Eigenschaften und die Spracherkennung wird gegen sich ändernde Umgebungseinflüsse robust. Die Wirksamkeit des vorgeschlagenen Verfahrens wurde in verschiedenen Experimenten mit unterschiedlichen Hintergrundgeräuschen und Mikrophonen bestätigt. Bei einem Einzelworterkennungssystem verringerte sich die Fehlerrate um mehr als 70% im Vergleich zur herkömmlichen Vorhegensweise. Bei Verwendung unterschiedlicher Mikrophone wurde eine Verringerung der Fehlerrate um 75% erreicht. Schliesslich wurde die Fehlerrate bei sprecherunabhängiger Ziffernerkennung um mehr als 16% gesenkt.

Résumé

Les systèmes de reconnaissance vocale sont maintenant utilisés dans des applications où ils doivent fournir une reconnaissance satisfaisante dans différentes conditions de bruit. Cependant, un écart entre les conditions d'entraînement et de test est souvent à l'origine d'une serieuse baisse de performance des systèmes. Nous décrivons dans cet article

* Corresponding author. Tel.: +358 3 272 5479; fax: +358 3 272 5897; e-mail: olli.viikki@research.nokia.fi.

0167-6393/98/$ – see front matter
PII: S0167-6393(98)00033-8

une technique de normalisation du vecteur des caractéristiques segmentaires qui rend un système de l'environnement de reconnaissance vocale automatique plus robuste aux changements de l'environnement. A cette fin, après traitement initial du signal, la sortie est normalisée de façon a obtenir les mêmes paramètres statistiques de segmentation dans toutes les conditions de bruit. Le succès de cette technique a été verifiée pour plusieurs expériences utilisant différents bruits de fond et microphones. Dans une tâche de reconnaissance de mots isolés, la technique de normalisation proposée a reduit le taux d'erreur de plus de 70% en présence de bruit par rapport aut tests de référence. Dans le cas d'un écart au niveau du microphone, une réduction de plus de 75% a été obtenue. La méthode proposée appliquée à la reconnaissance de chiffres indépendante du locuteur et dans un environement multiple a réduit le taux d'erreur de plus de 16%.

Keywords: Speech recognition; Noise robustness; Feature vector normalization

1. Introduction

Noise robustness is one major requirement for practical automatic speech recognition systems. Typically, a high recognition accuracy can be obtained if there is a good matching between the training and testing conditions. In real world applications, however, the acoustically same training and testing environments can only seldom be guaranteed. Different communication channel characteristics and ambient background noise usually degrade the recognition performance in practical systems. Even though noise robustness has recently gained a great interest in speech recognition research, current practical speech recognition systems still provide fairly moderate recognition performance if they are used under realistic noise conditions.

The effects of noise on the clean speech distribution are widely known (Openshaw and Mason, 1994). In the case of additive noise, the distribution becomes non-Gaussian and bimodal. As the noise starts to dominate in the signal, the corrupted distribution eventually becomes unimodal, but it still remains non-Gaussian. Moreover, the distribution mean changes and variance decreases. In the presence of convolutional noise, it has been found that only the means of the clean speech distributions have been shifted if certain simplifications are assumed to be valid.

The easiest and most straightforward way to improve noise robustness is to attempt to collect large amounts of data from a wide range of acoustic environments. Although this so-called multi-environment training works reasonably well up to a certain limit, it is obvious that the problem of noise robustness cannot be solved by simply collecting a huge training set. It is impossible to collect such a database that would cover all possible usage environments. Furthermore, the use of large amounts of data often leads to HMMs with large variances, and hence, the model set does not provide a high recognition accuracy in any environment. In addition to multi-environment training, more sophisticated techniques have also been developed for improving the noise robustness of speech recognition systems. At the feature representation level, various normalization methods (van Compernolle and Claes, 1996; Acero and Stern, 1992), and noise robust feature extraction techniques (Hermansky, 1990) have been developed. The mismatch effects can also be considered in the recognition unit. In the technique called Parallel Model Combination (PMC) (Gales and Young, 1993), the HMMs estimated in a clean environment are transformed to characterize the current noise conditions. The main weakness of most compensation techniques is that they require a good noise estimate for performing adaptation to the new environment. For this purpose, a reliable Voice Activity Detector (VAD) is needed to make the decision whether the input frame is noise or speech. As the classification accuracy of VAD depends heavily on the noise level, many normalization or compensation techniques tend to fail in adverse 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 much depending on noise conditions, we propose a normalization technique which converts the

output of the feature extraction unit to have equal segmental parameter statistics in all noise conditions in order to reduce mismatches between training and testing conditions. Due to the pure segmental nature of the proposed normalization algorithm, adaptation to new background noise conditions is fast provided that the length of the used normalization segment is short enough.

The viability of the proposed normalization method is studied in various speaker-dependent name recognition and speaker-independent connected digit recognition experiments. These tasks are the key speech recognition modes in voice dialling applications. In speaker-dependent speech recognition, the proposed technique clearly outperforms other noise compensation techniques in terms of recognition accuracy. In the speaker-independent case, the multi-environment training procedure is still in practice very competitive with known noise compensation techniques such as PMC if recognition is carried out in restricted noise conditions, e.g., in a car environment. We show further that the proposed normalization method is capable of significantly improving the performance of the multi-environment training scheme.

2. Segmental feature vector normalization

Our initial motivation for the proposed normalization scheme was created by the speaker-dependent name recognition task in voice dialling where model training is done in a noise-free environment using only one training utterance. Despite the clean training data, recognition must be possible in any environment, even at low Signal-to-Noise Ratios (SNR). Due to the large mismatch between training and testing environments, one can expect a poor performance when using a recognizer relying on the standard MFCC front-end.

To cope with the mismatch, we first tried alternative type of front-ends, including Line Spectral Frequencies (LSF) (Paliwal, 1990) and Perceptual Linear Prediction (PLP) (Hermansky, 1990). Since we could not significantly improve the performance with these techniques, we resorted to noise removal schemes, such as Nonlinear Spectral Subtraction (NSS) (Lockwood and Boudy, 1992) and Cepstral Mean Normalization (CMN) (Rosenberg et al., 1994). Even though many researchers have reported significant improvements using these techniques, we could not reproduce any substantial performance gains. We ended up with two possible reasons. At first, experiments with noise removal or speech enhancement algorithms reported in the literature have mostly been limited to unrealistically easy background noises. It is much easier to remove highly stationary Gaussian noise than a real world, slowly, or rapidly varying noise signal. Secondly, tests reported in the literature are often conducted in an off-line manner in which good noise estimates have been available. In practical recognition systems, a voice activity detector is needed to classify the input signal to speech and non-speech frames after which a noise estimate can be computed in an iterative way. In noisy environments, the speech/non-speech classification does not work very well which results in poor estimates of noise and causes a poor performance for an an overall system.

With disappointing experiments in the front-end side, we still decided to test PMC (Yang et al., 1995, Yang and Haavisto, 1996). PMC is widely known as the state-of-the-art noise compensation scheme. The method operates in the recognition unit and also requires a noise estimate. PMC is computationally complex and makes the standard assumption that speech and noise are additive in the spectral domain. In our experiments, we were able to improve the results with PMC. However, PMC still left some room for improvements, both in terms of recognition accuracy and computational complexity.

Since none of the tested techniques did not meet the performance requirements, we decided to list the properties that are the most important in our recognition task. We ended up with the following list:

- environment-independent parameter statistics;
- fast or immediate noise adaptation;
- unimodal parameter distributions (one-mixture Gaussian densities);
- independency of VAD.

Starting from these requirements, we created a method which we call segmental feature vector

normalization. In the proposed segmental normalization technique, the MFCCs are normalized to have a zero mean and unit variance within a segment of interest. A similar type of normalization has previously been applied in the context of neural network based classifiers (Riis and Krogh, 1996) to speed up the network parameter estimation and to avoid to be stuck in the local minima, 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 cannot be used in real-time applications because the processing delay equal to the length of the utterance is introduced. In our approach, the normalization coefficients are calculated over a relatively short sliding window, and therefore, the technique can be utilized in a real-time speech recognizer.

2.1. Algorithm

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

$$\hat{x}_{t-D}(i) = \frac{x_{t-D}(i) - \mu_t(i)}{\sigma_t(i)}, \tag{1}$$

where $x_{t-D}(i)$ is the ith component of the original feature vector at time $t - D$, $\hat{x}_{t-D}(i)$ is its normalized version, and D denotes the delay (in terms of feature vectors) associated with buffering, respectively. One can also perform simplified normalization where division with standard deviation is omitted. Here, this type of normalization is called Segmental Cepstral Mean Normalization (SCMN). The normalization coefficients, mean $\mu_t(i)$ and standard deviation $\sigma_t(i)$, for each feature vector component i, are calculated over the sliding finite length normalization window as

$$\mu_t(i) = \frac{1}{N}\sum_{j=1}^{N} x_j(i) \tag{2}$$

and

$$\sigma_t(i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(x_j(i) - \mu_t(i)\right)^2}, \tag{3}$$

where N denotes the normalization segment length in terms of feature vectors. The feature vector to be normalized and recognized is located at the center of the window. To minimize the buffering delay at the end of recognition, a shorter length normalization buffer is used for the first vectors until the feature vector to be processed is the centermost vector in the buffer, i.e., the buffer is full. At the beginning of recognition, the Viterbi decoding is omitted until there are $N/2$ vectors stored in the normalization segment. Once the buffer contains $N/2$ feature vectors, the normalization coefficients are computed, the first vector in the buffer is normalized and decoded, and a new feature vector is inserted to the buffer. Next, the normalization coefficients are re-computed and the second vector in the buffer is normalized and decoded. This procedure is continued until the normalization buffer 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 the normalization segment. Now, the normalization segment can be slided over feature vectors until the end of utterance is detected. The last $N/2$ feature vectors in the buffer are normalized and Viterbi decoded without updating the normalization coefficients.

For the proposed normalization technique, it is characteristic that the feature vector statistics are computed over the current utterance to be recognized. In (Cook et al., 1996; Tibrewala and Hermansky, 1997), it is suggested that the parameter statistics should be computed over several past utterances. This approach is nevertheless feasible only in certain type of applications where the recognizer can be running continuously and the usage environment is fairly stationary. If consecutive utterances are spoken in completely different noise conditions, the normalization coefficient computation over several past utterances is not meaningful. In the case of portable low-power consumption devices, such as mobile phones, it is also not even possible to compute feature vectors continuously. Due to these reasons, we see it vital that the parameter statistics are computed separately for each utterance to be recognized.

With the proposed approach, the delay due to feature vector buffering at the end of recognition corresponds to the half of the normalization segment length. The major advantage of the

segmental approach is that no VAD is needed, and normalization is done in the same way both for speech and noise. Hence, the VAD inaccuracy in noisy conditions does not decrease the performance. Consequently, there is no delay associated with normalization coefficient computation and adaptation to new noise conditions is very fast.

2.2. Effects of normalization segment length

The shape of the feature vector trajectories can fully be maintained if the normalization coefficients are calculated over the entire utterance. This means that the normalization coefficients are constant for all feature vectors. In the proposed segmental normalization scheme, mean and standard deviation are nevertheless determined over a finite length window (segment) individually for each feature vector, and thus, the shape of the original feature vectors is altered. Clearly, one cannot use too short normalization segment since the trajectory would then be too severely destroyed. 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. Our goal is to find the minimum segment length which provides reliable enough normalization coefficients and which does not alter the trajectory too much. Figs. 1 and 2 illustrate the time-domain trajectory of the first cepstral coefficient (C1) in clean and noisy (stationary car noise) environments using the whole utterance (file) based normalization and the segmental approach with the segments of 50 and 100 feature vectors (0.5 and 1.0 s, respectively).

Due to the finite length sliding normalization window, zero mean and unit variance are not exactly obtained over the entire utterance and the overall trajectory shape is slightly modified. It can however be seen that the time-domain trajectory is located around zero in all noise conditions. Moreover, in a clean environment, the background noise portions of the utterance are significantly emphasized because the variance is constrained to be unity within a segment.

The highest recognition performance can be expected to be achieved in the optimal normalization case where the normalization segment

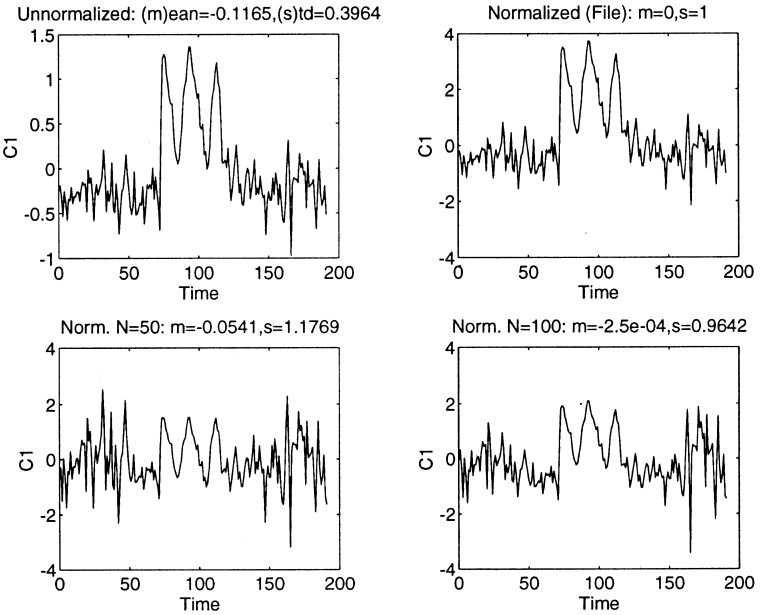

Fig. 1. Original C1 time trajectory of clean speech (top left) and its segmental normalized versions.

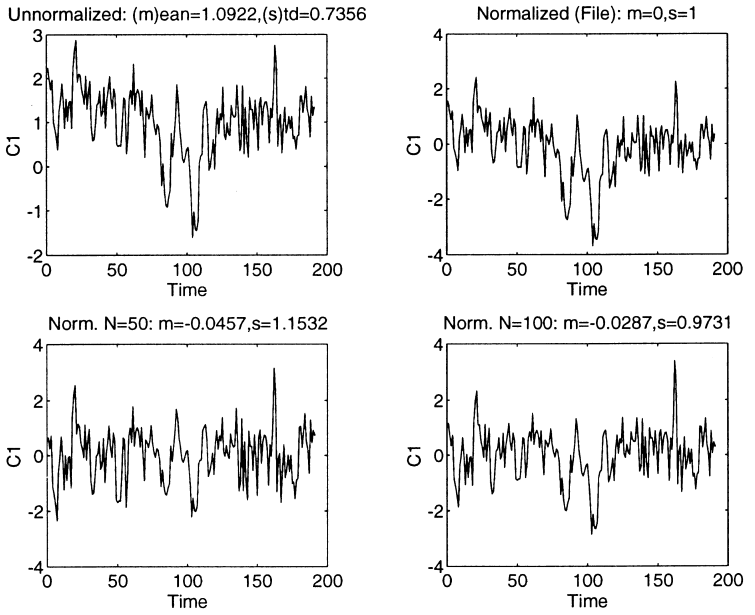

Fig. 2. Original C1 time trajectory of noisy (SNR = 0 dB) speech and its segmental normalized versions.

extends over the entire utterance. The effect of the normalization segment length on the recognition rate is illustrated in Fig. 3 where the recognition accuracy is given as a function of the normalization window length in an isolated word recognition task both in a clean environment and in the presence of car noise at two different SNRs. The horizontal lines show the baseline recognition accuracy with original MFCCs. The baseline performance was 96.9% in a clean environment, 89.3% at −5 dB SNR, and 70.1% at −10 dB SNR, respectively. The segment length of 500 frames corresponds to the case where the constant normalization coefficients were computed over the entire utterance.

Fig. 3 indicates that the normalization segment must extend over 30 frames in noisy conditions and over 50 frames in a clean environment in order to achieve some performance gain. It can further be seen that the recognition performance saturates around the segment length of 100 frames. Fig. 3 also shows that the whole utterance length normalization buffer is not necessarily required to achieve the highest possible recognition performance, but the best results are obtained with the buffer whose length is about 1.0 s. It has to be noted that a 100-frame normalization buffer is not the global optimum solution. In other tasks than isolated word recognition, e.g., dictation, a different length normalization buffer may be required.

3. Normalization analysis

It is very difficult to analyze the segmental normalization approach as a single block. However, it is possible to split the algorithm into two different parts:

1. Filtering,
2. Automatic Gain Control (AGC).

With this approach, it is easier to understand how the normalization algorithm works. The mean removal can be regarded as a linear high-pass filter and division by standard deviation acts as an AGC. Being non-linear, the effects of the AGC step are far more difficult to explain analytically. Basically, the order of these two blocks is not restricted, so the AGC function can be performed before or after filtering.

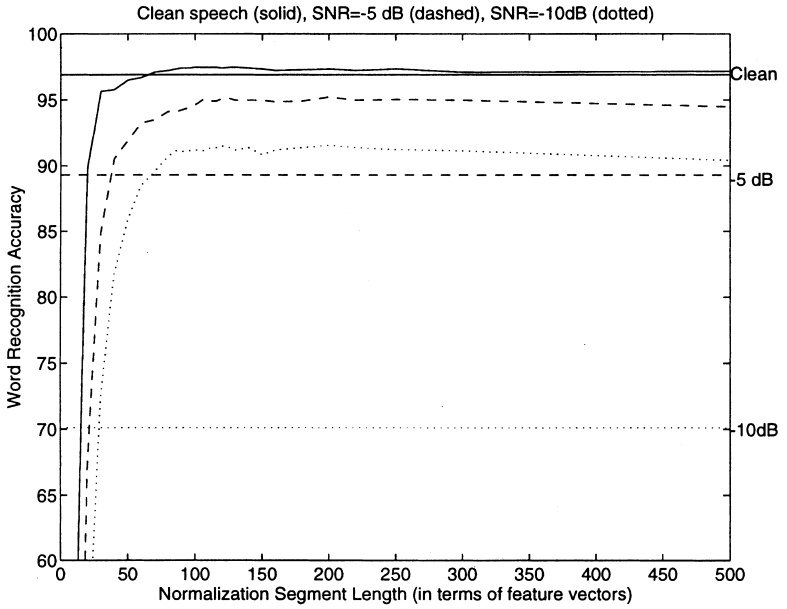

Fig. 3. Dependency of the recognition accuracy on the normalization window length.

3.1. Effect of normalization in que-frency domain

Figs. 1 and 2 illustrated the time-domain trajectories of the first cepstral coefficient after normalization. To obtain some knowledge on normalization, it is worthwhile also to examine the overall transfer function of the segmental normalization block. In Figs. 4–6, the responses of the entire normalization block have been calculated for the first static, delta, and delta–delta coefficients in a clean and noisy car environments with $N = 100$.

In general, Figs. 4–6 show that the normalization block normally amplifies the cepstral coefficients over the whole que-frency band so that the variance over the entire utterance would be unity. The first cepstral coefficients (C1) in a clean environment have nevertheless been attenuated due to large variation of original C1. For the static coefficients, it can be seen that the mean is removed by a high-pass filter, whereas for the time derivatives, the whole que-frency band has been amplified with some evident peaks that are due to time derivative computation. The proposed normalization approach attempts to compensate the linear regression filter response by emphasizing the frequency components which are attenuated by the regression filter used in time derivative computation. When comparing the different environments, it can be seen that in noise the cepstral coefficients have been amplified with respect to a clean environment. This is due to the fact that in the presence of noise the signal variation is smaller than in a clean environment. Furthermore, regardless of the noise conditions, the higher-order coefficients are always more emphasized than the low-order coefficients.

3.2. Spectral representation of segmental normalized MFCCs

It is also important to study how the spectral representation of feature vectors changes due to segmental normalization. In the following, the Mel-log power spectrograms have been used to display the spectral characteristics of the original and segmental normalized feature vectors. Figs. 7 and 8 show the 24 channels of Mel-log power spectrogram with and without feature vector normalization in clean and noisy environments. The vertical white lines denote the start- and endpoints of the utterance.

The spectrograms presented in Figs. 7 and 8 illustrate the noise robustness of segmental

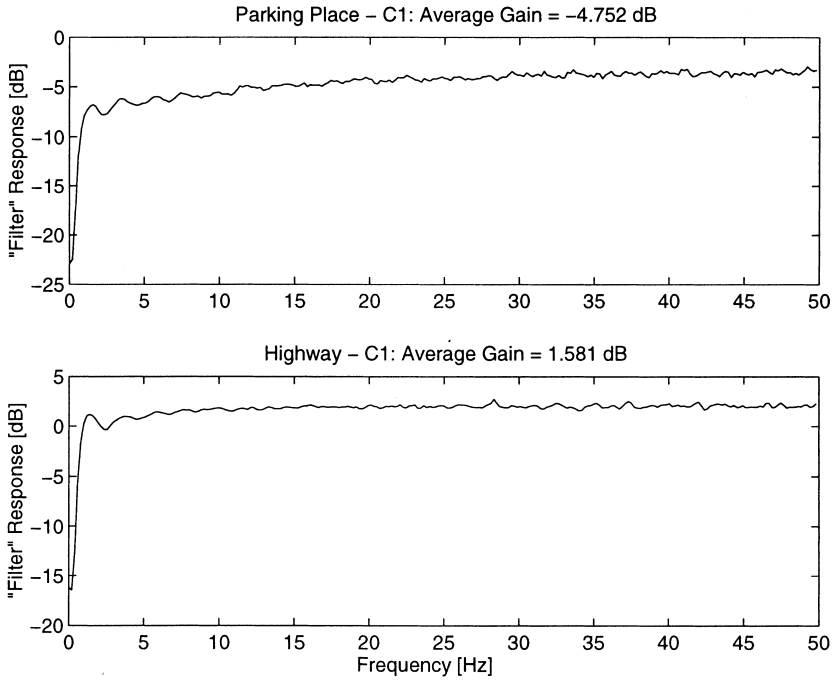

Fig. 4. Segmental normalization filter responses for the first cepstral coefficient in parking place and highway environments.

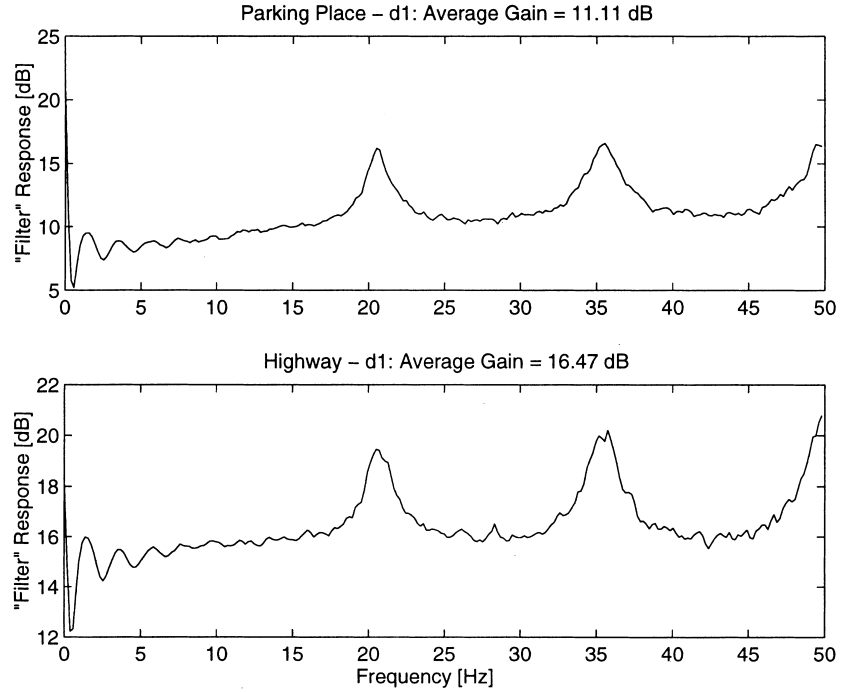

Fig. 5. Segmental normalization filter responses for the first delta coefficient in parking place and highway environments.

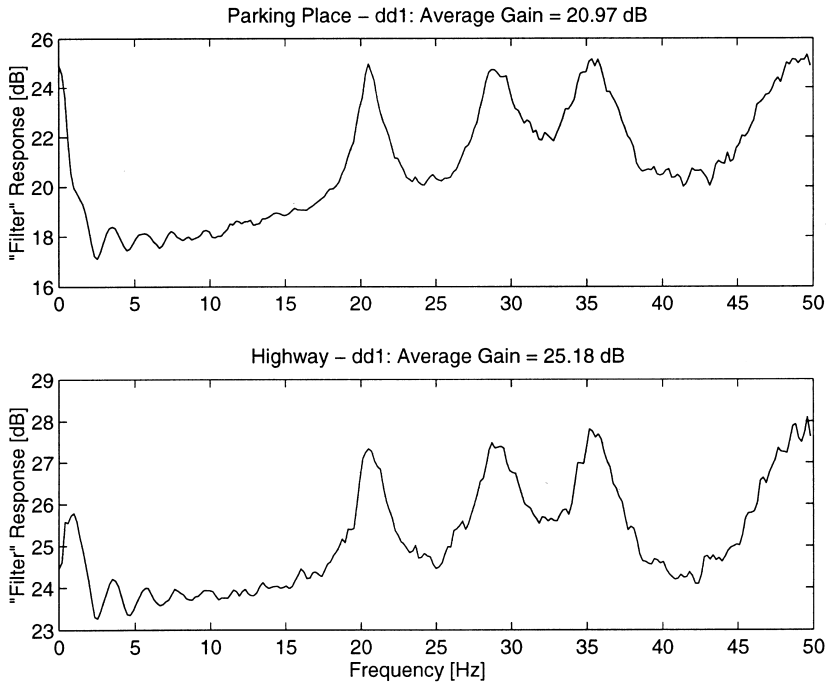

Fig. 6. Segmental normalization filter responses for the first delta–delta coefficient in parking place and highway environments.

normalized MFCCs. By looking into Fig. 7, it can be distinguished that in the case of segmental normalization (lower display) the boundary region between speech and background noise has been blurred, whereas in the case of original MFCCs, there is a distinct border between the speech and noise portions of the utterance. In a clean environment, the variance of background noise is smaller than the variance of the speech regions. Due to the unity variance objective, the noisy portions of the utterance are emphasized. This means that speech and background noise start to remind each other in spectral domain as shown in the lower display of Fig. 7. In spite of the fuzzy boundary region between speech and noise, some speech events can nevertheless be well detected in the segmental normalized MFCC spectrogram.

Fig. 8 shows the spectrograms of the same digit from the same speaker as in Fig. 7, but now spoken in a car environment when driving 120 km/h in highway. By comparing the spectrograms of original MFCCs in the presence of noise and in a clean environment (upper displays in Figs. 7 and 8), it can easily be noticed why the recognition performance decreases in the case of environmental mismatch. Since ambient background noise masks speech very effectively, the clean and noisy Mel-log power spectrograms look very different to each other. When comparing the spectrograms of normalized MFCCs in a clean and noisy case (lower displays in Figs. 7 and 8), it can be noticed that the use of segmental normalization reduces this masking. A visual examination shows that the segmental normalized MFCC spectrograms are more similar to each other than those of original MFCCs. Thus, in the case of environmental mismatch, the use of segmental normalized feature vectors enables a higher recognition performance. Recognition experiments described in the next chapter verify that imprecise boundary between the noise and speech regions does not have any negative effect on the recognition accuracy.

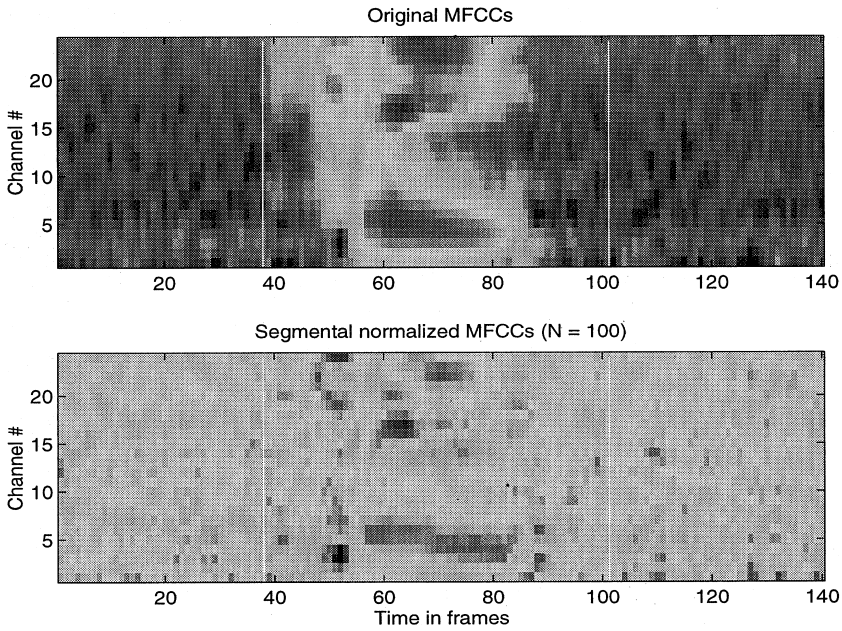

Fig. 7. Mel-log power spectrograms of digit "three" for original and segmental normalized MFCCs in a clean environment.

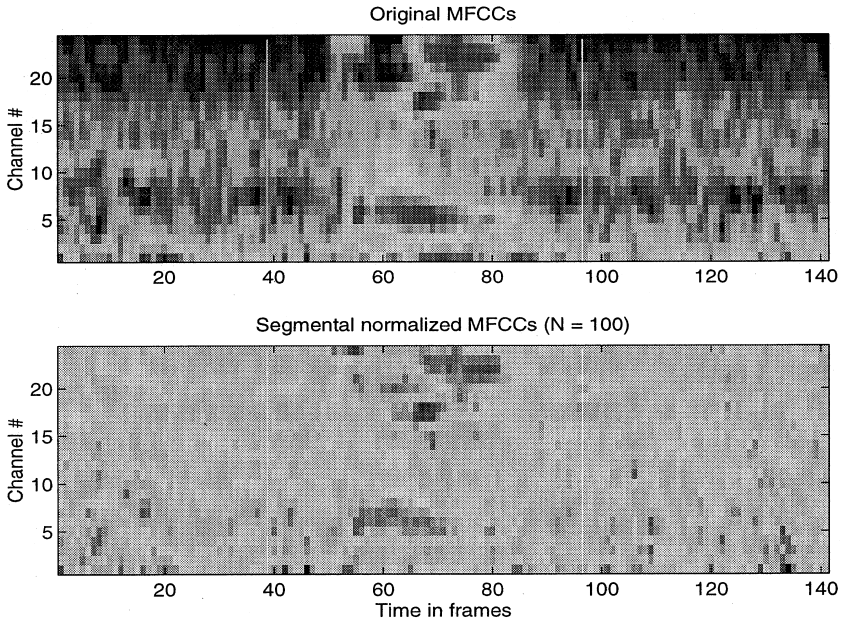

Fig. 8. Mel-log power spectrograms of digit "three" for original and segmental normalized MFCCs in a noisy environment (car noise).

4. Performance evaluation

The viability of the proposed normalization method was tested in various experiments with different noise types and levels. In all experiments, the original feature vectors consisted of the FFT-based MFCCs, log-energy, and their first and second order time-derivatives. Experiments were conducted in speaker-dependent isolated word recognition and in speaker-independent connected digit recognition.

First, the performance of the proposed method was studied in speaker-dependent isolated word recognition against two other widely used compensation techniques, namely, CMN and PMC. Both mean and variance compensation for static and dynamic feature vector components were carried out in the used real-time PMC implementation (Yang et al., 1995, Yang and Haavisto, 1996).

In speaker-independent connected digit recognition, the performance of the proposed normalization scheme was tested against the original MFCCs using a set of multi-environment HMMs. An equal amount of data from all selected noise conditions were used in training and testing. All tests were restricted to a car environment. The performance of segmental normalized cepstral coefficients was compared to the original MFCCs.

4.1. Test databases

The test database for the speaker-dependent experiments consisted of isolated words spoken by five different speakers in a clean office environment. The vocabulary size was 30 words (Finnish first names). In the testing phase, different background noise types were subsequently added to the clean speech waveforms at various SNRs. Training utterances were not used in testing.

In speaker-independent connected digit experiments, the test and training utterances were extracted from the TIDIGITS speech database. Car noise was artificially added to the clean speech waveforms at desired SNRs.

4.2. Recognition system

In speaker-dependent isolated word recognition experiments, for each speaker, speaker-dependent, single mixture, state duration constrained HMMs (Laurila, 1997) were estimated from a *single* training utterance spoken in a *noise-free* environment. An automatic endpointing algorithm based on frame powers and zero crossings was used to determine the starting and ending points of the training utterances. Each training utterance was uniformly segmented into states and the HMM mean vectors were computed over all those feature vectors which were assigned to a certain state. Due to the lack of training data, in the experiments with original feature vectors, all HMMs shared the same diagonal covariance matrix (grand variance) estimated from all training utterances, whereas in the case of segmental normalized MFCCs, a unity covariance matrix was used in all states. During recognition, background noise was modelled with a garbage modelling technique presented by Bourlard et al. (1994).

Speaker-independent recognition system was similar to speaker-dependent one. The only difference was that the single and multi mixture digit HMMs were estimated from a large amount of data according to the Maximum Likelihood principle.

4.3. Speaker-dependent isolated word recognition in various noise conditions

The performance of the proposed segmental normalization technique was evaluated in speaker-dependent isolated word recognition against the baseline system, CMN, SCMN and PMC methods in different noise conditions. In the testing phase, 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.

When using the segmental normalized feature vectors, the normalization segment length was set to 100. SCMN was only used in the experiments with car noise in order to demonstrate what happens if the division by standard deviation is not carried out. No normalization or compensation schemes were used in the baseline tests. The PMC

Table 1
Recognition performance in car noise

SNR	Baseline	CMN	PMC	SCMN	SG_NORM
Clean	96.9	96.8	96.9	97.0	97.5
5	95.3	95.6	95.5	95.8	96.3
0	94.0	93.5	94.8	94.4	96.1
−5	89.3	87.7	91.6	89.6	94.6
−10	70.1	74.5	83.3	74.9	91.2

and CMN techniques relied on the VAD (Freeman et al., 1989). 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, an adaptive single mixture noise model was updated every frame when VAD did not detect speech. Only in the presence of impulsive machine gun noise, possibly due to poor classification accuracy of VAD, PMC could not improve the recognition accuracy. Otherwise PMC performed much better than CMN. However, the performances of both CMN and PMC were still rather poor at lower SNRs. Results are summarized in Tables 1–4 (word recognition percentage). In Table 1, the row "clean" corresponds to the experiment in which both training and testing were carried out in the same noise-free environment.

Irrespective of noise type or level, the proposed segmental normalization approach clearly outperformed PMC, CMN, and baseline results. For example, at −10 dB SNR car noise and at +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. By studying the results with and without gain normalization in Table 1, we can clearly observe the importance of variance compensation.

4.4. Tests with different microphones

In the last set of speaker-dependent experiments, the effect of different microphones in training and testing was studied. Test settings were the same as in the experiments described in Section 4.3. Noise-free test utterances (isolated names) were now spoken by six different speakers, and each of them had a different vocabulary (vocabulary size was still 30 words). 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 microphone (OC);
- CMP-202 "Fico" – Multimedia 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. Training and testing utterances were separated from each other. Experiments were

Table 2
Recognition performance in babble noise

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

Table 3
Recognition performance in music

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

Table 4
Recognition performance in machine gun noise

SNR	Baseline	CMN	PMC	SG_NORM
10	95.2	95.3	95.2	96.2
5	93.6	92.4	93.9	95.8
0	88.9	88.1	89.4	94.6
−5	83.7	83.9	83.6	90.9
−10	77.9	78.4	76.9	84.7

Table 6
Recognition accuracy with different microphones using the segmental normalized 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**

restricted in noise-free conditions. Tables 5 and 6 show the recognition results with original and segmental normalized MFCCs. The recognition accuracy in the matched microphone cases are located on the diagonal of the tables (in bold).

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. These rates are regarded as the baseline results. Table 6 shows that with segmental feature vector normalization the average matched microphone recognition rate was 99.4% and the average mismatch microphone recognition rate was 99.5%, which actually means a surprising decrease of 15% in the error rate as compared to the matched case. When the average microphone mismatch rates for the original MFCCs and segmental normalized MFCCs are compared, one can notice 75% reduction of the error rate due to normalization.

4.5. Speaker-independent connected digit recognition with multi-environment HMMs

Finally, a speaker-independent connected digit recognition experiment was carried out. In this experiment, multi-environment digit models were trained using the male training part of the TIDIGITS database. Noise recorded in a moving car on a highway was added to the clean speech files at two different SNRs. Training was done utilizing all training set utterances, both clean and corrupted ones. Each digit was modelled with a 14-emitting-state HMM, mixture counts ranging from 1 to 3. For each mixture count, two separate model sets were estimated, one with conventional MFCCs and one with segmental feature vector normalized MFCCs.

The test set consisted of all 7-digit utterances taken from the male test part of the TIDIGITS database. Table 7 summarizes the string level recognition rates at different SNRs. One can notice that the segmental feature vector normalization scheme resulted in an improved performance in all cases. The average error rate reduction was over 16% compared to the standard MFCC case.

As shown in Table 7, the highest error rate reduction in the single mixture case was obtained in the most noisy condition. However, as the mixture count increased the SNR = 0 dB environment achieved the highest error rate reductions. In addition, the average reduction of the error rates increased as the mixture count got higher. This phenomenon was not the objective of the proposed

Table 5
Recognition accuracy with different microphones using the standard MFCC representation

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 7
Speaker-independent connected-digit recognition test

Model set/environment	Baseline	SG_NORM	Error rate reduction
1 mixture case, clean	85.53	86.50	6.7%
1 mixture case, SNR = 0	85.18	85.53	2.4%
1 mixture case, SNR = −10	63.25	71.38	22.1%
2 mixture case, clean	88.46	89.27	7.0%
2 mixture case, SNR = 0	88.44	90.41	17.0%
2 mixture case, SNR = −10	75.12	80.49	21.6%
3 mixture case, clean	88.94	90.89	17.6%
3 mixture case, SNR = 0	89.41	93.01	34.0%
3 mixture case, SNR = −10	78.21	82.28	18.7%

normalization technique. By the normalization technique we wanted to get the single mixture case to perform well in speaker-dependent case, and possibly also in the speaker-independent case. According to the recognition experiments in speaker-independent case, it seems to be quite the opposite: the smallest performance gain is achieved in the single mixture case.

Even though we could decrease the error rates with the proposed normalization technique, we are not fully satisfied with the observations in the speaker-independent connected digit recognition case. Perhaps speaker-variability and co-articulation effects together make the proposed technique less effective and to work in a way that we currently cannot explain well.

5. Conclusions

This paper proposes a segmental feature vector normalization technique for noise robust speech recognition. In the proposed technique, the MFCC feature vectors are normalized to have a zero mean and unit variance over a sliding finite length normalization segment. One of the main objectives of the proposed approach is to provide environment-independent parameter statistics in all noise conditions. Indeed, spectral analysis indicates that segmental normalization makes the spectrograms of clean and noisy utterances to be more similar than in the case of original MFCCs. In the performed speaker-dependent experiments, substantial performance improvements were achieved. With all noise types and levels, the proposed approach provided a better recognition performance than what were achieved in the baseline, CMN and PMC experiments. It was also observed that the segmental normalization technique is capable of significantly decreasing the performance degradation due to microphone mismatch.

The proposed normalization technique was also found to increase the recognition accuracy in a multi-environment speaker-independent connected digit recognition task. However, we observed somewhat strange behavior and the exact effects of the segmental normalization are difficult to understand in this more complex task. We are currently looking for more sophisticated and enhanced solutions for improving the performance of normalization in speaker-independent speech recognition.

Acknowledgements

The authors would like to thank David Bye at Nokia Mobile Phones UK in Camberley for valuable comments and for providing the software routines to display the Mel-log power spectrograms.

References

Acero, A., Stern R.M., 1992. Cepstral normalization for robust speech recognition. Proceedings of the Speech Processing in Adverse Conditions, pp. 89–92.

Bourlard, H., D'hoore, B., Boite, J.-M., 1994. Optimizing recognition and rejection performance in wordspotting systems, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Vol. 1, pp. 373–376.

Cook, G.D., Christie, J.D., Clarkson, P.R., Hochberg, M.M., Logan, B.T., Robinson, A.J., 1996. Real-time recognition of broadcast radio speech, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Vol. 1, pp. 141–144.

Freeman, D.K., Cosier, G., Southcott, C.B., Boyd, I., 1989. The voice activity detector for the Pan-European digital cellular mobile telephone service, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Vol. 1, pp. 369–372.

Gales, M.J.F., Young, S., 1993. Cepstral parameter compensation for HMM recognition in noise, Speech Communication 12 (3), 231–239.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

Laurila, K., 1997. Noise robust speech recognition with state duration constraints, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Vol. 2, pp. 871–874.

Lockwood, P., Boudy, J., 1992. Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. Speech Communication 11 (2–3), 215–228.

Openshaw, J.P., Mason, J.S., 1994. On the limitations of cepstral features in noise, Proceedings of the International Conference of Acoustics and Speech, Signal Processing, Vol. 2, pp. II-49–II-52.

Paliwal, K.K., 1990. A study of LSF representation for speaker-dependent and speaker-independent HMM-based speech recognition systems, Proceedings of the International Con-

ference of Acoustics, Speech and Signal Processing, Vol. 2, pp. 801–804.

Riis, S.K., Krogh, A., 1996. Joint estimation of parameters in hidden neural networks, Proceedings of the IEEE Nordic Signal Processing Symposium, pp. 431–434.

Rosenberg, A., Lee, C.-H., Soong, F., 1994. Cepstral channel normalization techniques for HMM-based speaker verification, Proceedings of the International Conference of Spoken Language Processing, Vol. 4, pp. 1835–1838.

Tibrewala, S., Hermansky, H., 1997. Multi-band and adaptation approaches to robust speech recognition, Proceedings of the EUROSPEECH, Vol. 5, pp. 2619–2622.

van Compernolle, D., Claes T., 1996. SNR-normalisation for robust speech recognition, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Vol. 1, pp. 331–334.

Yang, R., Haavisto, P., 1996. An improved noise compensation algorithm for speech recognition in noise, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Vol. 1, pp. 49–52.

Yang, R., Majaniemi, M., Haavisto, P., 1995. Dynamic parameter compensation for speech recognition in noise, Proceedings of the EUROSPEECH, Vol. 1, pp. 469–472.