Spectral Normalisation MFCC Derived Features for Robust Speech Recognition

Carlos S. Lima (1), Adriano C. Tavares (1), Carlos A. Silva (1), Jorge F. Oliveira (2)

(1) Department of Industrial Electronics of University of Minho, Campus de Azurém, Guimarães, Portugal
{carlos.lima, adriano.tavares, carlos.silva}@dei.uminho.pt

(2) Department of Electrical Engineering, Polytechnic Institute of Leiria, Leiria, Portugal
oliveira@estg.ipleiria.pt

Abstract

This paper presents a method for extracting MFCC parameters from a normalised power spectrum density. The underlined spectral normalisation method is based on the fact that the speech regions with less energy need more robustness, since in these regions the noise is more dominant, thus the speech is more corrupted. Less energy speech regions contain usually sounds of unvoiced nature where are included nearly half of the consonants, and are by nature the least reliable ones due to the effective noise presence even when the speech is acquired under controlled conditions. This spectral normalisation was tested under additive artificial white noise in an Isolated Speech Recogniser and showed very promising results [1].

It is well known that concerned to speech representation, MFCC parameters appear to be more effective than power spectrum based features. This paper shows how the cepstral speech representation can take advantage of the above-referred spectral normalisation and shows some results in the continuous speech recognition paradigm in clean and artificial noise conditions.

1. Introduction

Noise robustness can be accomplished either at the feature representation level using robust parameterisation or at the model compensation level. Some approaches maintain that the corrupting noise is by nature unknown, thus it is meaningless trying to compensate for it. Therefore, the search for a robust speech representation that diminishes the distortions caused by the environment seems to be the most promising solution to deal with noise conditions. However, in spite of the effort dedicated in these last years in the robust parameterisation field, conceiving systems with acceptable performance in environments for which they were not trained has been far too difficult.

In [1] it is argued that a proper spectral normalization, which concentrates essentially on the speech regions of less energy, could improve significantly the robustness of speech recognition systems when operating under additive noise conditions. Spectral regions with small energy would need a large degree of noise robustness since, assuming that the noise is speech independent, they are more corrupted. The spectral regions of small energies usually correspond to unvoiced sounds regions, which are spectrally not very well defined. Roughly speaking nearly half of the consonants can be classified as unvoiced, while the other half and the vowels are generally classified as voiced. Generally the importance of the vowels in classification and representation of written text is very low; however, most practical automatic speech recognition systems rely heavily on vowel recognition to achieve high performance, forgetting the speech regions of small energy, which perhaps contains the most important degraded information regarding to speech recognition tasks.

However, speech representation motivated by the human auditory system knowledge has been the approach more successful used for robust speech representation. This paper proposes to incorporate the spectral normalisation suggested in [1] in the MFCC parameters extraction, in order to take advantage of both the effectiveness of the MFCC speech representation and the additive noise robustness of the spectral normalisation. In order to join these two potential advantages we propose a minor change in the MFCC extraction scheme, which consists in normalising the mel-scale filter bank outputs according to the algorithm proposed in [1]. We call this technique SNMFCC given that the MFCC parameters are extracted from a Spectral Normalisation instead of from the conventional power spectrum density. This paper also proposes a method for compensating additive noise distortions, which improves the performance under additive artificial noise. This compensation is performed in the spectral normalization domain thus before the MFCC parameters computation. Hence the effectiveness of the spectral normalisation can be complemented by the good behaviour of the MFCC parameters concerned to speech parameterisation.

2. Baseline Spectral Normalisation

The distribution of the amplitudes of the relative spectral energy seems to be more useful for speech classification than its absolute counterpart, since this last one becomes very dependent on the speech level. This is undesirable given that we are only interested in the classification of the speech message independently of the speech level. The above comment, regarding to the distribution of the relative energy suggests that a feature extraction method based on a mathematical division can be adequate in order to emphasize the relative spectral variations relatively to its absolute counterpart. Concentrating on unvoiced speech segments, which can be roughly characterized by white noise we can make some assumptions, for instance the signal power can be considered as a constant by considering the segment large enough and the process stationary in the segment duration. This fact is a direct consequence of the variance of the sum of k random variables independents and identically distributed is reduced approximately by $1/k$. This result is well known in the estimation theory and was used by Bartlett to reduce the variance of the periodogram. Therefore, concerned to the unvoiced speech regions, the energy of a speech segment, being almost constant can be used for extracting only the spectral energies relative to the speech energy.

Hence, the proposed baseline normalisation process consists in a division of the frequency band in sub-bands given that usually a very fine detail in frequency is not required for western languages speech recognition applications. However, if the intonation is relevant, which occurs for example in conversational speech, this approach must be reconsidered. This issue is not however addressed in this paper.

The features extraction method is based on the power spectral density components and consists in dividing the speech power inside each sub-band by the total short-time speech power. The power in each sub-band is obtained by summing the power spectrum components inside the sub-band. All the sub-bands have the same number of spectral components and no one is shared by different sub-bands, thus avoiding increases of statistical dependence between sub-bands (feature components). This kind of normalisation seems to be also adequate for dealing with additive distortions since the numerator and denominator of the features are both increased, though by different values, however this fact contributes for stabilising the feature values, which means increasing the robustness.

To best understand this reasoning, consider S_i denoting the speech power in sub-band i and S denoting the short time speech signal power of the considered segment. Similarly, let N_i and N denote the power of the interfering noise in sub-band i and the short time noise power, respectively. So, the i^{th} component of the observation vector for clean speech is given by

$$c_i = \frac{S_i}{S} \qquad (1)$$

Similarly for noisy speech the next equation holds

$$c_{in} = \frac{S_i + N_i}{S + N} \qquad (2)$$

where the index n stands for noisy speech. Equations (1) and (2) are computed in the same way without concerning to the noise existence, so they can be viewed as the same equation. The denominators of equations (1) and (2) represent respectively the power of the speech segment in clean and noisy conditions and can be both computed by summing all the components of the power spectrum density.

Figure 1 shows the clean speech and noisy speech spectral power normalisation features for 240 ms of the word “zero” where each sub-band has 16 power spectral components. The SNR is 0 dB.

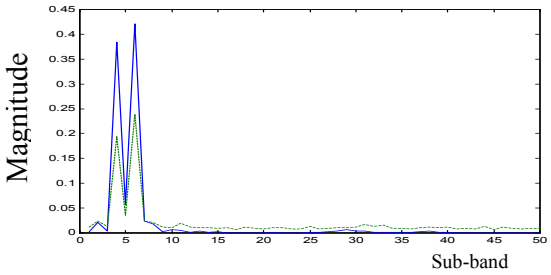

Figure 1. White noise effect in the power spectrum density normalization domain in the beginning of digit “zero”. Dashed line represents noisy speech features.

If the interfering noise has white noise characteristics the environment will shift the clean speech vector by a noise dependent vector $C_i(N)$, which can be calculated by subtracting equation (1) from equation (2).

If the noise is stationary then its short time power equals its long time power. Note that this does not occur for the speech due to its non-stationary property, but as an approximation we will consider that the short time speech signal power equals the long time speech signal power. This seems to be approximately true for unvoiced speech segments, where we want to concentrate in order to try increasing the noise immunity. Under this constraint, S and N can be related by the signal to noise ratio (SNR). Therefore the next expression holds

$$S + N = S\left(1 + \frac{1}{10^{\frac{SNR}{10}}}\right) \qquad (3)$$

Let l, the number of components in each sub-band and L the FFT length. Then N and N_i, considering flat noise

spectrum, are related by the quotient l/L. By using these considerations, the calculation of the shift vector imposed by the environment to the observed vector component *i* is noise dependent and is accomplished by subtracting equation (1) from equation (2)

$$C_i(N) = \frac{S_i + N_i}{S\left(1 + \frac{1}{10^{\frac{SNR}{10}}}\right)} - \frac{S_i}{S} = \frac{S_i - kS_i}{kS} + \frac{N_i}{kS}$$

$$= \left(\frac{S_i}{S} - \frac{N_i}{N}\right)\frac{1-k}{k} \qquad (4)$$

where k is given by

$$k = 1 + \frac{1}{10^{\frac{SNR}{10}}} \qquad (5)$$

and in terms of mean the next expression holds

$$N_i = \frac{l}{L}N \qquad (6)$$

Equation (4) shows that if the speech has a flat power spectrum density, which roughly occurs outside the voiced regions, the means of $C_i(N)$ become null as Si/S equals l/L. Thus, this normalisation process becomes optimal in the sense that the environment does not affect the means of the speech features, while the variances are strongly reduced by the intrinsic mechanism of speech energy normalization, which consists of the mathematical division of the power in each sub-band by the short-time power. This means that this normalisation procedure provides some noise robustness to unvoiced speech segments, where neither the speech nor the noise are spectrally well defined. As a conclusion of this section we can state that the advantages of the proposed baseline spectral normalisation procedure are:

1) As a white noise process does not corrupt in terms of means other white noise process, which is shown by equation (4), this parameterization increases the robustness of the unvoiced speech regions, which are the most corruptible ones, becoming the robustness of the recognizer less independent of the voiced regions where are included the vowels. This approach can also be useful concerned to conversational speech where usually the vowels are reduced or even deleted, thus a method that relies on the consonants to increase the recognition accuracy can be adequate not only for noisy recognition but also for spontaneous speech recognition.

2) The normalisation helps to extract only the relative variations of the spectrum instead of the absolute variations, which can help to classify linguistic messages independently of the speech or background level.

However, the main drawback associated with the proposed baseline normalization method is that it has not been possible to develop an algorithm that can help to preserve the spectral peaks structure against additive distortions. As the speech segments frequently contain parts of voiced and unvoiced sounds a post-processing algorithm which goal is to restore the spectral peaks structure of the speech spectrum is needed. This algorithm is described in the two following sections.

3. Accounting for Additive Distortions in the Power Spectral Density Domain

The environmental distortions that frequently occurs in speech recognition applications are frequently considered of two different nature; additive and convolutional. Convolutional noise is mainly due to the frequency response of communication channels and the different frequency response of the microphone used for collecting the training and testing speech. By considering the frequency response of these components sufficiently smooth, which roughly occurs for the most common cases, we can assume that its effect on the peak structure of the speech spectrum is not very significant.

Additive distortions are essentially due to the background, which includes competitive speech, or noise induced in the communication channel by both external sources such as electromagnetic induced noise, or internal sources such as crosstalk. The additive noise effect on speech has been studied with some detail and one of its most undesirable effects is the changing on the peaks structure of the speech spectrum. Hence we propose trying to restore the peak structure of the degraded speech spectrum by assuming that the majority of these changes are due to additive distortions. A second type of spectral normalization also independent on the corrupting noise and based on certain characteristics of the baseline spectral normalization can constitute a reasonable solution to deal with additive distortions in general.

Figure 1 shows that the noise effect, in the proposed power spectral baseline normalisation domain, is raising the "flat" spectral zones while the "peaked" spectral ones are "flatten". In fact equation (2) (in noisy conditions) shows that, for sub-bands with high speech power, as the amount of noise in the sub-band is much smaller than the total amount of noise, the speech features in that regions are decreased proportionally to the amount of contaminating noise. For sub-bands with small speech power the opposite happens, given that the sum of all the coefficients extracted in each segment is unitary. As the spectral flattening is proportional to the amount of contaminating noise, for low signal to noise ratios the "peaked" spectral regions almost disappear,

which is the main origin of degradation in performance under noisy conditions.

The main goal of a robust features extraction method is providing robustness against noise or other sources of variability by ignoring its presence. Although the noise can be compensated, the effectiveness of this approach becomes very dependent on the accuracy of the noise estimate, which is a very hard task in practical situations. Hence our main goal was searching for a robust feature extraction process, which must be ideally independent of the noise level or characteristics, although the proposed baseline normalisation assumes a wide band additive noise for maximal performance. More details can be found in [1].

In this context we propose the following two steps approach:

1) For task uniformity in clean and in noisy conditions the clean database must be considered lightly contaminated. Trying to clean completely the database, which can be viewed as another kind of normalisation, represents a procedure compatible with the noise compensation paradigm, however if the procedure is not particularised for any kind of noise, it can be used without concerning to the noise existence. Hence, under noisy conditions the features extraction method can compensate for the noise existence taking into account the noise level, which can be estimated in a frame-by-frame basis, becoming the procedure compatible with real time applications. We propose estimating the noise power in each segment, which can be viewed as a second normalisation factor (the first normalisation factor is behind the normalisation procedure in the baseline system) by taking the value of the lowest component of the power spectrum density in each speech frame. Our reasoning is based on the heuristic rule that the smaller spectral components have minor speech dependence. By considering wide band noise all the spectral components are roughly equally dependents on the noise process. Therefore a speech spectrum normalization, which preserves as most as possible the speech spectral content must be based on the smaller spectral component of the speech spectrum.

2) We propose alleviating the noise effect by using the estimated noise level (smaller spectral component) and taking into consideration the kind of distortion caused by the noise in the spectral normalisation of the baseline system, that is taking into account that the "peaked" spectral regions are "flattened" and the "flat" spectral regions are "raised" by the noise effect. This type of procedure presumes an efficient peak detector.

An efficient peak detector must be able to distinguish peaks of voiced nature (pitch) from weak peaks occurring in the speech regions of low energy, where the baseline system is efficient concerned to the attenuation of the additive noise effect. The upper part of figure 2 shows strong peaks due to the pitch, which can be classified as peaks by the peak detector, given that they occur in voiced regions just the regions "forgotten" by the baseline system, while the lower part of the figure shows weak peaks (right side of the figure) proceeding from unvoiced regions that must be ignored. This peak classification suggests the use of thresholds, where the key question is how to calculate the threshold level?

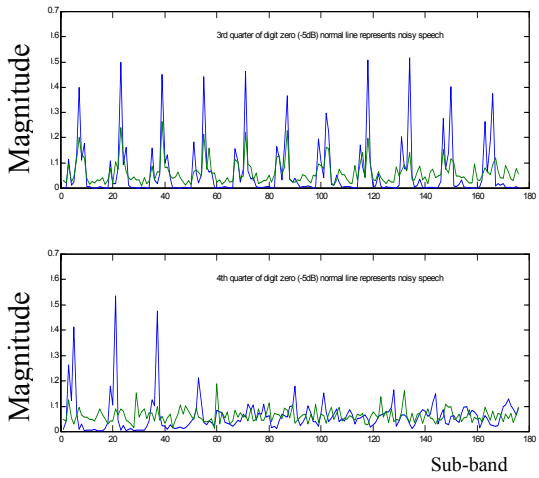

Figure 2. White noise effect in the power spectrum density normalization domain in a voiced segment (upper part) and in an unvoiced segment (last 2/3 of the lower part of the figure. Dashed line represents noisy speech features.

Based only in practical considerations especially in the inspection of the selected peaks we concluded that roughly speaking a peak which energy is above at least three times the mean of the rest of components in the frame must be classified as a true peak. Otherwise the selected peak must be ignored in order to preserve the benefits of the baseline normalisation on low energy segments.

4. Proposed Noise Compensation

To cope simultaneously with the noise effect on the "peaked" and on the "flat" spectral regions we have to consider two types of compensation procedures, once that the distortions caused by the noise are of different nature for the two types of considered regions.

The "flat" spectral regions are raised by the noise effect, so we suggest subtracting to each component of the observed vector the lowest one, according to the second normalisation procedure. Of course we are implicitly considering wide band noise and the procedure must be improved in the future to account for narrow band noise. To account for the second type of normalisation maintaining however compatibility between the two types of normalisation equations (1) and (2) must be changed respectively so that

$$c_i = \begin{cases} \dfrac{S_i - \min\{S_i\}}{S}, S_i \neq \min\{S_i\} \\ \dfrac{S_i}{S}, otherwise \end{cases} \quad (7)$$

or in noisy situations

$$c_{in} = \begin{cases} \dfrac{S_{in} - \min\{S_{in}\}}{S+N}, S_{in} \neq \min\{S_{in}\} \\ \dfrac{S_{in}}{S+N}, otherwise \end{cases} \quad (8)$$

where S_{in} stands for the noisy speech power contained in sub-band i, thus can be written as $S_{in} = S_i + N_i$. Regarding to equations (7) and (8) it is important to note that they are computed in the same way without concerning to the current background conditions. Thus under noisy conditions we are implicitly computing equation (8) while in clean speech conditions we are implicitly computing equation (7), which means that no knowledge concerning background conditions is needed in practical situations.

For wide band noise distortion, N_i is approximately constant and the mean of the clean speech coefficient equals the mean of the noisy speech coefficient. As in [1] this signifies that a white noise process does not deteriorate in terms of means another white noise process, which means good behaviour of the normalisation process in speech regions characterised by low energy level. It is important to note that some compensation algorithms assume that the compensation of the means has a better contribution to the recognition performance than the compensation of the variances [2]. In the context of the baseline normalisation we have automatic compensation of the means.

The noise compensation in the "peaked" spectral regions is performed by increasing the speech coefficient that was decreased (flattened) by the noise effect. Assuming clean speech (not lightly contaminated speech) equation (1) holds and the speech features are related by

$$\sum_{i=1}^{B} c_i = 1 \quad (9)$$

where B is the number of sub-bands. For a speech frame where n peaks are detected, these peaks have to be increased by a noise dependent factor so that

$$\sum_{j=1}^{n} c_j = 1 - \sum_{\substack{i=1 \\ \forall i \neq j}}^{B} c_i \quad (10)$$

where each c_j was previously decreased as shown by equation (8). Assuming that each spectral sub-band was decreased proportionally to its magnitude, which seems to be true by analysing figures 1 and 2 the noise compensation can be made by computing c_j as follows

$$c_j = \frac{\left(S_j + N_j\right)\left(1 + \dfrac{(B-n)}{S_n}\min\{S_i + N_i\}\right)}{S+N} \quad (11)$$

where S_n is given by

$$S_n = \sum_{j=1}^{n}\left(S_j + N_j\right) \quad (12)$$

Therefore, the energy subtracted in the "flat" spectral regions is restored in the "peaked" zone in order to invert the additive noise effect whereas the sum of all the speech features for each frame is maintained unitary as supposed by the baseline spectral normalisation.

The proposed algorithm for feature extraction works as follows:

First equation (7) or (8) is used to compute all the observed vector components. Then as a post-processing stage equation (11) is used to recalculate the vector components, which are selected as "peaks" by the peak detector. Hence the undesirable effect of the additive noise in the peak structure of the speech spectrum is partially alleviated while a kind of spectral power adaptation is automatically achieved, since the computation of the feature vector components involves a division by the speech power. This division means that the speech features are proportional to the relative energy contained in the sub-band while state-of-the-art feature extraction methods are essentially based on the absolute energy contained in a frequency range. Therefore alternative approaches, which compensates for the difference between speech energy levels in the clean and unknown utterances is needed. Usually a scaling factor, which is an SNR-matching gain, is used. The estimation of this factor can be achieved based on samples of the background noise, training data and a few samples of the noisy test data [3], which is a very hard task in real time applications and needs some initial estimate, which then is updated, based on the unknown input to the system. This approach, however, is not applicable when only noisy signals are available for recognition. A more practical approach is suggested in [4] where the gain contour of the clean signal is estimated from the noisy signal by using HMM's for the gain-normalised clean signals. This algorithm is a little bit computationally expensive, however is beneficial even when similar gain conditions exist in training and test recordings. In any way adapting spectral power is a real challenge in practical situations, though often ignored by the scientific community since regarding the existing databases the gain conditions are similar in clean and noisy speech recordings. The algorithm proposed in this paper tries to adapt spectral power automatically.

5. Markov Models Composition in the Baseline Spectral Normalisation Domain.

Previous section describes a spectral normalization method, which attenuates the additive distortions effect on speech over a baseline spectral normalisation described in section 2. The suggested normalisation method was obtained by only assuming the spectral characteristics of both the corrupting and corrupted processes. In other words the goal was to lessen the noise effect on the speech features given some previous knowledge of the speech spectral properties, namely those concerned to the voiced and unvoiced segments. This method seems to be adequate for the most common practical situations where the noise is not known and can't be accurately estimated. However, concerned to the noise compensation in robust speech recognition, it is frequently common to assume that the noise is additive and can be usually accurately estimated in a segment without speech or by an accurate speech/pause detector which permits to separate signal segments containing only the background. If the noise is known it can be compensated usually by two different ways; compensating the incoming feature vectors or compensating the internal distributions of the HMM's. It is reported in [5] that compensating the internal distributions is superior than compensating the income vectors.

Compensating the internal distributions is achieved by using the HMM composition technique, where the HMM's for noisy speech are derived from the HMM's of the clean speech and the HMM which models the background. Without loss of generality it is frequently assumed that the HMM for the background has only one state, which means stationary environments, however practical situations are frequently characterized by non-stationary environments.

The basic idea of the HMM composition is to recognise concurrent signals simultaneously. Parallel HMMs are used to model the concurrent signals while the composite signal is modelled as a function of their combined outputs. To perform Markov models composition one has to know the composite signal distribution and the statistical model of the corrupting environment. It is usually assumed that the speech and noise are additive in the linear power domain and we will consider stationary noise, thus a single state noise model is sufficient. Mel-frequency cepstral coefficients derived from the baseline spectral normalisation presented in section 2 are used in the recognition system. The front end is shown in figure 3.

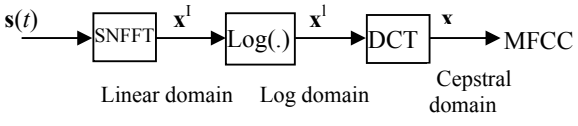

Figure 3. Front-end of the recognition system

Let x^c, x^l and x^I represent the observation vectors in the cepstral, log, and linear domain, respectively. Suppose the signal modelled in the cepstral domain by a Gaussian mixture, thus with density function

$$f(\mathbf{x}) = \sum_{c=1}^{C} p_c G(\mathbf{x}, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{13}$$

where G(.) denotes the Gaussian distribution, with $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ as its mean vector and covariance matrix for the c-th component of the mixture.

The mapping from the cepstral domain to the log domain is the inverse discrete cosine transform, which is a linear transformation represented by

$$\mathbf{x}^l = \mathbf{C}^{-1}\mathbf{x} \tag{14}$$

Then the distribution in the log domain is still a Gaussian mixture, i. e.,

$$f^l(\mathbf{x}^l) = \sum_{c=1}^{C} p_c G(\mathbf{x}^l, \boldsymbol{\mu}_c^l, \boldsymbol{\Sigma}_c^l) \tag{15}$$

with

$$\begin{aligned} \boldsymbol{\mu}_c^l &= \mathbf{C}^{-1}\boldsymbol{\mu}_c \\ \boldsymbol{\Sigma}_c^l &= \mathbf{C}^{-1}\boldsymbol{\Sigma}_c(\mathbf{C}^{-1})^T \end{aligned} \tag{16}$$

When transforming x^l to x^I in the linear domain, it can be shown that the density function in the linear domain is a lognormal mixture i. e.,

$$f^I(\mathbf{x}^I) = \sum_{c=1}^{C} p_c L(\mathbf{x}^I, \boldsymbol{\mu}_c^I, \boldsymbol{\Sigma}_c^I) \tag{17}$$

The distributions parameters are related by the following equations

$$\begin{aligned} \mu_{c,i}^I &= \exp\left(\mu_{c,i}^l + \frac{\sigma_{c,i,i}^l}{2}\right) \\ \sigma_{c,i,j}^I &= \mu_{c,i}^I \mu_{c,j}^I \left(\exp\left(\sigma_{c,i,j}^l\right) - 1\right) \end{aligned} \tag{18}$$

The above mapping process suggests that if the signal can be modelled by a Gaussian mixture in the cepstral domain, then in the linear domain its distribution is a lognormal mixture.

Let f_x^I and f_n^I denote the density functions for the clean speech and the noise in the linear domain,

$$f_x^I(\mathbf{x}^I) = \sum_{c=1}^{C_x} p_c L(\mathbf{x}^I, \boldsymbol{\mu}_{x,c}^I, \boldsymbol{\Sigma}_{x,c}^I) \tag{19}$$

$$f_n^I(\mathbf{n}^I) = \sum_{c=1}^{C_n} e_c L(\mathbf{n}^I, \boldsymbol{\mu}_{n,c}^I, \boldsymbol{\Sigma}_{n,c}^I) \qquad (20)$$

According to the assumption that in the linear domain the speech and noise are additive and considering that they are independent, then Si/S and Ni/N are also independent. If the convolution of two log normal functions is assumed to be approximately log normal as is assumed in the single mixture PMC [2], then the distribution of noisy speech ***y*** will be the convolution of equations (19) and (20), the number of mixture components for noisy speech is $C=C_x \times C_n$ and the density is

$$f_y^I(\mathbf{y}^I) = \sum_{c=1}^{C} h_c L(\mathbf{y}^I, \boldsymbol{\mu}_{y,c}^I, \boldsymbol{\Sigma}_{y,c}^I) \qquad (21)$$

Noisy speech parameters can be derived from equation (4), and for the mean the next expression holds

$$\begin{aligned} E\left\{\frac{S_i + N_i}{S+N}\right\} &= E\left\{\frac{S_i}{S} + C_i(N)\right\} \\ &= E\left\{\frac{S_i}{S}\frac{1-k}{k} - \frac{N_i}{N}\frac{1-k}{k} + \frac{S_i}{S}\right\} \\ &= \frac{1}{k}E\left\{\frac{S_i}{S}\right\} - \frac{N_i}{N}\frac{1-k}{k} \end{aligned} \qquad (22)$$

where k is given in equation (5). The noisy speech means are then

$$\boldsymbol{\mu}_{y,c}^I = \frac{1}{k}\boldsymbol{\mu}_{x,i}^I + \frac{k-1}{k}\boldsymbol{\mu}_{n,j}^I \qquad (23)$$

The covariance matrix of the corrupted process can be similarly calculated as

$$\begin{aligned} \mathrm{var}\left\{\frac{S_i + N_i}{S+N}\right\} &= \mathrm{var}\left\{\frac{S_i}{S} + C_i(N)\right\} \\ &= \mathrm{var}\left\{\frac{1}{k}\frac{Si}{S} + \frac{N_i}{N}\frac{k-1}{k}\right\} \\ &= \frac{1}{k^2}\mathrm{var}\left\{\frac{S_i}{S}\right\} + \left(\frac{k-1}{k}\right)^2 \mathrm{var}\left\{\frac{N_i}{N}\right\} \end{aligned} \qquad (24)$$

The noisy speech covariance matrix can be then calculated as

$$\boldsymbol{\Sigma}_{y,c}^I = \frac{1}{k^2}\boldsymbol{\Sigma}_{x,i}^I + \left(\frac{k-1}{k}\right)^2 \boldsymbol{\Sigma}_{n,j}^I \qquad (25)$$

Therefore, the noise compensation process is straightforward. Given the HMM's for clean speech and noise in the cepstral domain, their model parameters in the linear domain can be calculated using equations (16) and (18). Then compensation of the clean speech model by the noise model in the linear domain according to equations (23) and (25) is performed to get the model for noisy speech. The model parameters in the cepstral domain can be calculated by inverting equation (18) and (16).

6. Experimental Results

This paper suggests two methods for alleviating the noise effect in speech recognition applications. Both methods are based on MFCC parameterisation obtained from a normalised spectrum.

The first method described in sections 2, 3 and 4 tries to alleviate the noise effect mainly based on heuristic rules and on the most known properties of both speech and noise. The algorithm does not assume noise existence and the extraction parameters procedure is optimised for wide band noise. The main drawback of this method can be the need of a peak detector, which is based on a threshold level obtained from inspection and heuristic rules.

The second method described in sections 2 and 5 is the well-known PMC [2] but adapted to the suggested spectral normalisation proposed in section 2.

The purpose of this section is to evaluate comparative results among a baseline system with state-of-the-art MFCC parameters and our first and second methods. The experiments were evaluated in the continuous speech recognition framework by using the HTK and the TIMIT database. The train procedure and data management for all the experiments is described in chapter 3 of the HTK book. Computer generated (speech independent) white noise was added to the clean speech in the noisy experiment set. Table 1 shows the system performance.

Table 1 – Comparative performance of the spectral based multi-normalisation algorithms.

SNR (dB)	Baseline % Error (D, S, I)	Method 1 % Error (D, S, I)	Method 2 % Error (D, S, I)
Clean	0.23 (1, 1, 1)	0.23 (0, 0, 1)	0.23 (1, 0, 1)
15	32.9 (53, 47, 66)	1.3 (2, 1, 3)	1.6 (3, 4, 2)
10	72.7 (132, 241, 256)	5.6 (6, 10, 8)	2.8 (16, 5, 8)
5		16.3 (32, 26, 41)	6.7 (18, 21, 12)

Table 1 shows that method 1 is more effective than the PMC based method even under Gaussian noise for high SNR. The performance degradation of method 1 when compared with method 2 as SNR is decreasing can be due to inaccurate peak detection, which perhaps is becoming worse as the noise is increasing. However method 1 shows an interesting potential concerned to practical applications where the noise is frequently unknown and non-stationary, however more investigation is needed specially concerned to the peak

detector which is too much based on heuristics and inspection of the selected peaks. If the noise is known and stationary the PMC method (method 2) is adequate, however this is not frequently the case in practical applications. One important note is that none of the algorithms was experimented yet under real noise added to speech nor under real noisy speech situations, thus these results are very limitative yet concerned to the baseline spectral normalisation effectiveness. These are the main objectives of the near future developments.

7. References

[1] Lima, C., Almeida, Luís B. and Monteiro, João L. (2002). Improving the Role of Unvoiced Speech Segments by Spectral Normalisation in Robust Speech Recognition. 7th International Conference on Spoken Language Processing (ICSLP'2002).

[2] Galles, M. J. F. and Young, S. J (1993). PMC for Speech Recognition in Additive and Convolutional Noise. Technical Report 154.

[3] Beattie, V. L. and Young, S. J. (1992). Hidden Markov Model State-Based Noise Cancelation. Technical Report (TR 92). Cambridge University Engineering Department.

[4] Ephraim, Y. (1992). Gain-Adapted Hidden Markov Models for Recognition of Clean and Noisy Speech. IEEE Transactions on Signal Processing, VOL. 40 pp. 1303-1316.

[5] Moreno, P. J., (1996). Speech Recognition in Noisy Environments. Ph. D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University.