SPECTRAL BI-NORMALISATION FOR SPEECH RECOGNITION IN ADDITIVE NOISE

C. S. Lima[1], J. F. Oliveira [2]
[1]Department of Industrial Electronics, Universidade do Minho, Guimarães, Portugal
[2]Department of Electrical Engineering, Instituto Politécnico de Leiria, Leiria, Portugal

Abstract: **The changing on peaks structure of the speech spectrum is perhaps the most important cause of degradation of speech recognition systems under adverse conditions. Another drawback concerned to the additive noise effect occurs on the flat spectral zones which are usually raised proportionally to the noise level. These combined effects on both the peaked and the flat spectral zones can be alleviated by trying to restore its original structure, which assumes noise knowledge. However, the random nature and the variability of the noise, the difficulty in discriminating speech pauses, among others, discourage the use of noise estimates as the basis of robust speech recognition algorithms. Alternative approaches based on normalisation procedures become very promising since the noise effect can be alleviated without any knowledge regarding to its existence. This paper suggests a spectral normalisation that though being different can be viewed as a noise estimation procedure in a frame by frame basis, so assuming the clean database as lightly corrupted. This speech normalisation is used to restore the normalised speech spectrum. This normalised spectrum is then re-normalised by a baseline spectrum normalisation method, which concentrates essentially in the speech regions of small energy, since in these regions the noise is more dominant, so they require a better degree of robustness.**

I. INTRODUCTION

In [1] it is argued that a proper spectral normalisation, 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. From a theoretical point of view, the spectral regions with small energy would need more noise robustness, given that for the same noise level 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. Consequently, the spectral regions which contains higher speech energy seems to be usually more important in speech recognition under difficult conditions once they are generally less corrupted. On the other hand, the spectral regions with small energy are more corrupted, thus they need a larger degree of robustness.

Others authors [2] have also given an increasing importance to the spectral regions of small energy of the speech signal, although by using alternative approaches.

The algorithm proposed in [1] does not take into consideration the properties of the voiced speech regions, which are usually characterised by “peaked” spectral zones. These portions of spectrum are flattening, as the noise becomes more and more dominant which degrades the system performance.

The algorithm proposed in [3] tries to cope with this limitation by restoring partially both the original spectral “peaks” and the flat spectral regions where the signal power is increased by the wide band noise effect. This approach assumes the clean database lightly contaminated and the noise power is estimated in a frame-by-frame basis by the lowest power of all the sub-bands in each segment. The algorithm does not assume noise existence, in the sense that the features are extracted exactly in the same way in both noisy and noise free conditions. One drawback associated with this algorithm is concerned to the noise estimate which includes a significant amount of speech characteristics that is proportional to the number of spectral components that constitute a sub-band. This can mean that to many speech characteristics can be disregarded in the restoration of the clean speech normalised features. Another drawback of the algorithm proposed in [3] is that the spectral peaks classification is based on heuristics, which is obviously undesirable. In order to overcome these drawbacks the algorithm proposed in this paper differs from the algorithm proposed in [3] essentially in the following aspect:

The frame by frame spectral normalisation is done before the baseline normalisation instead of after it, assuring that the spectrum that will be processed by the baseline spectral normalisation is always the normalised spectrum (by the small spectral component), which is not very dependent on the noise level.

The results show a significant improvement in performance when compared with the baseline method when used alone [1] and an interesting improvement in performance when compared with the algorithm proposed in [3].

II. Baseline spectral normalisation

The baseline spectral normalisation defined in [1] is motivated by the fact that the additive noise is not a narrow band noise, thus its spectrum is reasonably dispersed in frequency. Additionally a mechanism adequate to dealing with non-stationary additive noise situations, which frequently occurs in practical situations, is needed. One solution can be trying to extract the distribution of the speech energy along the spectrum, normalised by the total energy of the speech within the segment. Therefore noise variations can be attenuated once that which is really measured is the relative and not the absolute distribution of the spectral energy of the speech signal.

The 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. The 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 components of the power spectral density inside the sub-band. All the sub-bands have the same number of spectral components and any spectral component is shared by different sub-bands, thus avoiding increases of statistical dependence between sub-bands (feature components). The background noise contributes simultaneously to increase the sub-band and total power, which contributes for stabilising the amplitudes of the feature vectors.

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 noise in sub-band i and the short time noise power, respectively. So, the i^{th} component of the observation vector for clean and noisy speech are given respectively by

$$c_i = \frac{S_i}{S}\ ,\quad c_i = \frac{S_i + N_i}{S + N} \qquad (1)$$

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.

If the noise is stationary then its short time power equals its long time power. Note that this is not true 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. Under this constraint, S and N can be related by the signal to noise ratio (SNR). Therefore the next expression holds

$$S + N = S\left(1 + \frac{1}{10^{\frac{SNR}{10}}}\right) \qquad (2)$$

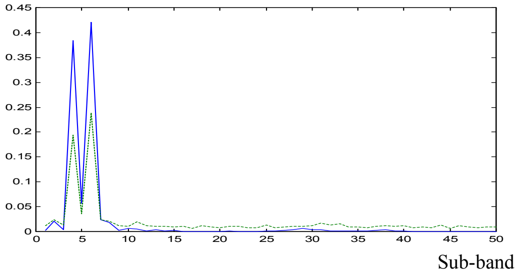

Figure 1. White noise effect in the power spectrum density normalization domain in the beginning of digit "zero". Dashed line represents noisy speech features.

If the 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 equations (1).

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 is accomplished by subtracting equations (1) and becomes [1]

$$C_i(N) = \left(\frac{S_i}{S} - \frac{l}{L}\right)\frac{1-k}{k}, \quad k = 1 + \frac{1}{10^{\frac{SNR}{10}}} \qquad (3)$$

Equation (3) shows that if the speech has a flat power spectrum density, the means of $C_i(N)$ become null as S_i/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. 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. More details can be found in [1]

III. Additive white noise effect and pre-processing approach

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 (1) in noisy conditions (equation shown on the right) 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 compensation process 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.

2) The estimated noise level, which really constitutes a spectral normalisation by the smallest spectral component. This speech component, which has small significance and is proportional to the amount of noise must be used to alleviate the noise effect. Then the baseline spectral normalisation algorithm [1] can be more efficient since the noise effect was *a priori* reduced.

IV. Proposed noise compensation

To cope with the additive noise effect we propose estimating the noise power in each segment, which can be viewed as a secondary normalisation procedure (the first normalisation procedure is behind the normalisation proposed in the baseline system [1]) by taking the value of the lowest component of the power spectrum density in each speech frame.

We propose alleviating the noise effect by subtracting the estimated noise level from all the others components of the feature vector. Therefore the power spectral components of the speech must be changed so that

$$c_i = \begin{cases} P_i - \min\{P_i\}, P_i \neq \min\{P_i\} \\ P_i, otherwise \end{cases} \qquad (4)$$

where P_i denotes the amplitude of the i^{th} component of the power spectral component of the speech, and c_i denotes the i^{th} component of the normalised spectrum (observation vector) that will be processed by the baseline spectral normalisation algorithm proposed in [1]. The spectral normalisation procedure described by equation (4) reduces clearly the noise effect since a factor (lower spectral component in each segment) that is proportional to the noise level is subtracted from all the others spectral components. Additionally the speech characteristics described by the smallest spectral component are maintained since this component is included in the observation vector. However these mathematical operations involving all the spectral components can increase the statistical dependence among them, which is undesirable regarding to the HMM modelling. In this context the baseline spectral normalisation procedure helps to decorrelate the data since the data are grouped and processed inside the group independently of the data inside the other groups.

Therefore considering wide band noise its effect is reduced in terms of means. It is obvious from equation (1) that the variance effect is also reduced by the baseline normalisation procedure once that each observation is divided by the power of the speech segment.

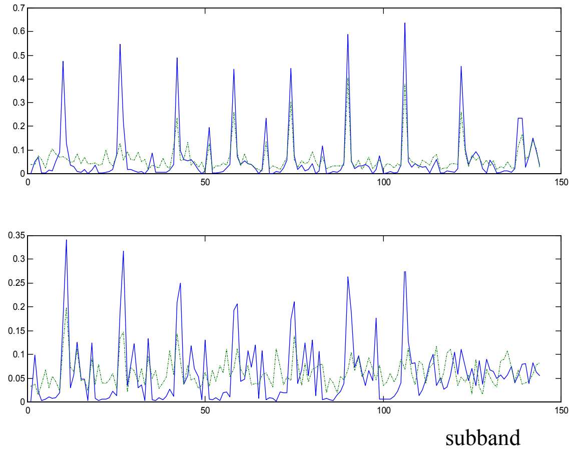

Figure 2. Spectral speech structure recovered by the algorithm proposed in [3] for the first half of the word "eight" at an SNR of 0 dB. Normal line stands for clean speech.

This *a priori* noise effect attenuation obtained by spectral normalisation in each frame shows better effectiveness than the *a posteriori* noise effect attenuation described in [3] as can be observed by comparing figure 2 and figure 3. It is clear that in figure 3 the recovered peak structure is more closed to the peak structure of the clean speech than the recovered peak structure in figure 2.

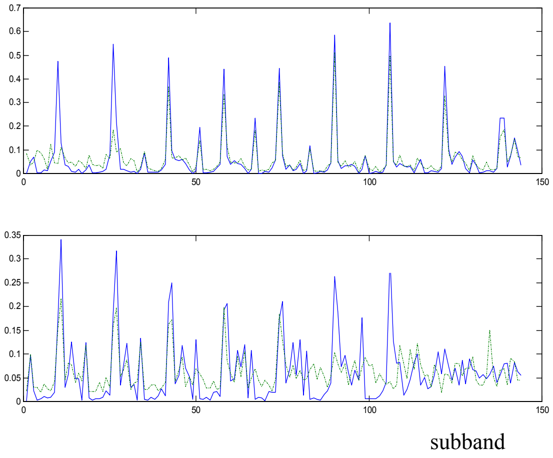

Figure 3. Spectral speech structure recovered by the algorithm proposed in this paper for the first half of the word "eight" at an SNR of 0 dB. Normal line stands for clean speech.

V. Experimental results

The proposed algorithm was tested in an Isolated Word Recognition system using Continuous Density Hidden Markov models. The database of isolated words used for training and testing is from AT&T Bell. The used speech was acquired under controlled environmental conditions band-pass filtered from 100 to 3200 Hz, sampled at a 6.67 kHz and analysed in segments of 45 ms duration at a frame rate of 66.67 windows/sec. Only the decimal digits were used. The noise has white noise characteristics, is speech independent and computationally generated at various SNR as shown in table 1. The goal is to compare the performance of the proposed and contemporary speech robust features. Some of these robust features are the OSALPC (One-Sided Autocorrelation Linear Predictive Coding), the conventional cepstrum with liftering (CEPS + liftering) and the well known MFCC (Mel-Frequency Cepstral Coefficients). In table 1, MMC stands for conventional Markov model composition in the power spectrum density domain, Norm. stands for the baseline normalisation procedure described in [1], N. + MMC stands for Markov model composition in the baseline power normalisation domain [1], PR stands for the post-processing spectral restoration procedure proposed in [3] and BN stands for the bi-normalisation proposed in this paper. Table 1 shows that the suggested spectral multi-normalisation features are more effective against additive white noise than both the baseline normalisation, which is more effective than some robust features used nowadays, and the PR algorithm proposed in [3]. For SNR greater than or equal to 5 dB the baseline spectral normalisation outperforms the conventional Markov model composition (MMC) when the noise parameters are learned from the periodogram method in a data segment of 100ms without speech. As in the Parallel Model Combination, the distortion can be integrated (compensated) in the composite model increasing thus the recogniser performance [1]. On the first six entries of the table 1, all the features are 8 static, energy and dynamic features excepting * (12 static + energy + dynamics) and ** (13 static + energy + dynamics).

Table 1 – Performance of the spectral normalisation

SNR (dB)	15	10	5	0	-5
LP	56.5	39.5	30	16.25	
OSALPC	98.25	92	65.75	32.25	
CEPS *	97.5	95	72	34.5	
+liftering	98.25	95	75.25	39	
MFCC **	97.75	94.75	72.25	37.5	
OSALPC*	98.5	96.25	74.25	32.5	
MMC	98	96.75	92.5	91	78.5
Norm.	98.5	97.75	93.75	88	42.5
PR	99.25	98.25	95	89.75	61.5
BN	99.25	98.5	95.75	90.75	64.25
N.+ MMC	99.5	98.75	97.25	92.25	84.75

VI. Discussion

The main advantage of this bi-normalisation process is the recognition performance obtained when no knowledge of the noise statistics exists. As a robust extraction features, the suggested method seems to be superior to the most used nowadays. Additionally, for white noise and at SNR greater than or equal to 5 dB it presents better performance than a standard noise compensation technique, which assumes integral noise knowledge. In fact for high Signal to Noise Ratios the spectral normalisation where the distortion is ignored outperforms the Markov model composition where the distortion is learned from a small amount of isolated noise samples and incorporated into the system. If isolated noise samples exist, the noise can be estimated and this knowledge can be incorporated into the system, and consequently increasing the recogniser performance.

References

[1] C. Lima, Luís B. Almeida, and João L. Monteiro, "Improving the Role of Unvoiced Speech Segments by Spectral Normalisation in Robust Speech Recognition," *proceedings of the 7th International Conference on Spoken Language Processing,* vol. pp 1573 - 1576 , 2002.

[2] Biksha Raj, "Reconstruction of Incomplete Spectrograms for Robust Speech Recognition", *Ph. D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University*, 2000.

[3] C. Lima, Luís B. Almeida, A. Tavares, and C. Silva, "Spectral Multi-Normalisation for Robust Speech Recognition", *IEEE & ISCA Workshop on Spontaneous Speech Processing and Recognition,* pp 39 - 42, 2003.