SPECTRAL MULTI-NORMALISATION FOR ROBUST SPEECH RECOGNITION

Carlos Lima, Luís B. Almeida, Adriano Tavares and Carlos Silva*

Department of Industrial Electronics of University of Minho, Portugal
{carlos.lima, adriano.tavares, carlos.silva}@dei.uminho.pt

*Department of Electrical and Computers Engineering, IST, Technical Univ. of Lisbon, Portugal
lba@speech.inesc.pt

ABSTRACT

This paper presents an improved version of a spectral normalisation based method for extraction of speech robust features in additive noise. The baseline normalisation method was developed by taking into consideration that, while the speech regions with less energy need more robustness, since in these regions the noise is more dominant, the "peaked" spectral regions which are the most reliable due to the higher speech energy must also be preserved as much as possible by the feature extraction process.

The additive noise effect tends to flatten the "peaked" spectral zones while the spectral zones of less energy are usually raised.

The algorithm proposed in this paper showed to alleviate the noise effect by emphasising the voiced nature of the speech signal by raising the spectral "peaks", which are "flatten" by the noise effect. The clean speech database is assumed as lightly contaminated, the additive noise is estimated in a frame by frame basis and then used to restore both the "peaked" and the flat spectral zones of the speech spectrum.

1. INTRODUCTION

In our last paper [1] we 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 this paper 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. The results show a significant improvement in performance when compared with the baseline method concerned to the situation where the noise is ignored.

2. 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 summing the components of the power spectral components 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 feature values.

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

$$c_i = \frac{S_i}{S}, \qquad c_i = \frac{S_i + N_i}{S + N} \tag{1}$$

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.

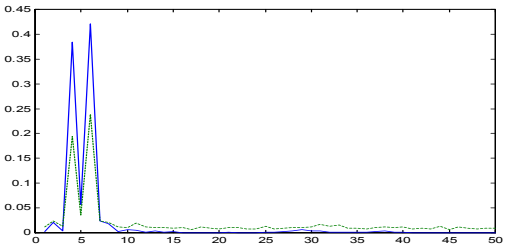

Figure 1. White noise effect in the power spectrum density normalization domain in the beginning of digit "zero". Dash[illegible] represents noisy speech features.

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).

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

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

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. Using these considerations, the calculation of the shift vector imposed by the environment is accomplished by subtracting equations (1) and becomes [1]

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

Equation (3) shows that if the speech has a flat power spectrum density, 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. 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]

3. ADDITIVE WHITE NOISE EFFECT IN THE POWER SPECTRAL NORMALISATION DOMAIN

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:

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 [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 using the estimated noise level in 1) 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.

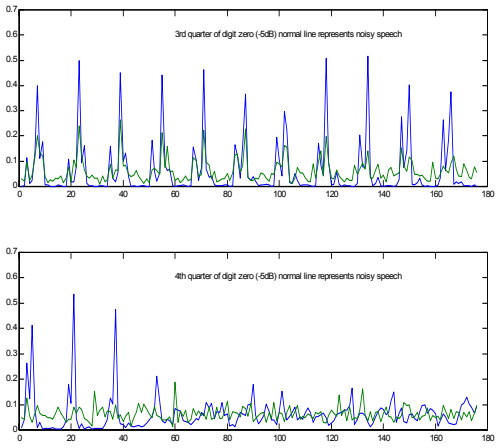

Figure 2. White noise effect in the power spectrum density normalization domain in a voiced segment (upper part) and in an unvoiced segment (last 2/3 of the lower part of the figure. Dashed line represents noisy speech features.

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?

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 component, 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 equation (1) must be changed so that

$$c_i = \begin{cases} \dfrac{S_i - \min\{S_i\}}{S}, S_i \neq \min\{S_i\} \\ \dfrac{S_i}{S}, otherwise \end{cases} \quad (4)$$

or in noisy situations

$$c_i = \begin{cases} \dfrac{S_i + N_i - \min\{S_i + N_i\}}{S + N} \\ \dfrac{S_i + N_i}{S + N} \end{cases} \quad (5)$$

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 means that a white noise process does not deteriorates 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. In the context of the baseline normalisation we have automatic compensation of the means.

The noise compensation in the "peaked" spectral regions is made 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

$$\sum_{i=1}^{B} c_i = 1 \quad (6)$$

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

$$\sum_{j=1}^{n} c_j = 1 - \sum_{\substack{i=1 \\ i \neq j}}^{B} c_i \quad (7)$$

where each c_j was previously decreased as shown by equation (5). Assuming that each spectral sub-band was decreased proportionally to its value, which seems to be true by analysing figures 1, 3 and 4 the noise compensation can be made by computing c_j as follows

$$c_j = \frac{\left(S_j + N_j\right)\left(1 + \frac{(B-n)}{S_n}\min\left\{S_i + N_i\right\}\right)}{S + N} \quad (8)$$

where S_n is given by

$$S_n = \sum_{j=1}^{n}\left(S_j + N_j\right) \quad (9)$$

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.

5. 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, N. + MMC stands for Markov model composition in the baseline power normalisation domain [1] and MN stands for the multi-normalisation procedure proposed in this paper. Table 1 shows that the suggested spectral multi-normalisation features are more effective against additive white noise than the baseline normalisation, which is more effective than some robust features used nowadays. 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
MN	99.25	98.25	95	89.75	61.5
N.+ MMC	99.5	98.75	97.25	92.25	84.75

6. DISCUSSION

The main advantage of this multi-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. 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] 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] Raj, Biksha (2000). Reconstruction of Incomplete Spectrograms for Robust Speech Recognition. Ph. D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University.