Improving the Role of Unvoiced Speech Segments by Spectral Normalisation in Robust Speech Recognition

Carlos Lima, Luís B. Almeida* and João L. Monteiro

Department of Industrial Electronics of University of Minho, Portugal
{carlos.lima, joao.monteiro}@dei.uminho.pt

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

Abstract

This paper presents a spectral normalisation based method for extraction of speech robust features in additive noise. The method has two main goals:

1) The "peaked" spectral zones, where the most speech energy is concentrated must be preserved (from clean to noisy speech features) as much as possible by the feature extraction process. Usually, these spectral regions are the most reliable due to the higher speech energy, and the frequently assumption of independence between speech and noise.

2) The speech regions with less energy need more robustness, since in these regions the noise is more dominant, thus the speech is more corrupted. Usually these speech regions correspond to unvoiced speech where are included nearly half of the consonants. The proposed normalisation will be optimal if the corrupted and the noise process have both white noise characteristics. Optimal normalisation means that the corrupting noise does not change at all the means of the observed vectors of the corrupted process.

For Signal to Noise Ratio greater than 5 dB the results show that for stationary white noise, the proposed normalisation process where the noise characteristics are ignored, outperforms the conventional Markov models composition where the noise must be known. Additionally, if the noise is known, a reasonable approximation of the inverted system can easily be obtained by performing noise compensation and still increasing the recogniser performance.

1. Introduction

Noise robustness can be accomplished either at the feature representation level using robust parameterisation or at the model compensation level. Generally, in the feature analysis process, only a lightly knowledge about the noise characteristics is needed. Some approaches consider that the corrupting noise is by nature unknown, thus it is meaningless 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. Noise pre-filtering [1] [2], projection based distortion measures [3], vector space mapping [4] [5], all pole modelling of the autocorrelation sequence [6] [7] [8] [9] [10], speech representation motivated by the human auditory system knowledge (Perceptually Linear Prediction analysis (PLP)) [11] [12], and more recently, complementing the PLP technique with a band-pass filter (RASTA-PLP) [13], have been the more successful techniques used for robust speech representation. However, in spite of the effort dedicated last years in the field of the robust parameterisation, conceiving systems with acceptable performance in environments for which they were not trained, has been far too difficult.

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. This approach was followed in the development of the spectral normalisation method suggested in this paper.

2. Spectral Normalisation

This spectral normalisation is motivated by the fact that the additive noise is not a narrow band noise, thus its spectrum is reasonably dispersed in frequency. The goal is preserving as much as possible the speech features against noise. The 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 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.

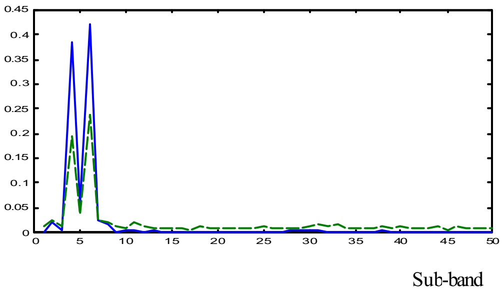

Figure 1. White noise effect in the power spectrum density normalization domain in the beginning of digit "zero". Dashed line represents noisy speech features.

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)

$$C_i(N) = \frac{S_i + N_i}{S + N} - \frac{S_i}{S} \tag{2}$$

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

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

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. Using these considerations, the calculation of the shift vector imposed by the environment is accomplished by subtracting equations (1) and becomes [14]

$$C_i(N) = \left(\frac{S_i}{S} - \frac{l}{L}\right)\frac{1-k}{k}, \quad k = 1 + \frac{1}{10^{\frac{SNR}{10}}} \tag{4}$$

Equation (4) 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.

Figure 2 shows the relative deviation caused by the environment (additive white noise at 0 dB) in the suggested power spectrum normalisation domain and in the power spectrum density domain. The relative deviation was computed as

$$D_i = \frac{Z_i - X_i}{X_i}$$

where Z_i is the i^{th} component of the observation vector for noisy speech and X_i has the same meaning but for clean speech. It is evident by comparing figure 1 and figure 2 that the "peaked" spectral regions of the clean speech are more robust against additive white noise than the rest of the band. Additionally the proposed normalisation shows more robustness (less deviation in the features) for all the frequency sub-bands.

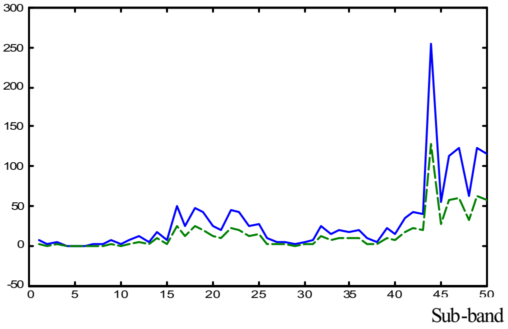

Figure 2. Relative deviation caused by additive white noise at 0 dB at the beginning of digit "zero" when working in the power spectral density domain (normal line) and in the power spectral density normalisation domain (dashed line).

3. Markov models composition in the spectral normalisation domain

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.

1) *Distribution of the composite signal (noisy speech)*: Usually the corrupting Gaussian additive white noise process is considered in the time domain. As the Fourier Transform is a linear operation then the distribution is maintained from time to frequency domain. It is well known from the statistics theory that if a random variable has a Gaussian distribution then the square of its modulus (power spectral density) has a chi-square distribution with two degrees of freedom, also known by exponential distribution. As the speech and noise are considered additive in the time domain, the additivity is maintained in the power spectrum density (PSD) domain. The clean speech is modelled as Gaussian in the PSD domain and the distribution of the noisy speech becomes the convolution between a Gaussian and an exponential function. Reference [14] shows that the noisy speech distribution $f_z(z)$ is

$$f_z(z)=\frac{e^{-\frac{4\lambda z-4\lambda\mu_y-2\sigma_y^2}{4\lambda^2}}}{2\lambda}\left(1+erf\left(\frac{\lambda z-\lambda\mu_y-\sigma_y^2}{\sqrt{2\sigma_y^2}\lambda}\right)\right) \tag{5}$$

where the **y** vector refers to the clean speech signal, λ is the parameter of the exponential distribution and *erf* stands for the well known error function.

To reduce the observed vector dimensionality when working in the spectral density space, it is commun grouping by sum some contiguous components. The number of components considered must be a compromise between the training database size and the required frequency resolution. In our case we used 16 components in each sub-band. Therefore equation (5) holds for the noisy speech distribution, and it would be still necessary to develop the distribution of the sum of 16 random variables each one with the distribution given by this equation. As equation (5) is complex to handle by convolution, an easier solution is to develop the probability density function of the sum of 16 exponential distributed random variables (noise in sub-bands) and perform the convolution of this function with a Gaussian function which models the sum of 16 spectral components of the clean speech. By mathematical induction it is easy to prove [14] that the distribution of the sum of 16 random independents and identically distributed variables is

$$f_W(w)=\frac{w^{15}\exp\left\{-\frac{w}{\lambda}\right\}}{15!\lambda^{16}}U(w) \tag{6}$$

The integral of convolution between the above equation and a Gaussian function becomes very difficult to calculate due to the w^{15} term. Using the Central Limit Theorem, equation (6) can be approximated by a Gaussian function with mean equal to 16λ and variance equal to $16\lambda^2$.

The nature of the Central Limit theorem approximation and the required number of variables for a specified error bound, depend on the form of the densities of the summed random variables. For most applications a number of 30 random variables is adequate, however, for smooth distributions a number as low as 5 can be used. In our case we have 16 random variables and no smooth distributions, so a considerable difference between the real and approximated function can be expected. This difference is shown in figure 4 for λ=10. However, in real situations λ is greater, (order of 10^7 at 10dB), the variance is in order of the square of λ and the Gaussian function fits better to the function defined by equation (6) than is expected by the inspection of figure 4.

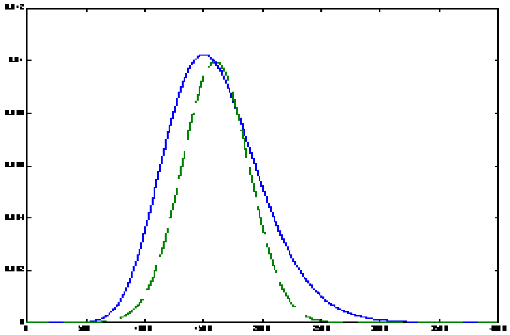

Figure 4. Approximation of the sum of 16 random i. i. d. variables with λ=10, by a Gaussian.

Under this approximation the noisy speech distribution becomes

$$f_z(z)=\frac{1}{4\sqrt{2\pi}\sqrt{\sigma_y^2+16\lambda^2}}e^{-\frac{(z-\mu_y-16\lambda)^2}{2(\sigma_y^2+16\lambda^2)}} \tag{7}$$

2) *Noisy speech distribution in the spectral normalisation space*: As shown above, the noise can be approximately Gaussian modelled in the sub-band PSD domain and so, the noisy speech has also a Gaussian distribution. Similarly, we can consider that if the clean speech spectral normalisation can be Gaussian modelled then the noisy speech spectral normalisation also follows a Gaussian distribution. So C_i(N) has a Gaussian distribution given the distribution of the speech features is Gaussian and all the other terms involved in the equation (2) are considered constants for white noise. The knowledge of the C_i(N) statistics is then reduced to the knowledge of its mean and variance. Using equation (4) the next expression holds for the mean

$$E\left\{\frac{S_i+N_i}{S+N}\right\}=\frac{1}{k}E\left\{\frac{S_i}{S}\right\}-\frac{l}{L}\frac{1-k}{k} \tag{8}$$

where k is given in equation (4).

The variance of the corrupted process can be similarly calculated, considering white noise and that each sub-band is composed by summing 16 power spectral density components

$$\mathrm{var}\left\{\frac{S_i+N_i}{S+N}\right\}=\frac{1}{k^2}\mathrm{var}\left\{\frac{S_i}{S}\right\}+16\left(\frac{k-1}{k}\right)^2 \tag{9}$$

Updating the clean speech HMM distributions according to equations (8) and (9), that is, performing Markov model composition for stationary white noise in the spectral normalisation domain, the recognition accuracy was increased as shown in table 1.

4. 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 proposed normalisation procedure and N. + MMC stands for Markov model composition in the proposed power normalisation domain. Table 1 shows that the suggested spectral normalisation features are more effective against additive white noise than some robust features used nowadays. For SNR greater than or equal to 5 dB the 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. 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
N.+ MMC	99.5	98.75	97.25	92.25	84.75

5. Discussion

The main advantage of this 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] Boll, S.F. (1979). Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust, Speech, Signal Processing, Vol. ASSP–27, n.º 2, pp. 113–120.

[2] Compernolle, D.Van. (1989). Noise adaptation in a hidden Markov model speech recognition system. Comput, Speech, Language, Vol. 3, pp.151–167.

[3] Mansour, D. and Juang, B.H. (1989). A family of distortion operators based on projection operation for robust speech recognition. IEEE Trans. Acoust., Signal Processing, pp. 1659–1671.

[4] Gish, H., Chow, Y. and Rohlicek, J.R. (1990). Probabilistic vector mapping of noisy speech parameters for HMM word spotting. In Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 117–120.

[5] Juang, B.H. and Rabiner, L.R. (1987). Signal restoration by spectral mapping. In Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 2368–2371.

[6] Mansour, D. and Juang, B. (1988b). The short-time modified coherence representation and its application for noisy speech recognition. In. ICASSP, pages 525- 528.

[7] Mansour, D. and Juang, B. (1989a), The short-time modified coherence representation and its application for noisy speech recognition. IEEE Trans., ASSP, ASSP – 37 (6) : 795 – 804.

[8] Hernando, J. and Nadeu, C. (1991). A comparative study of parameters and distances for noisy speech recognition. In EUROSPEECH, pages 91 - 94.

[9] Hernando, J., Nadeu, C. and Lleida, E. (1992). On the AR modelling of the one-sided autocorrelation sequence for noisy speech recognition. In ICSLP, pages 1593 - 1596.

[10] Hernando, J., Nadeu, C. (1994). Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques. In ICASSP, pages II.69 – II.72.

[11] Hermansky, H., Hanson, B. and Wakita, H. (1985). Low–dimensional representation of vowels based on all-pole modelling in the psychophysical domain. Speech Communication, 4(13):181-187.

[12] Hermansky, H. (1990). Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87(4): 1738–1752.

[13] Hermansky, H., Morgan, N. Bayya, A. and Kohn, P. (1991). Compensation for the effect of the communication channel in auditory like analysis of Speech (RASTA–PLP). In EUROSPEECH, pages 1367–1370.

[14] Lima, Carlos. Speech Recognition in non-Stationary Environments. Ph. D. Thesis to submit in the Department of Industrial Electronics of University of Minho (Portugal).

[15] Gales, M. and Young, S. (1994). Parallel model combination on a noise corrupted resource management task. In ICSLP, pages 255-258.