HMM MODELLING OF ADDITIVE NOISE IN THE WESTERN LANGUAGES CONTEXT

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*: This paper is concerned to the noisy speech HMM modelling when the noise is additive, speech independent and the spectral analysis is based on sub-bands. The internal distributions of the noisy speech HMM's were derived when Gaussian mixture density distributions for clean speech HMM modelling are used, and the noise is normally distributed and additive in the time domain. In these circumstances it is showed that the HMM noisy speech distributions are not Gaussians, however, fitting these distributions as a Gaussian mixture, only a little bit of loss in performance was obtained at very low signal to noise ratios, when compared with the case where the real distributions were computed using Monte Carlo methods.**

I. INTRODUCTION

In the western languages the intonation does not make part of the linguistic message, so a very fine detail in frequency is not necessary concerning to speech recognition applications, becoming the signal envelope of the most importance. Therefore some spectral components are frequently grouped, for example by sum and each group is known as sub-band.

Recently the importance given to the field of environmental/speaker adaptation has been increased in part to the difficulties in the obtaining of a feature extraction method sufficiently robust against these types of speech variability. The contemporary adaptation algorithms are mostly based on the MLLR algorithm [1], which can't be able to separate speaker mismatch from environmental (additive and convolutional) mismatch. Alternative approaches can deal separately with an additive noise model and a convolutional noise model in both stationary [2] and non-stationary [3] noise conditions in order to separate these two types of distortions. However these algorithms are essentially based on cepstrum based features, which contributes to increase significantly the computational load once that a mapping between the cepstral and linear domains is required. In [4] [5] it is suggested that a proper spectral normalisation can be more useful than the cepstrum derived features in the noisy speech modelling, while [6] proposes an incremental adaptation algorithm based on spectral derived features. The next step is to investigate the drawbacks of using a gaussian mixture to model the internal distributions of the noisy speech HMM's when using power spectral density based features jointly with additive noise in the linear (not cepstral) domain. This is the purpose of this paper.

II. NOISE AND NOISY SPEECH STATISTICS

The use of continuous observation density in HMMs is not restricted to the use of Gaussian mixtures. Although some restrictions must be placed on the form of the model probability density function (pdf) to ensure that the parameters of the pdf can be re-estimated in a constant way, any log-concave or elliptically symmetric density [7] can be used.

Typically the clean speech features are modelled as a Gaussian mixture and generally the existing speech recognisers perform well in clean speech conditions. In noisy conditions the performance degrades in part due to inaccuracies in noise modelling, given that in some situations the noise is artificially generated thus, known. Using power spectral density features and Gaussian distributed additive noise strong evidences exist that the noisy speech distribution can't be Gaussian. In fact if the noise is Gaussian distributed in the time domain it is well known from the statistics theory that it becomes exponentially (chi-square with two degrees of freedom) distributed in the power spectral density domain, which is the feature domain where the distribution of the noisy speech must be computed. Additionally, as usual, some power spectrum density components have to be grouped anyway (in our case by sum) in order to reduce the feature vector dimensionality, which will also be taken into consideration in the obtaining of the noisy speech statistics.

An exponential distribution of parameter λ is defined by the following probability density function

$$f_x(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} U(x) \qquad (1)$$

where U(x) is the unit step function. The exponential distribution is characterised by the fact that its mean is equal to its standard deviation, which is equal to λ. So, the periodogram distribution of a white noise Gaussian stochastic process with zero mean is a white noise exponential stochastic process with zero mean and $\lambda=N\sigma^2$, where σ^2 is the signal variance and N the signal length.

Supposing HMMs with Gaussian sources then the clean speech y has a Gaussian mixture distribution where the distribution of each component of the mixture is given by

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \quad (2)$$

Let y=(y[0], …,y[N-1])^T, x=(x[0], …,x[N-1])^T and z=(z[0], …,z[N-1])^T be, respectively, vectors of clean, noise and noisy signals. If the noise is additive, y[n] is given by

$$z[n] = y[n] + x[n], \text{ n=0, …, N-1.}$$

The autocorrelation function of the noisy speech can be obtained from the autocorrelation functions of the clean speech, the noise the respective cross correlation as follows

$$\begin{aligned}\varphi_{zz}(m) &= E\{z[n]z^*[n+m]\} \\ &= E\{(x[n]+y[n])(x[n+m]+y[n+m])^*\} \\ &= E\{x[n]x^*[n+m]+x[n]y^*[n+m]+ \\ &\quad y[n]x^*[n+m]+y[n]y^*[n+m]\} \\ &= E\{x[n]x^*[n+m]\}+E\{x[n]y^*[n+m]\}+ \\ &\quad E\{y[n]x^*[n+m]\}+E\{y[n]y^*[n+m]\} \\ &= \varphi_{xx}(m)+\varphi_{xy}(m)+\varphi_{yx}(m)+\varphi_{yy}(m)\end{aligned}$$

As the noise is speech independent, the two processes are non-correlated, so the cross-correlations in the above equation are null. Consequently the autocorrelation function of the noisy speech is simply the sum of the autocorrelation functions of the clean speech and noise.

Let Y=(Y(0), …,Y(K-1))^T, X=(X(0), …,X(K-1))^T and Z=(Z(0), …,Z(K-1))^T denote, respectively, vectors of spectral components of clean, noise and noisy signals. As the Fourier transform is a linear operation and the power spectral density is the Fourier transform of the autorrelation sequence, then for additive noise, and considering the analysis window too large the next expression holds

$$|Z(k)|^2 = |Y(k)|^2 + |X(k)|^2$$

Accounting to the nature of the speech signal $|Y(k)|^2$ in the above equation does not represent the true autocorrelation sequence of the speech once that the autocorrelation sequence of an autoregressive process is theoretically infinite. The segment analysis truncates the autocorrelation sequence. However, as this occurs in both the test and training and the autocorrelation of the noise is finite the above equation stays approximately valid.

Therefore each component of the clean speech distribution generates jointly with the Gaussian noise a noisy speech distribution component (z) given by

$$f_z(z) = \int_{-\infty}^{+\infty} f_x(z-y)U(z-y)f_y(y)dy \quad (3)$$

In reference [8] it is proved that the solution for the above integral is

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

where erf stands for error function which is defined by the integral

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \quad (5)$$

For high SNRs equation (4) roughly fits the Gaussian distribution given that the noise distribution approaches the impulse function.

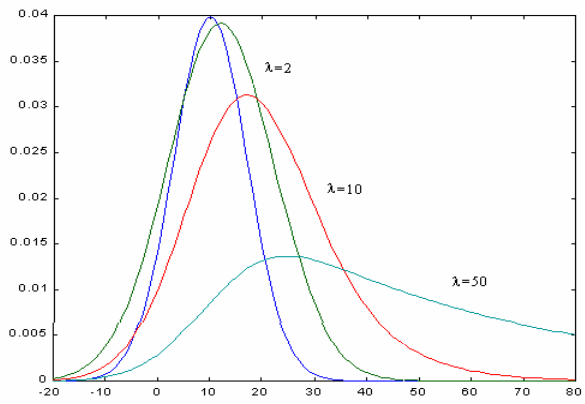

Figure 1. Distribution of the clean and noisy speech for λ=2, 10 (SNR=0dB), 50.

Figure 1 shows the difference between equation (4) and the Gaussian function for λ=2, 10, 50; mean of the Gaussian equals to 10 and variance equals to 100, therefore simulating a SNR=0 dB when λ=10.

For high noises the noisy speech distribution is clearly non-Gaussian and so, the noisy speech distributions have to be changed from Gaussians as usually used, to the function defined by equation (4).

By analysing in the sub-bands context the HMMs for clean speech model the sum of n power spectral contiguous components, instead of only one power spectral component. Therefore, the solution for the noisy

speech sub-band distribution can be obtained from equation (4) taking into account that the means and variances in each model must be divided by *n*, once that they model the sum of *n* random variables with Gaussian distribution and all with the same parameters. Therefore equation (4) holds for the noisy speech distribution, and it would still be necessary develop the distribution of the sum of *n* random variables each one with the distribution given by equation (4).

An easier and equivalent solution is to develop the probability density function of the sum of *n* exponential distributed random variables as shown in equation (1) and perform the convolution of this function with a Gaussian function which models the sum of *n* power spectral components of the clean speech.

Reference [8] shows that the distribution of the sum of *n* random independents and identicaly distributed (according equation (1) variables is

$$f_x(x) = \frac{x^{n-1}\exp\left\{-\frac{x}{\lambda}\right\}}{(n-1)!\lambda^n}U(x) \qquad (6)$$

Equations (2) and (6) allow to derive the probability density function as usual by convolving the two probability density functions

$$f_z(z) = \int_{-\infty}^{+\infty} f_x(x)U(x)f_y(z-x)dx$$

$$= \int_{0}^{+\infty} f_x(x)f_y(z-x)dx$$

$$= \frac{1}{(n-1)!\lambda^n\sqrt{2\pi\sigma_y^2}}\int_0^{+\infty} x^{n-1}e^{\frac{-x}{\lambda}}e^{\frac{-(z-x-\mu_y)^2}{2\sigma_y^2}}dx \qquad (7)$$

The above integral is difficult to calculate due to the term x^{n-1} where *n* is of the order of more than ten, once that the recognition systems nowadays use observation vectors dimensionality from tipically ten to forty (with dinamical characteristics) thus, much smaller than the normally used as FFT length.

III. Aproximated Distribution of the Noise and Noisy Speech

By using the Central Limit theorem equation (6) can be approximated by

$$f_x(x) = \frac{1}{4\sqrt{2\pi\lambda}}e^{-\frac{(x-16\lambda)^2}{16\lambda^2}} \qquad (8)$$

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 thus, a considerable difference between the real and approximated function can be expected. This difference is shown in figure 2 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 function defined by equation (8) fits best to the function defined by equation (6), what is expected by the inspection of figure 2.

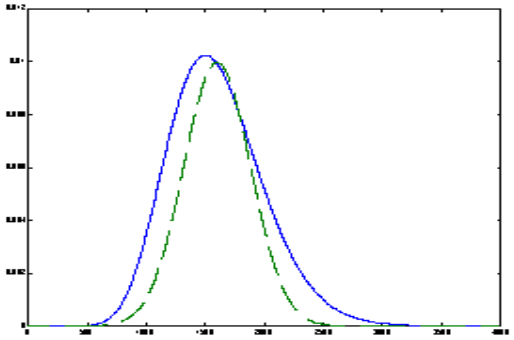

Figure 2. Approximation of the sum of 16 random i. i. d. variables with λ=10, by a Gaussian.

Under this approximation the noisy speech distribution (equation (6)) becomes

$$f_z(z) = \frac{1}{4\sqrt{2\pi}\sqrt{\sigma_y^2+16\lambda^2}}e^{-\frac{(z-\mu_y-16\lambda)^2}{2(\sigma_y^2+16\lambda^2)}} \qquad (9)$$

given that the convolution between two Gaussian functions is still a Gaussian function which mean and variance are equal to the sum of the initial means and variances, respectively.

IV. Experimental Results

The loss in performance due to the using of equation (9) instead of equation (7), which was computed by numerical integration (exact method), 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 approximation, exact solution 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 by using the suggested approximation while NI stands for the numerical integration. Table 1 shows that the suggested approximation is as effective against additive white noise as the exact solution except for very low signal to noise ratios (-5db), where the loss in performance is even so very low. In both cases the noise parameters were learned from the periodogram method in a data segment of 100ms without speech. 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 proposed approximation

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
NI	98	96.75	92.5	91	80.25

V. Discussion

The main advantage of using spectral based features instead of cepstral based features is the decreasing of computational load given that the mapping between the linear and cepstral domains becomes not necessary. In fact, as the noise is considered additive in the linear domain and the features adaptation is performed in the cepstral domain, a mapping from cepstral to linear domain and then an inverse mapping from linear to cepstral domain are needed (Parallel Model Combination). This decreasing in computational load is particularly important on environmental/speaker incremental adaptation where recently some effort has been made in order, for example, to separate speaker mismatch from environmental mismatch or adapting to non-stationary additive noise situations where the channel distortion is stationary. This situation requires training, of the combined HMM's of the clean speech and noise, on the recognising speech (incremental adaptation) which becomes more easy if the internal distributions remain Gaussians. Additionally a proper spectral normalisation [4][5] can be more effective concerned to speech modelling than the cepstral based features, at least for some types of noise. However, the main drawback associated with cepstral based features is related with the difficulty in the modelling of speech dynamics. In fact the adaptation of the dynamic coefficients is not possible, although some approximate solutions have been suggested.

References

[1] C. J. Leggetter, and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech & Language, vol 9, pages 171-185, 1995.

[2] M. J. F. Gales and S. J. Young, "PMC for speech recognition in additive and convolutional noise," Technical Report 154, Cambridge University, 1993.

[3] Z. Wang, P. Kenny, and O'Shaughnessy, "Robust Speech Recognition in Nonstationary Adverse Environments," Proceedings of ICASSP, Seattle, vol. 1, pp 265-268, 1998

[4] C. Lima, L. B. Almeida and J. L. Monteiro, "Improving the Role of Unvoiced Speech Segments by Spectral Normalisation in Robust Speech Recognition," 7th International Conference on Spoken Language Processing (ICSLP'2002), pp 1573-1576, 2002.

[5] C. Lima, L. 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.

[6] C. Lima, L. B. Almeida and J. L. Monteiro, "Continuous Environmental Adaptation of a Speech Recogniser in Telephone Line Conditions," 7th International Conference on Spoken Language Processing (ICSLP'2002), pp 1401-1404, 2002.

[7] S. E. Levinson., L. R. Rabiner and M. M. Sondhi, "An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition," Bell System Tech. J., 62(4): 1035-1074, 1983.

[8] C. Lima, "Speech Recognition in Non-stationary Environments," Ph. D. Thesis, Department of Industrial Electronics, University of Minho, Portugal, 2002.