GENERALIZED GAUSSIANS FOR CONTINUOUS OBSERVATION DISTRIBUTIONS IN SPEECH RECOGNITION

Antonio Miguel, Eduardo Lleida, Alfonso Ortega

Communication Technologies Group (GTC), I3A, University of Zaragoza, Spain.

ABSTRACT

One of the most successful models for speech recognition has been the HMM with mixture of Gaussians in the states to generate/capture observations. In this work we show how the addition of a parameter to model higher order moment statistics, such us the kurtosis, can provide improvements to the system. The distributions in which this degree of freedom is integrated are the generalized Gaussians. It is shown a method to estimate the parameters of these distributions even if they are embedded in a HMM or mixture of distributions. Some experimental results are obtained with this method compared to baseline systems of full and diagonal covariance matrices.

1. INTRODUCTION

This paper offers a new approach to model more accurately the observation generation process in the states of the HMM. The working hypothesis for the approach of this paper is that there is information in the speech signal which is not accurately captured by standard models in the states of the HMMs, usually GMMs with diagonal covariance matrix.

In this paper we propose to increase the complexity of the pdfs which are the components of the mixture in the states of the HMM by adding a degree of freedom to control a higher order moment, the kurtosis. The basic idea is that the new pdfs should be able to capture or generate data with statistics beyond the normal distribution.

The goal of models in speech recognition is to keep the maximum of information that we think it is useful to recognize speech, not to synthesize a speech waveform. In this work we try to improve the quality of the statistics captured from the speech signal in order to capture the maximum of information. To do so, a modification in the nature of the Gaussian distribution is proposed. The proposed probability density function is the Generalized Gaussian distribution [1, 2], this distribution has an additional parameter over the normal distribution which controls the fourth order moment. For selected values of this parameter the distribution can adapt its shape to many symmetric distributions as the normal, the Laplacian, or even the uniform and Dirac's delta for extreme values. The generalized Gaussian distribution provides a richer mechanism to adapt to feature statistics.

Some authors have contributed to similar lines of research to enhance the models ability to generalize but the application of the generalized Gaussian distributions in the generation of observations is novel. The generalized Gaussian is an interesting distribution but to be useful in speech recognition, two additional mechanisms are proposed to complete those models. The first one is related to the fact that the generalized Gaussian, as the Gaussian, is not a multimodal distribution. In order to adapt to the complex statistics of the speech signal a hidden variable mechanism is needed to explain the observation in a more accurate way. This is achieved with the mixtures of generalized Gaussians. The parameter estimation will demand a modification in the standard EM algorithm based in the method of moments. The second one is to consider a method to reduce the amount of correlation in the features that we want to model. The features we want to model are usually vector valued. The simplest approach to face multivariate distributions is to use a Naive Bayes approach assuming independence between the features. The proposal to overcome this simplicity in order to find more complex dependences is to assume a linear transformation of the vector by means of a rotation which preserves the scale of the projected vectors.

This paper is organized as follows. In Section 1 there is an introduction. In Section 2 the generalized Gaussian is described. In Section 3 the parameter estimation is discussed. In Section 4 a rotation is included to model covariance.In Section 5, the estimation in hidden variable structures is presented. Experimental results are shown in Section 6 and finally conclusions are in Section 7.

This work has been supported by the national project TIN 2005-08660-C04-01.

2. GENERALIZED GAUSSIAN DISTRIBUTION

The Gaussian distribution is adequate for many problems in speech technologies but is still limited in the sense of modeling accurately distributions with a wide range of high order moments, greater to the second order. The Gaussian distribution has a fixed value of 3 for the kurtosis, which is related to the fourth order moment.

Along this paper we modify the Gaussian fundamental distribution to include an additional parameter which controls the kurtosis of the distribution, This distribution is called a generalized Gaussian (GG) [1, 2] and has the following definition: a continuous real valued random variable is assumed to follow a generalized Gaussian distribution if the probability density function takes the form:

$$x \sim GG(\mu, \sigma, \alpha), \quad x \in \mathbb{R} \tag{1}$$

$$f(x|\mu, \sigma, \alpha) = \frac{\beta(\alpha)}{2\Gamma\left(1+\frac{1}{\alpha}\right)\sigma} e^{-\left|\beta(\alpha)\frac{x-\mu}{\sigma}\right|^{\alpha}}, \tag{2}$$

where α is called the shape parameter and $\beta(\alpha)$ and $\Gamma(x)$ are defined as:

$$\beta(\alpha) = \sqrt{\frac{\Gamma\left(\frac{3}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)}}, \quad \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \tag{3}$$

The parameter α in (2) controls the kurtosis of the distribution. We can see that for some values of α the distribution equals some well known distributions. For $\alpha = 1$ the expression (2) equals the Laplacian distribution, for $\alpha = 2$ the Gaussian distribution and as α tends to infinity the distributions gets closer to the uniform distribution, $U(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$ and if alpha tends to 0^+ the distribution is closer to a degenerated function, the dirac's delta. This is exemplified in Figure 1, where some examples of the pdf (2) are plot varying the value of the parameter α for a fixed value of $\mu = 0$ and $\sigma = 1$.

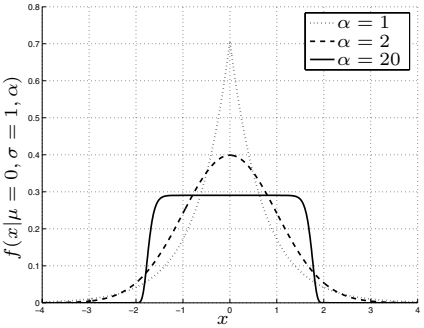


Figure 1. Examples of GG probability density functions for some values of the shape parameter $alpha$.

In [3] a general study about the GGs and its moments is performed. The even order moments of the distribution can be expressed in terms of the parameters . The following is a general expression of a centered moment of even orders r:

$$M_r^* = E[(x-\mu)^r] = \left(\frac{\sigma^2\Gamma\left(\frac{1}{\alpha}\right)}{\Gamma\left(\frac{3}{\alpha}\right)}\right)^{\frac{r}{2}} \frac{\Gamma\left(\frac{r+1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)} \quad (4)$$

The central moments of odd r orders are equal to zero due to the symmetry of the pdf with respect to the mean.

From expression (12) we can see that if the order is $r = 2$:

$$M_2^* = E[(x-\mu)^2] = \sigma^2, \quad (5)$$

where it interesting to note that the variance only depends on the parameter σ in the model pdf, which makes expression (2) a very convenient parametrization. Also it will be of interest for the estimation process the moment of order 4 as will see in next section.

3. ESTIMATION OF PARAMETERS OF GG DISTRIBUTIONS

In [1] the estimation of moments method was proposed, this a simple alternative since in order to fix the three parameters we only have to propose a system of equations with three moments, the mean, the variance and the fourth order as will see. In [2] similar method was proposed to estimate the parameters μ and σ from the moment estimator and the shape parameter using a expression that related the variance, the mean of the absolute values and the shape parameter. In our work we are going to focus on the moment estimation method, we will argue some reasons for this selection due to the nature of the model, the HMM, that the pdf is going to be embedded into. Firstly we will compare the methods in [2] and [3] with the moment estimation method in [1].

In order to establish the notation, let us consider a random variable X with outcomes $x \in \mathbb{R}$, which is assumed to follow a GG distribution as expressed in (2). The training is set is defined as $\mathbf{X} = \{x_1, \ldots, x_n, \ldots, x_N\}$. The pdf is going to be estimated from a i.i.d.(identically distributed) sequence of samples from the variable X.

The parameters μ and σ in all those methods are estimated with equal expressions, the standard moments mean and variance. The expressions are:

$$\hat{\mu} = M_1 = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \hat{\sigma}^2 = M_2^* = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2, \quad (6)$$

The shape parameter can be estimated using the moment estimation method. Since all odd order centered moments are zero, and the shape parameter only is present in the centered moments of order $r \geq 4$, then this method is based on the calculation of a moment in this range of orders and also the sample estimator for this moment.

$$M_r^* = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^r, \quad (7)$$

for any r even and $r \geq 4$.

Then, considering the simplest case $r = 4$, we can obtain the value of α from the following expression:

$$M_4^* = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^4 = \left(\frac{\hat{\sigma}^2\Gamma\left(\frac{1}{\alpha}\right)}{\Gamma\left(\frac{3}{\alpha}\right)}\right)^2 \frac{\Gamma\left(\frac{5}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)}, \quad (8)$$

In order to reduce the sensitivity of the estimation compared to the calculation of the fourth order moment, an alternative method was proposed in [2]. It was demonstrated the following dependency of a function of the shape parameter with the mean of the absolute value of the random variable:

$$\frac{\hat{\sigma}^2}{(E[|x-\hat{\mu}|])^2} = \frac{\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{3}{\alpha}\right)}{\left(\Gamma\left(\frac{2}{\alpha}\right)\right)^2}. \quad (9)$$

which involves lower order moments computations and more estimation accuracy.

Now we discuss which is the best method to estimate the parameters of the generalized Gaussian distribution in order to be easily embedded in an HMM and as a component of a mixture of models. A first approach to the previously presented estimation methods shows that the method based on the absolute value mean is more robust and accurate than the method of moments, which is based on the estimation of moments of orders $r = 4$ and $r = 2$.

Nevertheless, there is an important argument in favour of the method of moments which is the implementation convenience. The estimators in the method of moments with orders $r = 4$ and $r = 2$ can be implemented in one pass over the training data. Therefore the integration in a multiple iteration training procedure as the EM algorithm as an HMM or a mixture of models will be more natural.

The estimators of the centered moments cannot be directly computed in one pass in the training process since while we have access to the samples x_n in the training process we have not calculated the mean for that iteration. In the following expression we perform some algebraic manipulations and the expressions are transformed to:

$$M_2^* = \frac{1}{N}S_2 - \frac{1}{N^2}(S_1)^2 \quad (10)$$

and

$$M_4^* = \frac{1}{N}S_4 - \frac{4}{N^2}S_1 \cdot S_3 + \frac{6}{N^3}(S_1)^2 \cdot S_2 - \frac{3}{N^4}(S_1)^4 \quad (11)$$

Therefore, the moment estimators can be implemented in one pass and the only operations needed during the iteration are the accumulation of powers of the samples. The accumulators are defined as $S_r = \sum_n x_n^r$ for orders $r = 1, 2, 3, 4$.

4. DIAGONALIZATION OF GG DISTRIBUTIONS

Usually, the first approach to a multivariate model is to use the independence assumption of the Naive Bayes approach. Let us suppose that we are modeling the probability density function of a D-dimensional feature vector, $\mathbf{x} = (x_1, \ldots, x_d, \ldots, x_D)$. We can assume that each individual component of the vector follows a GG.

The likelihood of that model can be expressed as:

$$\mathbf{x} \sim GG_D(\mu, \sigma, \alpha), \quad x \in \mathbb{R}^D \tag{12}$$

$$f(\mathbf{x}|\mu, \sigma, \alpha) = \prod_d^D f(x_d|\mu_d, \sigma_d, \alpha_d) \tag{13}$$

$$= \prod_d^D \frac{\beta(\alpha_d)}{2\Gamma\left(1+\frac{1}{\alpha_d}\right)\sigma_d} e^{-\left|\beta(\alpha_d)\frac{x_d-\mu_d}{\sigma_d}\right|^{\alpha_d}}, \tag{14}$$

where each component of the vector is modeled by an independent GG distribution, with parameters μ_d, σ_d and α_d. It is clear from the independence approximation that the estimation of the parameters can be performed with the method of moments for each component of the feature vector separately. The limitation of this technique is similar to the limitation of a Gaussian with diagonal covariance matrix, which is a special case of the previous model, where all the values $\alpha_d = 2$.

We propose the use of a linear transformation to reduce the correlation between the variables in the vector. The method consists in the estimation of a linear transformation matrix, $\mathbf{A}$, to transform the random vector $\mathbf{x}$ to a vector $\mathbf{y}$ as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{15}$$

where the objective is that the components of random vector $\mathbf{y}$ can be considered independent and its covariance matrix be diagonal. The problem of diagonalizing a covariance matrix is the classic problem of the principal component analysis (PCA).

The probability density function of the resulting random variable which is obtained applying a function $\mathbf{y} = g(\mathbf{x}) = \mathbf{A}\mathbf{x}$ over the existing variable $\mathbf{y}$ is:

$$f_y(\mathbf{y}) = f_x(g^{-1}(\mathbf{y}))\left|\frac{\delta g^{-1}(\mathbf{y})}{\delta \mathbf{y}}\right| = f_x(\mathbf{A}^{-1}\mathbf{y})\left|\mathbf{A}^{-1}\right| \tag{16}$$

One of the properties that it would be desiderable for the linear transformation is that $|\mathbf{A}^{-1}| = 1$, so that we can easily apply the Naive Bayes GG distribution to the transformed vectors $\mathbf{x}$. This is also desiderable since no further re-scaling of the likelihood is needed, this provides an important simplicity if the likelihood is going to be compared or operated with other likelihoods as in a mixture of models or in HMMs.

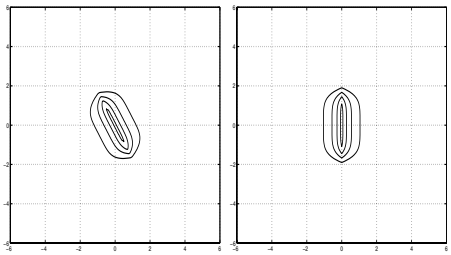
The question that remains is how to calculate the transformation $\mathbf{A}$. This problem can be solved considering two expressions: the relationship of the covariance matrices of variables in a linear transformation and the decomposition theorem. Given a random vector $\mathbf{x}$ with a full covariance matrix $\mathbf{\Sigma}_x$, then the covariance matrix of the random vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ is $\mathbf{\Sigma}_y = \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^T$. The eigendecomposition a semidefinite positive matrix $\mathbf{V}$ can be obtained as $\mathbf{V} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where the matrix $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of $\mathbf{V}$ as the diagonal elements and the matrix $\mathbf{U}$ are the eigenvectors of $\mathbf{V}$ as columns.

If we use both results we can find the linear transformation of the variable $\mathbf{y}$ which makes the covariance matrix $\mathbf{\Sigma}_x$ diagonal as the eigen vectors of the matrix $\mathbf{\Sigma}_y$ by columns.

In the Figure 2 there is an example, we can see the pdf of some artificially generated data of a random vector of dimension $D = 2$. We can see that after the linear transformation, for the pdf in Figure 2b the main variation axis are the cartesian coordinate system which makes possible the application of a Naive Bayes GG distribution.

5. MIXTURES OF GG DISTRIBUTIONS

Similarly to the case of multivariate Gaussian distributions, an unimodal pdf is not an accurate model for the complexity of the speech signal. The mixture model is a natural solution to increase the modes of a pdf and to adapt to higher complexities in the data.



(a) (b)

Figure 2. Example of 2D GG distributions. a) A rotated space GG distributrion b) Naive Bayes multivariate GG

A mixture of GG distributions of C components is defined as a weighted sum of the pdfs of the components in the following way:

$$f(x) = \sum_{c=1}^{C} p_c \cdot \frac{\beta(\alpha_c)}{2\Gamma\left(1+\frac{1}{\alpha_c}\right)\sigma} e^{-\left|\beta(\alpha_c)\frac{x-\mu_c}{\sigma_c}\right|_c^{\alpha}}, \tag{17}$$

where for simplicity all the derivations in this section are expressed in terms of an univariate GG.

The previous expression can be considered the marginalization of hidden discrete variable Z that can take values $z \in \{1, \ldots, C\}$ which selects a from a pull of GG distributions as shown in [4]. This variable is assumed to follow a Multinomial distribution. The pdf of the variable Z is a Multinomial of size $z+ = 1$ and prototype vector $\mathbf{p} = (p_1, \ldots, p_c, \ldots, p_C)$,

$$Z \sim Mult(1, \mathbf{p}), \quad z \in \{1, \ldots, C\}, \quad p(Z = z) = \prod_{c=1}^{C} p_c^{\delta_{z,c}}, \tag{18}$$

The estimation of the parameters of a probability model in which there are hidden variables involved is usually solved with the EM algorithm [5] The E step auxiliary function is:

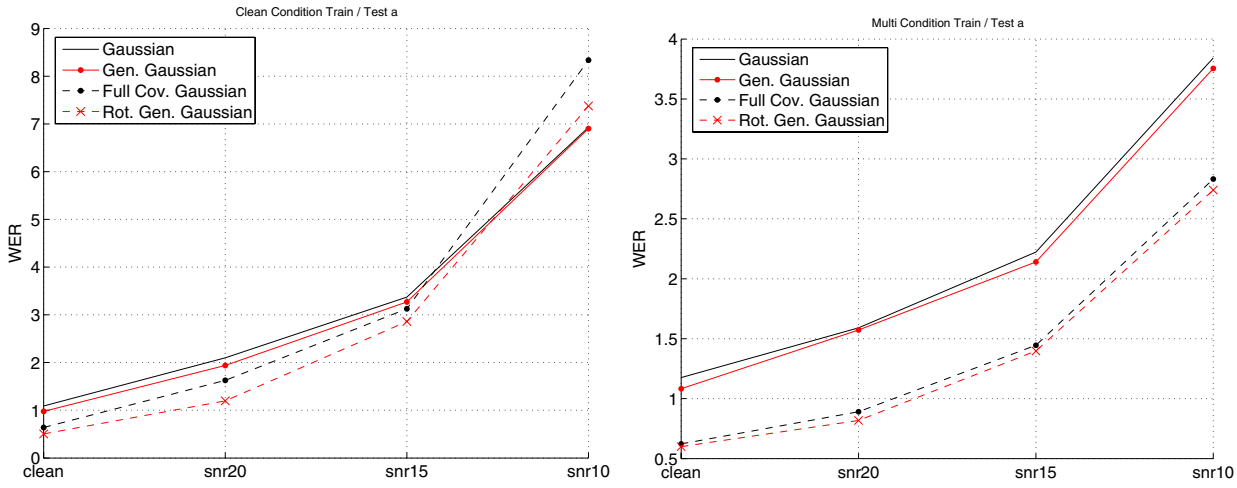
$$\begin{aligned} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) &= E[\log p(\mathbf{X}, \mathbf{Z}|\mathbf{\Theta})|\mathbf{X}, \mathbf{\Theta}^{(k)}] \\ &= \sum_n \sum_c \langle \delta_{Z,c} \rangle_{Z|_{x_n}}^{(k)} \cdot (\log p_c + \\ &\quad + \log\left(\beta(\alpha_c) 2\Gamma\left(1+\frac{1}{\alpha_c}\right)\sigma_c\right) - \left|\beta(\alpha_c)\frac{x_n-\mu_c}{\sigma_c}\right|^{\alpha_c}\Bigg) \end{aligned} \tag{19}$$

where $\langle \delta_{Z,c} \rangle_{Z|_{x_n}}^{(k)}$ is a short notation for the expected value of the function $\delta_{Z,c}$ of the variable Z conditioned to $X = x_n$:

$$\begin{aligned} \langle \delta_{Z,c} \rangle_{Z|_{x_n}}^{(k)} &= E_Z[\delta_{Z,c}|x_n, \mathbf{\Theta}^{(k)}] \\ &= \sum_{\forall z} \delta_{z,c} \cdot p(Z = z, |x_n, \mathbf{\Theta}^{(k)}) \\ &= p(Z = c|x_n, \mathbf{\Theta}^{(k)}) \\ &= \frac{p(Z = c|\mathbf{\Theta}^{(k)}) \cdot f(x_n|Z = c, \mathbf{\Theta}^{(k)})}{\sum_{c'} p(Z = c'|\mathbf{\Theta}^{(k)}) \cdot f(x_n|Z = c', \mathbf{\Theta}^{(k)})} \end{aligned} \tag{20}$$

where $f(x_n|Z = c, \mathbf{\Theta}^{(k)})$ is the component specific GG pdf.

It is possible to find an alternative to the EM direct estimation where the maximization step is substituted by a special moment estimation. This algorithm which is called expectation moment estimation (EME). It can be applied in general to any distribution for which we have defined moments of the distribution in terms of the parameters.



(a) (b)

Figure 3. Experimental WER results for Aurora2 database test set a, for moderate noisy conditions for the different techniques a) clean condition train, b) multi condition train.

The main result of the algorithm is the moment estimation ME step. The ME step, provides estimates for the moments of the distributions which are components of the mixture as a function of the expected values previously calculated $\langle \delta_{Z,c} \rangle^{(k)}_{Z|_{x_n}}$. The first order moment for the component c of the mixture is computed as:

$$M_{1,c}^{(k+1)} = \mu_c^{(k)} = \frac{\sum_{n=1}^{N} \langle \delta_{Z,c} \rangle^{(k)}_{Z|_{x_n}} \cdot x_n}{\sum_{n=1}^{N} \langle \delta_{Z,c} \rangle^{(k)}_{Z|_{x_n}}}, \qquad (21)$$

And the centered moments of order r for the component c of the mixture:

$$M^{*}{}_{r,c}^{(k+1)} = \frac{\sum_{n=1}^{N} \langle \delta_{Z,c} \rangle^{(k)}_{Z|_{x_n}} \left(x_n - \mu_c^{(k+1)} \right)^r}{\sum_{n=1}^{N} \langle \delta_{Z,c} \rangle^{(k)}_{Z|_{x_n}}}. \qquad (22)$$

Depending on the number of free parameters in the model a number of these EME expressions will be needed. For a generalized Gaussian distribution, the number of equations needed is three: $M_{1,c}^{(k)}$, $M^{*}{}_{2,c}^{(k)}$ and $M^{*}{}_{4,c}^{(k)}$.

The procedure to estimate in a single pass the EME $r = 2$ and $r = 4$ order moments can be written in a similar way to the previously shown for the moment estimation method.

6. EXPERIMENTS

The different proposals in this paper have been evaluated on the Aurora 2 task [6] which is a connected digit strings recognizing task in different noise environments. The feature set are the adv ETSI front-end features [7], and the baseline system has been trained with HMM word models of 14 states and 3 component Gaussian mixtures for the digits, a 1 state with 6 components model for the inter-word silence unit and a 3 state with 6 components model for the begin-end silence unit. The models were trained with 20 iterations of the EM algorithm.

In Figure 3 we can see experiments performed in Aurora 2 corpus. There are two baseline systems the observation distribution is a Gaussian mixture in both of them but in a case there are diagonal covariance matrices and in the other full covariance matrices. The results for the GG distributions are also shown where we can see that the error is below the corresponding baseline system in all the cases. We compare the mixture of Gaussians system with the mixture of GG system, and the full covariance Gaussian system with rotated GG, performed as shown in Section 4. The mean WER (word error rate) reduction for noise free condition training (clean) in clean test set a is 10.3% of GGs with respect to Gaussians and 21.1% of rotated GGs with respect to full covariance Gaussians.

7. CONCLUSIONS

In this work it has been shown a method to model high order moments with a distribution in which a parameter related with the kurtosis can be configured. Some methods are provided to estimate the parameters of these distributions, and also the solution to the estimation when the distributions are the outputs of a model with hidden variables. We have seen that the models with the additional degree of freedom perform better than the baseline system specially in noise free conditions.

8. REFERENCES

[1] M.K. Varanasi y B. Aazhang, "Parametric generalized gaussian density estimation," *J. Acoustical Society America*, vol. 86 (4), pp. 1404–1415, 1989.

[2] K. Sharifi y A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distribution in subband decomposition of video," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

[3] R. M. Rodríguez-Dagnino J. A. Domínguez-Molina, G. González-Farías, "A practical procedure to estimate the shape parameter in the generalized gaussian distribution," Tech. Rep., Centro de Investigación en Matemáticas, México, 2000.

[4] A. Juan y E. Vidal, "On the use of Bernoulli mixture models for text classification," *Pattern Recognition*, vol. 35, no. 12, pp. 2705–2710, December 2002.

[5] A. P. Dempster, N. M. Laird, y D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–21, 1977.

[6] H. G. Hirsch y D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 2000, pp. 18–20.

[7] "ETSI ES 202 050 v1.1.1 Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms," July 2002.