Voice Activity Detection using Generalized Gamma Distribution

George Almpanidis and Constantine Kotropoulos [1]

[1] Department of Informatics, Aristotle University of Thessaloniki,
Box 451 Thessaloniki, GR-54124, Greece
{galba, costas}@aiia.csd.auth.gr
http://www.aiia.csd.auth.gr/index.html

Abstract. In this work, we model speech samples with a two-sided generalized Gamma distribution and evaluate its efficiency for voice activity detection. Using a computationally inexpensive maximum likelihood approach, we employ the Bayesian Information Criterion for identifying the phoneme boundaries in noisy speech.

1 Introduction

A common problem in many areas of speech processing is the identification of the presence or absence of a voice component in a given signal, especially the determination of the beginning and ending boundaries of voice segments. In many cases voice activity detection (VAD), endpoint detection, speaker segmentation, and audio classification can be seen as similar problems and they share a common methodology. In this work, we are interested in off-line VAD algorithms that are suitable for applications such as automatic transcription and speech segmentation in broadcast news. Our goal is to implement and evaluate a robust VAD and end-point detection algorithm under noisy conditions, various classes of noise, and short frames. Categorisation of audio signal at such a small scale has applications to phoneme segmentation and consequently, to speech recognition and speech synthesis. In particular, speech synthesis requires accurate knowledge of phoneme transitions, in order to obtain a naturally sounding speech waveform from stored parameters.

The detection principles of conventional VADs are usually based on the signal full-band energy levels, subband short-energy, Itakura linear prediction coefficient (LPC) distance measure [1], spectral density, zero crossing rate, Teager energy operator [2], cepstral coefficients, line spectral frequencies, etc. Energy-based approaches have been proved to work relatively well in high signal to noise ratios (SNR) and for known stationary noise [1]. Moreover, these methods have been proved to be computationally efficient to such an extent that they allow real-time signal processing [3]. But in highly noisy environments the performance and robustness of energy-based voice activity detectors are not optimal. An important disadvantage is that they rely on simple energy thresholds, so they are not able to identify unvoiced speech segments like fricatives satisfactorily, because the latter can be masked by noise. They

may also incorrectly classify clicking and other non-stationary noise as speech activity. Furthermore, they are not always very efficient in real-world recordings where speakers tend to leave "artifacts" including breathing/sighing, teeth chatters, and echoes. Many efforts have been made to use adaptive schemes for noise estimation but these are usually heuristic-based or limited to stationary noise. A discussion for classical and geometrically adaptive energy threshold methods can be found in [4].

Recently, much research discussion has been done regarding the exploration of the speech and noise signal statistics for VAD and endpoint detection. Statistical model-based methods/approaches typically employ a statistical model with the decision rule being derived from the Likelihood Ratio Test (LRT) applied to a set of hypotheses [5]. These approaches can be further improved by incorporating soft decision rules [6] and High Order Statistics (HOS) [7]. The main disadvantage of statistical model-based methods is that they are more complicated than the energy-based detectors regarding computation time and storage requirements, so they have limited appeal in online applications. Furthermore, model-based segmentation does not generalize to unseen acoustic conditions.

In this paper, we present a statistical model-based method for VAD using the generalised version of the Gamma distribution and evaluate its performance in phoneme boundary identification under noisy environments.

2 Voice Activity Detection using the BIC criterion

Bayesian Information Criterion (BIC) is an asymptotically optimal method for estimating the best model using only an in-sample estimate [8], [9]. It is an approximation to minimum description length (MDL) and can be viewed as a penalized maximum likelihood technique [10]. BIC can also be applied as a termination criterion in hierarchical methods for clustering of audio segments: two nodes can be merged only if the merging increases the BIC value. BIC is more frequently applied for speaker-turn detection. However, nothing precludes its application to VAD.

In BIC, adjacent signal segments are modelled using different multivariate Gaussian distributions (GD) while their concatenation are assumed to obey a third multivariate Gaussian pdf, as in Fig.1. The problem is to decide whether the data in the large segment fits better a single multivariate Gaussian pdf or whether a two-segment representation describes more accurately the data. A sliding window moves over the signal making statistical decisions at its middle using BIC. The step-size of the sliding window indicates the resolution of the system.

The problem is formulated as a hypothesis testing problem. For the purpose of VAD, we need to evaluate the following statistical hypotheses:

- H_0: $(x_1,x_2,...,x_B) \sim N(\mu_Z,\Sigma_Z)$: the sequence data comes from one source Z (i.e., noisy speech)
- H_1: $(x_1,x_2,...,x_A) \sim N(\mu_X,\Sigma_X)$ and $(x_{A+1},x_{A+2},...,x_B) \sim N(\mu_Y,\Sigma_Y)$: the sequence data comes from two sources X and Y, meaning that there is a transition from speech utterance to silence or vice versa

where x_i are K-dimensional feature vectors in a transformed domain such as Mel Frequency Cepstral Coefficients (MFCCs). In the example of Fig.1 Σ_X, Σ_Y, Σ_Z are respectively the covariance matrices of the complete sequence Z and the two subsets X and Y while A and B-A are the number of feature vectors for each subset.

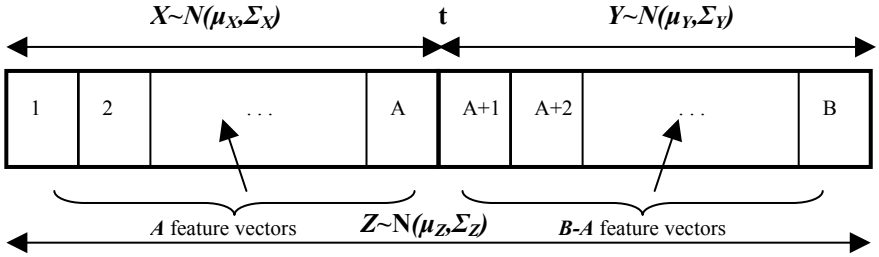

Fig. 1. Models for two adjacent speech segments.

The GLR associated with the defined hypothesis test is

$$R=\frac{L(z,\mu_z;\Sigma_z)}{L(x,\mu_x;\Sigma_x)L(y,\mu_y;\Sigma_y)} \tag{1}$$

where, for example, $L(x,\mu_x;\Sigma_x)$ represents the likelihood of the sequence of feature vectors X given the multi-dimensional Gaussian process $N(x,\mu_x;\Sigma_x)$. $L(y,\mu_y;\Sigma_y)$ and $L(z,\mu_z;\Sigma_z)$ are similarly defined. The distance between the two segments in Fig.1 is obtained using the log-value of this ratio,

$$d_R = -\log R \tag{2}$$

The variation of the BIC value between the two models is given by [11]

$$\Delta BIC = -\left(\frac{B}{2}\log|\Sigma_z| - \frac{A}{2}\log|\Sigma_x| - \frac{B-A}{2}\log|\Sigma_y|\right) + \lambda P \tag{3}$$

$$P = \frac{1}{2}\left[K+\frac{1}{2}K(K+1)\right]\times\log B \tag{4}$$

where P is the penalty for model complexity and λ a tuning parameter for the penalty factor. Negative values of ΔBIC indicate that the multi-dimensional Gaussian mixtures best fit the data, meaning that t is change point from speech to silence or vice versa.

One advantage of BIC is that it does not involve thresholds, but there is still the penalty factor λ which is depended on the type of analysed data and must be estimated heuristically [12]. Also, BIC tends to choose models that are too simple due to heavy penalty on complexity. Nevertheless, BIC is a consistent estimate and various algorithms based on this criterion have extended the basic algorithm combining it with

other metrics such as Generalized Likelihood Ratio (GLR), Kullback-Leibler (KL) distance, sphericity tests and other HOS [7].
A variant of BIC that attempts to deal with some of the problems mentioned above is DISTBIC [11]. The algorithm performs two steps. The first step uses a distance computation to choose the possible candidates for a change point. Different criteria such as KL and GLR can be applied to this pre-segmentation step. The second step uses the BIC to validate or discard the candidates determined in the first step.

3 Speech distributions

One common assumption for most VAD algorithms that operate in the DFT domain, such as DISTBIC, is that both noise and speech spectra can be modelled satisfactorily by GDs. Furthermore, using a transformed feature space, it is possible to assume that these two Gaussian random processes are independent of each other. When both speech and noise are Gaussians the Minimum Mean-Squared Error Estimator (MMSE) estimator is linear and the spectral coefficients of the filter are real. Consequently, the clean speech and noisy coefficients differ only in magnitude. In such cases, maximum a posteriori estimators can be used to determine the signal parameters. Another reason for selecting GD in parametric modelling is the central limit theorem. This justifies the GD for speech signals in the DFT domain. Generally, GDs are simple to use, their properties are theoretically clear and consequently, they have been used extensively for statistical modelling in VAD and generally in speech processing.

Nevertheless, previous work has demonstrated that Laplacian (LD), Gamma (ΓD), and the Generalized Gaussian Distribution (GGD) can yield better performance than GD. Using likelihoods, coefficients of variation (CVs), and Kolmogorov-Smirnov (KS) tests, [13] has concluded that LD is more suitable for approximating active voice segments in most cases and for different frame sizes. More specifically, LD fits well the highly correlated univariate space of the speech amplitudes as well as the uncorrelated multivariate space of the feature values after a Karhunen- Loeve Transformation (KLT) or Discrete Cosine Transformation (DCT) [14].

While some reports attest that LD offers only a marginally better fit than GD, this is not valid when silence segments are absent from the testing [13]. The reason is that while clean speech segments best exhibit Laplacian or Gamma statistical properties the silence segments are Gaussian random processes. [15] and others have also asserted that Laplacian and Gamma distributions fit better the voiced speech signal than normal distributions.

Both LD and GD are members of the family of exponential distributions and can be considered as special cases of ΓD. They are specified by the location and the scale parameters. But the LD compared to GD has a sharper peak and a wider spread so that samples far from the mean are assigned a higher likelihood than in GD. Due to this property of the decreased likelihood, in segments that include periods locally deviating from the model the correct hypothesis is less likely to be rejected. On the other hand, both distributions have insufficient representation powers with respect to the distribution shape [16].

4 DISTBIC using Generalized Gamma Distribution

Our goal is to accomplish the task of phoneme boundary identification without any previous knowledge of the audio stream while achieving a robust performance under noisy environments. For this purpose, we try to improve the performance of the DISTBIC algorithm in VAD. In order to model the signal in a more theoretically complete manner and to conform to the experimental findings mentioned in Sect. 3, we need to consider generalized versions of distributions that have GD and LD as special cases. A generalized LD (GLD) [16] or the two-sided generalized gamma distribution (GΓD) [17] for an efficient parametric characterization of speech spectra can be used. In this paper, we modify the first step of the DISTBIC algorithm by assuming a GΓD distribution model for our signal in the analysis windows. Let us call the proposed method DISTBIC-Γ from now on.

The GΓD is an extremely flexible distribution that is useful for reliability modelling. Among its many special cases are the Weibull and exponential distributions. It is defined as

$$f_X(x)=\frac{\gamma\beta^{\eta}}{2\Gamma(\eta)}|x|^{\eta\gamma-1}e^{-\beta|x|^{\gamma}} \tag{5}$$

where $\Gamma(z)$ denotes the gamma function and γ, η, β are real values corresponding to location, scale and shape parameters. GD is a special case for γ=2 and η=0.5, for (γ=1 and η=1) it represents the LD, while for (γ=1 and η=0.5) it represents the common ΓD.

The parameter estimation of this family of distributions with both known and unknown location parameters can be achieved using the maximum likelihood estimation (MLE) method. Unfortunately, estimating the parameters of these distributions with MLE in an analytic way is difficult because the maximised likelihood results in nonlinear equations. Regarding GLD, the location parameter can be numerically determined by using the gradient ascend algorithm according to the MLE principle. Using a learning factor we can then reestimate the location value that locally maximizes the logarithmic likelihood function L, until L reaches convergence. Using this value and the data samples we can determine the scale and shape parameters. It must be stated that in MLE it is difficult to apply the same analytical procedure for the shape parameter and moreover the convergence of the gradient method is not always good [16]. The principal parameters of GΓD are estimated similarly according to the MLE principle. A computationally inexpensive on-line algorithm based on the gradient ascent algorithm is introduced in [17]. Here, we use the online ML algorithm proposed by [18]. Given N data $x=\{x_1,x_2,...x_N\}$ of a sample and assuming the data is mutually independent, we iteratively update the statistics

$$S_1(n)=(1-\lambda)S_1(n-1)+\lambda\left|x_n\right|^{\hat{\gamma}(n)} \tag{6}$$

$$S_2(n)=(1-\lambda)S_2(n-1)+\lambda\log\left|x_n\right|^{\hat{\gamma}(n)} \tag{7}$$

$$S_3(n) = (1-\lambda)S_3(n-1) + \lambda \left|x_n\right|^{\hat{\gamma}(n)} \log \left|x_n\right|^{\hat{\gamma}(n)} \quad (8)$$

over the frame N values, updating each time the parameter γ as

$$\hat{\gamma}(n+1)=\hat{\gamma}(n)+\mu\left(\frac{1}{\hat{\eta}(n)}+S_2(n)-\frac{S_3(n)}{S_1(n)}\right) \quad (9)$$

where λ is a forgetting factor and μ is the learning rate of the gradient ascent approach. Using appropriate initial estimates for the parameters (e.g. $\hat{\gamma}(1)=1$, which corresponds to GD or LD), we are able to recursively estimate the remaining parameters by solving the equations:

$$\psi_0(\hat{\eta}(n)) - \log \hat{\eta}(n) = S_2(n) - \log S_1(n) \quad (10)$$

$$\hat{\beta}(n)=\frac{\hat{\eta}(n)}{S_1(n)} \quad (11)$$

where ψ_0 is the digamma function. The left part of (10) is monotonically increasing function of $\hat{\eta}(n)$, so we are able to uniquely determine the solution by having an inverse table.

The proposed algorithm, DISTBIC-Γ, is implemented in two steps. First, using a sufficiently big sliding window and modelling it and its adjacent sub-segments using GΓD instead of GD, we calculate the distance associated with the GLR. Here, as in [14], we are making the assumption that both noise and speech signals have uncorrelated components in the DCT domain. Depending on the window size, this assumption gives a reasonable approximation for their multivariate PDFs using the marginal PDFs. A potential problem arouses when using MLE for short segments [17]. Nevertheless, we can relax the convergence conditions of the gradient ascend method and still yield improved results. Then, we create a plot of the distances as output with respect to time and filter out insignificant peaks using the same criteria as [11].

In the second step, using the BIC test as a merging criterion we compute the ΔBIC values for each change point candidate in order to validate the results of the first step. Because small frame lengths suggest a GD according to [13] and due to the length limitation of the [18] method for GΓD parameter estimation, we choose to use Gaussians in this step.

5 Experiments

The testing of automatic phoneme boundary detection requires an expansive, annotated speech corpus. In order to evaluate the performance of the proposed method, two sets of preliminary experiments on VAD were conducted on two different corpora. In the first experiment we compare the efficiency of the proposed method using samples from the M2VTS audio-visual database [19]. In our tests we used 15 audio recordings that consist of the utterances of ten digits from zero to nine in French. We

measured the mismatch between manual segmentation of audio performed by a human transcriber and the automatic segmentation. Samples were manually segmented or concatenated using sound editing software. The human error and accuracy of visually and acoustically identifying break points were taken into account. In the second set of experiments we used samples from the TIMIT dataset [20] totaling 100 seconds of speech time. For both experiments we used the same set of parameter values and features were used (500ms initial window, 5ms shift of analysis window, first 12 MFCCs for GD, 10 DCTs for GΓD, λ=7). White and babble noise from the NOISEX-92 database [21] was added to the clean speech samples at various SNR levels ranging from 20 to 5 dB.

The errors that can be identified in VADs are distinguished by whether speech is misclassified as noise or vice versa, and by the position in an utterance in which the error occurs (beginning, middle or end). A point incorrectly identified as a change point gives a type-2 error (false alarm) while a point totally missed by the detector is a type-1 error (missed detection). The detection error rate of the system is described by False Alarm Rate (FAR) and Missed Detection Rate (MDR) defined below. ACP stands for the Actual Change Points in the signal as determined by human in our case.

$$FAR = \frac{\text{number of FA}}{\text{number of AST} + \text{number of FA}} * 100\% \quad (12)$$

$$MDR = \frac{\text{number of MD}}{\text{number of AST}} * 100\% \quad (13)$$

A high value of FAR means that an over-segmentation of the speech signal is obtained while a high value of MDR means that the algorithm does not segment the audio signal properly. The results for the VAD error rates are illustrated in Table 1, 2, 3, and 4.

Comparing the four tables, we can deduce that there is notable improvement especially at low SNR levels. These results showed that the VAD accuracy of the proposed method DISTBIC-Γ is improved by an average of 16% in FAR and 23.8% in MDR compared to that of the conventional DISTBIC method. We can also indicate the improvement in the recognition of unvoiced speech elements. The improved results denote the higher representation power of the GΓD distribution.

Table 1. Performance of VAD in M2VTS (voiced phonemes)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR (%)**	**MDR(%)**	**FAR (%)**	**MDR(%)**
(clean speech)	-	22.6	16.4	27.5	19.2
white	20	24.8	19.7	29.4	23.4
white	10	25.1	20.3	30.5	24.4
white	5	28.2	23.5	35.4	29.8
babble	20	27.5	21.3	32.7	24.7
babble	10	28.8	24.1	34.9	28.4
babble	5	31.4	26.8	38.5	32.7

Table 2. Performance of VAD in TIMIT (voiced phonemes)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR (%)**	**MDR(%)**	**FAR (%)**	**MDR(%)**
(clean speech)	-	25.5	17.1	31.4	19.6
white	20	26.9	18.5	32.2	23.2
white	10	29.4	22.8	33.3	26.9
white	5	32.4	25.0	38.8	31.9
babble	20	30.5	20.6	33.8	25.6
babble	10	31.8	24.4	36.6	29.3
babble	5	34.9	27.7	40.8	35.3

Table 3. Performance of VAD in M2VTS (voiced + unvoiced phonemes)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR (%)**	**MDR(%)**	**FAR (%)**	**MDR(%)**
(clean speech)	-	27.5	18.2	29.9	21.5
white	20	28.9	19.4	32.5	23.9
white	10	30.3	22.5	35.4	28.0
white	5	34.1	25.9	39.8	33.1
babble	20	28.8	21.6	33.5	26.2
babble	10	31.6	24.2	37.6	31.4
babble	5	36.1	28.0	42.4	37.5

Table 4. Performance of VAD in TIMIT (voiced + unvoiced phonemes)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR (%)**	**MDR(%)**	**FAR (%)**	**MDR(%)**
(clean speech)	-	28.3	18.7	32.1	24.5
white	20	31.2	20.9	34.5	25.7
white	10	33.1	24.4	37.5	30.5
white	5	35.7	28.5	40.4	36.4
babble	20	33.1	22.1	35.0	27.8
babble	10	34.0	25.8	38.6	32.5
babble	5	36.5	29.1	43.5	38.7

6 Conclusions

The identification of phoneme boundaries in continuous speech is an important problem in areas of speech synthesis and recognition. As we have seen, there are numerous combinations worth exploring for offline 2-step speech activity detection. We have demonstrated that by representing the signal samples with a GΓD we are able to yield improved results than simple normal distributions. We concluded that the Generalised Gamma model is more adequate characterising noisy speech than the Gaussian model. This conforms with the findings of [22]. Also, despite making assumptions on the correlation of distribution components for the computation of the likelihood ratio in GΓD, the proposed algorithm, DISTBIC-Γ, yielded better results than DISTBIC especially in noisy signals. Using the KL distance instead of GLR we could further improve the system performance since the first offers better discriminative ability [23]. Nevertheless, the size of experiments was limited and the computation time was multiple times greater for the method with GΓDs than using the simple

DISTBIC algorithms. An open problem worth investigating is the criterion for convergence in the gradient ascend algorithm.

7 Acknowledgements

G. Almpanidis was granted a basic research fellowship "HERAKLEITOS" by the Greek Ministry of Education.

References

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances", *Bell Syst. Tech. Journal,* vol. 54, no. 2, pp. 297-315, 1975.

[2] G. S.Ying, C. D. Mitchell, and L. H. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement", In Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp.732-735, 1992.

[3] A. Ganapathiraju, L. Webster, J. Trimble, K. Bush, and P. Kornman., "Comparison of Energy-Based Endpoint Detectors for Speech Signal Processing", in Proc. *IEEE Southeastcon Bringing Together Education, Science and Technology*, pp. 500-503, Florida, April 1996.

[4] S. Tanyer and H. Ozer "Voice activity detection in nonstationary noise", *IEEE Trans. Speech and Audio Processing,* vol. 8, no. 4, pp.478-482, April 2000.

[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model based voice activity detection", *IEEE Signal Processing Letters*, vol. 6, no. 1 pp.1-3, January 1999.

[6] J. Chang, J. Shin, and N. S. Kim, "Likelihood ratio test with complex Laplacian model for voice activity detection", in Proc. *European Conf. Speech Communication Technology*, 2003.

[7] E. Nemer, R. Goubran, and S. Mahmould, "Robust voice activity detection using higher-order statistics in the LPC residual domain", *IEEE Trans. Speech and Audio Processing*, vol.9, no. 3, pp. 217-231, March 2001.

[8] G. Schwartz, "Estimating the dimension of a model", *Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[9] S. Chen and P. Gopalakrishnam, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion", in *DARPA Broadcast News Workshop,* 1998.

[10] P. Grunwald, "Minimum description length tutorial", *Advances in Minimum Description Length: Theory and Applications.* pp. 23-80. Cambridge, MA: MIT Press.

[11] P. Delacourt and C. J. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing", *Speech Communication*, vol. 32, no. 1-2, pp. 111-126, September 2000.

[12] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion", in Proc. *1999 European Speech Processing*, vol. 2, pp. 679-682, 1999.

[13] S. Gazor and W. Zhang, "Speech probability distribution", *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204-207, July 2003.

[14] S. Gazor and W. Zhang "A soft voice activity detector based on a Laplacian-Gaussian model", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 498-505, 2003.

[15] R. Martin, "Speech enhancement using short time spectral estimation with Gamma distributed priors", in Proc. *IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, vol. 1, pp. 253-256, 2005.

[16] A. Nakamura, "Acoustic modeling for speech recognition based on a generalized Laplacian mixture distribution", *Electronics and Communications in Japan Part II: Electronics*, vol. 85, no. 11, pp. 32-42, October 2002.

[17] W. -H. Shin, B. –S. Lee, Y. –K. Lee, and J. –S. Lee, "Speech/non-speech classification using multiple features for robust endpoint detection", in Proc. *IEEE Intl Conf. Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1399-1402, 2000.

[18] J.W. Shin and J-H. Chang, "Statistical Modeling of Speech Signals Based on Generalized Gamma Distribution", *IEEE Signal Processing Letters*, vol. 12, no. 3 pp.258-261, March 2005.

[19] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database", in *Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication,* (J. Bigun, G. Chollet, and G. Borgefors, Eds.), vol. 1206, pp. 403-409, 1997.

[20] TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech. Disc 1-1.1, NTIS Order No. PB91-505065, 1990.

[21] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the affect of additive noise on automatic speech recognition", Technical Report, DRA Speech Research Unit, Malvern, England, 1992.

[22] J.W. Shi, J-H Chang, H.S. Yun, and N.S. Kim, "Voice Activity Detection based on Generalized Gamma Distribution", in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 781-784, 2005.

[23] J. Ramirez, C. Segura, C. Benitez, A. Torre, and A. Rubio, "A new Kullback-Leibler VAD for speech recognition in noise", *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 266-269, February 2004.