VOICE ACTIVITY DETECTION WITH GENERALIZED GAMMA DISTRIBUTION

George Almpanidis and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki,
Box 451 Thessaloniki, GR-54124, Greece
{galba, costas}@aiia.csd.auth.gr

ABSTRACT

In this work, we model speech samples with the generalized Gamma distribution and evaluate the efficiency of such modelling 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 starting and ending boundaries of voice segments. In this work, we are interested in offline multi-pass voice activity detection (VAD) algorithms suitable for speech communication applications applications such as automatic transcription and speech segmentation. Our goal is to implement and evaluate a robust algorithm for noisy environments, various classes of noise, and short frames. Categorisation of audio signal at a small scale has applications to phoneme segmentation and consequently, to speech recognition and synthesis.

The detection principles of conventional VADs are usually energy-based approaches, which have been proved computationally efficient to such an extent that they allow real-time signal processing [1]. Moreover, these methods work relatively well in high signal to noise ratios (SNR) and for known stationary noise. But in very noisy environments the performance and robustness of energy-based voice activity detectors are not optimal. Because they rely on simple energy thresholds, they are not able to identify unvoiced speech segments like fricatives satisfactorily, since the latter can be masked by noise. They may also misclassify non-stationary noise such as clicking as speech activity. Furthermore, they are inefficient in real-world recordings where speakers tend to leave "artifacts" including breathing/sighing, teeth chatters, and echoes.

Recently, much research has been done with respect to the exploration of the speech and noise signal statistics for VAD. Statistical model-based methods typically employ a decision rule derived from the likelihood ratio test (LRT) applied to a set of hypotheses [2]. These approaches can be further improved by incorporating soft decision rules [3] and high-order statistics (HOS) [4]. The main disadvantage of statistical model-based methods is that they are more complicated than the energy-based detectors with respect to the computation time and storage requirements, so they have a limited appeal in online applications.

In this paper, we present a statistical model-based method for VAD using the generalised version of the Gamma distribution, which offers more efficient parametric characterisation of the speech spectra than the Gaussian distribution (GD). We evaluate the system performance in the identification of phoneme boundaries. From the results presented in Sect. 5 we attest that the proposed method yields significant improvements in noisy environments.

2. BAYESSIAN INFORMATION CRITERION

The Bayesian Information Criterion (BIC) is an asymptotically optimal method for estimating the best model using only an in-sample estimate [5]. It can be viewed as a penalized maximum likelihood technique. BIC can also be applied as a termination criterion in hierarchical methods for clustering of audio segments: two nodes can be merged only if this increases the BIC value.

In BIC, adjacent signal segments are modelled using different multivariate GDs while their concatenation is assumed to obey a third multivariate GD, as in Fig. 1. The problem is to decide whether the data in the large segment fit better a single Gaussian or whether a two-segment representation describes it more accurately. A sliding window moves over the signal making statistical decisions at its middle. The step-size of the sliding window indicates the resolution of the system. 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 data sequence 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 data sequence 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). Let Σ_Z, Σ_X, and Σ_Y be the covariance matrices of the feature vectors over the complete sequence Z and the two subsets X and Y, while A and B-A are the numbers of feature vectors in X and Y, respectively.

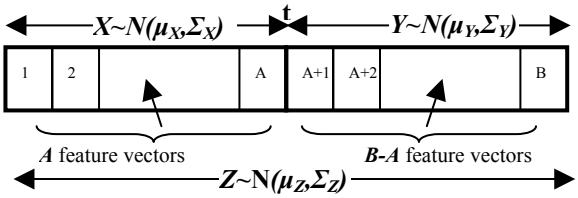

Fig.1. Models for two adjacent speech segments.

Using the log-value of the Generalized Likelihood Ratio Test (GLRT), associated with the defined hypothesis test the distance between the two segments in Fig. 1 is:

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

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 $L(z,\mu_z;\Sigma_z)$ can be similarly defined. The variation of the BIC value between the two models is given by [6]

$$\Delta BIC = -\left(\frac{B}{2}\log|\Sigma_Z| - \frac{A}{2}\log|\Sigma_X| - \frac{B-A}{2}\log|\Sigma_Y|\right) + \lambda P \quad (2)$$

$$P = \frac{1}{2}\left[K + \frac{1}{2}K(K+1)\right]\times \log B \quad (3)$$

where P is the penalty for model complexity and λ a tuning parameter for the penalty factor. Negative ΔBIC values indicate that the multi-dimensional Gaussian mixtures best fit the data, meaning that t is a change point from speech to silence or vice versa. BIC does not involve thresholds, but there is still the factor λ that depends on the type of analysed data and must be estimated heuristically [7]. Also, BIC tends to choose oversimplistic models due to the heavy penalty on the complexity. Nevertheless, BIC is a consistent estimate and various algorithms have extended the basic method combining it with other metrics such as Kullback-Leibler (KL) distance, sphericity tests, and HOS [4], [6].

Such a variant of BIC that attempts to deal with some of the problems mentioned above is DISTBIC [6]. The algorithm performs two steps. First, it uses a distance computation to choose the possible candidates for a change point. Different criteria such as KL or GLRT can be applied to the first step of DISTBIC. In the second step, ΔBIC values are used in order to validate or discard the candidates determined in the first step.

3. SPEECH DISTRIBUTIONS

A common assumption for most VAD algorithms that operate in the DFT domain is that both noise and speech spectra can be modelled satisfactorily by GDs. Using a transformed feature space, it is possible to assume that these two Gaussian random processes are independent of each other and maximum a posteriori estimators can be used to determine the signal parameters. Nevertheless, previous work in speech processing has demonstrated that Laplacian (LD) and Gamma (ΓD) distributions are more suitable than GD for approximating active voice segments for most frame sizes [8]. In specific, 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) [9]. 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 [8]. The reason is that while clean speech segments best exhibit LD or ΓD properties the silence segments are Gaussian random processes. [10] have also asserted that LDs and ΓDs fit better the voiced speech signal than normal distributions.

4. DISTBIC USING GENERALIZED GAMMA DISTRIBUTION

Our goal is to identify phoneme boundaries without any previous knowledge of the audio stream while achieving a robust performance under noisy environments. For this purpose we propose an improved version of the DISTBIC algorithm, DISTBIC-Γ, where the signal is modelled using the generalized ΓD (GΓD) [11]. Considering the experimental findings mentioned in Sect. 3 we modify the first step of the DISTBIC algorithm by assuming a GΓD distribution model for our signal in the analysis windows.

The GΓD is an extremely flexible distribution that is useful for reliability modelling. It is defined as

$$f_x(x) = \frac{\gamma\beta^{\eta}}{2\Gamma(\eta)} |x|^{\eta\gamma-1} e^{-\beta|x|^{\gamma}} \quad (4)$$

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 can be achieved using the maximum likelihood estimation (MLE) method. Unfortunately, estimating the parameters in an analytic way is difficult because the MLE results in nonlinear equations. A computationally inexpensive online MLE algorithm for GΓD, based on the gradient ascent algorithm, is introduced in [12]. The location parameter is numerically determined by using the gradient ascend algorithm according to the MLE principle. Using a learning factor, we can then re-estimate the location value that locally maximizes the logarithmic likelihood function L, until L convergences. Using this value and the data samples, we can determine the scale and shape parameters. Given N data $x = \{x_1, x_2, ...x_N\}$ of a sample and assuming the data are

mutually independent, we iteratively update the following statistics over the frame N values

$$S_1(n) = (1-\xi)S_1(n-1) + \xi \left|x_n\right|^{\hat{\gamma}(n)} \quad (5)$$

$$S_2(n) = (1-\xi)S_2(n-1) + \xi \log\left|x_n\right|^{\hat{\gamma}(n)} \quad (6)$$

$$S_3(n) = (1-\xi)S_3(n-1) + \xi \left|x_n\right|^{\hat{\gamma}(n)} \log\left|x_n\right|^{\hat{\gamma}(n)} \quad (7)$$

updating each time the parameter γ as

$$\hat{\gamma}(n+1) = \hat{\gamma}(n) + \mu\left(\frac{1}{\hat{\eta}(n)} + S_2(n) - \frac{S_3(n)}{S_1(n)}\right) \quad (8)$$

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

$$\psi_0(\hat{\eta}(n)) - \log\hat{\eta}(n) = S_2(n) - \log S_1(n) \quad (9)$$

$$\hat{\beta}(n) = \frac{\hat{\eta}(n)}{S_1(n)} \quad (10)$$

where ψ_0 is the digamma function. The left part of (9) is monotonically increasing function of $\hat{\eta}(n)$, so we can determine uniquely the solution by having an inverse table.

The proposed algorithm, DISTBIC-Γ, is implemented in two steps. First, we select a sufficiently big sliding window, model it and its adjacent sub-segments using GΓD instead of GD, and calculate the distance d_R associated with the GLRT using (1). Here, as in [9], 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 marginals. A potential problem arouses when using MLE for short segments but we can relax the convergence conditions of the gradient ascend method and still yield improved results [11]. Then, we create a distance plot as output with respect to time and filter out insignificant peaks using the same criteria as [6]. 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 [8] and due to the length limitation of the gradient ascend method for GΓD parameter estimation, we can use GDs in this step.

5. EXPERIMENTS

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 [13]. 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 the manual segmentation of audio performed by a human transcriber and the automatic segmentation. The human error and accuracy of visually and acoustically identifying break points were taken into account. In the second set of experiments we used 25 utterances from the TIMIT dataset [14] totalling 100 seconds of speech time. For both experiments we used the same set of parameter values and features (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 [15] was added to the clean speech samples at SNR levels ranging from 20 to 5 dB.

The detection performance of the system is described by precision (*PRC*) and recall (*RCL*) rates

$$PRC = \frac{CFC}{DET}100\% \quad RCL = \frac{CFC}{ACP}100\% \quad (11)$$

where *CFC* denotes the number of correctly found changes, *DET* is the number of changes detected by the system, and *ACP* is the actual change points. The overall effectiveness of the system can be measured by the F_1-measure:

$$F_1 = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL} \quad (12)$$

The results of our tests are illustrated in Table 1, 2, 3, and 4. For each case, we calculate the average *PRC*, *RCL*, and F_1 rates over the test samples. Examining the average F_1-measure for each case using a two-sample one-tailed *t* test we see that the DISTIBIC-Γ performance is superior to DISTBIC at a confidence level of 0.05 for every case since the *t* values are larger than the corresponding critical values (t_0=1.701 for M2VTS and t_0=1.677 for TIMIT). The improvement is notable especially at low SNR levels. We can also indicate the improvement in the recognition of unvoiced speech elements. [16] have also used GΓD recently for modeling speech and noise in the DFT domain. Their LRT-based VAD obtained better results under vehicular and office noise than conventional methods. In our work we have presented a more robust threshold-tuning method based on DISTBIC and asserted that we can operate in the DCT domain as well with similar success. In both works, the improved experimental results over existing GD-based methods denote the higher representation power of the GΓD.

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

		DISTBIC-Γ			DISTBIC			
Noise	**SNR**	**PRC**	**RCL**	**F_1**	**PRC**	**RCL**	**F_1**	**t**
(clean)	-	76.3	83.5	79.6	71.2	80.8	75.6	4.65
white	20	73.1	80.4	76.5	68.0	76.5	71.9	5.52
white	10	72.6	79.6	75.9	67.0	75.7	70.8	6.04
white	5	68.8	76.5	72.3	61.0	70.2	64.5	9.11
babble	20	70.4	78.8	74.3	65.0	75.3	69.7	5.52
babble	10	68.0	75.7	71.6	62.0	71.8	66.3	6.52
babble	5	65.0	73.3	68.8	57.0	67.5	61.8	8.98

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

		DISTBIC-Γ			DISTBIC			
Noise	**SNR**	**PRC**	**RCL**	**F_1**	**PRC**	**RCL**	**F_1**	**t**
(clean)	-	71.4	82.2	76.4	68.4	78.5	73.1	6.57
white	20	70.7	81.0	75.0	65.0	75.9	70.3	5.68
white	10	68.0	76.9	72.2	60.7	71.3	66.3	7.31
white	5	62.9	74.4	68.3	54.5	65.1	60.6	9.84
babble	20	69.9	79.0	73.5	63.7	73.8	68.4	6.25
babble	10	67.4	76.9	70.5	57.4	67.7	62.3	9.46
babble	5	60.9	71.8	66.0	51.0	62.1	56.6	12.3

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

		DISTBIC-Γ			DISTBIC			
Noise	**SNR**	**PRC**	**RCL**	**F_1**	**PRC**	**RCL**	**F_1**	**t**
(clean)	-	73.6	82.9	77.9	68.3	80.4	73.8	4.86
white	20	72.0	81.4	76.4	66.1	76.8	70.1	6.48
white	10	68.2	77.3	72.4	63.5	73.2	67.9	5.51
white	5	64.9	75.0	69.6	57.1	68.1	62.1	9.44
babble	20	68.3	79.3	73.3	63.8	74.4	68.6	5.75
babble	10	65.6	75.5	70.2	60.6	70.6	65.2	6.24
babble	5	61.9	72.3	66.7	54.0	64.7	58.8	10.1

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

		DISTBIC-Γ			DISTBIC			
Noise	**SNR**	**PRC**	**RCL**	**F_1**	**PRC**	**RCL**	**F_1**	**t**
(clean)	-	71.0	81.3	76.3	68.1	77.7	73.1	6.37
white	20	69.7	80.5	74.9	64.7	75.1	70.3	5.54
white	10	67.3	77.4	72.4	60.5	71.2	66.3	7.51
white	5	62.4	72.8	68.3	54.6	65.7	60.9	9.41
babble	20	68.7	77.7	73.5	62.7	72.7	68.3	6.26
babble	10	65.2	74.8	70.5	57.5	66.5	63.3	9.06
babble	5	60.0	71.7	65.6	51.2	61.8	56.7	12.1

6. CONCLUSIONS

The identification of phoneme boundaries in continuous speech is an important problem in areas of speech synthesis and recognition. We have demonstrated that by representing the signal samples with a GΓD we are able to yield statistically more significant results than the normal distribution for offline 2-step VAD.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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, Sept. 2000.

[7] 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.

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

[9] 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.

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

[11] 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 Int. Conf. Acoustics, Speech, and Signal Processing,* vol. 3, pp. 1399-1402, 2000.

[12] 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.

[13] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database", in *Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication,* vol. 1206, pp. 403-409, 1997.

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

[15] 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.

[16] J.W. Shin, 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.