Phoneme Segment Boundary Detection Based on the Generalized Gamma Distribution

George Almpanidis and Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki,
{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 segment boundaries in noisy speech.**

I. 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 voice activity detection (VAD) and acoustic change detection algorithms suitable for applications such as automatic transcription and speech segmentation in video editing. Our goal is to implement and evaluate a robust detection algorithm for noisy signals, 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.

Conventional systems follow energy-based approaches, which have been proved computationally efficient, almost allowing real-time signal processing [1]. These methods work relatively well in high signal to noise ratios (SNR) and for known stationary noise. But in low SNRs, the performance and robustness of energy-based voice activity detectors is not optimal. Since they rely on simple energy thresholds, they are not able to identify unvoiced speech segments like fricatives satisfactorily, as 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 artefacts such as 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 higher order statistics (HOS) [4]. Since model-based methods are more complicated than the energy-based detectors with respect to the computation time and storage requirements, 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 segment boundaries. From the results presented in Sect. 5 we attest that the proposed method yields significant improvements in noisy environments.

II. Speech Modelling

A critical parameter that affects the performance of statistical-based VAD methods is the choice of the distribution for the modelling of clean speech and noise/silence. A common assumption for most VAD algorithms is that both noise and speech spectra can be modelled satisfactorily by GDs. Furthermore, using a transformed feature space, it is also possible to assume that these two Gaussian random processes are independent of each other and the spectral coefficients of the clean speech and noise differ only in magnitude. In such cases, maximum a posteriori estimators can be used to determine the signal parameters.

Nevertheless, the choice of the GD is generally justified on its simplicity and its nice theoretical properties (e.g. central limit theorem). 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 many frame sizes [5], [6]. 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) [7]. While some reports attest that LD offers only a marginally better fit than GD, this is not valid when silence is absent from the test [5]. The reason is that while clean speech segments best exhibit LD or ΓD properties silence is more efficiently modelled by a Gaussian random process.

Recently, it has been asserted that the generalized ΓD (GΓD) fits the voiced speech signal even better than Gamma, Laplacian, and normal distributions [8]. GΓD is defined as

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

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 of (1) for γ=2 and η=0.5. For γ=1 and η=1 (1) yields the LD, while for γ=1 and η=0.5 it represents the common ΓD.

Although the GΓD is an extremely flexible distribution it has been used mostly in reliability modelling and life data analysis. Until recently, little interest was shown in speech processing literature, the main reason being its complexity. Estimating the parameters of GΓD in analytically using the maximum likelihood estimation (MLE) method is difficult, because the maximised likelihood results in nonlinear equations involving numerical integrations. A computationally inexpensive on-line algorithm for GΓD, based on the gradient ascent algorithm, has been introduced in [8]. The location parameter is 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 convergences. Using this value and the data samples we can determine the scale and shape parameters. Given N mutually independent data $x = \{x_1, x_2, ... x_N\}$, we iteratively update the following statistics over the N frame values

$$S_1(n) = (1-\xi)S_1(n-1) + \xi \left|x_n\right|^{\hat{\gamma}(n)} \tag{2}$$

$$S_2(n) = (1-\xi)S_2(n-1) + \xi \log\left|x_n\right|^{\hat{\gamma}(n)} \tag{3}$$

$$S_3(n) = (1-\xi)S_3(n-1) + \xi \left|x_n\right|^{\hat{\gamma}(n)} \log\left|x_n\right|^{\hat{\gamma}(n)} \tag{4}$$

updating each time the parameter γ as

$$\hat{\gamma}(n+1) = \hat{\gamma}(n) + \mu\left(\frac{1}{\hat{\eta}(n)} + S_2(n) - \frac{S_3(n)}{S_1(n)}\right) \tag{5}$$

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 are able to recursively estimate the remaining parameters by solving the equations:

$$\psi_0(\hat{\eta}(n)) \text{ - } \log\hat{\eta}(n) = S_2(n) \text{ - } \log S_1(n) \tag{6}$$

$$\hat{\beta}(n) = \frac{\hat{\eta}(n)}{S_1(n)} \tag{7}$$

where ψ_0 is the digamma function. The left part of (6) is monotonically increasing function of $\hat{\eta}(n)$, so we can determine uniquely the solution by having an inverse table.

III. BAYESSIAN INFORMATION CRITERION

The Bayesian Information Criterion (BIC) is an asymptotically optimal method for estimating the best model using only sample estimates [9]. 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 the merger increases the BIC value.

In BIC, adjacent signal segments are modelled using different multivariate GDs while their concatenation are 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 $V(m)$ 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: $(x_1,x_2,...,x_B)$~$N(\mu_Z,\Sigma_Z)$: the data sequence comes from one source Z (i.e., noisy speech)
- H_1: $(x_1,x_2,...,x_A)$~$N(\mu_X,\Sigma_X)$ and $(x_{A+1},x_{A+2},...,x_B)$~$N(\mu_Y,\Sigma_Y)$: 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 the Mel Frequency Cepstral Coefficients (MFCCs). Let Σ_X, Σ_Y, and Σ_Z be 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.

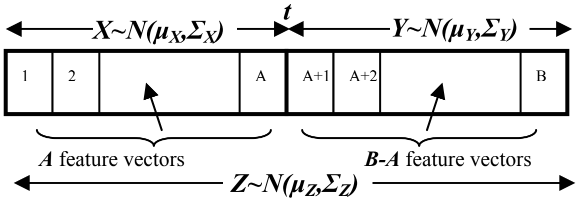

Fig. 1. Models for two adjacent speech segments.

The variation of the BIC value between the two models is given by [10]

$$\Delta BIC = -\log R + \lambda P \tag{8}$$

$$R = \frac{L(z,\mu_Z;\Sigma_Z)}{L(x,\mu_X;\Sigma_X)L(y,\mu_Y;\Sigma_Y)} \tag{9}$$

$$P = \frac{1}{2}\left[K + \frac{1}{2}K(K+1)\right] \times \log B \tag{10}$$

where R is the Generalized Likelihood Ratio Test (GLRT), P is the penalty for model complexity and λ is a tuning parameter for the penalty factor. In (9) $L(x,\mu_x;\Sigma_x)$ 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. Assuming multivariate GD modelling and we can easily calculate ΔBIC from sample values:

$$\Delta BIC = -\left(\frac{B}{2}\log\left|\Sigma_Z\right| - \frac{A}{2}\log\left|\Sigma_X\right| - \frac{B-A}{2}\log\left|\Sigma_Y\right|\right) + \lambda P \tag{11}$$

Negative ΔBIC values indicate that the multi-dimensional Gaussian mixture best fits 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 penalty factor λ that depends on the type of analysed data and must be estimated heuristically [11]. 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], [10].

Such a variant of BIC that attempts to deal with some of the problems mentioned above is DISTBIC. This is a two-pass distance-based algorithm that searches for change point candidates at the maxima of the distances computed between adjacent windows over the entire signal [10]. 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 this pre-segmentation step. Using the log-value of the GLRT, associated with the defined hypothesis test, the dissimilarity between the two subsequent segments of Fig. 1 is measured by

$$d_R = -\log R = -\log \frac{L(z,\mu_Z;\Sigma_Z)}{L(x,\mu_X;\Sigma_X)L(y,\mu_Y;\Sigma_Y)} \tag{12}$$

In the second step, ΔBIC values are used in order to validate or discard the candidates determined in the first step. The last step can be iterated and serves as a refinement step in order to avoid over-segmentation.

IV. Distbic Using Generalized Gamma Distribution

Our goal is to detect the boundaries of phoneme segments without any previous knowledge of the audio stream while achieving a robust performance under noisy environments. For this purpose we introduce an improved version of the DISTBIC algorithm, DISTBIC-Γ, where the signal is modelled using the generalized ΓD (GΓD) instead of GD. Considering the experimental findings mentioned in Sect. 2 we modify the presegmentation and refinement steps of the DISTBIC algorithm by assuming a GΓD distribution model for our signal in the analysis windows.

The proposed algorithm, DISTBIC-Γ, works in two steps. First, using a sufficiently big sliding window and modelling it and its adjacent sub-segments using GΓDs instead of GDs, we calculate the distance d_R associated with the GLRT using (12). Once the parameters for each of the K variables have been estimated with the online gradient ascent algorithm, the likelihood ratio can be calculated. Here, as in [7], we are making the assumption that both noise and speech signals have uncorrelated components $\{x\}_{i=1}^{K}$ in the DCT domain. Depending on the window size, this assumption gives a reasonable approximation for their multivariate probability distribution functions (PDF) using the marginal PDFs. Using (2) and (3), the average log-likelihood function for each univarate marginal distribution in the given hypothesis test is:

$$L(x;\eta,\beta,\gamma) = \log \frac{\gamma\beta^{\eta}}{2\Gamma(\eta)} + (\eta - \frac{1}{\gamma})S_2 - \beta S_1 \tag{13}$$

Since the multivariate equivalents are simple products over the K components the dissimilarity distance in (12) can be easily calculated. A potential problem arises when using MLE for short segments, but we can relax the convergence conditions of the gradient ascend method and still yield improved results [8]. Then, we create a plot of the distances as output with respect to time and filter out insignificant peaks using the same heuristic criteria as [10]. 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 [5] and due to the length limitation of the gradient ascend method for GΓD parameter estimation, we can use Gaussians in this step.

V. Experiments

The performance of the proposed method is evaluated using two sets of experiments on two different corpora. In the first experiment, we compare the efficiency of the proposed method using samples from the M2VTS audio-visual database [12]. 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. 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 [13] totalling 100 seconds of speech time. The performance of the detector was evaluated against the pre-existing phoneme labelling. 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 [14] 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 the false alarm rate (*FAR*) and the missed detection rate (*MDR*) defined below. *ACP* stands for the actual change points in the signal as determined by human in our case.

$$FAR = \frac{number\ of\ FA}{number\ of\ ACP + number\ of\ FA} 100\% \tag{14}$$

$$MDR = \frac{number\ of\ MD}{number\ of\ ACP} 100\% \tag{15}$$

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. An important aspect inherited from original DISTBIC is that segmentation results can be refined by an iterative operation. Also, by tuning the system parameters (e.g. frame size) it is possible to search for an optimal *FAR* after the required *MDR* has been met. The results for the VAD error rates are illustrated in Table I, II, III, and IV. Just as in DISTBIC, it is possible to fine-tune the performance and limit the over-segmentation by changing the penalty factor λ.

The detection performance of the system can also be described by precision (PRC) and recall (RCL) rates

$$PRC = \frac{CFC}{DET}100\% \quad RCL = \frac{CFC}{ACP}100\% \tag{16}$$

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

$$F_1 = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL} \tag{17}$$

The Recall-Precision (R-P) results of our tests are illustrated in Table V, VI, VII, and VIII. 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 significance test results are depicted in tables IX and X. We can deduce that there is notable improvement especially at low SNRs. We also notice the improvement in the recognition of unvoiced speech elements. The improved results demonstrate the higher representation power of the GΓD. Likewise, [15] have also yielded improved VAD performance by modeling speech and noise with a two-sided GΓD in the DFT domain. Their LRT-based detector 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.

TABLE I
ERROR RATES OF VAD IN M2VTS (VOICED PHONEMES)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR**	**MDR**	**PRC**	**RCL**
(clean)	-	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 II
ERROR RATES OF VAD IN TIMIT (VOICED PHONEMES)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR**	**MDR**	**PRC**	**RCL**
(clean)	-	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 III
ERROR RATES OF VAD IN M2VTS (VOICED + UNVOICED)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR**	**MDR**	**PRC**	**RCL**
(clean)	-	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 IV
ERROR RATES OF VAD IN TIMIT (VOICED + UNVOICED)

		DISTBIC-Γ		DISTBIC	
Noise	**SNR**	**FAR**	**MDR**	**PRC**	**RCL**
(clean)	-	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

TABLE V
PERFORMANCE OF VAD IN M2VTS (VOICED PHONEMES)

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

TABLE VI
PERFORMANCE OF VAD IN TIMIT (VOICED PHONEMES)

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

TABLE VII
PERFORMANCE OF VAD IN M2VTS (VOICED + UNVOICED)

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

TABLE VIII
PERFORMANCE OF VAD IN TIMIT (VOICED + UNVOICED)

		DISTBIC-Γ			**DISTBIC**		
Noise	**SNR**	**PRC**	**RCL**	**F$_1$**	**PRC**	**RCL**	**F$_1$**
(clean)	-	71.0	81.3	76.3	68.1	77.7	73.1
white	20	69.7	80.5	74.9	64.7	75.1	70.3
white	10	67.3	77.4	72.4	60.5	71.2	66.3
white	5	62.4	72.8	68.3	54.6	65.7	60.9
babble	20	68.7	77.7	73.5	62.7	72.7	68.3
babble	10	65.2	74.8	70.5	57.5	66.5	63.3
babble	5	60.0	71.7	65.6	51.2	61.8	56.7

TABLE IX
TWO-SAMPLE POOLED T-TEST FOR M2VTS

phonemes		**voiced**	**voiced+invoiced**
Actual Change Points		17	37
# recordings		15	15
critical value at 0.05		t_0=1.701	t_0=1.701
Noise	**SNR**	***t* values**	***t* values**
(clean)	-	4.65	4.86
white	20	5.52	6.48
white	10	6.04	5.51
white	5	9.11	9.44
babble	20	5.52	5.75
babble	10	6.52	6.24
babble	5	8.98	10.1

TABLE X
TWO-SAMPLE POOLED T-TEST FOR TIMIT

phonemes		**voiced**	**voiced+invoiced**
Actual Change Points		13	41
# recordings		25	25
critical value at 0.05		t_0=1.677	t_0=1.677
Noise	**SNR**	***t* values**	***t* values**
(clean)	-	6.57	6.37
white	20	5.68	5.54
white	10	7.31	7.51
white	5	9.84	9.41
babble	20	6.25	6.26
babble	10	9.46	9.06
babble	5	12.3	12.1

VI. CONCLUSIONS

The identification of phoneme boundaries in continuous speech is an important problem in areas of speech synthesis and recognition. In this paper, we have demonstrated that by representing the signal samples with a GΓD we are able to obtain improved results compared to the normal distribution for offline 2-step VAD. We concluded that the GΓD model is more adequate to characterise noisy speech than the Gaussian model. Despite making assumptions on the correlation of distribution components for the computation of the likelihood ratio in GΓD, we generally improved on the VAD performance.

ACKNOWLEDGMENT

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

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, Apr. 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, Jan. 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, Mar. 2001.

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

[6] 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.

[7] 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.

[8] 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, Mar. 2005.

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

[10] 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, Sep. 2000.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.