Phonemic Segmentation Using the Generalised Gamma Distribution and Small Sample Bayesian Information Criterion

George Almpanidis and Constantine Kotropoulos*

Aristotle University of Thessaloniki, Department of Informatics,
Box 451, Thessaloniki, 54124, GREECE
Tel: +30 2310 996361, Fax: +30 2310 998453
email: {galba, costas}@aiia.csd.auth.gr

ABSTRACT

In this work, we present a text-independent automatic phone segmentation algorithm based on the Bayesian Information Criterion. Speech segmentation at a phone level imposes high resolution requirements in the short-time analysis of the audio signal; otherwise the limited information available in such a small scale would be too restrictive for an efficient characterisation of the signal. In order to alleviate this problem and detect the phone boundaries accurately, we employ an information criterion corrected for small samples while modelling speech samples with the generalised Gamma distribution, which offers a more efficient parametric characterisation of speech in the frequency domain than the Gaussian distribution. Using a computationally inexpensive maximum likelihood approach for parameter estimation, we evaluate the efficiency of the proposed algorithm in M2VTS and NTIMIT datasets and we demonstrate that the proposed adjustments yield significant performance improvement in noisy environments.

Keywords: phonemic segmentation, Bayesian information criterion, generalised Gamma distribution, small sample.

1. INTRODUCTION

The identification of the starting and ending boundaries of voice segments in continuous speech is an important problem in many areas of speech processing. For example, it can benefit segment-based speech recognition systems, which have the ability to integrate the dynamics of speech better than frame-based ones (Glass, 2003). Phone segmentation can also assist the creation of speech database used for concatenative speech synthesis (Adell and Bonafonte, 2004). Conventional speech detection and segmentation systems that follow energy-based approaches work relatively well only in high signal to noise ratios (SNR) and for known stationary noise. But for low SNRs, the performance and robustness of energy-based voice activity detectors is not optimal. Since they rely on simple energy

* Corresponding author.

thresholds, they may misclassify non-stationary noise as speech activity. Furthermore, they are not able to identify unvoiced speech segments like fricatives satisfactorily, as the latter can be masked by noise. Consequently, they are inefficient in real-world recordings where speakers tend to leave artefacts such as breathing/sighing, mouth clicks, teeth chatters, and echoes (Pickett and Morris, 1999). Recently, much research discussion has been done regarding the exploration of the speech and noise signal statistics for speech segmentation and voice activity detection. Statistical methods typically employ a decision rule derived from the likelihood ratio test (LRT) applied to a set of hypotheses (Sohn et al., 1999). These approaches can be further improved by incorporating soft decision rules (Chang et al., 2003) and higher order statistics (Nemer et al., 2001). Since these 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. However, real-time applications are not the target of our work.

In this paper, we propose an automatic acoustic change detection algorithm that identifies phone boundaries in speech, using an information criterion for statistical inference. Avoiding the need for linguistic constraints and training data, the algorithm is suitable for speaker segmentation, speech enhancement in telecommunications, speech transcription in computer-aided systems, and for multilingual speech recognition and synthesis applications. For this purpose, we also consider the robustness of the proposed algorithm to various or unknown noise conditions. The novelty of this work is that it suggests alternatives to Gaussian distribution (GD) for modelling the speech and noise signal and considers the limited availability of information in short-frame speech processing, employing small-sample approximations to model selection. Specifically, we exploit the representation power of generalised gamma distribution (GΓD) in order to model the noisy speech signal efficiently. GΓD has been rarely discussed in speech processing literature until recently. Shin and Chang (2005) have considered using GΓD in voice activity detection with encouraging results. Their method of parameter estimation has been used in our paper and it is discussed in Sect. 4 and in Appendix A. In (Huy Dat et al., 2006), GΓD is applied for MAP speech spectral magnitude estimation, yielding superior results under non-stationary noise conditions. Kokkinakis and Nandi (2006) have also demonstrated GΓD-based score yields a consistent increase in convergence speed and separation performance for the task of blind source separation of speech signals.

The remainder of this paper is organised as follows. Sect. 2 shortly describes the problem of phonemic segmentation and discusses the existing challenges. In the next section, we examine the *Bayesian information criterion* (BIC) as a parametric method for identifying segments in speech signal, we present the baseline algorithm for phone boundary detection and review alternative model selection methods. In Sect. 4 we assert that speech modelling in very small frame sizes impels to consider alternative distributions instead of Gaussians. In Sect. 5, we propose an algorithm for text-independent phone boundary detection by making necessary adjustments to the baseline algorithm and in the next section we present experimental results for evaluating phone boundary detection performance in two speech corpora. We demonstrate that the proposed method yields significant improvements in noisy environments. Finally, Sect. 7 concludes the paper and points out future research topics.

2. PHONEMIC SEGMENTATION

Speech may be roughly considered as the result of sequential linking of phonemes. A *phoneme* is defined in linguistics as the minimal information bearing distinct unit. Its acoustic realization, the *phone,* represents a distinct target configuration of the speech tract (regarding articulation as well as form of excitation). In the acoustic realization, however, phone boundaries are mostly ambiguous. In the short-time representation, the speech signal is typically considered stationary and the voiced segments quasi-periodic. Consequently, in statistical phone segmentation, it is assumed that the properties of the speech signal change instantly in the transition from one phone to the next. This is not always true, because, in reality, most signal structures change smoothly, not abruptly due to the mutual influence of adjacent phones as a result of articulatory conditions and restrictions (coarticulation). Therefore, phone boundaries can only serve as pointers to approximately the time instants where one sound ends and another begins.

Speech segmentation algorithms can be typically categorised into decoder-guided, model-based, and metric-based approaches (Chen and Gopalakrishnam, 1998; Chen et al., 2002). In the decoder-guided method, the speech changes are determined according to information provided by a speech recognition system, which decodes the spoken audio stream at first (Woodland et al., 1997). In model-based approaches, it is assumed that a trained model of each component in the signal exists before the speech change detection algorithm starts. The acoustic changes are then determined as the instants when it is necessary to change the model in order for it to match the signal (Bakis et al., 1997). In the metric-based approach, acoustic changes occur when a dissimilarity distance measure between two adjacent windows, shifted along the speech signal, reaches a local maximum. A large number of metric-based approaches are combined with BIC, which is discussed in next section.

Many works in the phonemic segmentation employing statistical methods exist in the literature. Pellom and Hansen (1998) investigate various segmentation, speech enhancement, and parameter compensation techniques in noisy channel corrupted environments. Using a linguistically constrained hidden Markov model (HMM) based method, they yield over 85% boundary detection rate in noise-free environments at 20msec boundary misalignment (tolerance). Aversano et al. (2001) introduce a novel approach for text-independent speech segmentation where the preprocessing is based on critical-band perceptual analysis. It results in 74% segmentation accuracy while limiting over-segmentation to a minimum. Glass (2003) examines a maximum a posteriori (MAP) decoding strategy for segment-based speech recognition where landmarks are modelled in addition to phonemic acoustic units. In (Zhao et al., 2005; Wang et al., 2006), a two-step HMM-based approach is proposed, where a well-trained context dependent boundary model for segment boundary refinement is adapted using a MAP approach. The segmentation accuracy within a 20ms tolerance exceeds 90%. Schwarz P. et al. (2006) also deal with phoneme recognition using a hierarchical structure of multilayer perceptrons, where a block of spectral vectors is split into several blocks processed separately (a hierarchical structure of multilayer perceptrons with separate classification of input patterns in frequency bands, and split temporal context system). An overview of machine learning techniques exploited for phone segmentation is done by Adell and Bonafonte (2004). Evaluating HMMs, artificial neural networks, dynamic time warping, Gaussian mixture models, and pronunciation modelling, it is concluded that

they yield 85-90% detection accuracy when training data is available and a 20ms tolerance is assumed. Toledano and Gomez (2003) also use a modified HMM recogniser and propose a statistical correction procedure to compensate for the systematic errors produced by context-dependent HMMs. The algorithm is evaluated using the percentage of boundaries with errors smaller than 20ms as a figure of merit and attest that over 90% accuracy is possible. An evaluation of phoneme segmentation for unit selection synthesis showed that dynamic time warping is prone to gross labelling errors while HMM modelling exhibits a systematic bias of 15ms (Kominek et al., 2003).

Compared to generative methods based on HMMs, phonemic segmentation methods based on spectral distortion measures are independent of linguistic constraints and computationally inexpensive. Many algorithms of this class that have been proposed in the literature define a *change function* as a function that directly measures the spectral variation of the acoustic signal and utilise this function as a transition penalty. Mitchell et al. (1995) have proposed the Delta Cepstral Function (DCF), which estimates spectral change by summing the normalised time derivative of each cepstral dimension, $DCF_q(t) = C_q(t+1) - C_q(t-1),\ \ q = 1,...,Q$ where $C_q(t)$ is the q^{th} cepstral coefficient for the frame t and Q is the number of cepstral coefficients. $DCF_q(t)$ is then used to compute a cost function $c(t)$ that is able to detect spectral changes associated with phoneme transitions.

$$c(t)_{DCF} = \frac{\sum_{q=1}^{Q} \frac{DCF_q(t)}{\max_q \left| DCF_q(t) \right|}}{\max_q \sum_{q=1}^{Q} \frac{DCF_q(t)}{\max_q \left| DCF_q(t) \right|}}. \tag{1}$$

Brugnara et al. (1992) and Mitchell et al. (1995) have used the Spectral Variation Function (SVF), which estimates spectral change as the angle between two normalised cepstral vectors that are separated in time by a fixed number of frames.

$$SVF(t) = \frac{\widehat{C}(t-1) \bullet \widehat{C}(t+1)}{\left\| \widehat{C}(t-1) \right\| \cdot \left\| \widehat{C}(t+1) \right\|}. \tag{2}$$

In (2), $\widehat{C}(t)$ is the difference between the t^{th} cepstral vector and the time average of the cepstral vectors that lie within a window centred at t and the "•" indicates the scalar dot product operation. The change function is calculated as below

$$c(t)_{SVF} = \frac{1}{2}\left(1 - \frac{SVF(t)}{\max_q \left| SVF(t) \right|} \right). \tag{3}$$

Since both these change functions are derived directly from the observation vectors, little signal processing overhead is required.

3. BAYESIAN INFORMATION CRITERION

3.1 Hypothesis testing and model selection

Common statistical methodology in speech processing areas, such as voice activity detection and speech segmentation, embraces binary decision-making strategies. When statistical parameters are estimated from random samples these parameters can also be considered as random variables. This additional level of uncertainty can be represented by a posterior distribution over parameter values (Bayesian perspective) or by the sampling distribution of the unknown true parameters (frequentist perspective) (Van Allen, 2000). While traditional significance tests are useful for interpreting data that arise from controlled experimental designs, in non-experimental settings, the utilisation of significance testing for pairwise model comparison and interpretation of effects within specific models has been criticised as ill-advised (Dayton, 2003; Raftery, 1999). In contrast, increasingly popular techniques such as Akaike information criterion (AIC) (Akaike, 1974) and BIC (Schwarz G., 1978) have been successfully employed for model selection and hypothesis testing procedures.

In general, information criteria can be classified into two distinct classes (Anderson and Burnham, 1999). The first type of criteria are estimates of Kullback-Leibler (KL) information/distance and attempt to select a good approximating model for inference, based on the principle of parsimony by penalising model complexity. The prominent member of this class is AIC which was introduced in 1973 and allowed major practical and theoretical advances in model selection and the analysis of complex data sets (Bozdogan, 1987). AIC is defined as

$$AIC(M) = -2\ln l(\boldsymbol{x}, M) + k \tag{4}$$

where $\boldsymbol{x}$ are the sample data, $l(\boldsymbol{x}, M)$ is the maximised likelihood function under a model M and k is the number of estimated parameters of the model.

Because AIC performs poorly if there are too many parameters in relation to the size of the sample, many variations of AIC were introduced later. AICC is a second order, small sample approximation to AIC derived under Gaussian assumptions (Sugiura, 1978), consistent AIC (CAIC) compensates for the overestimating nature of AIC using a correction factor based on the sample size (Bozdogan, 1987), QAIC adjusts for over-dispersion or lack of fit, while QAICC introduces a variance inflation factor and is based on quasi-likelihood theory (Anderson and White, 1994). Takeuchi's information criterion (TIC), as a generalised AIC, extends AIC to be applicable to unfaithful models using the maximum likelihood estimation (MLE) as a learning method (Takeuchi, 1976). TIC is defined as

$$TIC(M) = -2\ln l(\boldsymbol{x}, M) - tr(\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}) \tag{5}$$

where $\boldsymbol{\theta}_0$ is the optimal set of model parameters for a given family which minimises the KL divergence to the true model and $\boldsymbol{I}(\boldsymbol{\theta}_0)$ denotes the information matrix (A.3)

and $\boldsymbol{\Sigma} = E(\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}_0)^{\mathrm{T}}$. By approximating the trace of $\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}$ with the number of free parameters, *k* of the model *M*, we derive AIC, which is a good approximation when $\boldsymbol{\theta}_0$ is substituted by the MLE and the data are suitable replaced by their expectations taken with respect to the true model.

AIC-derived approaches can be regarded as *holistic*, in the sense that a variety of competing models can be assessed simultaneously and the best model is selected by applying a single rule. The underlying goal is to select the best approximating model, which is best supported from data among the models under consideration without being imposed to assume that one of these models is "true". In fact, it is not necessary to assume that such a "true" model even exists. An apparent advantage of holistic approaches is that the considerations related to underlying distributions for random variables can be incorporated into the decision-making process rather than being regarded as an assumption whose robustness must also be considered (Dayton, 2003).

The second type of criteria includes BIC, minimum description length principle (MDL) (Rissanen, 1978), Hannan and Quinn criterion (Hannan and Quinn, 1978), which are regarded as *dimension consistent* in that they attempt to consistently estimate the dimension of the true model (Anderson and Burnham, 1999). The goal of model selection here is to find the true model provided that such a model exists and it is in the set of candidate models. Assuming that implicitly, a dimension consistent criterion such as BIC will select the true model asymptotically with probability 1. As a consequence, this consistency imposes large sample size requirements in order to achieve efficient statistical inference.

In statistics, the use of *Bayes Factors* (BF) is a Bayesian alternative to classical hypothesis testing. Given a model selection problem in which we have to choose between two models M_0 and M_1, on the basis of a data vector $\boldsymbol{x}$, the BF is given by the ratio of integrated likelihoods for the two competing models:

$$BF = \frac{P(\boldsymbol{x}|M_0)}{P(\boldsymbol{x}|M_1)}. \tag{6}$$

We should notice that (6) is similar to a likelihood-ratio test, but instead of maximising the likelihood, we average it over the parameters. Generally, if M_0 and M_1 are parameterised by parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ the BF is then given by

$$BF = \frac{P(\boldsymbol{x}|M_0)}{P(\boldsymbol{x}|M_1)} = \frac{\int_{\boldsymbol{\theta}_0} p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, M_0)\, p(\boldsymbol{\theta}_0 \mid M_0)\, d\boldsymbol{\theta}_0}{\int_{\boldsymbol{\theta}_1} p(\boldsymbol{x} \mid \boldsymbol{\theta}_1, M_1)\, p(\boldsymbol{\theta}_1 \mid M_1)\, d\boldsymbol{\theta}_1}. \tag{7}$$

The Bayesian approach to model comparison and selection is based on posterior model probabilities. In a classical binary-hypothesis test, the value of BF measures the strength of evidence meaning that it is more appropriate in the context of inference rather than decision-making under uncertainty. When

comparing models M_0 and M_1, we can choose the model with the higher posterior probability, using Bayes' theorem, by calculating their marginal likelihood ratio:

$$\frac{P(M_0|\boldsymbol{x})}{P(M_1|\boldsymbol{x})} = \frac{P(\boldsymbol{x}|M_0)P(M_0)}{P(\boldsymbol{x}|M_1)P(M_1)} = BF \times prior\ odds\ . \tag{8}$$

BIC, also known as Schwarz's criterion, is an asymptotically optimal method for estimating the best model using only sample estimates (Schwarz G., 1978). It can be viewed as a penalised maximum likelihood (ML) technique, because it imposes a penalty for including too many terms in a regression model in order to battle overfitting. BIC is defined as

$$BIC(M) = -2\ln l(\boldsymbol{x}, M) + k\ln(n) \tag{9}$$

where $\boldsymbol{x}$ are the sample data, $l(\boldsymbol{x}, M)$ is the maximised likelihood function under a model M, k is the number of estimated parameters, and n is the sample size (i.e. the number of observations; that is, the cardinality of $\boldsymbol{x}$). For sufficiently large n, the best model for the data is the one that maximises the BIC. So, considering a binary hypothesis test we could indicate that model M_1 best fits the data, if its BIC value is greater than that of the reference model M_0 that is assumed to stand by default:

$$\Delta BIC = BIC(M_1) - BIC(M_0)\ . \tag{10}$$

BIC can be derived as an approximation of BF, because $2\ln BF \approx BIC(M_1) - BIC(M_0) = \Delta BIC$. It provides a close approximation to BF when the prior over the parameters is the *unit information prior*. This is a multivariate normal prior with mean at the MLE and variance equal to the expected information matrix for one observation. This can be regarded as a prior distribution that contains the same amount of information as a single observation (Kass and Wasserman, 1995; Raftery, 1999)

$$\boldsymbol{\theta}_k \sim N\left(\boldsymbol{\theta}_{k0}, \boldsymbol{I}_{\theta\theta}^{-1}(\boldsymbol{\theta}_{k0})\right) \tag{11}$$

where $\boldsymbol{I}_{\theta\theta}^{-1}$ is the inverse of the block of the Fisher information matrix corresponding to $\boldsymbol{\theta}_k$.

Historically, there has been much critique in the literature over the BIC methodology and its efficiency. According to Weakliem, BIC depends on implicit prior beliefs (the unit information prior) on the model parameters, something that might disagree with the observer's prior information or belief (Weakliem, 1999). It is also argued that statistical consistent criteria like BIC are inclined to favour excessively simple models in practice (where the true model is not on the hypothesis candidate set) and thus it has little benefit over frequentist approaches or other criteria such as AIC (Anderson and Burnham, 1999). Raftery on the other hand, states that the unit information prior in BF and BIC provides reasonable representation when there is little prior information (Raftery, 1999). The hypothesis testing procedure defined by choosing the model with the higher posterior probability minimises the total error rate (the sum of Type I and Type II error rates). Even so, Bayesian model

averaging, i.e. averaging the posterior densities of the quantity of interest across the different models with weights proportional to their posterior probabilities (derived from BFs) yield optimal predictive performance in situations that there is model uncertainty. Kuha (2004) has asserted that despite their different foundations, similarities between AIC and BIC can be observed, for example, in analogous interpretations of their penalty terms and argued that useful information for model selection can be obtained from using the two statistics together. Also, Burnham and Anderson (2004) have shown that BIC can be derived as a non-Bayesian result and AIC from a Bayesian standpoint to BF, making their distinction philological.

3.2 Phonemic segmentation using the Bayesian Information Criterion

In a typical statistical speech segmentation system, feature vectors are produced by moving a window/frame over the audio recording in a series of fixed length steps. Then, statistical decisions are made on a frame basis by calculating likelihood ratios. The choice of the frame length is a compromise between having sufficient data for calculating the feature vectors and limiting the influence of surrounding parts of the recording. The frame shift determines the time resolution for the boundaries. For an accurate segmentation, the frame shift should be as small as possible. Such an acoustic change detection system based on BIC has been proposed by Chen and Gopalakrishnam (1998). In this approach, 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. The decision problem is undertaken by using BIC as a model selection criterion. A sliding window moves over the signal, making statistical decisions at its middle. When the window is sufficiently small, its adjacent sub-windows can be assumed as homogenous segments. The offset of the sliding window indicates the resolution of the system. For the purpose of phonemic segmentation, we would need to evaluate the following statistical hypotheses:

- H_0: $(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n) \sim N(\boldsymbol{\mu}_Z,\boldsymbol{\Sigma}_Z)$: the data sequence comes from one source $\boldsymbol{Z}$ (i.e., noisy speech/silence, the same phone) and is described by model M_0
- H_1: $(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m) \sim N(\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X)$ and $(\boldsymbol{x}_{m+1},\boldsymbol{x}_{m+2},\ldots,\boldsymbol{x}_n) \sim N(\boldsymbol{\mu}_Y,\boldsymbol{\Sigma}_Y)$: the data sequence comes from two sources $\boldsymbol{X}$ and $\boldsymbol{Y}$, meaning that there is a transition from speech utterance to silence, or a transition between two different phones or vice versa. We denote that the data come from model M_1

$\boldsymbol{x}_i$ are Q-dimensional feature vectors in a transformed domain, typically using the Mel Frequency Cepstral Coefficients (MFCC) representation. Let $\boldsymbol{\Sigma}_Z$ and $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_Y$, respectively be the covariance matrices of the complete sequence $\boldsymbol{Z}$ and the two subsets $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{\mu}_Z$, $\boldsymbol{\mu}_X$, $\boldsymbol{\mu}_Z$ the corresponding mean vectors, while m and n-m are the number of feature vectors for each subset. When the hypothesis is tested at the time instant t, corresponding to the middle of the sliding window, then it is clearly $m=n/2$.

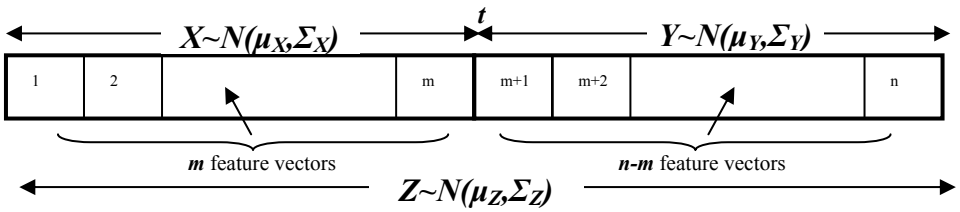

Fig. 1. Models for two adjacent speech segments.

Since $\boldsymbol{x}_i \in R^Q$, Gaussian model M_0 has $k_0 = Q + \frac{Q(Q+1)}{2}$ parameters, while mixture model M_1 has $k_1 = 2k_0$ parameters. The variation of the BIC value between the two models is given by Tritschler and Gopinath (1999)

$$\Delta BIC = 2\ln R - \lambda P_0 \tag{12}$$

$$R = \frac{l(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)l(\boldsymbol{y};\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)}{l(\boldsymbol{z};\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)} \tag{13}$$

$$P_0 = \frac{1}{2}\left[Q + \frac{Q(Q+1)}{2}\right]\ln(n) \tag{14}$$

where R is the Generalised Likelihood Ratio Test (GLRT), P_0 is the penalty for model complexity regarding the two models and λ is a tuning parameter for the penalty factor. The complexity penalty P_0 depends on the degrees of freedom associated with the test. In (13), $l(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)$ represents the likelihood of the sequence of feature vectors $\boldsymbol{X}$ given the multivariate Gaussian process $N(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)$. $l(\boldsymbol{y};\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)$ and $l(\boldsymbol{z};\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)$ are similarly defined. Assuming multivariate GD modelling, we can easily calculate ΔBIC from sample values:

$$\Delta BIC = n\ln\left|\boldsymbol{\Sigma}_z\right| - m\ln\left|\boldsymbol{\Sigma}_x\right| - (n-m)\ln\left|\boldsymbol{\Sigma}_Y\right| - \lambda P_0 \tag{15}$$

Positive ΔBIC values indicate that the multivariate Gaussian mixture best fits the data, meaning that t corresponds to a segment boundary.

Regarding the application of BIC to the detection of phone boundaries, certain assumptions must be considered. First, the incoming signal must either originate from a single speaker or, equivalently, no multiple speakers must talk simultaneously. Second, since the phone durations are relatively small, we need to operate on very small frame sizes in order to correctly assume that the speech signal is stationary. The BIC scheme presented here, while inherently threshold-free can also be viewed as a dynamic thresholding scheme on the log-likelihood distance. There is still a penalty factor λ that depends on the type of analysed data and must be estimated heuristically (Chen et al., 2002; Tritschler and Gopinath, 1999). This allows for reducing Type II errors without increasing the number of Type I errors, but it is experimentally tuned. Also, BIC tends to choose oversimplistic models due to the heavy penalty on the complexity. Nevertheless, BIC is a consistent estimate and various speech processing algorithms have extended the basic method combining it with other metrics such as Kullback-Leibler

(KL) distance, sphericity tests, and higher-order statistics (Delacourt and Wellekens, 2000; Nemer et al., 2001).

Tritschler and Gopinath (1999) has proposed an improved segmentation algorithm for speaker detection using a variable window scheme instead of considering a fixed amount of data when no segment boundary is found in the current sliding window. The new algorithm uses relatively small window sizes in areas where boundaries are very likely to occur, while increasing the window size more generously when boundaries are not very likely to occur. Incorporating long-term speech information (i.e. more observations) to the decision rule benefits the speech/pause discrimination and phone transition detection.

3.3 DISTBIC

DISTBIC is a two-pass distance-based algorithm that searches for change point candidates at the maxima of distances computed between adjacent windows over the entire signal (Delacourt and Wellekens, 2000). First, it uses a distance computation to choose the possible candidates for a change point. Different criteria such as KL distance 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 the distance D

$$D = \ln R = \ln \frac{l(z;\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)}{l(x;\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)l(y;\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)} = L(z;\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z) - L(x;\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x) - L(y;\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y) \tag{16}$$

where $L(\cdot) = \ln l(\cdot)$ denote the log-likelihood. Next, a plot of distances is created and significant local peaks are selected as candidate change points. Thresholds that filter out the insignificantly small distance values in the plot (horizontal threshold) or thresholds that constrain the minimum horizontal/time distance between consecutive local maxima corresponding to candidate segmentation points (vertical threshold) can be applied. In the last step, using the variable window scheme of Tritschler and Gopinath (1999), ΔBIC values are used in order to validate or discard the candidates determined in the first step. A free parameter constrains the maximum window length. The last step of DISTBIC can be iterated and serves as a refinement step in order to avoid over-segmentation.

DISTBIC has been found efficient in detecting acoustic changes that are relatively close one another but at the price that a lot of false changes are detected too. Nevertheless, by tuning the parameters of the algorithm it is possible to fix the over-segmentation (false alarms) on a minimum value and then try to maximise the detection rate.

3.4 Modification of the Bayesian Information Criterion for small samples

A central problem in statistical inference is dealing with situations where little information from data is given. This is clearly the case with phone segmentation where the duration of a single phone can be as small as a few milliseconds (Villiers and Preez, 2001). Most researchers agree that the number of estimated parameters must be substantially fewer than the number of observations in order to obtain a valid optimum (Olsthoorn, 1995). As stated in Sect. 3.2, an estimator is statistically consistent if it

converges on the true value given enough data and it is considered efficient if the rate of convergence is as fast as possible. BIC asymptotically favours the correct model with a probability that tends to 1 as sample size increases. But the difference between BIC and twice the logarithm of BF has a nonvanishing asymptotic error of constant order *O(1)* (Volinsky and Raftery, 2000). While the error vanishes for an implicit choice of prior on the parameters (the overall unit information prior), for general priors of the parameters, the error tends to zero as a proportion of BIC, since the absolute value of BIC increases with *n*. Therefore, BIC has the undesirable property that for any constant *C*, *BIC*+*C* also approximates twice the log BF to the same order of approximation as BIC itself. This implies that the BIC approximation is rudimentary, and the performance might suffer for small samples.

Furthermore, while the BIC target model does not depend on sample size *n*, the number of parameters that can be estimated reliably from finite data does depend on *n*. For small *n*, the BIC-selected model can be quite biased as an estimator of its target model. Due to this limitation, the application of the BIC measure to target domains containing an insufficiently small number of samples requires caution (Lee, 2002). There has been much discussion in the literature about the challenges in determining *n* in the BIC equation (9), especially for dependent data (Visser et al., 2005). It is stated that for multivariate observations, *n* in (9) should be replaced by the number of observed statistics per parameter. In effect, *n* may depend on the values of specific parameters, meaning that different parameters may have different values of *effective sample sizes* associated with them. According to Raftery (1999) *n* should be the rate at which the Hessian matrix of the log-likelihood function grows.

Using the Laplace approximation for calculating the integrals in (10) and substituting $\boldsymbol{\theta}$ with the MLE parameter estimates $\hat{\boldsymbol{\theta}}$ there, it is possible to derive approximations to the BF by suggesting alternative model complexity penalties (Raftery and Richardson, 1996). Subsequently, it is possible to derive modified BIC criteria that perform better than ordinary BIC for model selection and for different sample sizes. From Appendix A, we have

$$\begin{aligned} -2\ln P(\boldsymbol{x} \mid M_k) = {} & -2\ln P(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k) - 2\ln P(\hat{\boldsymbol{\theta}}_k \mid M_k) - d_k \ln 2\pi + d_k \ln(n) + \ln\left|\overline{I}_E(\hat{\boldsymbol{\theta}}_k)\right| \\ & + O(n^{-1/2}) \end{aligned} \tag{17}$$

where $P(\boldsymbol{x} \mid M_k)$ is the posterior distribution, $P(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k)$ is the prior distribution of $\boldsymbol{X}$ given the model $(M_k, \hat{\boldsymbol{\theta}}_k)$, $\ln P(\hat{\boldsymbol{\theta}}_k \mid M_k) = L(\hat{\boldsymbol{\theta}}_k)$ is the log-likelihood at the MLE $\hat{\boldsymbol{\theta}}_k$, $d_k = \dim(\boldsymbol{\theta}_k)$ is the dimension of the parameter space of the model M_k and $\overline{I}_E$ is the expected information matrix per observation.

Accordingly, Kashyap's modification (KBIC) (Kashyap, 1977), is defined as

$$KBIC = -2L(\hat{\boldsymbol{\theta}}_k) + d_k \ln(n) + \ln|\overline{I_E}(\hat{\boldsymbol{\theta}}_k)| \tag{18}$$

and can be derived by (17) by discarding $2\ln P(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k)$ and $d_k \ln(2\pi)$. Haughton's BF modification (HBF) (Haughton, 1988) is defined as

$$HBF = -2L(\hat{\boldsymbol{\theta}}_k) + d_k \ln \frac{n}{2\pi} \tag{19}$$

where the terms $2\ln P(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k)$ and $\ln\left|\overline{I}_E(\hat{\boldsymbol{\theta}}_k)\right|$ are discarded from (17). Likewise, Bollen's BF approximation ABF1 (Bollen et al., 2005) eliminates $2\ln P(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k)$ from (17):

$$ABF1 = -2L(\hat{\boldsymbol{\theta}}_k) + d_k \ln \frac{n}{2\pi} + \ln|\overline{I}_E(\hat{\boldsymbol{\theta}}_k)| \tag{20}$$

Bollen's ABF2 is derived by considering a specific implicit unit information prior that is more flexible than BIC (Bollen et al., 2005). By choosing special scaled unit information prior $P(\boldsymbol{\theta}_k \mid M_k)_{ABF2}$

$$P(\boldsymbol{\theta}_k \mid M_k)_{ABF2} \sim N\left(\hat{\boldsymbol{\theta}}_k, \left[w\frac{I_0(\hat{\boldsymbol{\theta}}_k)}{n}\right]^{-1}\right) \tag{21}$$

where the optimum value of scale w maximises $P(\boldsymbol{x} \mid M_k)$, and using a Laplace approximation only for the likelihood, we get

$$ABF2 = \begin{cases} -2L(\hat{\boldsymbol{\theta}}_k) + d_k\left(1 + \ln\dfrac{d_k}{\hat{\boldsymbol{\theta}}_k^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k}\right), & \text{if } d_k > \hat{\boldsymbol{\theta}}_k^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k \\ -2L(\hat{\boldsymbol{\theta}}_k) - \hat{\boldsymbol{\theta}}_k^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k & \text{, otherwise} \end{cases} \tag{22}$$

It has been attested that the performance gain using ABF2 instead of BIC in small samples (n<60) is significant (Bollen et al., 2005). Another BF modification is Rissanen's Shortest Data Description criterion (SSD) (Rissanen, 1978) that was originally derived as a solution to the problem of minimum-bit representation of a signal and is defined as

$$SSD = -2L(\hat{\boldsymbol{\theta}}_k) + d_k \ln \frac{n+2}{24} + 2\ln(d_k + 1)\,. \tag{23}$$

Alternative penalty factors for BIC have also been proposed by Visser et al. (2005) (generalised BIC, genBIC), Bogdan et al. (2004), Gheissari and Bab-Hadiashar (2003) (Geometric BIC, GBIC), and Tremblay and Wallach (2004) (BIC corrected, BICC). BICC has been described as *BIC corrected for small samples*. It considers the case where the residual errors of the linear regression in the sample are independent and identically distributed as a Gaussian process. It is defined as:

$$BICC = -2L(\hat{\boldsymbol{\theta}}_k) + \frac{d_k n \ln(n)}{n - d_k - 1} \tag{24}$$

where n is the sample size, and d_k the number of estimated parameters. It must be noticed that for a given number of parameters d_k, the penalization term is constant regardless of which parameters are estimated. Tremblay and Wallach (2004) have attested that BICC performs better than classic BIC both in terms of mean squared error of the parameter estimates and in terms of prediction error.

For the purpose of this paper, we are going to use BICC and ABF2 approximations as model selection criteria and we will evaluate them as viable components for automatic phonemic segmentation systems. When $\boldsymbol{X}$ are Q-variate data it is common to assume that the statistical components are independent. This simplifies the computations greatly, since we have to deal with diagonal covariance matrices. Consequently, the marginal probabilities multiply and the log-likelihoods add, so we can presume univariate statistics. If for example, each independent component q is modelled by a univariate GD $P(M_q;\boldsymbol{\theta}_q) \sim N(\mu_q,\sigma_q)$, then $\boldsymbol{\theta}_q = (\mu_q,\sigma_q)^T$, where μ_q and σ_q are the mean and standard deviation respectively. Clearly, the number of the parameters for each component which need to be estimated for model M_k is $d_q = \dim(\boldsymbol{\theta}_q) = 2$ and the total number of estimated parameters is $d_k = 2Q$. Working independently, we can approximate the multivariate joint probability distribution functions (PDF) using the marginal PDFs.

4. SPEECH MODELLING WITH THE GENERALISED GAMMA DISTRIBUTION

A critical parameter that affects the performance of statistical speech segmentation methods is the choice of distribution for modelling clean speech and noise/silence. A common assumption for most algorithms in speech processing 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 (Gazor and Zhang, 2003; Martin, 2005). More specifically, LD fits well the highly correlated univariate space of speech amplitudes as well as the uncorrelated multivariate space of feature values after a Karhunen-Loeve Transformation (KLT) or Discrete Cosine Transformation (DCT). It is also asserted that DCT coefficients of clean speech samples are better modelled using the generalised Gaussian distribution (GGD) since they have a distribution which is more peaky than a LD (Gazor and Zhang, 2003). Nakamura (2000) proposes a generalised Laplace distribution (GLD) for speech recognition, demonstrating significant improvement in word accuracy when MFCC is used.

Recently, it has been asserted that the generalised Gamma, GΓD, fits the voiced speech signal even better than Gamma, Laplacian, and Gaussian distributions, especially in small frame sizes, and consequently it offers great perspectives in very short-time speech analysis and in phone boundary

detection (Shin and Chang, 2005). Fig.2 illustrates that GΓD has the better fit than GD with the empirical PDF of the speech spectra in MFCC domain.

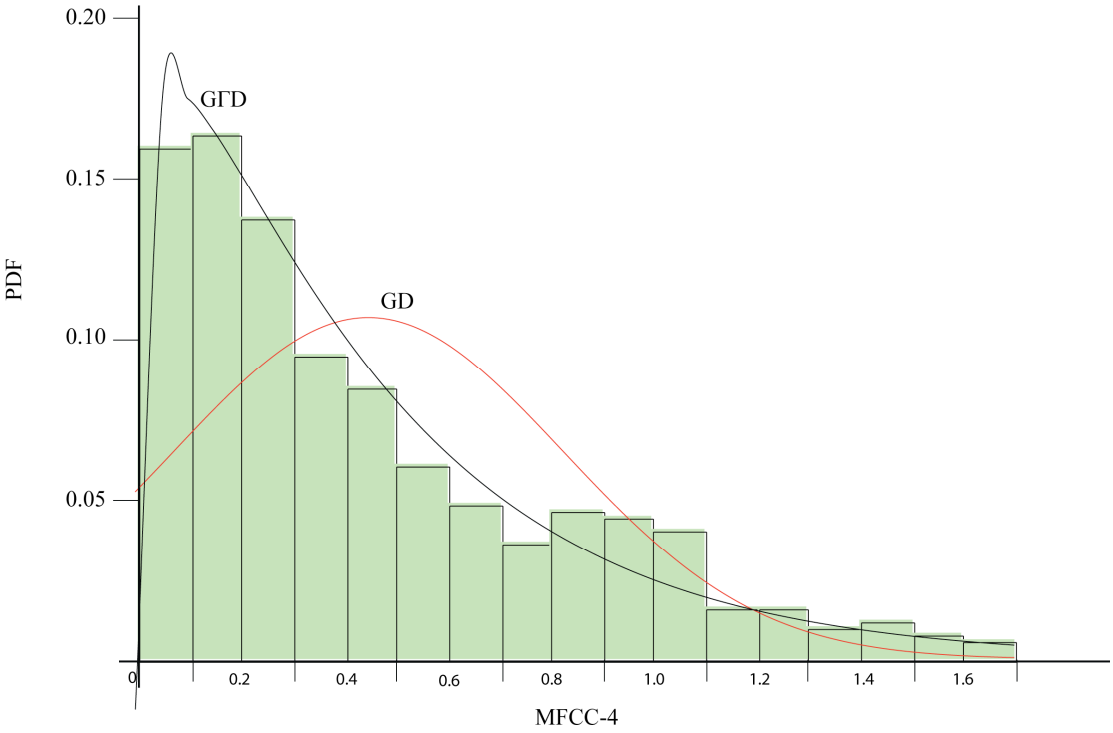

Fig. 2. Empirical histogram and GΓD, GD PDFs for the real part of the 4th MFCC coefficient in speech data (100ms NTIMIT sample).

The four-parameter GΓD was first proposed by Stacy (1962) and it is defined with the PDF

$$f_x(x,q,r,s,t) = \frac{t}{r^{st}\Gamma(s)}(x-q)^{st-1}e^{-\left(\frac{x-q}{r}\right)^t} \tag{25}$$

where $\Gamma(s)$ denotes the gamma function, q, r, s, and t are positive real values corresponding to location (q), scale (r), and shape/power (s), and power (t) parameters of GΓD respectively. Although the GΓD is of considerable importance in theoretical and applied statistics, 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. It is likely that two GΓDs which have quite different parameters can look alike. Furthermore, estimating the parameters of GΓD analytically using the MLE method is difficult, because the maximised likelihood results in nonlinear equations involving numerical integrations. Nevertheless, various numerical methods have been discussed in the literature. These require nonlinear equations to be solved, but under certain approximations, these equations may only involve simple functions such as digamma and trigamma, which are well behaved and available in many software packages where they can be accurately computed using algorithms by Bernando (1976) and Schneider (1978), respectively. A discussion on a MLE for the four-parameter GΓD is given by Harter (1967), while Cohen and Whitten (1986) have proposed a method based on moment estimation.

Parr and Webster (1965), Hager and Bain (1987), and Lawless (1980) have considered MLE for the GΓD but reported problems with the iterative solution of the nonlinear equations implied by this method, unless the sample size is large and the shape parameter is also relatively large. Numerical methods organised around Gibbs sampling have been shown to perform well for Bayesian inference (Tsionas, 2001).

In their proposal for voice activity detection, Shin and Chang (2005) use a simpler, reparameterised, three-parameter, two-sided version of the GΓD, with the location parameter discarded

$$f_x(x) = \frac{cb^a}{2\Gamma(a)} |\, x \,|^{ac-1} \; e^{-b|x|^c} \tag{26}$$

where b, and a, c are positive real values corresponding to scale (b) and shape (a, c) parameters respectively. GD is a special case of (26) for c=2 and a=0.5. For c=1 and a=1, (26) yields the LD, while for c=1 and a=0.5, it represents the common ΓD. GΓD also includes GGD, log-normal, exponential and Weibull, distributions as special cases. This special property of GΓD (the ability to simulate either LD/ΓD or GD) allows the modelling of both clean speech and noise/silence respectively.

A computationally efficient on-line algorithm for the three-parameter GΓD in (26), based on the MLE and the gradient ascent algorithm, has been introduced by Shin and Chang (2005). The shape parameter c is numerically determined by using the gradient ascend algorithm according to the MLE principle. Using a learning factor we can then reestimate the value of c that locally maximizes the log-likelihood function L, until L convergences. Using this value and the data samples, we can determine the scale (b) and shape (a) parameters. The rationale for the parameter estimation algorithm is given in Appendix C.

5. PHONEMIC SEGMENTATION USING GENERALISED GAMMA DISTRIBUTION

Our goal is to detect the boundaries of phone 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 GΓDs instead of GDs. Considering the discussion in Sect. 4, we modify the pre-segmentation 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 with GΓDs instead of GDs, we calculate the distance D associated with the GLRT using (13). It must be noticed here that zero inputs are ignored because the GΓD in (26) cannot be specified when the argument equals zero exactly. Once the parameters for each of the Q components have been estimated with the online gradient ascent algorithm, the likelihood ratio can be calculated. Here, as in (Gazor and Zhang, 2003), we are making the assumption that the noisy speech signal has uncorrelated components in the MFCC domain. Depending on the window size, this assumption gives a reasonable approximation for the multivariate probability densities using the marginal PDFs. Using the equations (C.4) and (C.9) in Appendix C, the

average log-likelihood function for each univariate marginal distribution in the given hypothesis test can be calculated. Since the multivariate equivalents of likelihood are simple products over the Q components, the dissimilarity distance in (16) can be easily calculated. While KL distance offers better discriminative ability than GLR and other similarity measures (Delacourt and Wellekens, 2000), especially in noisy signals, it is unrealistic to use maximum likelihood estimates, such as those derived for GΓD in Appendix C, with the KL-divergence loss function. As the parameters are estimated on relatively few data it becomes increasingly likely that KL-divergence for a maximum-likelihood estimate is infinite (Van Allen, 2000).

Next, we create a plot of the distances as output with respect to time and filter out insignificant peaks using the same heuristic criteria as (Delacourt and Wellekens, 2000). In the second step, using the BIC test as a merging criterion, we compute the ΔBIC values for each segmentation point candidate in order to validate the results of the first step. A block diagram depicting the processing stages of the proposed phonemic segmentation system is given in Fig. 3. 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 satisfactory results (Shin and Chang, 2005). In order to further improve performance we also propose alternative versions of the algorithm, DISTBICC-Γ and DISTABF2-Γ, where the BIC is replaced by BICC and ABF2 respectively, as in Sect. 3.4. The model complexity penalty P_0 for ABF2, assuming GΓD modelling, is given in Appendix B. These four algorithms, DISTBIC, DISTBIC-Γ, DISTBICC-Γ, and DISTABF2-Γ are tested in the setting of phonemic segmentation in the next section.

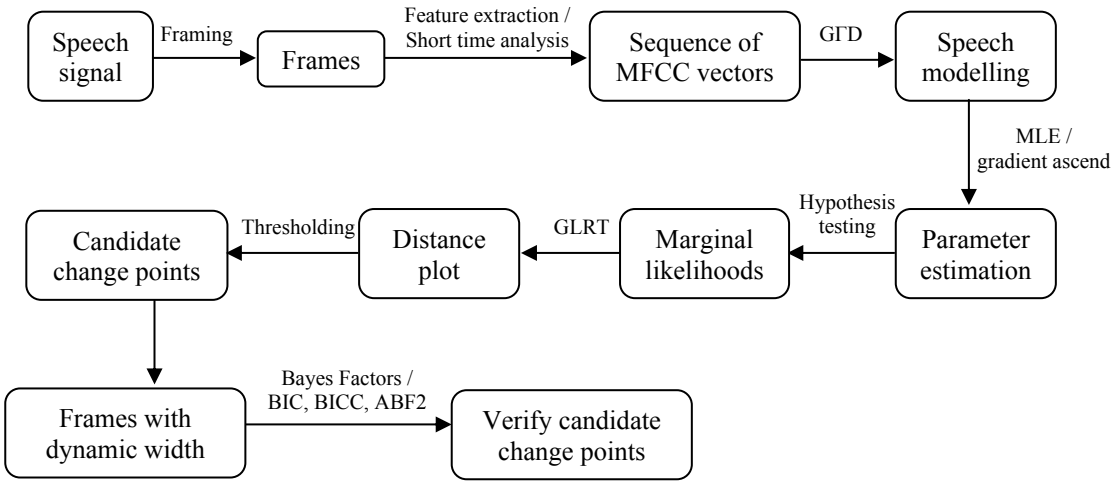

Fig. 3. Block diagram of the proposed phonemic segmentation system.

6. EVALUATION

The performance of the proposed methods is assessed using two sets of experiments on two different datasets. In the first experiment, we compare the efficiency of the proposed methods using samples from the M2VTS audio-visual database (Pigeon and Vandendorpe, 1997). In our tests we used 25 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 provided from the DISTBIC, DISTBIC-Γ, DISTBICC-Γ, and DISTABF2-Γ algorithms. The human error and accuracy of visually and acoustically identifying segmentation points were taken into account. A phone boundary identified by the system is considered "correct" if it is placed within a range of ±20ms from a true (hand-labelled) segmentation point, which implies a 20ms tolerance. White and babble noise from the NOISEX-92 database (Varga et al., 1992) was added to the clean speech samples at various SNR levels ranging from 20 to 5 dB.

In the second set of experiments we used samples from the Test Set of the NTIMIT corpus (Jankowski et al., 1990). NTIMIT is a telephone bandwidth version of the well-known TIMIT dataset (Garofolo et al., 1990), where its utterances are transmitted over telephone channels through either local or long-distance calls. The dataset in our experiments consisted of 2 male and 2 female speakers from each of the 8 dialects (32 speakers in total) with 8 utterances per speaker. Common utterances sa1 and sa2 were not used. This accounts for 256 unique utterances totalling 787 seconds of speech time. The performance of the detector was evaluated against the manual phonetic transcription that is provided in the dataset. In this set of experiments, we made performance comparisons against segmentation algorithms reported in the literature by including DCF, SVF, and TIC in the tests. TIC is being considered as an alternative to BIC-based criteria in the candidate point verification step of the algorithm as described in Sect. 5. We call the resulting algorithm DISTTIC. For both experiments we used the same set of parameter values and features (20ms analysis window, 10ms overlap/shift, first 12 MFCCs excluding the energy component, λ=7 for DISTBIC, λ=1 for the rest, 20ms tolerance).

The errors that can be identified in acoustic change point detection are distinguished by whether the point corresponds to a phone boundary, 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 II error (false alarm, *FA*) while a point totally missed by the detector is a Type I error (missed detection, *MD*). The detection error rate of the system is described by the missed detection rate (*MDR*) and the false alarm rate (*FAR*) as defined in (Esposito and Aversano, 2004). *ACP* stands for the actual change points in the signal as determined by human in our case and A*P* is the overall points in the system (i.e. the total number of the frames).

$$FAR = \frac{FA}{AP - ACP} 100\% \tag{27}$$

$$MDR = \frac{MD}{ACP} 100\% \tag{28}$$

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 equations (27) and (28) imply that a higher detection performance (lower *MDR*) comes at expense of a higher *FAR*. 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. Just as in DISTBIC, it is possible to fine-tune the performance and limit the over-segmentation by changing the penalty factor λ. Fig. 4 depicts type I and type II errors for the NTIMIT sentence sx404.

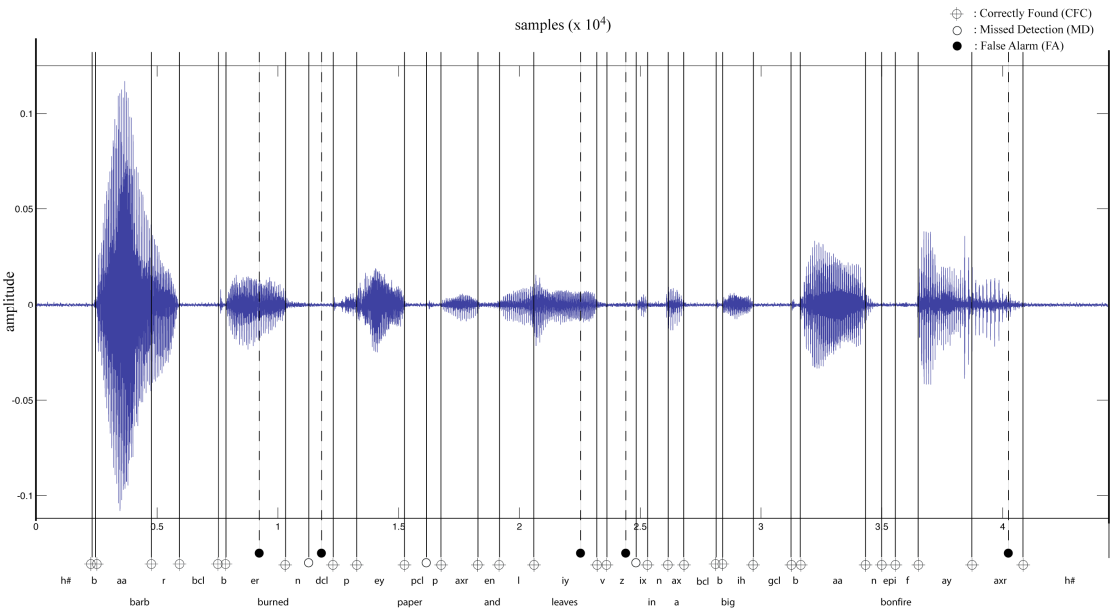

Fig. 4. Phonemic segmentation using the DISTABF2-Γ algorithm for utterance sx404 of NTIMIT female speaker PAS0 ("Barb burned paper and leaves in a big bonfire").

The detection performance of the system can be assessed by precision (*PRC*) and recall (*RCL*) rates:

$$PRC = \frac{CFC}{DET}100\% \qquad RCL = \frac{CFC}{ACP}100\% \tag{29}$$

where *CFC* is the number of correctly found changes and *DET* is the number of changes detected by the system. The overall objective effectiveness of the system can be evaluated by the F_1-measure:

$$F_1 = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL} \tag{30}$$

The results of the system error rates for the M2VTS dataset are demonstrated in Table 1, while the recall-precision results are illustrated in Table 2, where we calculate the average *PRC*, *RCL*, and F_1 rates over all recordings. Examining the average F_1-measure 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=1.677 corresponding to one-tailed t-test with 25+25-2=48 degrees of freedom). The improved results demonstrate the higher representation power of the GΓD. Similarly, using different model selection criteria than BIC, we yield even better results than DISTIBIC-Γ under all SNRs. Fig. 5 depicts the overall performance of the four algorithms under various classes of additive noise at different SNRs. We can easily deduce that DISTABF2-Γ performs best, followed by DISTBICC-Γ. The performance improvement is most notable at low SNRs. The significance test results that are depicted in Table 3 validate the experimental findings.

For the NTIMIT dataset, we compare the segmentation accuracy of the DISTBIC-based algorithms regarding noisy speech against other phonemic segmentation algorithms reported in the

literature using the Receiver Operating Characteristic (ROC) curves in Fig. 6. Different points in the ROC are calculated by varying the free parameters mentioned in Sect. 3.4. The closer a curve is to the bottom left in the diagram, the more accurate is the segmentation. The overall performance of the segmentation algorithms is shown in Table 4. The results show that DISTABF2-Γ yields significantly better detection results than DISTBIC and DISTBIC-Γ at all FAR rates. DCF has comparable segmentation accuracy with DISTBIC-Γ while DISTTIC and SVF perform worse than DISTBIC. Overall, DISTABF2-Γ performs best, followed by DISTBICC-Γ. The performance of DISTABF2-Γ compares favourably with the results of Esposito and Aversano (2004). In their work, they proposed a fitting procedure, which places segmentation boundaries into the barycentre of each group of quasi-simultaneous sharp transitions. Using a relative thresholding procedure and a Melbank front-end, they yielded over 80% recall rate in TIMIT and NTIMIT with a ±20ms tolerance while keeping over-segmentation at a minimum.

Table 1. Error rates in M2VTS.

		DISTBIC		**DISTBIC-Γ**		**DISTBICC-Γ**		**DISTABF2-Γ**	
Noise	**SNR (dB)**	**MDR**	**FAR**	**MDR**	**FAR**	**MDR**	**FAR**	**MDR**	**FAR**
(clean)	-	18.9	2.01	16.4	1.53	15.2	1.47	13.5	1.30
white	20	21.9	2.09	18.1	1.76	16.3	1.57	15.8	1.36
white	10	26.5	2.36	22.6	2.03	19.1	1.64	17.1	1.54
white	5	29.7	2.79	25.1	2.26	21.0	1.93	20.0	1.67
babble	20	25.4	2.29	20.3	1.97	17.7	1.78	16.8	1.64
babble	10	28.4	2.5	22.9	2.15	20.2	1.92	19.8	1.77
babble	5	33.4	2.79	27.0	2.38	23.8	2.01	22.4	1.88

Table 2. Recall, Precision, and F_1 measure in M2VTS.

		DISTBIC			**DISTBIC-Γ**			**DISTBICC-Γ**			**DISTABF2-Γ**		
Noise	**SNR (dB)**	**PRC**	**RCL**	**F1**	**PRC**	**RCL**	**F1**	**PRC**	**RCL**	**F1**	**PRC**	**RCL**	**F1**
(clean)	-	70.4	81.1	75.3	76.3	83.6	79.7	77.2	84.8	80.7	79.6	86.5	82.8
white	20	68.8	78.4	73.2	73.2	81.9	77.3	75.8	83.7	79.5	78.4	84.2	81.2
white	10	64.6	73.5	68.7	69.1	77.4	72.9	74.5	80.9	77.5	76.0	82.9	79.3
white	5	59.7	70.3	64.5	66.1	74.9	70.2	70.6	79.0	74.5	73.8	80.0	76.7
babble	20	65.7	74.6	69.8	70.5	79.7	74.7	73.1	82.3	77.3	74.9	83.2	78.8
babble	10	62.8	71.6	66.8	67.8	77.1	72.1	70.9	79.8	75.0	72.7	80.2	76.2
babble	5	58.3	66.6	62.1	64.3	73.0	68.4	69.0	76.2	72.4	70.7	77.6	73.9

Table 3. One-sided two-sample pooled t-test comparing the F_1 values of DISTBIC-Γ against DISTBIC, DISTBICC-Γ against DISTBIC-Γ, and DISTABF2-Γ against DISTBICC –Γ.

Total Actual Change Points	925			
Number of recordings	25			
Critical value at 0.05	t_0=1.677			
		DISTBIC-Γ vs DISTBIC	DISTBICC-Γ vs DISTBIC-Γ	DISTABF2-Γ vs DISTBICC -Γ
Noise	**SNR (dB)**	***t* values**	***t* values**	***t* values**
(clean)	-	8.76	1.95	4.03
white	20	8.12	4.31	3.23
white	10	8.67	9.02	3.53
white	5	11.95	8.88	4.37
babble	20	9.98	5.21	2.83
babble	10	11.00	5.93	2.35
babble	5	13.37	8.30	3.17

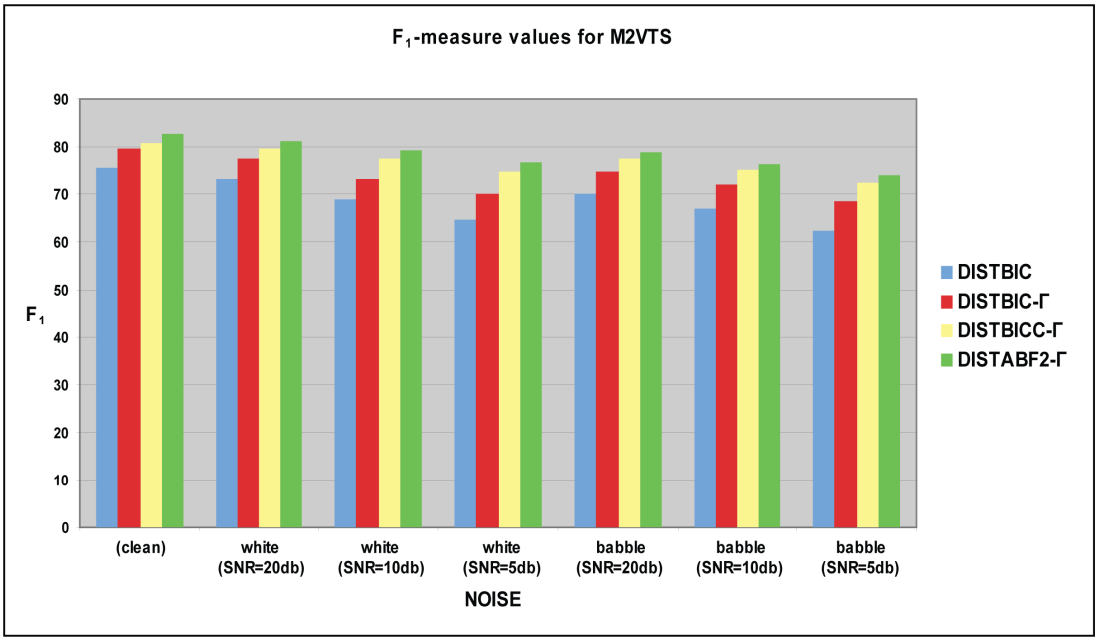

Fig. 5. Overall system evaluation using the F_1 measure for the M2VTS dataset.

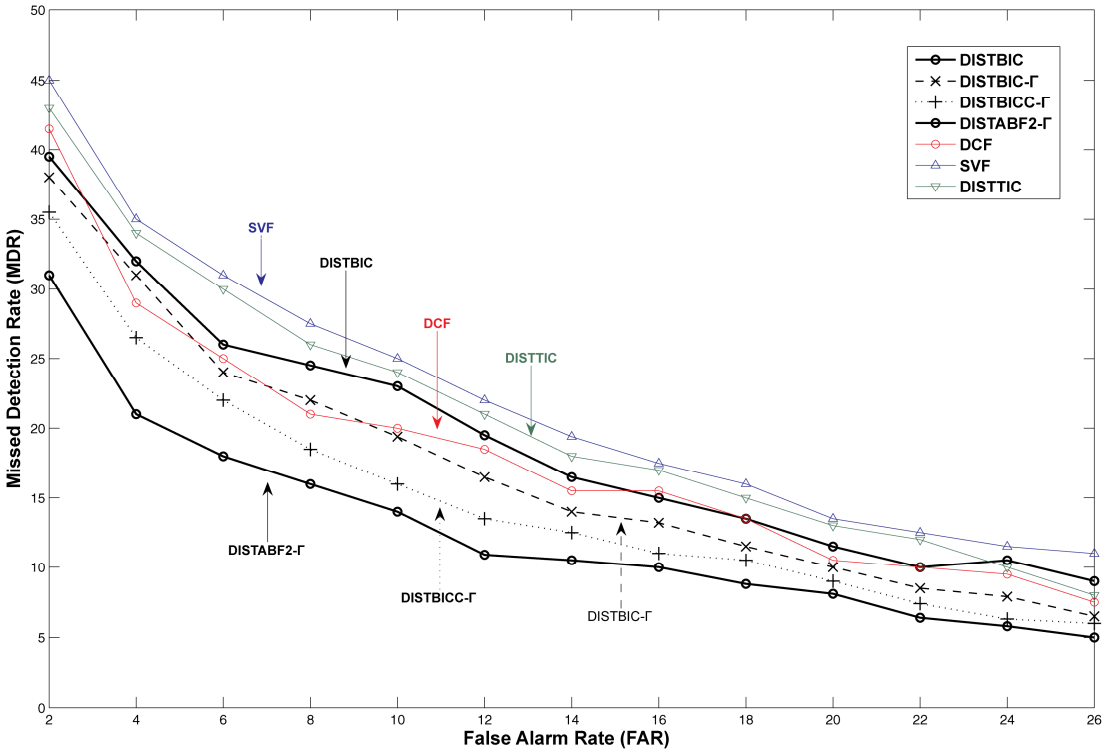

Fig. 6. ROC curves for different segmentation algorithms in NTIMIT.

Table 4. Average Error rates in NTIMIT.

SNR (dB)	MDR	FAR	PRC	RCL	F1
DISTBIC	28.40	5.30	65.86	71.6	68.59
DISTBIC-Γ	26.13	5.02	67.76	73.87	70.65
DISTBICC-Γ	23.74	4.69	69.89	76.26	72.9
DISTABF2-Γ	19.28	4.38	72.52	80.72	76.37
DCF	26.57	4.57	69.62	73.43	71.46
SVF	31.17	4.88	66.85	68.83	67.76
DISTTIC	28.59	5.49	64.99	71.41	68.02

7. CONCLUSIONS

The identification of phone boundaries in continuous speech is an important problem in areas of speech synthesis and recognition. In this work, we examine the applicability of the DISTBIC algorithm for the text-independent detection of phone boundaries in continuous speech under noisy conditions. DISTBIC and its variants have been used primarily for speaker-based segmentation (Delacourt and Wellekens, 2000; Kadri et al., 2006; Wand and Cheng, 2004; Zhang et al., 2006). Because the performance of the baseline algorithm is sensitive to sample sizes, we proposed modifications on the original algorithm so that it serves better for the phonemic segmentation task. By considering small-sample corrections to Bayesian information criterion, we were able to improve the segmentation accuracy by replacing BIC with BICC or ABF2.

We have also demonstrated that by representing the signal samples with a GΓD, we are able to yield improved results compared to the normal distribution for an offline two-pass algorithm. 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-based algorithms (DISTBIC-Γ, DISTBICC-Γ, DISTABF2-Γ), we generally improved the segmentation accuracy. While GΓD offers great flexibility and can represent accurately the speech signal especially in a small scale, its MLE is problematic since it has large bias in moderate and small sample sizes. Nevertheless, judging from the results in the present application, this inefficiency is not destructive and we are able to relax the convergence requirements of the online gradient ascent algorithm for the parameter estimation by assuming initial values close to ΓD or LD. These have been found to be good approximations of the speech signal in various domains. Using alternative parameter estimation methods for GΓD, such as the method of moments, would be worth exploring since many have asserted that the extracted estimates introduce less bias than MLE when small sample sizes are considered (Hwang and Huang, 2002). Likewise, Shin et al. (2005) have also yielded improved performance by modelling both speech and noise with a two-sided GΓD in the DFT domain. An assumption is made there that the real and imaginary parts of the DFT coefficient are statistically independent and distributed according to the same GΓD. Their LRT-based detector obtained better results under vehicular and office noise than conventional methods. In our work, we have asserted that we can operate in the MFCC domain as well with similar success.

Evaluation of automatic phone segmentation in two different datasets and for different noise conditions confirmed that ABF2 performed best overall, followed by BICC. The superior results in small sample sizes can be credited to the fact that ABF2 has a more flexible implicit unit information prior than BIC. Performance comparisons with other speech segmentation algorithms reported in the literature show that DISTABF2-Γ yields improved results.

In the present work, the size of experiments was limited and the computation time in the proposed approach was multiple times greater for the method with GΓDs than using the simple DISTBIC algorithm. We assert that by refining the set of algorithm parameters (i.e. analysis window size/shift, penalty factor λ, etc.) it is possible to limit both type-I and type-II errors. There are numerous combinations worth exploring for offline two-step speech segmentation at phone level. In future, it will be worth investigating alternative corrected versions of BIC and evaluate them against other criteria, such as the deviation information criterion (DIC) (Spiegelhalter et al., 2002), MDL, and ICOMP (Bozdogan, 1988), similarly corrected for small samples. The advantage of DIC, for Bayesian model selection, is that DIC is easily calculated from the samples generated by a Markov chain Monte Carlo simulation while AIC and BIC require calculating the likelihood at its maximum over parameters $\boldsymbol{\theta}$, which is not readily available from the Markov Chain Monte Carlo (MCMC) simulation. Sugiyama and Ogawa (2001) have proposed the subspace information criterion (SIC), which gives an unbiased estimate of the generalisation error in model selection when finite samples are considered. They have attested that SIC outperforms other criteria, including AIC, BIC, and CAIC, in small samples. A more detailed discussion of different information criteria, including BIC, is given in (Chickering and Heckerman, 1997) and (Gheissari and Bab-Hadiashar, 2003). An evaluation of model selection criteria for acoustic segmentation is given in (Cettolo and Federico, 2000).

APPENDIX A. Bayesian Information Criterion small sample approximation assuming a generalised Gamma distribution prior.

A method widely used in probabilistic modelling to approximate the value of marginal likelihoods in model comparison is Laplace's method. This approach approximates the integral of a function $\int f(x)dx$ by fitting a Gaussian at the maximum $\hat{x}$ of $f(x)$, and computing the volume under the Gaussian. The covariance of the fitted Gaussian is determined by the Hessian matrix $\boldsymbol{H}$ of $\ln f(x)$ at the maximum point $\hat{x}$. In other words, Laplace's method is a way of approximating the posterior distribution with a Gaussian centred at the MAP estimate. This can be justified by the fact that under certain regularity conditions, the posterior distribution approaches a GD as the number of samples grows (Gelman et al., 1995). Despite using a full distribution to approximate the posterior, Laplace's method still suffers from most of the problems of MAP estimation. Estimating the variances does not help if the procedure has already lead to an area of low probability mass. Laplace's method employs a Taylor series expansion of $\boldsymbol{H}$ around the MAP estimate $\hat{\boldsymbol{x}}$, taken to second order

$$\int f(\boldsymbol{x})d\boldsymbol{x} \approx f(\hat{\boldsymbol{x}})(2\pi)^{k/2} \,|-\nabla\nabla \ln(f(\hat{\boldsymbol{x}}))|^{-1/2}. \tag{A.1}$$

In Bayesian statistics, the marginal likelihood $P(\boldsymbol{X} \mid M_k)$ which indicates the evidence for model M_k is computed by integrating over the unknown parameter values in that model. Considering the hypothesis testing of Sect. 3.2, we need to calculate the BF for the model M_1 against the model M_0 given data X. The BF is the ratio of posterior to prior odds, i.e. the ratio of the marginal likelihoods. The likelihoods for M_0 and M_1 can be written down explicitly, so once the prior has been fully specified, the marginal likelihood in (7) can be computed using Laplace' approximation

$$P(\boldsymbol{x} \mid M_k) = \int_{\boldsymbol{\theta}_k} p(\boldsymbol{\theta}_k \mid M_k)\, p(\boldsymbol{x} \mid \boldsymbol{\theta}_k, M_k) d\boldsymbol{\theta}_k \approx (2\pi)^{d_k/2} \left|\tilde{I}_0(\tilde{\boldsymbol{\theta}}_\kappa)\right|^{-1/2} P(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}}_\kappa, M_k) P(\tilde{\boldsymbol{\theta}}_\kappa \mid M_k) \tag{A.2}$$

where $p(\boldsymbol{\theta}_k \mid M_k)$ is the prior distribution of $\boldsymbol{\theta}$ given the model M_k, $\tilde{\boldsymbol{\theta}}_\kappa$ corresponds to the mode of $p(\boldsymbol{\theta}_k \mid M_k)$, $d_k = \dim(\boldsymbol{\theta}_k)$ is the dimension of the parameter space, $\tilde{I}_0(\cdot)$ is Fischer's observed information matrix of the posterior distribution $p(\boldsymbol{x} \mid M_k)$ and $|\cdot|$ denotes the determinant of a matrix. By substituting $\tilde{\boldsymbol{\theta}}_\kappa$ with the MLE estimated value $\hat{\boldsymbol{\theta}}_k$, and replacing $\tilde{I}_0$ by the expected information matrix per observation $\bar{I}_E$ with i,j –th element

$$\bar{I}_E(\hat{\boldsymbol{\theta}}_k)_{ij} = \frac{1}{n} I_E(\hat{\boldsymbol{\theta}}_k)_{ij} = -\frac{1}{n} E\left[\frac{\partial^2 \ln[P(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k)]}{(\partial \boldsymbol{\theta}_k)_i (\partial \boldsymbol{\theta}_k)_j}\right]_{\hat{\boldsymbol{\theta}}_k} = -\frac{1}{n} E\left[\frac{\partial^2 L(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_k, M_k)}{(\partial \boldsymbol{\theta}_k)_i (\partial \boldsymbol{\theta}_k)_j}\right]_{\hat{\boldsymbol{\theta}}_k} \tag{A.3}$$

where n is the number of observations (i.e. the sample size), we get:

$$-2\ln P(\boldsymbol{x}\mid M_k) = -2\ln P(\boldsymbol{x}\mid \hat{\boldsymbol{\theta}}_k, M_k) - 2\ln P(\hat{\boldsymbol{\theta}}_k \mid M_k) - d_k \ln 2\pi + d_k \ln(n) + \ln\left|\overline{I}_E(\hat{\boldsymbol{\theta}}_k)\right| + O(n^{-1/2}) \tag{A.4}$$

APPENDIX B. Calculation of ABF2 model complexity penalty when using a generalised Gamma distribution prior for data.

As stated in Sect. 3.4, Bollen et al propose a special unit information prior where the variance of the parameter $\boldsymbol{\theta}_k$ is scaled by w (Bollen et al., 2005). By using the optimum value for w that maximises the posterior probability $P(\boldsymbol{X}\mid M_k)$ and approximating only the likelihood $p(\boldsymbol{\theta}_k \mid M_k)$ with Laplace's method, we can derive (Bollen et al., 2005)

$$ABF2 = \begin{cases} -2L(\hat{\boldsymbol{\theta}}_k) + d_k\left(1+\ln\dfrac{d_k}{\hat{\boldsymbol{\theta}}_k^{\,T}\overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k}\right), & \text{if } d_k > \hat{\boldsymbol{\theta}}_k^{\,T}\overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k \\ -2L(\hat{\boldsymbol{\theta}}_k) - \hat{\boldsymbol{\theta}}_k^{\,T}\overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k, & \text{otherwise} \end{cases} \tag{B.1}$$

Looking at (B.1), it is clear that the expected Fisher's information matrix must be calculated explicitly. When we have univariate data and a single parameter θ to estimate, Fisher's information is defined as:

$$I_E(\theta_k) = -E\left[\frac{\partial^2 \ln[p(x\mid\theta_k, M_k)]}{\partial\theta^2}\right] = -\int \frac{\partial^2 \ln[p(x\mid\theta_k, M_k)]}{\partial\theta^2}\, p(x\mid\theta_k, M_k)dx\,. \tag{B.2}$$

If we have n independent observations $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ then the probability densities multiply and log-likelihoods add. Thus, by linearity of expectation

$$I_E(\theta\mid\boldsymbol{x}) = nI_E(\theta\mid x) \tag{B.3}$$

where x is any one of the x_i. This means that n times as many observations should give us n times as much information about the value of an unknown parameter. Accordingly, for multiple parameters and observations,

$$I_E(\boldsymbol{\theta}\mid\boldsymbol{x}) = nI_E(\boldsymbol{\theta}\mid x)\,. \tag{B.4}$$

Assuming observation comes from an i.i.d distribution, the average (per observation) expected information matrix per observation at the MLE $\overline{I}_E(\hat{\boldsymbol{\theta}}_k)$ is given in (A.3). For the 3-parameter GΓD PDF of (30) $\boldsymbol{X} \sim \text{GΓD}(a,b,c)$, it is clearly $\boldsymbol{\theta}_k = (\theta_1,\theta_2,\theta_3)^T = (a,b,c)^T$. Under the MLE approximation, each element in the expected information matrix can be analytically calculated, e.g.

$$I_E(\boldsymbol{\theta}_k)_{11} = -E\left[\frac{\partial^2 \ln[P(\boldsymbol{x}\mid\boldsymbol{\theta}_k, M_k)]}{\partial a \partial a}\right] = -\int\left[\frac{\partial^2 \ln[P(\boldsymbol{x}\mid\boldsymbol{\theta}_k, M_k)]}{\partial a \partial a} P(\boldsymbol{x}\mid\boldsymbol{\theta}_k, M_k)dx\right] =$$
$$= -\int\left[\frac{\partial^2 \ln\left(\frac{cb^a}{2\Gamma(a)} \mid x\mid^{ac-1} e^{-b|x|^c}\right)}{\partial a \partial a} \frac{cb^a}{2\Gamma(a)} \mid x\mid^{ac-1} e^{-b|x|^c} dx\right] = \psi_1(a) \tag{B.5}$$

Similarly,

$$I_E(\boldsymbol{\theta}_k)_{12} = I_E(\boldsymbol{\theta}_k)_{21} = -\frac{1}{b}$$
$$I_E(\boldsymbol{\theta}_k)_{13} = I_E(\boldsymbol{\theta}_k)_{31} = \frac{-\psi_0(a) + \ln b}{c}$$
$$I_E(\boldsymbol{\theta}_k)_{23} = I_E(\boldsymbol{\theta}_k)_{32} = \frac{a\psi_0(a) - a\ln b + 1}{bc} \tag{B.6}$$
$$I_E(\boldsymbol{\theta}_k)_{22} = \frac{a}{b^2}$$
$$I_E(\boldsymbol{\theta}_k)_{33} = \frac{a\psi_0(a)^2 - 2(a\ln b - 1)\psi_0(a) + a\psi_1(a) - 2\ln b + a\ln(b)^2 + 1}{c^2}$$

where ψ_0 and ψ_1 are the digamma and trigamma functions respectively:

$$\psi_0(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{B.7}$$

$$\psi_1(x) = \frac{d\psi_1(x)}{dx} = \frac{d^2\ln\Gamma(x)}{dx^2} \tag{B.8}$$

The expected information matrix is then

$$I_E(\boldsymbol{\theta}_k) = n\begin{pmatrix} \psi_1(a) & -\frac{1}{b} & \frac{-\psi_0(a)+\ln b}{c} \\ -\frac{1}{b} & \frac{a}{b^2} & \frac{a\psi_0(a) - a\ln b + 1}{bc} \\ \frac{-\psi_0(a)+\ln b}{c} & \frac{a\psi_0(a) - a\ln b + 1}{bc} & \frac{a\psi_0(a)^2 - 2(a\ln b - 1)\psi_0(a) + a\psi_1(a) - 2\ln b + a\ln(b)^2 + 1}{c^2} \end{pmatrix} \tag{B.9}$$

Since we have i.i.d. data, the average expected information matrix per observation $\overline{I}_E(\boldsymbol{\theta}_k)_{11}$ is given by

$$\overline{I}_E(\boldsymbol{\theta}_k) = \frac{1}{n} I_E(\boldsymbol{\theta}_k)\,. \tag{B.10}$$

We can now calculate the complexity penalty in (B.1) by first calculating the term

$$\hat{\boldsymbol{\theta}}_k^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k = \begin{pmatrix}\hat{a} & \hat{b} & \hat{c}\end{pmatrix}\begin{pmatrix} \psi_1(\hat{a}) & -\frac{1}{\hat{b}} & \frac{-\psi_0(\hat{a})+\ln\hat{b}}{\hat{c}} \\ -\frac{1}{\hat{b}} & \frac{\hat{a}}{\hat{b}^2} & \frac{\psi_0(\hat{a})-\hat{a}\ln\hat{b}+1}{\hat{b}\hat{c}} \\ \frac{-\psi_0(\hat{a})+\ln\hat{b}}{\hat{c}} & \frac{\psi_0(\hat{a})-\hat{a}\ln\hat{b}+1}{\hat{b}\hat{c}} & \frac{\hat{a}\psi_0(\hat{a})^2-2(\hat{a}\ln\hat{b}-1)\psi_0(\hat{a})+\hat{a}\psi_1(\hat{a})-2\ln\hat{b}+a\ln\hat{b}^2+1}{\hat{c}^2} \end{pmatrix}\begin{pmatrix}\hat{a}\\ \hat{b}\\ \hat{c}\end{pmatrix} \quad \text{(B.11)}$$

$$= \hat{a}\psi_0(\hat{a})^2 - 2(\hat{a}\ln\hat{b}-1)\psi_0(\hat{a}) + (\hat{a}^2+\hat{a})\psi_1(\hat{a}) + \hat{a}\ln\hat{b}^2 - \hat{a} - 2\ln\hat{b} + 3$$

APPENDIX C. Parameter estimation of Generalised Gamma Distribution.

The complexity of GΓD imposes difficulties in the estimation of its parameters. Various existing methods include the method of moments, MLE, minimum mean square error, Gibbs sampling, etc. Shin and Chang (2005) have suggested an inexpensive MLE approach using the gradient ascend method. Gradient ascent is an optimisation method for finding the nearest local minimum of a univariate or multivariate function. Starting at a point P_0, the algorithm moves from P_i to P_{i+1} by maximising along the line extending from P_i in the direction of the local uphill gradient $\nabla f(P_i)$. Taking steps proportional to the gradient (or the approximate gradient) of the function at the current point, *f* increases fastest, reaching a local maximum, assuming that the sequence convergences.

Given *n* data $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ and assuming the data are univariate and mutually independent, the log-likelihood function is

$$L(\boldsymbol{x}; a, b, c) = \ln f(\boldsymbol{x}; a, b, c) = n \ln \frac{cb^a}{2\Gamma(a)} + (ac-1)\sum_{i=1}^{n} \ln |x_i| - b\sum_{i=1}^{n} |x_i|^c \,. \quad \text{(C.1)}$$

In order to calculate the MLE of the parameters, we need to differentiate the log-likelihood function with respect to *a*, *b*, and *c*. Setting the partial derivatives of the function to zero simultaneously

$$\frac{\partial L(x; a, b, c)}{\partial a} = \frac{\partial L(x; a, b, c)}{\partial b} = \frac{\partial L(x; a, b, c)}{\partial c} = 0 \quad \text{(C.2)}$$

we get the following equations

$$\begin{aligned} \psi_0(a) &= \ln b + \frac{1}{n}\sum_{i=1}^{n} \ln |x_i|^c \\ b &= \frac{an}{\sum_{i=1}^{n} |x_i|^c} \\ \frac{1}{a} + \psi_0(a) - \ln b - \frac{b}{an}\sum_{i=1}^{n} |x_i|^c \ln |x_i|^c &= 0 \end{aligned} \quad \text{(C.3)}$$

where ψ_0 is the digamma function which is the first-order derivative of $\ln\Gamma(x)$ (B.7). In order to estimate the three parameters, we iteratively update the following statistics by computing them over the given data

$$S_1(i) = (1-\xi)S_1(i-1) + \xi \left|x_i\right|^{\hat{c}(i)}$$
$$S_2(i) = (1-\xi)S_2(i-1) + \xi \ln\left|x_i\right|^{\hat{c}(i)} \quad \text{(C.4)}$$
$$S_3(i) = (1-\xi)S_3(i-1) + \xi \left|x_i\right|^{\hat{c}(i)} \ln\left|x_i\right|^{\hat{c}(i)}$$

updating each time the parameter c as

$$\hat{c}(i+1) = \hat{c}(i) + \mu\phi(\hat{a}(i), \hat{c}(i), x) = \hat{c}(i) + \mu\left(\frac{1}{\hat{a}(i)} + S_2(i) - \frac{S_3(i)}{S_1(i)}\right) \quad \text{(C.5)}$$

where ξ is a forgetting factor and $\mu>0$ is small enough number that corresponds to the learning rate of the gradient ascent approach. The online version of the gradient of the average log-likelihood function with respect to c, $\phi(\hat{a}(i), \hat{c}(i), x)$, is derived by (C.3) and (C.4)

$$\phi(\hat{a}(i), \hat{c}(i), x) = \frac{1}{\hat{a}(i)} + S_1(i) - \frac{S_3(i)}{S_1(i)}. \quad \text{(C.6)}$$

Using appropriate initial estimates for the parameter c (e.g. $\hat{c}(1) = 1$, which corresponds to ΓD or LD), we are able to recursively estimate the remaining parameters by solving the equations:

$$\psi_0(\hat{a}(m)) - \ln\hat{a}(m) = S_2(m) - \ln S_1(m) \quad \text{(C.7)}$$
$$\hat{b}(n) = \frac{\hat{a}(m)}{S_1(m)} \quad \text{(C.8)}$$

The left part of (C.7) is monotonically increasing function of $\hat{a}(m)$, which means that we can determine uniquely the solution by having an inverse table. Once the algorithm has converged, or after a sufficient number of iterations, the log-likelihood can be computed

$$L(x_i; a, b, c) = \ln\frac{\hat{c}\hat{b}^{\hat{a}}}{2\Gamma(\hat{a})} + (\hat{a} - \frac{1}{\hat{c}})S_2 - \hat{b}S_1. \quad \text{(C.9)}$$

A weakness of the gradient ascent algorithm is that it may require many iterations to converge towards a local maximum/minimum, if the curvature in different directions is very different or for functions which have long, narrow valley structures. As a result, finding the optimal c per step can be time-consuming. Conversely, using a fixed c can yield poor results. Methods based on Newton's method and inversion of the Hessian using conjugate gradient techniques are often a better alternative (Shewchuk, 1994).

8. ACKNOWLEDGMENTS

G. Almpanidis was granted a basic research fellowship co funded by the European Union and the Hellenic Ministry of Education in the framework of the program "HERAKLEITOS" of the Operational Program for Education and Initial Vocational Training within the 3rd Community Support Framework.

9. REFERENCES

Adell, J., Bonafonte, A., June 2004. Towards phone segmentation for concatenative speech synthesis. In: Proc. 5th ISCA Speech Synthesis Workshop (SSW5), 139-144.

Akaike, H., Dec. 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control. 19 (24), 716-723.

Anderson, D., Burnham, K., 1999. Understanding information criteria for selection among capture-recapture or ring recovery models. Bird Study. 46, 514-521.

Anderson, D., White, G., 1994. AIC model selection in overdispersed capture-recapture data. Ecology. 75, 1780-1793.

Aversano, G., Esposito, A., Marinaro, M., 2001. A new Text-Independent Method for Phoneme Segmentation. In: Proc. 44th IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 516-519.

Bakis, R., Chen, S., Gopalakrishnan, P., Gopinath, R., Maes, S., Polymenakos, L., Franz, M., 1997. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. In: Proc. Speech Recognition Workshop, pp. 67-72.

Baudoin, G., Capman, F., Cernocky, J., El-chami, F., Charbit, M., Chollet, G., Petrovska-Delacrétaz, D., Sep. 2002. Advances in Very Low Bit Rate Speech Coding using Recognition and Synthesis Techniques. In: Proc. 5th Int. Conf. Text, Speech and Dialogue (TSD 2002), pp. 269-276.

Bernardo, J., 1976. Algorithm AS 103: Psi (digamma) function. Applied Statistics. 25, 315-317.

Bogdan, M., Ghosh, J., Doerge, R., June 2004. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. Genetics. 167, 989-999.

Bollen, K., Ray, O., Zavisca, J., Nov. 2005. A Scaled Unit Information Prior Approximation to Bayes Factor. SAMSI LVSSS Workshop Transition Workshop: Latent Variable Models in the Social Sciences.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika. 52, 345-370.

Bozdogan, H., Apr. 1988. ICOMP: A New Model Selection Criterion. In: Proc. Classification and Related Methods in of Data Analysis, pp. 599-608.

Brugnara, F., De Mori, R., Giuliani, D., Omologo, M., Oct. 1992. Improved Connected Digit Recognition Using Spectral Variation Functions. In: Proc. Int. Conf. Spoken Language Processing (ICSLP92), vol. 1, pp. 627-630.

Burnham, K., Anderson, D., 2004. Multimodel Inference Understanding AIC and BIC in Model Selection. Sociological Methods & Research. 33(2), 261-304.

Cettolo, M., Federico, M. 2000. Model selection criteria for acoustic segmentation. In: Proc. ISCA ITRW ASR2000 Automatic Speech Recognition, pp. 221-227.

Cohen, A., Whitten, B., 1986. Modified moment estimation for the three-parameter gamma distribution. Journal of Quality Technology. 17, 147-154.

Chang, J., Shin, J., Kim, N.S., 2003. Likelihood ratio test with complex Laplacian model for voice activity detection. In: Proc. European Conf. Speech Communication Technology, pp. 1065-1068.

Chen, S., Gopalakrishnam, P., 1998. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: DARPA Speech Rec. Workshop.

Chen, S., Eide, E., Gales, M., Gopinath, R., Kanvesky, D., Olsen, P., May 2002. Automatic transcription of Broadcast News. Speech Communication archive - Special Issue on Automatic Transcription of Broadcast News Data. 37(1-2), 69-87.

Chickering, D., Heckerman, D., 1997. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. Machine Learning. 29(2-3), 181-212.

Dayton, C., 2003. Model comparisons using information measures. Modern Applied Statistical Methods. 2, 281-292.

Delacourt, P. Wellekens, C.J., Sep. 2000. DISTBIC: a speaker-based segmentation for audio data indexing. Speech Communication. 32(1-2), 111-126.

Esposito A., Aversano, G., 2004. Text Independent Methods for Speech Segmentation. In: Proc. Summer School on Neural Networks, pp. 261-290.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1990. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech. Disc 1-1.1, NTIS Order No. PB91-505065.

Gazor, S. Zhang, W., 2003. Speech probability distribution. IEEE Signal Processing Letters. 10(7), 204-207.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton.

Gheissari, N., Bab-Hadiashar, A., Dec. 2003. Model Selection Criteria in Computer Vision Are they different?. In: Proc. Digital Image Computing Techniques and Applications, vol. 1, pp. 185-204.

Glass, J., 2003. A Probabilistic Framework for Segment-Based Speech Recognition. Computer Speech and Language. 17, 137-152.

Hannan, E., Quinn, B., 1978. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B. 41, 190-195.

Harter, H., 1967. Maximum likelihood estimation of the parameters of a four-parameter generalized gamma population from complete and censored samples. Technometrics. 9, 159-165.

Hager, H., Bain, L., 1987. Inferential procedures for the generalized gamma distribution. Journal of the American Statistical Association. 82, 528-550.

Haughton, D., 1988. On the choice of a model to fit the data from an exponential family. Annals of Statistics. 16, 342-355.

Hwang, T., Huang, P., 2002. On New Moment Estimation of Parameters of the Gamma Distribution Using its Characterization. Annals of the Institute of Statistical Mathematics. 54(4), 840-847.

Huy Dat, T., Takeda, K., Itakura, F., May 2006. Multichannel Speech Enhancement Based On Speech Spectral Magnitude Estimation Using Generalized Gamma Prior Distribution. In: Proc. 2006 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 4, pp. 1149-1152.

Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J., 1990. NTIMIT: A Phonetically Balanced, Continous Speech, Telephone Bandwidth Speech Database. In: Proc. 1990 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 109-112.

Kadri, H., Lachiri, Z., Ellouze, N., 2006. Hybrid Approach for Unsupervised Audio Speaker Segmentation. In: Proc. European Conf. Signal Processing.

Kashyap, R., 1977. A Bayesian Comparison of Different Classes of Dynamic Models Using Empirical Data. IEEE Trans. Auto Control. AC-22(5), 715-727.

Kass, R., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association. 90, 928-934.

Kokkinakis, K., Nandi, A., May 2006. Flexible Score Functions for Blind Separation of Speech Signals Based on Generalized Gamma Probability Density Functions. In: Proc. 2006 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1, pp. 1217-1220.

Kominek, J., Bennet, C., Black, A., Sep. 2003. Evaluating and correcting phoneme segmentation for unit selection synthesis. In: Proc. 8th European Conf. Speech Communication and Technology (EUROSPEECH2003), pp. 313-316.

Kuha, J., 2004. AIC and BIC: Comparisons of Assumptions and Performance. Sociological Methods & Research. 33(2), 188-229.

Lawless, J., 1980. Inference in the generalized gamma and log gamma distributions. Technometrics. 33, 409-419.

Lee, M., 2002. On the complexity of additive clustering models. Mathematical Psychology. 45(1), 131-148.

Martin, R., 2005. Speech enhancement using short time spectral estimation with Gamma distributed priors. In: Proc. 2005 IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 1, pp. 253-256.

Mitchell, C., Harper, M., Jamieson, L., Using Explicit Segmentation to Improve HMM Phone Recognition, 1995. In: Proc. 1995 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol.1, pp. 229-232.

Nakamura, A., Nov. 2000. Acoustic modeling for speech recognition based on a generalized Laplacian mixture distribution. IEICE Trans. Inf. & Syst. J83-D-II(11), 2118-2127.

Nemer, E., Goubran, R., Mahmould, S., Mar. 2001. Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans. Speech and Audio Processing. 9(3), 217-231.

Olsthoorn, T., 1995. Effective parameter optimization for groundwater model calibration. Ground Water. 33, 42-48.

Parr, V., Webster, J., 1965. A method for discriminating between failure density functions used in reliability predictions. Technometrics. 7, 1-10.

Pickett, J., Morris, S., 1999. The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology. Allyn & Bacon, Boston.

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

Raftery, A., 1999. Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". Sociological Methods & Research. 27(3), 411-427.

Raftery, A., Richardson, S., 1996. Model Selection for Generalized Linear Models via GLIB, With Application to Epidemiology. Bayesian Biostatistics. 321-354.

Rissanen, J., 1978. Modeling by the shortest data description. Automatica. 14, 465-471.

Schneider, B., 1978. Algorithm AS 121: Trigamma function. Applied Statistics. 27, 97-99.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics. 6, 461-464.

Schwarz, P., Matejka, P., Cernocky, J., May 2006. Hierarchical structures of neural networks for phoneme recognition. In: Proc. 2006 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1, pp. 325-328.

Shewchuk, J., Mar. 1994. An introduction to the conjugate gradient method without the agonizing pain. Technical Report CMU-CS-94-125, School of Computer Science, Carnegie Mellon University, Pittsburgh, Philadelphia.

Shin, J.W., Chang, J-H., Mar. 2005. Statistical modeling of speech signals based on generalized gamma distribution. IEEE Signal Processing Letters. 12(3), 258-261.

Shin, J.W., Chang, J-H., Yun, H.S., Kim, N.S., 2005. Voice activity detection based on generalized gamma distribution. In: Proc. 2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1, pp. 781-784.

Sohn, J., Kim, N.S., Sung, W., Jan. 1999. A statistical model based voice activity detection. IEEE Signal Processing Letters. 6(1), 1-3.

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A., Oct. 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology). 64(4), 583-639.

Stacy, E., 1962. A generalization of the gamma distribution. Annals of Mathematical Statistics. 33, 1187-1192.

Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics: Theory and Methods. 7(1), 13-26.

Sugiyama, M., Ogawa, H., Apr. 2001. Model selection with small samples. In: Proc. Artificial Neural Nets and Genetic Algorithms, pp. 418-421.

Takeuchi, K., 1976. Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematic Sciences). 153, 12-18.

Toledano, D., Gomez, A., Nov. 2003. Automatic phonetic segmentation. IEEE Trans. Speech Audio Process.. 11(6), 617-625.

Tremblay, M., Wallach, D., 2004. Comparison of parameter estimation methods for crop models. Agronomie. 24, 351-365.

Tritschler, A., Gopinath, R., 1999. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In: Proc. 1999 European Speech Processing, vol. 2, pp. 679-682.

Tsionas, E., 2001. Exact Inference in Four-Parameter Generalized Gamma Distributions. Communications in Statistics (Theory and Methods). 30(4), 747-756.

Van Allen, T., 2000. Handling uncertainty when you' re handling uncertainty: model selection and error bars for belief networks. Master of Science thesis, Department of Computer Science, University of Alberta. Available online at http://www.cs.ualberta.ca/~vanallen/thesis.ps

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

Villiers, E., Preez , J.A., Nov. 2001. The advantage of using higher order HMM's for segmenting acoustic files. In: Proc. 12th Annual Symposium of the Pattern Recognition Association of South Africa, pp. 120-122.

Visser, I., Ray, S., Jang, W., Berger, J., Nov. 2005. Effective sample size and the Bayes factor. SAMSI LVSSS Workshop Transition Workshop: Latent Variable Models in the Social Sciences.

Volinsky, C., Raftery, A., 2000. Bayesian Information Criterion for Censored Survival Models. Biometrics. 56, 256-262.

Weakliem, D., 1999. A Critique of the Bayesian Information Criterion for Model Selection. Sociological Methods & Research. 27(3), 359-397.

Wang, H., Cheng, S., Oct. 2004. METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation. In: Proc. Int. Conf. Spoken Language Processing (ICSLP2004), pp. 1617-1620.

Wang, L., Zhao, Y., Chu, M., Soong, F., Zhou, J., Cao, Z., 2006. Context-Dependent Boundary Model for Refining Boundaries Segmentation of TTS Units. IEICE Trans. Information and Systems. E89-D(3), 1082-1091.

Woodland, P., Gales, M., Pye, D., Young, S., 1997. The Development of the 1996 HTK broadcast news transcription system. In: Proc. Speech Recognition Workshop, pp. 73-78.

Zhang, S., Zhang, S., Xu, B., 2006. A Two-level Method for Unsupervised Speaker-based Audio Segmentation. In: Proc. 18th Int. Conf. Pattern Recognition (ICPR2006), pp. 298-301.

Zhao, Y., Wang, L., Chu, M., Soong, F., Cao, Z., 2005. Refining Phoneme Segmentations Using Speaker-Adaptive Context Dependent Boundary Models. In: Proc. Interspeech, pp. 2557-2560.

LIST OF FIGURES

LIST OF TABLES