Robust Detection of Phone Boundaries Using Model Selection Criteria With Few Observations

George Almpanidis, Margarita Kotti, and Constantine Kotropoulos, *Senior Member, IEEE*

***Abstract*—Automatic phone segmentation techniques based on model selection criteria are studied. We investigate the phone boundary detection efficiency of entropy- and Bayesian- based model selection criteria in continuous speech based on the DISTBIC hybrid segmentation algorithm. DISTBIC is a text-independent bottom-up approach that identifies sequential model changes by combining metric distances with statistical hypothesis testing. Using robust statistics and small sample corrections in the baseline DISTBIC algorithm, phone boundary detection accuracy is significantly improved, while false alarms are reduced. We also demonstrate further improvement in phonemic segmentation by taking into account how the model parameters are related in the probability density functions of the underlying hypotheses as well as in the model selection via the information complexity criterion and by employing M-estimators of the model parameters. The proposed DISTBIC variants are tested on the NTIMIT database and the achieved F_1 measure is 74.7% using a 20-ms tolerance in phonemic segmentation.**

***Index Terms*—Automatic phonetic segmentation, model selection, robust statistics.**

I. Introduction

MANY areas in speech processing require algorithms that automatically associate the speech signal to other annotation layers (orthographic, phonetic transcription, etc.) by means of time stamps [1]. Phonemic segmentation is commonly used in preprocessing and initial training steps of automatic speech/speaker recognition, speech enhancement, computer-aided speech transcription systems, and the development of corpus-based speech synthesis systems. The automatic detection of the start and end boundaries of phone segments in continuous speech by statistical methods is a challenging task due to the small sample size. While hand-labeling of continuous speech by listening to the sound and visually inspecting the speech waveform or spectrogram in order to determine the phone boundaries yields better accuracy than automatic methods, the speed of segmentation by expert phoneticians is over 130 times real-time [2]. Consequently, recent concatenative speech synthesizers, which rely on large phone-segmented speech databases to realize high-quality synthetic speech, require a huge amount of human effort. Clearly, even a semiautomatic approach to phone boundary detection, with the automatic annotation acting as reference for further human correction, can accelerate the segmentation procedure significantly. However, the savings in time depend on the accuracy of the automatic segmentation. The better the accuracy, the higher the time savings. Besides speech database automatic annotation, additional applications of phonemic segmentation can be found in language identification, cellular telephony technology, and lip-synchronization techniques in audiovisual communication [3]. In environments with low signal-to-noise ratio (SNR), energy-based phonemic segmentation algorithms often misclassify nonstationary noise as speech activity. They can not identify unvoiced phones, such as fricatives, satisfactorily as the latter can be masked by noise. Consequently, they are inefficient for real-world recordings where speakers tend to leave artifacts, such as breathing/sighing, mouth clicks, teeth chatters, and echoes. In contrast, approaches that explore speech and noise statistics and incorporate *model selection criteria* (MSC) are more robust in low SNRs [4].

In this paper, we propose an automatic acoustic change detection algorithm that identifies phone boundaries using information-theoretic approaches for statistical inference while avoiding the need for linguistic constraints and training data. Linguistically unconstrained approaches are useful in applications that require explicit speech segmentation, when a phonetic transcription is either unavailable or inaccurate. Phone segmentation using MSC has been studied in an earlier work, where speech and noise features were independently modeled using univariate generalized Gamma distributions [5]. The novelty of this paper is in suggesting various MSC, some not fully exploited in speech processing yet, deploying multivariate statistics without assuming variable uncorrelatedness, and considering the limited information available in phonemic segmentation as well as the presence of outliers whose influence is reduced by employing robust estimators of the model parameters.

Parsimony, working hypotheses, and strength of evidence are three principles that regulate the ability to make inferences [6]. Information-theoretic approaches adhere in part to all these concepts, which make them more attractive than classical pairwise significance testing. Akaike information criterion (AIC) [7] and Bayesian information criterion (BIC) [8] have been the most commonly adopted MSC as they have a straightforward implementation and yield reasonable results. AIC has a strong theoretical underpinning, based on Kullback–Leibler (KL) information and maximum-likelihood estimation (MLE) theory, while

Manuscript received December 11, 2007; revised September 23, 2008. Current version published January 14, 2009. This work was supported in part 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 Third Community Support Framework.
The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

The authors are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece (e-mail: galba@aiia.csd.auth.gr; mkotti@aiia.csd.auth.gr; costas@aiia.csd.auth.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2009162

BIC stems from Bayesian arguments and is strongly related to Bayes factors (BF). AIC and BIC have been widely used in several applied fields and have motivated a plethora of extensions and derivatives that attempt to overcome many of their shortcomings. The application of MSC to speech and audio segmentation has been studied by numerous researchers (cf. [9]–[12]).

In this paper, we build on the DISTBIC segmentation algorithm [10] to determine the phone boundaries in continuous speech. DISTBIC is a hybrid method that combines distance measures with BIC and has been applied successfully in many speech segmentation tasks. In order to deal with small sample sizes in phonemic segmentation, the basic algorithm needs to be modified. The various modifications of the baseline DISTBIC algorithm (cf. [13]–[15]) have been derived mostly from speech segmentation at the word/sentence scale or speaker segmentation. In this paper, we apply robust statistics corrections to BIC through M-estimation of the model parameters, feature vector transformation, and small sample corrections to the complexity penalty term of BIC. We also disregard the assumption of uncorrelated elements within a feature vector by employing full covariance matrices. Moreover, we study the possibility of using alternative MSC for estimating more efficiently the penalty due to model complexity. In particular, BIC and AIC do not consider the functional form of a probability model, or the way in which model parameters control estimated feature interdependence. In order to better represent feature interdependence, we propose to replace BIC with a version of the information complexity criterion (ICOMP) that is based on Fisher information matrix (FIM), abbreviated as ICOMP(IFIM), where IFIM stands for inverse FIM [16].

The outline of the paper is as follows. Section II describes the problem of speech and phonemic segmentation, surveys related past work, and presents the baseline hybrid statistical technique for speech segmentation, DISTBIC. In Section III, we assert that speech modeling in very small window sizes impels us to consider alternative MSC with corrections due to small sample sizes as well as to alleviate the presence of outliers. Section IV discusses robust techniques in phonemic segmentation, reviews complexity penalized MSC, and presents necessary adjustments to the baseline algorithm DISTBIC for text-independent phone boundary detection. Section V features experimental results for evaluating phone boundary detection performance in a noisy environment. It is shown that the proposed DISTIBIC variants yield significant reductions in boundary detection errors. Results are discussed in Section VI. Finally, Section VII concludes the paper.

II. Phonemic Segmentation

Phone segmentation methods can be classified into text-dependent and linguistically unconstrained approaches [4], [17]. The approaches of the first class typically adopt a generative top-down procedure estimating the likelihood of top-level linguistic hypotheses. In linguistically unconstrained segmentation, no prior knowledge about the text content is used and the acoustic information contained in the speech signal is only exploited in order to detect phone transitions.

A. Top-Down Approaches to Phonemic Segmentation

Most of the recent studies in phonemic segmentation are based on forced Viterbi phoneme recognition using hidden Markov models (HMMs). Pellom and Hansen investigate various segmentation, speech enhancement, and parameter compensation techniques in noisy environments using the TIMIT dataset, which is degraded by additive colored noise [3]. They propose a linguistically constrained HMM-based method, which yields over 85% boundary detection rate in noise-free environments (with 20-ms boundary misalignment tolerance), while achieving significant improvement in noisy environments, such as aircraft cockpits, automobile highways, etc. In [18], a two-step HMM-based approach is proposed, where a well-trained context dependent boundary model is adapted using a maximum *a posteriori* approach for segment boundary refinement. The segmentation accuracy exceeds 90% within a 20-ms tolerance in Mandarin Chinese and English in the Microsoft TTS speech corpora. An overview of machine learning techniques exploited for phone segmentation using the TALP Research Center corpus is done by Adell and Bonafonte [19]. The assessment of HMMs, artificial neural networks (ANNs), dynamic time warping, Gaussian mixture models (GMMs), and pronunciation modeling, indicates that 85%–90% detection accuracy can be achieved when training data are available at a 20-ms tolerance. Toledano and Gomez [20] use a modified HMM recognizer and propose statistical correction to compensate for the systematic errors produced by context-dependent HMMs. The algorithm is evaluated using the percentage of boundaries with errors smaller than 20 ms as a figure of merit and attest that over 90% accuracy is possible. Hosom [21] has proposed a hybrid HMM/ANN phoneme alignment method where distinctive phonetic features and transition-dependent observation probabilities are employed. The algorithm yields 92.57% accuracy within 20 ms on the TIMIT corpus. While HMM-based approaches yield over 90% phone boundary detection accuracy, they require training data and an orthographic transcription as well as precise modeling of the pronunciation variants in order to estimate observation sequence probabilities [4].

B. Linguistically Unconstrained Phonemic Segmentation

Linguistically unconstrained phone segmentation uses a bottom-up strategy that does not depend on phonetic transcription neither requires training data. Unlike generative approaches based on HMMs, methods using spectral distortion measures are model-free and thus computationally inexpensive and text-, speaker-, dialect-, and language-independent, although they yield worse accuracy. Thus, they are suitable for multilingual applications and online implementations that realize near real-time processing and low bit rate speech coding. Some phone boundaries are instantaneous (e.g., the burst of a fully closed plosive), others are not. Though instantaneous boundaries are not ubiquitous, their existence allows us to consider spectral changes as potential transition points that correspond to phone boundaries.

1) Nonparametric Techniques: Many algorithms in this class define a change function that directly measures the spectral variation of the acoustic signal and utilize this function as a transition penalty. Mitchell *et al.* [22] have proposed the Delta Cepstral Function (DCF), which estimates spectral change by summing the normalized time derivative of each cepstral coefficient, $\mathrm{DCF}_q(t) = C_q(t+1) - C_q(t-1), q = 1, \ldots, Q$, where $C_q(t)$ is the qth cepstral coefficient at the frame t and Q is the number of cepstral coefficients. $\mathrm{DCF}_q(t)$ is then used to compute a cost function $c(t)$ that detects spectral changes associated with phoneme transitions

$$c(t)_{\mathrm{DCF}} = \frac{\sum_{q=1}^{Q}(\mathrm{DCF}_q(t)/\max_t|\mathrm{DCF}_q(t)|)}{\max_t\sum_{q=1}^{Q}(\mathrm{DCF}_q(t)/\max_{t'}|\mathrm{DCF}_q(t')|)}. \tag{1}$$

Brugnara *et al.* [23] and Mitchell *et al.* [22] have used the Spectral Variation Function (SVF), which estimates spectral change as the angle between two normalized cepstral vectors that are separated by a fixed number of frames in time. The change function is

$$c(t)_{\mathrm{SVF}} = \frac{1}{2}\left(1 - \mathrm{SVF}(t)/\max_t|\mathrm{SVF}(t)|\right) \tag{2}$$

where

$$\mathrm{SVF}(t) = [\hat{\boldsymbol{C}}(t-1)]^T\hat{\boldsymbol{C}}(t+1)/(\|\hat{\boldsymbol{C}}(t-1)\|\cdot\|\hat{\boldsymbol{C}}(t+1)\|).$$

$\hat{\boldsymbol{C}}(t)$ is the difference between the tth cepstral vector and the time average of the cepstral vectors that lie within a window centred at t, and $\|\cdot\|$ indicates the vector norm. Since $c(t)_{\mathrm{DCF}}$ and $c(t)_{\mathrm{SVF}}$ are derived from the observations, little signal processing overhead is required. Other cost-effective nonparametric methods are zero crossing measurement, the convex-hull method, and the normal decomposition method [24].

2) Parametric Techniques: Distance-based parametric methods for segmentation use the KL divergence, the generalized likelihood ratio, and similarity measures based on the eigenanalysis of the sample covariance matrix [10]. Assuming feature vectors, which follow a known probability density function (pdf), e.g., the multivariate Gaussian distribution (GD), statistical distances are measured between adjacent windows and a distance plot is formed. Then, heuristics are applied in order to identify significant local peaks that presumably indicate spectral changes.

The identification of potential transition points between speech segments can be addressed as a statistical hypothesis testing problem. Instead of considering spectral changes, it is assumed that potential segmentation points correspond to sequential model changes [4]. This approach has been used for speech and speaker segmentation (cf., [5], [9]–[13]), but it can also be applied to phonemic segmentation, if properly modified.

More specifically, an acoustic change detection system based on BIC has been proposed in [9]. The sequence of feature vectors, typically Mel-frequency cepstral coefficients (MFCCs), in adjacent speech segments are modeled 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 a single GD better, or whether a two-segment representation describes it better.

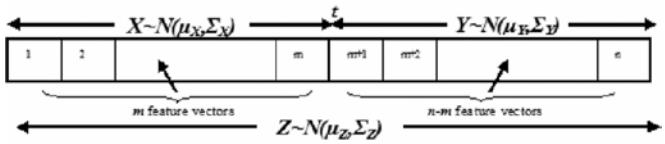

Fig. 1. Models for two adjacent speech segments.

A sliding window Z having $N_Z = n$ observations, moves along the signal making statistical decisions at time instant t. Let its sub-windows X and Y have $N_X = m$ and $N_Y = n-m$ observations, respectively, with $N_Z = N_X + N_Y = n$ as depicted in Fig. 1. Assume that the feature vectors follow a known pdf (e.g., multivariate GD). The offset of the sliding window, which is typically fixed, indicates the resolution of the system. For the purpose of phonemic segmentation, we must 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 Z (i.e., noisy speech/silence, the same phone) 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 X and Y, implying that there is a transition between two different phones or a transition from speech utterance to silence and vice versa. We denote that the data come from model M_1.

In Bayesian model selection, the comparison of two competing models M_0 and M_1, given $\boldsymbol{x}$ involves choosing the model with the higher posterior probability. The posterior odds ratio is

$$\frac{p(M_0\,|\,\boldsymbol{x})}{p(M_1\,|\,\boldsymbol{x})} = \frac{p(\boldsymbol{x}\,|\,M_0)}{p(\boldsymbol{x}\,|\,M_1)}\frac{P(M_0)}{P(M_1)} = \mathrm{BF}\times\text{ prior odds}$$
$$\mathrm{BF} = \frac{\int_{\Theta_0} p(\boldsymbol{x}|\boldsymbol{\theta}_0, M_0)p(\boldsymbol{\theta}_0|M_0)d\boldsymbol{\theta}_0}{\int_{\Theta_1} p(\boldsymbol{x}|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1} \tag{3}$$

where BF is the ratio of marginal likelihoods of the two competing models whose parameters are $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ and $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$, respectively. The exact calculation of marginal likelihoods in (3) is difficult to compute in closed form and requires numerical integration, such as Gaussian integration, Gibbs sampling, or the Laplace approximation. Schwarz assumed that no intrinsic linear structure existed in the parameter space [8]. Considering that the observations follow an exponential distribution, BIC results as an easily calculated and asymptotically optimal method for estimating the best model using only MLE of the parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. It presents a useful, and easily calculated, long-sample approximation to BF, assuming flat priors. If K is the number of the estimated free parameters, n is the sample size and $l(\hat{\boldsymbol{\theta}})$ is the likelihood function at the MLE, BIC is defined as $\mathrm{BIC} = -2\ln l(\hat{\boldsymbol{\theta}}) + K\ln(n)$.

Let $\boldsymbol{x}_i$ be Q-dimensional feature vectors, $\boldsymbol{\Sigma}_Z, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y$ be the full covariance matrices of the full window Z and the two

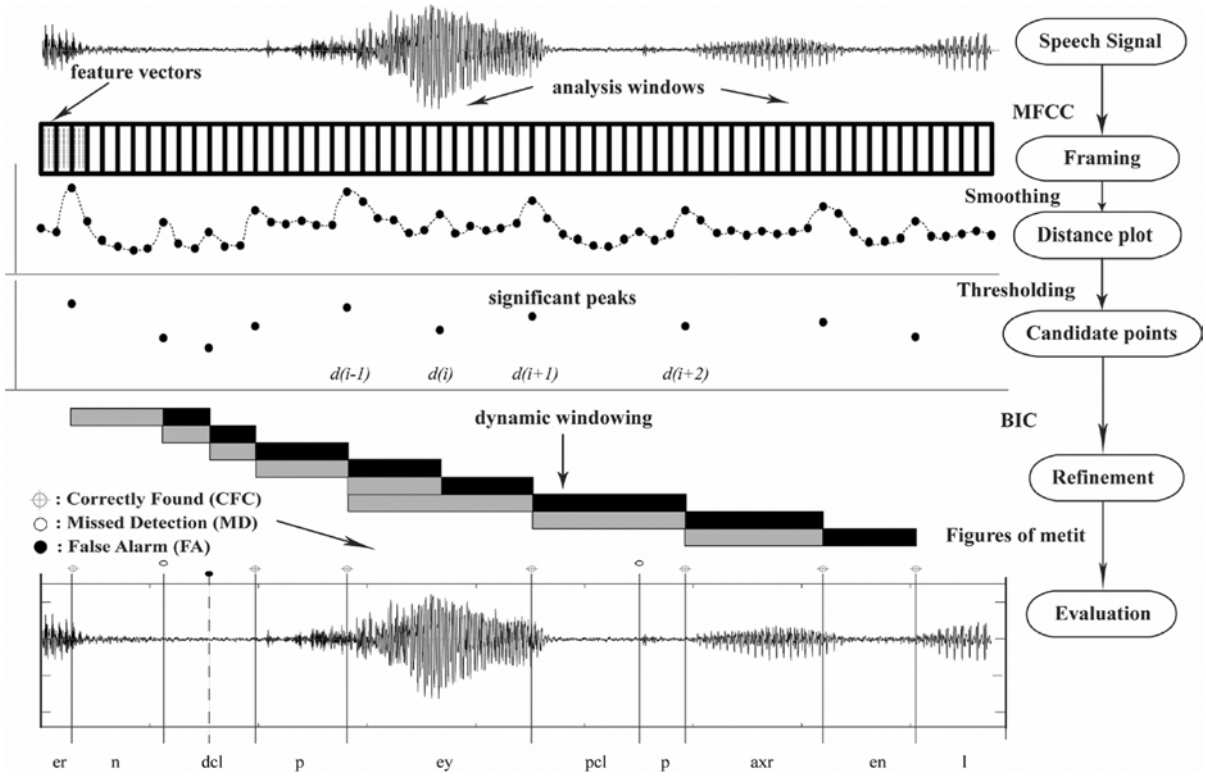

Fig. 2. Block diagram of DISTBIC phonemic segmentation algorithm.

sub-windows X and Y, respectively, and $\boldsymbol{\mu}_Z, \boldsymbol{\mu}_X, \boldsymbol{\mu}_Y$ be the corresponding mean vectors. Since $\boldsymbol{x}_i \in R^Q$, the parameter space of the multivariate Gaussian model M_0 is given by the composite vector $\boldsymbol{\Theta}_Z$ with $\dim(\boldsymbol{\Theta}_Z) \equiv K = Q + Q(Q+1)/2$ free parameters (Q for $\boldsymbol{\mu}_Y$ and another $Q(Q+1)/2$ for $\boldsymbol{\Sigma}_Z$), while the mixture model M_1 has $2K$ parameters. The variation of the BIC value between the two models, $\Delta\mathrm{BIC}$, determines whether the multivariate GD mixture $X \cup Y$ best fits the data, indicating that t corresponds to a segment boundary. If MLE is used, then

$$\begin{aligned}\Delta\mathrm{BIC}(M_1, M_0) &= \mathrm{BIC}(M_1) - \mathrm{BIC}(M_0) \\ &= -2\left(\sum_{i=1}^{N_X} \ln p(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X) + \sum_{i=N_X+1}^{N_Z} \ln p(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)\right) \\ &\quad + 2\left(\sum_{i=1}^{N_Z} \ln p(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)\right) + \left(Q + \frac{Q(Q+1)}{2}\right) \ln N_Z. \end{aligned} \tag{4}$$

Negative values of $\Delta\mathrm{BIC}$ indicate that there is a transition point in t, i.e., a phone boundary. In the next subsection, we focus on a hybrid parametric technique that combines distance- and model-based segmentation, which is taken as baseline for our developments.

C. Hybrid Distance and Model-Based Segmentation Using DISTBIC

DISTBIC is a two-pass segmentation algorithm that searches for change point candidates at the maxima of distances computed between adjacent windows over the entire signal [10]. DISTBIC is a hybrid method, in that it combines distance metrics with MSC. A block diagram depicting the processing stages of the DISTBIC segmentation algorithm is shown in Fig. 2. First, DISTBIC uses a distance computation between adjacent windows with fixed width, which slide across the signal using a shift value less than the fixed width, in order to determine possible candidates for a change point. Different criteria such as the KL divergence, the generalized likelihood ratio, the Bhattacharyya distance, the $\Delta\mathrm{BIC}$ values, and various second-order statistical measures, that evaluate the sphericity of the matrix $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_Y^{-1}$ or its deviation from the identity matrix using the arithmetic and geometric mean of the eigenvalues of $\boldsymbol{\Gamma}$, can be applied to this pre-segmentation step [10], [15]. In this paper, the symmetric version of KL divergence, denoted as $\mathrm{KL2}(X, Y)$ is used. Assuming multivariate GDs for the feature vectors, this distance can be estimated from the sample statistics

$$\begin{aligned}\mathrm{KL2}(X, Y) &= \mathrm{KL}(X, Y) + \mathrm{KL}(Y, X) \\ &= \frac{1}{2}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}_X)^T \left(\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_Y^{-1}\right)(\boldsymbol{\mu}_Y - \boldsymbol{\mu}_X) \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\Sigma}_Y^{-1/2}\right)\left(\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\Sigma}_Y^{-1/2}\right)^T\right) \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{\Sigma}_Y^{1/2}\right)\left(\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{\Sigma}_Y^{1/2}\right)^T\right) - Q. \end{aligned} \tag{5}$$

Next, a plot of distances is created and significant local peaks are selected as candidate change points to filter out the insignificantly small distance values. Peaks are selected

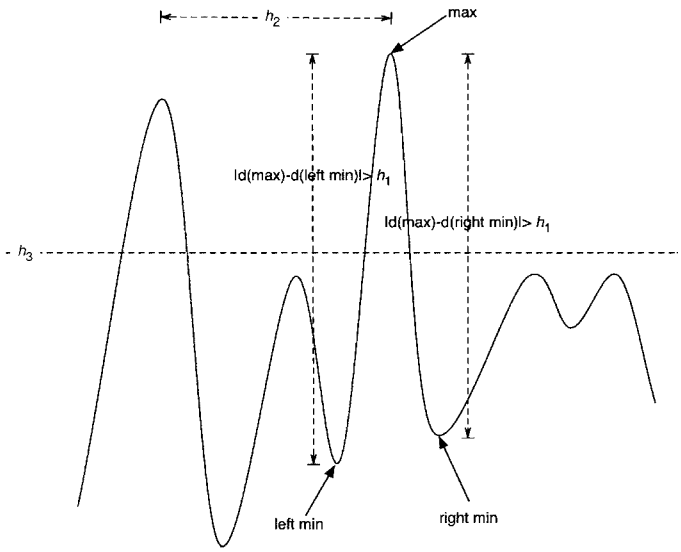

Fig. 3. Thresholds of DISTBIC algorithm.

as "significant" using a series of thresholds, as depicted in Fig. 3. In particular, h_2 constrains the minimum time distance between consecutive local maxima corresponding to candidate segmentation points, while h_2 and h_3 threshold the relative and absolute distance peaks, respectively. For example, a peak is marked as significant, if it is larger than its left and right local minima by threshold h_1, i.e., $|d(\max) - d(\text{left min})| > h_1$ and $|d(\max) - d(\text{right min})| > h_1$, respectively. Delacourt proposes the value of $h_1 = a\sigma$, where σ corresponds to the standard deviation of the distances along the plot. Threshold h_2 smooths out the plot: two local maximum points are not marked both as significant unless they occur at least h_2 ms apart. When two local maxima are close, they can be replaced by a virtual peak at the center of their time distance [13].

In the last step, a second window scheme with tunable width is used, where ΔBIC values validate or discard the candidates determined in the first step. A sliding window containing two sub-windows X, Y, is modeled according to the hypothesis test in Fig. 1. The boundaries of the adjacent sub-windows are determined by three consecutive (significant) candidate points $d(i-1), d(i)$, and $d(i+1)$ in the distance plot, where i is the index of the candidate points, that were chosen in the first step, i.e., $[t_{d(i-1)}, t_{d(i)}]$ and $[t_{d(i)}, t_{d(i+1)}]$, as shown in Fig. 2. If, according to BIC, the point $d(i)$, which occurs at the time instant $t_{d(i)}$, is discarded, then, in the next step of the analysis, the two sub-windows of the hypothesis test are set $[t_{d(i-1)}, t_{d(i+1)}]$ and $[t_{d(i+1)}, t_{d(i+2)}]$. This refinement scheme 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 more observations to the decision rule benefits the speech/pause discrimination and phone transition detection [25]. A free parameter winmax constrains the maximum window length. For example, $[t_{d(i-1)}, t_{d(i)}]$ becomes $[t_{d(i)} - \text{winmax}, t_{d(i)}]$, if $t_{d(i)} - t_{d(i-1)} > \text{winmax}$.

The last step of DISTBIC can be iterated and serves as a refinement step in order to avoid over-segmentation. While over-segmentation in speaker segmentation is usually compensated by speaker clustering in a following step, this is not the case in phoneme segmentation. Therefore, the efficient concatenation of homogenous speech segments in a short time scale, according to refinement capability of BIC, is particularly important to the identification of phone boundaries.

Similarly to the AIC generalization proposed in [26], a tuning parameter λ for the penalty term of BIC can be used, but λ must be estimated heuristically from data (cf., [10], [12], [27], [25]). If the value of λ is too high, the algorithm avoids many false alarms, but at the cost of ignoring genuine segmentation points. If it is too low, the number of missed detections is reduced at the cost of increasing the number of false alarms. Ajmera *et al.* [28] have proposed a method that avoids λ by modeling the data in the sliding window Z with a two-component GMM and estimating its parameters using the expectation maximization algorithm. DISTBIC is efficient in detecting acoustic changes that are relatively close to one another, but at the price of many false alarms. Nevertheless, by tuning the parameters of the algorithm it is possible to fix the over-segmentation (false alarms) to a minimum value and then try to maximize the detection rate.

Regarding the application of BIC to the detection of phone boundaries, certain assumptions must be considered. First, either the incoming signal must originate from a single speaker or no multiple speakers talk simultaneously. Second, since the phone durations are relatively small, we must operate on very small window sizes in order to be able to assume that the test windows correspond to homogenous segments or at most single transitions. The choice of the window length is a compromise between having enough data to calculate the feature vector statistics and limiting the influence of surrounding parts of the recording. The window shift determines the time resolution for the boundaries. For an accurate segmentation, this value must be as small as possible.

III. Phonemic Segmentation and Robustness

Commonly a set of assumptions is embraced in order to reduce the complexity of speech modeling and analysis. For example, it is often assumed that: the features have sufficient discriminative power; the underlying data distribution is Gaussian; the feature vector elements are uncorrelated; the feature vectors are independent and identically distributed (i.i.d.); the sample size is sufficiently large; and the statistical analysis is robust to noise and outliers. Whenever such assumptions do not hold, erroneous inference results. Clearly, the assumptions imply constraints that appear to be interrelated. For example, using multivariate features, like MFCCs, implies extra complexity and potential imprecision unless the sample size is large and/or independent variables are assumed.

When the features are obtained by sampling the short-time Fourier spectrum nearby frequencies within the same observation frame are most probably *highly correlated*. Thus, when filter bank features are employed, using conditional pdfs with diagonal covariance matrices is inappropriate. On the contrary, cepstral coefficients eliminate some of the correlation between coefficients extracted from a single observation frame, fitting the data more closely to the variable uncorrelatedness assumption. This allows the use of diagonal covariance matrices and

marginal likelihoods in (3). Therefore, closed-form analytical expressions of the integrated likelihoods in MSC may be possible. In practice, during phonemic segmentation, small sample sizes are involved, which corresponds to small number of observation frames in the short-time analysis windows, thus multicollinearity might be present. Furthermore, MFCC features are sensitive to distortion with additive white noise, which means that many multivariate statistical techniques are inadequate to deal with the inserted outliers [29]. The diagonal covariance assumption is also violated when heterogeneous features are merged, e.g., when combining MFCC features with their corresponding first differences Δ MFCCs. Instead of assuming variables modeled by univariate distributions, it would be more accurate to consider joint multivariate pdfs that capture the stochastic dependence between features. Consequently, MSC that take into account the functional form, i.e., the way the model parameters are interrelated in the pdf and in the estimation of model complexity, would be more suitable than simple criteria such as BIC and AIC [16].

Dealing with situations where *few data* (i.e., feature vectors) are available is a challenging problem in statistical inference that is often neglected for the sake of theoretical flexibility and computational simplicity. A sufficiently large sample size avoids the singularity of covariance matrices and allows asymptotic approximations to be applied to MSC, but in phonemic segmentation, where the duration of a single phone can be as small as a few milliseconds, it is important to consider small-sample approximations. MLE is biased for small-samples and consequently AIC and BIC, which resort to MLE, also suffer due to insufficient data. Overfitting effects of complex models become more dramatic as sample size decreases, thus alternative corrected versions must be used. Small sample size also has a critical effect in MSC employing a functional form, such as ICOMP, because the analytical equation for the asymptotic covariance matrices of the model parameters has to be replaced by unstable finite sample empirical estimators and the expectations have to be replaced with sample averages. Data limitation and model misspecification lead to singular asymptotic covariance matrices, and thus to noninvertible Hessian matrices, restraining the use of FIM-based MSC. In such cases, generalized inversion procedures based on Cholesky decomposition or quadratic approximation can be used. A feasible alternative is to estimate the covariance matrices by parametric bootstrap [30] or Monte Carlo methods [31].

IV. Robust Phonemic Segmentation With Model Selection

Past work on the application of MSC to speech segmentation has ignored many of the constraints discussed in Section III. In order to maintain robustness against outliers, the baseline algorithm DISTBIC (described in Section II-C) is extended, in this Section, by suggesting alternative MSC to BIC and incorporating robust statistics for the estimation of model parameters.

A. Alternative Model Selection Criteria

AIC was derived as a large sample approximation of the expected KL divergence between the pdf of the fitted model and that of the true model, with the expectation taken over all possible observations under the true pdf. In particular, AIC estimates the asymptotic bias between the average (over a set of candidate models) of the maximized log-likelihood and the expected one by twice the number of the free model parameters K, i.e., $\mathrm{AIC} = -2\ln l(\hat{\boldsymbol{\theta}}) + 2K$, where $l(\hat{\boldsymbol{\theta}})$ is the maximized likelihood function under a model. When MLE parameter estimators are used, the first term in AIC reflects the goodness of fit (GoF) of the model and measures the bias for model inaccuracy. The second term emerges from the parsimony principle and acts as a penalty for the increased unreliability of the fit bias when additional free parameters are included in the model.

Under the condition that the specified parametric family of pdfs contains the true distribution, which also implies large sample sizes, AIC provides an unbiased estimate of the KL divergence between the specified model and the true model when the parametric model is estimated by the method of ML. When the sample size is large and the dimension of candidate model is relatively small, AIC is an approximately unbiased estimator. If these conditions are not met, AIC introduces a large negative bias. According to [32], AIC tends to overestimate the parameters needed, even asymptotically, thus it offers a crude estimator of the expected discrepancy between the model generating the data and a fitted candidate model. A second-order small-sample corrected AIC (AICC) assuming that the data are generated by a fixed-effect linear model with homogenous, normally distributed errors has been proposed in [32], [33]. AICC is defined as

$$\begin{aligned}\mathrm{AICC} &= -2\ln l(\hat{\boldsymbol{\theta}}) + 2Kn/(n-K-1) \\ &= \mathrm{AIC} + 2K(K+1)/(n-K-1)\end{aligned}$$

where n is the sample size. When there are too many parameters in relation to the size of the sample (i.e., $n/K < 40$), AICC estimates the expected discrepancy with less bias than AIC. As sample size increases, AICC converges to AIC. Similarly to AIC, AICC is unbiased under the same conditions. On the other hand, AICC's justification depends upon the form of the candidate model, while AIC is more universally applicable [32].

Although 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 to domains containing small number of samples requires caution. The concern for small or moderate sample sizes is that BIC overvalues parsimonious models imposing a rather heavy penalty to model complexity and thus the BIC-selected model may be underfit and inappropriate for inference. A BIC corrected for small samples is BICC [34], which uses a complexity penalty inspired by AICC, i.e., $\mathrm{BICC} = -2\ln l(\hat{\boldsymbol{\theta}}) + Kn\ln(n)/(n-K-1)$. BICC performs better than classic BIC both in terms of mean squared error of the parameter estimates and the prediction error.

Apart from sample size, there is another independent factor which contributes to model complexity; the functional form. This refers to the way in which the model parameters are

related in the model pdf. Bozdogan proposed a generalization to information-based covariance complexity index by introducing the maximal information complexity of the asymptotic of the model parameters, $\boldsymbol{\Sigma}_\theta$ [35]. Bozdogan's information complexity criterion (ICOMP) is an entropic measure of statistical dependence between the parameter estimates and it is invariant under multiplication and orthonormal transformations [16], [35]. Instead of penalizing the number of free parameters directly like AIC, ICOMP also measures and penalizes the interdependence between the parameter estimates by estimating the covariance complexity of the model

$$\begin{aligned} \mathrm{ICOMP}(C_1(\hat{\boldsymbol{\Sigma}}_\theta)) &= -2\ln l(\hat{\boldsymbol{\theta}}) + 2C_1(\hat{\boldsymbol{\Sigma}}_\theta) \\ C_1(\cdot) &= \frac{K}{2}\ln\left(\frac{\mathrm{tr}(\cdot)}{K}\right) - \frac{1}{2}\ln|\cdot| \end{aligned} \tag{6}$$

where K is the number of free model parameters corresponding to the rank of the MLE of the expected asymptotic covariance matrix $\hat{\boldsymbol{\Sigma}}_\theta$ and $|\cdot|$ is the determinant of a square matrix. The maximal information complexity $C_1(\hat{\boldsymbol{\Sigma}}_\theta)$ is an information theoretic measure of complexity of the covariance matrix that reflects the Cramer–Rao lower bound of the model and gives a scalar combined measure of parameter redundancy and stability. The trace and the determinant in (6) are the arithmetic and geometric mean of the eigenvalues of $\hat{\boldsymbol{\Sigma}}_\theta$. Therefore, $C_1(\hat{\boldsymbol{\Sigma}}_\theta)$ penalizes ellipsoidal dispersion. Consequently, ICOMP offers a judicious balance between GoF, model complexity, and accuracy of the parameter estimates. As the sample size increases, the sensitivity of functional form gradually diminishes compared to the number of free parameters K. A refined version, ICOMP(IFIM) [16], uses the IFIM $\boldsymbol{F}^{-1}$ instead of $\hat{\boldsymbol{\Sigma}}_\theta$, where $\boldsymbol{F}$ is the FIM, i.e.,

$$\begin{aligned} \mathrm{ICOMP(IFIM)} &= -2\ln l(\hat{\boldsymbol{\theta}}) + 2C_1(\boldsymbol{F}^{-1}(\hat{\boldsymbol{\theta}})) \\ &= -2\ln l(\hat{\boldsymbol{\theta}}) + \frac{K}{2}\ln\left(\frac{\mathrm{tr}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\theta}})}{K}\right) \\ &\quad - \frac{1}{2}\ln|\boldsymbol{F}^{-1}(\hat{\boldsymbol{\theta}})|. \end{aligned} \tag{7}$$

ICOMP(IFIM) evaluates model complexity from the correlation structure of the parameter estimates through the IFIM. The diagonal elements of IFIM correspond to the estimated variances of the model parameters and indicate the parameter sensitivity, while the off-diagonal elements are the covariances between the parameters, which measure the degree of multicollinearity among the columns of IFIM and reveal the extent of parameter interdependence. It is clear that orthogonal models or linear models with no collinearity minimize (7). When measuring the model complexity with ICOMP, it is required to approximate the FIM. The estimated IFIM and its trace and determinant can be derived in a closed form for multivariate regression [16]. Practically, the asymptotic covariance matrix still has to be estimated. Therefore, the analytical calculation of ICOMP depends on the parameterization of the model and can be difficult or impossible for nonlinear models. ICOMP(IFIM) has been attested to provide better results than BIC and AIC in various settings [36].

B. Robust Statistics in Model Selection

The presence of outliers in the data is a common reason for lack of fit. Possible sources of outliers in speech processing are recording and measurement errors, extreme random effects and non-Gaussian noise, correlated observations, and an unknown data structure. Incorrect assumption about the data distribution can also lead to mislabeling data as outliers. Many statistical techniques are sensitive to the presence of outliers, because that has a strong influence on the estimation of the mean and the covariance. Robust techniques can reduce the effect of outliers without deleting them. M-estimators define a large class of robust estimators that includes the MLE as a subclass. M-estimation filters measurement noise efficiently, while its robustness properties bound the influence of outliers without totally rejecting them [37]. An efficient and robust location and scatter multivariate estimator with a high breakdown point is the minimum covariance determinant estimator (MCD). For a set of n Q-variate observations $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the MCD location and scatter estimates are given by $\hat{\boldsymbol{\mu}}_{\mathrm{MCD}} = \sum_{\iota\in J^*} t_i \boldsymbol{x}_i$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{MCD}} = \sum_{i\in J^*} t_i(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}})^{\mathrm{T}}$, respectively, where $t_i = 1/r, r \simeq n/2$, and J^* is the r-element set, where the determinant of $\hat{\boldsymbol{\Sigma}}_{\mathrm{MCD}}$ is minimized. Fast-MCD is a three-step MCD implementation that improves computational speed significantly [38].

Besides the robust estimates of the mean vectors and covariance matrices, robust versions of AIC and BIC criteria have also been investigated in the literature. Ronchetti [39] proposed an asymptotic unbiased criterion that is a robust extension of AIC. This criterion is an application of the generalization of MLE to M-estimation by following Huber's least favorable distribution. Hampel suggested a similar IC using a different penalty term based on heuristic arguments [37]. Shi and Tsai [40] derived a robust version of AICC, that also accommodates non-normally distributed errors. Machado [41] derived a robust version of BIC based on objective functions defining M-estimators for a parametric model. This simply replaces MLE with a robust estimate, while keeping the same penalty with BIC. Qian and Kunsch [42] presented a penalty term with more comprehensive information about the model, based on stochastic complexity.

C. Modification of DISTBIC Using Alternative Model Selection Criteria

BIC is inappropriate for phoneme segmentation, because of small-sample biases discussed in Section III. In this paper, we propose modified algorithms that use AICC, BICC, or ICOMP as alternatives to BIC in the candidate point verification step of the DISTBIC algorithm. The resulting algorithms are referred to as DISTAICC, DISTBICC, and DISTICOMP.

DISTAICC: AIC, in its original form, is not a reasonable choice for our hypothesis test, because the ad hoc penalty term, it imposes, does not depend explicitly on the data and the sample size. Since both frames and their concatenation are always modeled by GD, AIC will reduce to a simple GoF measure. Instead, a refined version, such as AICC, should be used since it takes into account the frame size and moreover works better in small-sample problems.

DISTBICC: Like AICC, BICC is a small-sample corrected version of a promising divergence estimator, therefore BICC is a second candidate for improved model selection.

DISTICOMP: ICOMP(IFIM) considers the interdependencies among the variables and both the linearity and nonlinearity of the model parameters. This makes ICOMP(IFIM) appealing, since it can distinguish among equivalent models and control the risk of underfitting and overfitting phenomena judiciously. We denote ICOMP(IFIM) by ICOMP from now on for simplicity.

Instead of estimating IFIM using the equivalent multivariate regression setting [16], which requires the calculation of the asymptotic covariance matrix, we approximate the estimated IFIM using a Monte Carlo resampling-based method [31]. This technique produces a number of efficient, almost unbiased, estimates of the Hessian matrix of the log likelihood using a Monte Carlo approach analogous to a bootstrap resampling scheme and then averages the negative of these estimates to obtain an approximation to FIM. For a sequence of n Q-dimensional feature vectors $\boldsymbol{X}$, we form S i.i.d., pseudovectors $\{\boldsymbol{z}_{\text{pseudo}}(1),\ldots,\boldsymbol{z}_{\text{pseudo}}(S)\}$ that are randomly generated from the pdf $p(\boldsymbol{X}\,|\,\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the pdf parameter vector estimated from the data. For every pseudovector $\boldsymbol{z}_{\text{pseudo}}(s), s=1,\ldots,S$ we calculate $M\geq 1$ estimates of the Hessian matrix, $\hat{\boldsymbol{H}}_m^{(s)}, m=1,\ldots,M$, and by taking their average, we calculate the mean $\hat{\boldsymbol{H}}^{(s)}$, which corresponds to the estimated FIM. Assuming a small real number $c>0$, we form a perturbation vector $\boldsymbol{u}_m=[u_{m1},\ldots,u_{mK}]^T$ that is randomly generated and is independent of $\boldsymbol{Z}_{\text{pseudo}}(s)$. The sth estimation of the gradient vector of the log-likelihood $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{X})=\partial \ln l(\boldsymbol{\theta}|\boldsymbol{X})/\partial\boldsymbol{\theta}$, can be calculated using the one-sided simultaneous perturbation

$$\begin{aligned}\hat{\boldsymbol{g}}&(\boldsymbol{\theta}\pm c\boldsymbol{u}_m\,|\,\boldsymbol{Z}_{\text{pseudo}}(s))\\&=\frac{1}{\tilde{c}}[\ln l(\boldsymbol{\theta}+\tilde{c}\tilde{\boldsymbol{u}}_m\pm c\boldsymbol{u}_m\,|\,\boldsymbol{Z}_{\text{pseudo}}(s))\\&\quad-\ln l(\boldsymbol{\theta}\pm c\boldsymbol{u}_m\,|\,\boldsymbol{Z}_{\text{pseudo}}(s))]\\&\quad\times\left[\tilde{u}_{m1}^{-1},\ldots,\tilde{u}_{mK}^{-1}\right]^T\end{aligned}\tag{8}$$

where $\tilde{c}>0$ is a small number (typically $\tilde{c}>c$), the random vector $\tilde{\boldsymbol{u}}_m=[\tilde{u}_{m1},\ldots,\tilde{u}_{mK}]^T$ is independent of $\boldsymbol{u}_m$ and $\boldsymbol{Z}_{\text{pseudo}}(s)$, and the random variables u_{sk} and $\tilde{u}_{sk}, s=1,\ldots,S, k=1,\ldots,K$ are zero mean, i.i.d., and identically bounded from the symmetric Bernoulli distribution [31]. The Hessian, which corresponds to the FIM, is the average of the M Hessians $\hat{\boldsymbol{H}}_m$ estimates

$$\begin{aligned}\hat{\boldsymbol{H}}_m^{(s)}=\frac{1}{2}\Bigg\{&\frac{\delta\hat{\boldsymbol{g}}_m^{(s)}}{2c}\left[u_{m1}^{-1},\ldots,u_{mK}^{-1}\right]\\&+\left(\frac{\delta\hat{\boldsymbol{g}}_m^{(s)}}{2c}\left[u_{m1}^{-1},\ldots,u_{mK}^{-1}\right]\right)^T\Bigg\}\end{aligned}\tag{9}$$

$$\begin{aligned}\delta\hat{\boldsymbol{g}}_m^{(s)}\equiv&\,\hat{\boldsymbol{g}}(\boldsymbol{\theta}+c\boldsymbol{u}_m\,|\,\boldsymbol{Z}_{\text{pseudo}}(s))\\&-\hat{\boldsymbol{g}}(\boldsymbol{\theta}-c\boldsymbol{u}_m\,|\,\boldsymbol{Z}_{\text{pseudo}}(s)).\end{aligned}\tag{10}$$

D. Modification of DISTBIC Using Robust Parameter Estimates

In order to deal with outliers and improve the detection accuracy of the baseline algorithm, we also examine robust model parameter estimation. Kotti *et al.* [11] have proposed an equivalent formulation of BIC, when the covariance matrix estimators are not limited to sample dispersion matrices (i.e., MLE). Assuming observations drawn from multivariate GDs, first they apply centering of data (around $\boldsymbol{\mu}_Z$) and then simultaneous diagonalization of covariance matrices. In the equivalent BIC formulation, robust covariance matrix estimates can be used. For the data $\boldsymbol{x}_i\in X\cap Z=X$ we make the transformations $\tilde{\boldsymbol{x}}_i=\boldsymbol{x}_i-\boldsymbol{\mu}_Z, \boldsymbol{\mu}_X^{'}=1/N_X\sum_{i=1}^{N_X}\boldsymbol{x}_i-\boldsymbol{\mu}_Z=1/N_X\sum_{i=1}^{N_X}\tilde{\boldsymbol{x}}_i$, while for $\boldsymbol{x}_i\in Y\cap Z=Y$ the transformations $\tilde{\boldsymbol{x}}_i=\boldsymbol{x}_i-\boldsymbol{\mu}_Z, \boldsymbol{\mu}_Y^{'}=1/N_X\sum_{i=N_X+1}^{N_Z}\boldsymbol{x}_i-\boldsymbol{\mu}_Z=1/N_Y\sum_{i=N_X+1}^{N_Z}\tilde{\boldsymbol{x}}_i$. Then we apply the simultaneous diagonalization transformation, i.e., $\tilde{\boldsymbol{w}}_i=\boldsymbol{\Psi}^T\boldsymbol{\Lambda}_Z^{-1/2}\boldsymbol{\Phi}^T\tilde{\boldsymbol{x}}_i=\boldsymbol{W}^T\tilde{\boldsymbol{x}}_i$ for $\boldsymbol{x}_i\in X\cap Z=X$ and $\tilde{\boldsymbol{u}}_i=\boldsymbol{\Xi}^T\boldsymbol{\Lambda}_Z^{-1/2}\boldsymbol{\Phi}^T\tilde{\boldsymbol{x}}_i=\boldsymbol{\Omega}^T\tilde{\boldsymbol{x}}_i$ for $\boldsymbol{x}_i\in Y\cap Z=Y$, respectively, where $\boldsymbol{\Lambda}_z$ is the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_Z, \boldsymbol{\Phi}$ is its modal matrix, $\boldsymbol{\Psi}$ is the modal matrix of $\boldsymbol{K}=\boldsymbol{\Lambda}_Z^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_X\boldsymbol{\Phi}\boldsymbol{\Lambda}_Z^{-1/2}$, and $\boldsymbol{\Xi}$ is the modal matrix of $\boldsymbol{H}=\boldsymbol{\Lambda}_Z^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_Y\boldsymbol{\Phi}\boldsymbol{\Lambda}_Z^{-1/2}$. If we note by $\boldsymbol{\Lambda}_K$ the diagonal matrix of the eigenvalues of $\boldsymbol{K}=\boldsymbol{\Lambda}_Z^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_X\boldsymbol{\Phi}\boldsymbol{\Lambda}_Z^{-1/2}$ and $\boldsymbol{\Lambda}_H$ is the diagonal matrix of the eigenvalues of $\boldsymbol{H}=\boldsymbol{\Lambda}_Z^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_Y\boldsymbol{\Phi}\boldsymbol{\Lambda}_Z^{-1/2}$, (4) becomes after analytical computations

$$\begin{aligned}\Delta\text{BIC}=&\sum_{i=1}^{N_X}\tilde{\boldsymbol{w}}_i^T\left(\boldsymbol{\Lambda}_K^{-1}-\boldsymbol{I}\right)\tilde{\boldsymbol{w}}_i+\sum_{i=N_X+1}^{N_Z}\tilde{\boldsymbol{u}}_i^T\left(\boldsymbol{\Lambda}_H^{-1}-\boldsymbol{I}\right)\tilde{\boldsymbol{u}}_i\\&-\gamma_{\text{BIC}}-N_X\boldsymbol{\mu}_X^{\prime T}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X^{\prime}-N_Y\boldsymbol{\mu}_Y^{\prime T}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\mu}_Y^{\prime}\end{aligned}\tag{11}$$

$$\begin{aligned}\gamma_{\text{BIC}}=&N_Z\ln|\boldsymbol{\Sigma}_Z|-N_X\ln|\boldsymbol{\Sigma}_X|-N_Y\ln|\boldsymbol{\Sigma}_Y|\\&+\left(Q+\frac{Q(Q+1)}{2}\right)\ln N_Z.\end{aligned}\tag{12}$$

In addition, we refine the penalty term as in BICC and apply Fast-MCD estimators for measuring GoF [38]. We denote the adjusted criterion BICCR. Similarly, we use Fast-MCD estimators to AICC and ICOMP(IFIM) and denote the new criteria as AICCR and ICOMPR, respectively.

V. Experimental Results and Analysis

A. Dataset

The performance of the proposed methods is assessed on the NTIMIT hand-checked reference phonetic transcription corpus, where utterances are transmitted over telephone channels through either local or long-distance calls [43]. While NTIMIT contains 61 phonetic symbols, these were clustered in 39 phonetic groups by folding allophones into single groups, i.e., {ah, ax, axh}, {aa, ao}, {uw, ux}, {axr, er}, {ih, ix}, {el, l}, {hh, hv}, {m, em}, {en, n, nx}, {ng, eng}, {sh, zh}, and a last group {q, *cl, sil, h#, epi, pau}, which includes the glottal

stop, the closures, and nonspeech symbols as in [44]. In our experiments, we used a subset of the NTIMIT dataset with recordings of two male and two female speakers from each of the eight dialects (32 speakers in total) and eight utterances per speaker. Common utterances sa1 and sa2 were not used. This accounts for 256 unique utterances totalling 13 min of speech time.

B. Experimental Setup

Detection performance was evaluated against the manual phonetic transcription that is provided in the NTIMIT dataset. BICC, AICC, and ICOMP were considered as alternatives to BIC in the candidate point verification step of the DISTBIC algorithm, as described in Section IV. The resulting algorithms are referred to as DISTBICC, DISTAICC, and DISTICOMP. In order to deal with outliers, Fast-MCD M-estimators were also used within the aforementioned algorithms yielding DISTBICCR, DISTAICCR, and DISTICOMPR, respectively. The IFIM in ICOMP was calculated using $c = 0.0001$ and $\tilde{c} = 0.001$ [31]. Performance comparisons against DCF [22] and SVF [22], [23] nonparametric segmentation algorithms were also made.

The mismatch between manual segmentation of audio performed by human transcribers and the automatic segmentation provided by the algorithms was measured. Possible errors in the annotation and accuracy were taken into account by introducing a tolerance. That is, a phone boundary identified by the system is considered correct if it is placed within a range of ± 10 ms from a true (hand-labeled) segmentation point, which implies a 20-ms tolerance. For the experiments, we used the same set of parameter values and features: 40-ms analysis window, 10-ms overlap/shift, 12 MFCCs excluding the energy computed every 4 ms, $\lambda = 1$ for DISTBIC, and 20-ms tolerance. We used the threshold values $h_1 = 0.1\sigma$ $(a = 0.1)$ and $h_2 = 20$ ms, while the threshold h_3 was not utilized. The choice of h_2 is explained by the fact that almost 90% of the phones in the dataset exceed 20 ms in duration. The value of the maximum sub-window in the second step of DISTBIC was selected as 130 ms, since this allows a significant number of observations to be included in the statistical processing, while only 10% of the phones in the dataset have longer duration than 130 ms. This means that if the applied MSC has joined portions of the signal forming windows that exceed 130 ms in the second step, then the union would have been erroneous with high probability.

C. Figures of Merit

Two kinds of errors can be identified in hypothesis testing. Type I error accounts for choosing hypothesis H_1, when H_0 is true; type II error accounts for choosing H_0, when H_1 is true. The probability of type I error defines the significance of the test, whereas 1-(probability of type II error) reflects the power of the test. In phone boundary detection, a point incorrectly identified as a phone boundary yields a type I error (false alarm, FA) while a boundary totally missed by the algorithm is a type II error (missed detection, MD). The detection error rate of the algorithm is described by the missed detection rate $\mathrm{MDR} = \mathrm{MD}/\mathrm{APB}$ and the false alarm rate $\mathrm{FAR} = \mathrm{FA}/(\mathrm{APB} + \mathrm{FA})$ as defined in [10], where APB stands for the actual phone boundaries identified by human annotators. 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 identify the phone boundaries properly. It is also implied that a higher detection performance (lower MDR) comes at expense of a higher FAR. The detection performance of the algorithm can also be assessed by precision $\mathrm{PRC} = \mathrm{CFB}/\mathrm{DET}$ and recall $\mathrm{RCL} = \mathrm{CFB}/\mathrm{APB}$ rates, while the overall objective effectiveness of the algorithm can be evaluated by the F_1-measure $F_1 = 2\mathrm{PRC} \cdot \mathrm{RCL}/(\mathrm{PRC} + \mathrm{RCL})$, where CFB is the number of correctly found boundaries and DET is the number of phone boundaries detected by the algorithm.

TABLE I
PERCENT AVERAGE ERROR RATES IN NTIMIT

		MDR	FAR	PRC	RCL	F_1
1	DISTBIC	28.2	27.1	65.9	71.8	68.7
2	DISTBICC	27.4	25.4	68.1	72.6	70.2
3	DISTBICCR	25.7	23.3	71.0	74.3	72.6
4	DISTAICC	29.4	26.8	65.8	70.6	68.1
5	DISTAICCR	29.1	24.6	68.5	70.9	69.6
6	DISTICOMP	25.5	21.8	72.7	74.5	**73.6**
7	DISTICOMPR	24.6	21.0	73.9	75.4	**74.7**
8	SVF	29.6	27.7	64.7	70.4	67.7
9	DCF	26.7	26.4	67.2	73.3	70.1

TABLE II
TUKEY'S HONESTLY SIGNIFICANT DIFFERENCES CRITERION ON F_1 RATES

1-2	[-2.2,-0.9]	2-4	[1.5,2.8]	3-7	[-2.7,-1.5]	5-7	[-5.6,-4.4]
1-3	[-4.5,-3.2]	2-5	[-0.1,1.2]	3-8	[4.6,5.8]	5-8	[1.7,2.9]
1-4	[-0.1,1.2]	2-6	[-4.0,-2.7]	3-9	[1.9,3.2]	5-9	[-1.0,0.3]
1-5	[-1.6,-0.4]	2-7	[-5.1,-3.8]	4-5	[-2.2,-1.0]	6-7	[-1.7,-0.5]
1-6	[-5.5,-4.3]	2-8	[2.2,3.5]	4-6	[-6.1,-4.9]	6-8	[5.6,6.8]
1-7	[-6.6,-5.4]	2-9	[-0.5,0.8]	4-7	[-7.2,-6.0]	6-9	[2.9,4.2]
1-8	[0.7,2.0]	3-4	[3.9,5.1]	4-8	[0.1,1.3]	7-8	[6.7,7.9]
1-9	[-2.0,-0.7]	3-5	[2.3,3.5]	4-9	[-2.6,-1.3]	7-9	[4.0,5.3]
2-3	[-3.0,-1.7]	3-6	[-1.7,-0.4]	5-6	[-4.5,-3.3]	8-9	[-3.3,-2.1]

95% Confidence intervals for all F_l pairwise comparisons between the 9 segmentation criteria (numbers correspond to indices of Table I).

D. Results and Analysis

The algorithm error rates for the NTIMIT dataset are demonstrated in Table I, where we calculate the average PRC, RCL, and F_1 rates over all recordings. We deduce that DISTICOMPR, DISTICOMP, and DISTBICCR yield the best three results with respect to F_1, while DISTBIC and nonparametric method SVF are the worst performers. Local segmentation using spectral distortion measures yields satisfactory results in DCF, considering its simple derivation. Yet, the improvement in hybrid parametric approaches, that refine model segmentation with statistical hypothesis testing, by using alternative MSC, small

TABLE III
AVERAGE F_1 RATES FOR THE TOP MOST FREQUENT PHONEME CLASS TRANSITIONS

		Percent of occurrences	F_1 (DISTBIC)	F_1 (DISTICOMPR)	F_1 (DIST)
1	silence -> stop	11.8	64.04	70.10	62.67
2	vowel -> silence	9.1	71.53	78.16	67.45
3	semivowel&glide -> vowel	8.9	65.48	71.11	65.08
4	stop -> vowel	8.4	69.44	77.75	67.72
5	vowel -> fricative	8.1	71.01	75.65	69.72
6	fricative -> vowel	7.5	70.76	77.02	67.78
7	vowel -> nasal	7.1	70.57	74.90	67.38
8	vowel -> semivowel&glide	4.8	63.37	68.77	63.22
9	fricative -> silence	4.0	67.16	72.74	64.44
10	nasal -> vowel	3.4	67.15	74.39	65.49

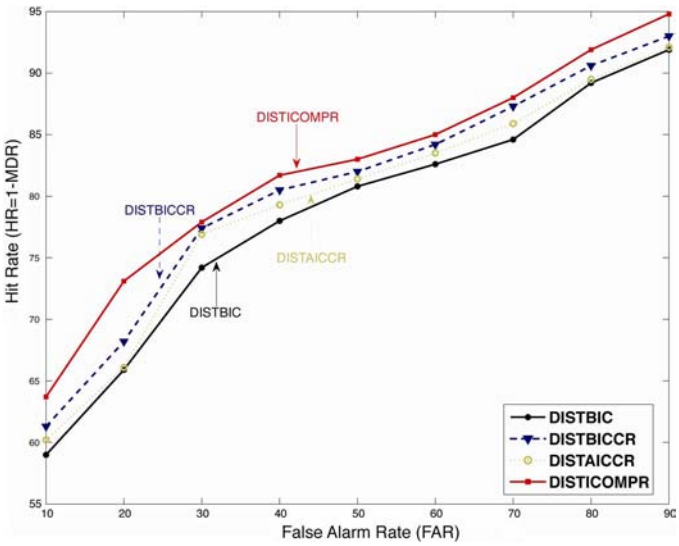

Fig. 4. ROC curves for different segmentation algorithms in NTIMIT, showing that substitution of BIC in DISTBIC with different model selection criteria yields better results.

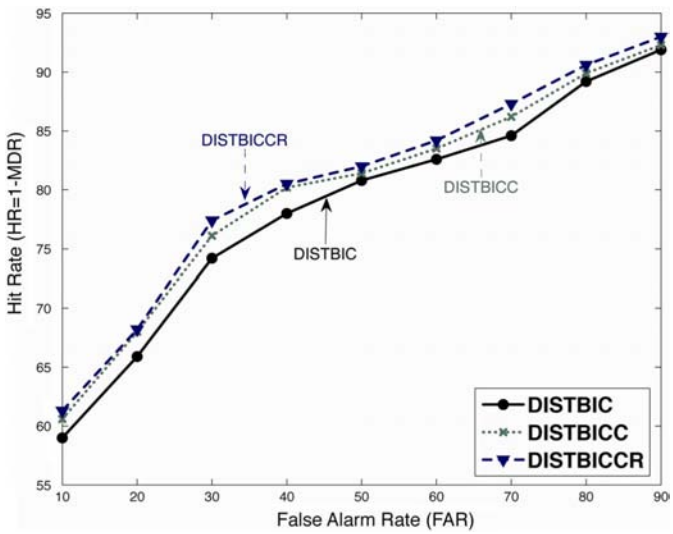

Fig. 5. ROC curves comparing DISTBIC, DISTBICC, and DISTBICCR algorithms in NTIMIT. Small-sample corrections and robust statistics lead to significant increase in segmentation performance.

sample corrections, and robust parameter estimation, is significant. The descriptive analysis results in Table II validate the experimental findings. Using one-way ANOVA for the F_1 rates, the p-value equals $1.432 \cdot 10^{-14}$, indicating that the aforementioned algorithms exhibit significant differences with respect to F_1. Applying post-hoc analysis via Tukey's honestly significant differences criterion, all pairwise comparisons are conducted, as can be seen in Table II. If the confidence interval contains zero, the difference is not significant. It is clear that zero is not included in most intervals, which implies that differences are due to systematic causes rather than random effects in most cases [45]. We deduce that by using alternative MSC to BIC, we obtain statistically significant improvements in the baseline method DISTIBIC. We observe that both small-sample and robust statistics corrections to AIC and BIC refine performance at a 95% confidence interval. We notice that the functional form plays an important role in MSC, since DISTICOMPR yields the best results. Receiver operating characteristic (ROC) curves for different segmentation algorithms are illustrated in Fig. 4. The ROC curves were created by varying the threshold values h_1, h_2, and the parameter winmax as well as the analysis window size. DISTICOMPR exhibits clearly lower FAR at all Hit Rates ($\mathrm{HR} = 1 - \mathrm{MDR} \equiv \mathrm{RCL}$). The performance gain due to small-sample and robust statistics corrections is depicted in Fig. 5, where DISTBIC is compared against DISTBICC and DISTBICCR.

While some phones have relatively stationary spectral properties, others do not. Table III illustrates the average F_1 rates of the most frequent phoneme class transitions, which account for 73.1% of the transitions in the dataset, for DISTBIC and DISTICOMPR. The former is the baseline algorithm whereas the latter is the best performing one. It is clear that DISTICOMPR yields better results than DISTBIC in every class. Since DISTBIC and DISTICOMPR share the same first step, i.e., they use the KL2 distance metric, the gain for each algorithm is due to the proposed alternative MSC.

As stated in Section II-A, top-down phonemic segmentation approaches perform better than linguistically unconstrained methods, since the former make use of additional context information. Concerning comparisons with previous works

on text-independent phonemic segmentation, Esposito and Aversano [4] introduced a novel approach for text-independent speech segmentation, where the preprocessing is based on critical-band perceptual analysis. The algorithm yields 74% hit rate in NTIMIT dataset using MelBank features, which is comparable to ours, and 76% hit rate using MFCCs in TIMIT dataset, while considering a tolerance of 20 ms and limiting over-segmentation to a minimum. Mporas *et al.* [17] have exploited prior knowledge of glottal pulse locations for the estimation of adjacent broad phonemic class boundaries. The algorithm yields 74.9% hit rate in TIMIT database with a 25-ms tolerance. Dusan and Rabiner [46] report 84.6% correct detection at 28.2% false alarm rate in TIMIT with a 20-ms tolerance by relating the maximum spectral transition positions with the perceptual critical points that contain the most important information for consonant perception.

VI. General Discussion

AIC and BIC appear to be naive methods for estimating model error, since both are justified in very general frameworks. Still, they have some attractive properties and practical advantages over their more complex derivatives: the bias correction term does not require any analytical derivation and it can be applied in an automatic way. ICOMP on the other hand, gives a quantitative integrated measure of model complexity, but relies on the estimation of FIM. It must be noted that robustness through M-estimation has been applied in a rudimentary way. We presume that model misspecification of the MFCC features might undervalue the functional form discrimination ability of ICOMP against GoF. Nevertheless, a more systematic assessment is left for future work.

It is well known that DISTBIC introduces a large number of false alarms. While we have attested these can be reduced by using alternative MSC and robust statistics, FAR remains significant (over 20%). Further reduction in FAR requires tuning of the parameters (e.g., larger analysis windows, larger value of h_1) and, consequently, leads to worse MDR. In speaker recognition, limiting missed detection is more important, even at the cost of introducing many false alarms, since these can be easily discarded next by using clustering algorithms. That is not always feasible in phonemic segmentation.

VII. Conclusion and Future Work

In this paper, we examined the applicability of hybrid bottom-up statistical approaches for the text-independent detection of phone boundaries in continuous speech. Because the performance of the baseline DISTBIC algorithm is sensitive to sample sizes and outliers, modifications of DISTBIC were proposed, so that it performs better for phonemic segmentation. By considering small-sample corrected and robust alternatives to BIC, we have improved the accuracy of MSC-based phone boundary detection. Moreover, robust estimators (M-estimators) improve the performance of MSC in phone segmentation. We also infer that complexity due to functional form should not be ignored in small sample sizes, since the criterion ICOMP performs best. While for many sounds the stationarity issue is clear, these conditions are not effective for some phone classes, such as plosives. In future, we intend to combine acoustic-phonetic, temporal, prosodic, and cepstral features and fuse the alternative MSC in order to enhance detection performance. Yet, extra attention should be paid, because the increase in dimensionality would lead to unstable covariance matrix estimation, since the number of observation is small.

Acknowledgment

The authors would like to thank most sincerely the anonymous reviewers for their invaluable comments and suggestions, which contributed a lot to the improvement of the manuscript.

References

[1] K. Demuynck and T. Laureys, "A comparison of different approaches to automatic speech segmentation," in *Proc. 5th Int. Conf. Text, Speech, Dialogue*, 2002, pp. 277–284.

[2] H. Kawai and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. 1, pp. 677–680.

[3] B. Pellom and J. Hansen, "Automatic segmentation of speech recorded in unknown noisy channel characteristics," *Speech Commun.*, vol. 25, no. 1–3, pp. 97–116, 1998.

[4] A. Esposito and G. Aversano, "Text independent methods for speech segmentation," in *Proc. Summer School on Neural Netw.*, 2004, pp. 261–290.

[5] G. Almpanidis and C. Kotropoulos, "Phonemic segmentation using the generalized gamma distribution and small sample Bayesian information criterion," *Speech Commun.*, vol. 50, no. 1, pp. 38–55, 2008.

[6] K. Burnham and D. Anderson, "Multimodel inference understanding AIC and BIC in model selection," *Soc. Methods Res.*, vol. 33, no. 2, pp. 263–304, 2004.

[7] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Inf. Theory*, 1973, pp. 267–281.

[8] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[9] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. Broadcast News Trans. Under. Workshop*, 1998, pp. 127–132.

[10] P. Delacourt and C. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Commun.*, vol. 32, no. 1–2, pp. 111–126, 2000.

[11] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust BIC-based speaker segmentation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 920–933, Jul. 2008.

[12] M. Cettolo, M. Vescovi, and R. Rizzi, "Evaluation of BIC-based algorithms for audio segmentation," *Comput. Speech, Lang.*, vol. 19, no. 2, pp. 147–170, 2005.

[13] P. Zochova and V. Radova, "Modified DISTBIC algorithm for speaker change detection," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 3073–3076.

[14] H. Kadri, Z. Lachiri, and N. Ellouze, "Hybrid approach for unsupervised audio speaker segmentation," in *Proc. 14th Eur. Signal Process. Conf.*, 2006, CD-ROM.

[15] J. Zibert and F. Mihelic, "Development, evaluation and automatic segmentation of Slovenian broadcast news speech database," in *Proc. 7th Int. Conf. Inf. Soc.*, 2004, pp. 72–79.

[16] H. Bozdogan, "Akaike's information criterion and recent developments in information complexity," *Math. Psychol.*, vol. 44, no. 1, pp. 62–91, 2000.

[17] I. Mporas, P. Zervas, and N. Fakotakis, "Evaluation of implicit broad phonemic segmentation of speech signals using pitchmarks," in *Proc. 14th Eur. Signal Process. Conf.*, 2006, CD-ROM.

[18] L. Wang, Y. Zhao, M. Chu, F. Soong, J. Zhou, and Z. Cao, "Context-dependent boundary model for refining boundaries segmentation of TTS Units," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 1082–1091, 2006.

[19] J. Adell and A. Bonafonte, "Towards phone segmentation for concatenative speech synthesis," in *Proc. 5th ISCA Speech Synth. Workshop*, 2004, pp. 139–144.

[20] D. Toledano and A. Gomez, "Automatic phonetic segmentation," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 617–625, Nov. 2003.

[21] J. Hosom, "Automatic phoneme alignment based on acoustic-phonetic modeling," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, vol. 1, pp. 357–360.
[22] C. Mitchell, M. Harper, and L. Jamieson, "Using explicit segmentation to improve HMM phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, vol. 1, pp. 229–232.
[23] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo, "Improved connected digit recognition using spectral variation functions," in *Proc. Int. Conf. Spoken Lang. Process.*, 1992, vol. 1, pp. 627–630.
[24] B. Li and J. Liu, "A comparative study of speech segmentation and feature extraction on the recognition of different dialects," in *Proc. IEEE Conf. Syst., Man, Cybern.*, 1999, vol. 1, pp. 538–542.
[25] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *Proc. Eur. Speech Process.*, 1999, vol. 2, pp. 679–682.
[26] R. Bhansali and D. Downham, "Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion," *Biometrika*, vol. 63, no. 3, pp. 547–551, 1977.
[27] S. Chen, E. Eide, M. Gales, R. Gopinath, D. Kanvesky, and P. Olsen, "Automatic transcription of broadcast news," *Speech Commun. Archive—Special Iss. Autom. Transcription of Broadcast News Data*, vol. 37, no. 1–2, pp. 69–87, 2002.
[28] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Process. Lett.*, vol. 11, no. 8, pp. 649–651, 2004.
[29] X. Zhou, Y. Fu, M. Liu, M. Hasegawa-Johnson, and T. Huang, "Robust analysis and weighting on MFCC components for speech recognition and speaker identification," in *Proc. 2007 IEEE Int. Conf. Multimedia and Expo.*, 2007, pp. 188–191.
[30] D. Dhaene and Hoorelbeke, "The information matrix test with bootstrap-based covariance matrix estimation," *Economics Lett.*, vol. 82, no. 3, pp. 341–347, 2004.
[31] J. Spall, "Cramer-rao bounds and Monte Carlo calculation of the Fisher information matrix in difficult problems," in *Proc. Amer. Control Conf.*, 2004, vol. 4, pp. 3140–3145.
[32] J. Cavanaugh, "Unifying the derivations for the Akaike and corrected Akaike information criteria," *Statist. Probab. Lett.*, vol. 33, no. 2, pp. 201–208, 1997.
[33] C. Hurvich and C. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.
[34] M. Tremblay and D. Wallach, "Comparison of parameter estimation methods for crop models," *Agronomie*, vol. 24, pp. 351–365, 2004.
[35] H. Bozdogan, "ICOMP: A new model selection criterion," in *Proc. Classification Rel. Methods of Data Anal.*, 1988, pp. 599–608.
[36] P. Bearse, H. Bozdogan, and A. Schlottmann, "Empirical econometric modelling of food consumption using a new informational complexity approach," *Appl. Econ.*, vol. 12, no. 5, pp. 563–586, 1997.
[37] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley, 1986.
[38] P. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.
[39] E. Ronchetti, "Robust model selection in regression," *Statist. Prob. Lett.*, vol. 3, pp. 21–23, 1985.
[40] P. Shi and C. Tsai, "A note on the unification of the Akaike information criterion," *R. Statist. Soc.*, vol. 60, pp. 551–558, 1998.
[41] J. Machado, "Robust model selection and M-estimation," *Econ. Theory*, vol. 9, pp. 478–493, 1993.
[42] G. Qian and H. Kunsch, "On model selection in robust linear regression," Seminar Fur Statistik, Eidgenossische Technische Hochschule, Zurich, Switzerland, Tech. Rep. 80.
[43] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 109–112.
[44] M. Antal, "Speaker independent phoneme classification in continuous speech," *Studia Univ. Babes-Bolyai Informatica*, vol. 49, no. 2, pp. 55–64, 2004.
[45] J. Barnette and J. McLean, "The Tukey honestly significant difference procedure and its control of the type I error-rate," in *Annu. Meeting Mid-South Educational Res. Assoc.*, 1998, CD-ROM.
[46] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2000, pp. 17–21.

George Almpanidis received the B.Sc. degree in physics from Aristotle University of Thessaloniki, Thessaloniki Greece, in 1997, the M.Sc. in information technology from the University of Glasgow, Glasgow, U.K., in 2000, and the Ph.D. degree in computer science from Aristotle University of Thessaloniki in 2008.

Since 2001 he has been a Research Assistant in the Artificial Intelligence and Information Analysis Laboratory, Department of Informatics,Aristotle University of Thessaloniki. His research areas include speech processing, information retrieval, and natural language processing.

Margarita Kotti received her B.Sc. degree (with honors) in informatics from the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2005. She is currently working towards the Ph.D. degree in the same department.

She has worked as a Research Assistant in the Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki. Her research interests lie in the areas of speech and audio processing: pattern recognition, speaker segmentation, dialogue detection, and emotion recognition. She has published five journal and 14 conference papers.

Ms. Kotti received scholarships and awards from the State Scholarships Foundation. She received an IBM award as one of the top Ph.D. students in multimedia research worldwide. She is a member of the Greek Computer Society.

Constantine Kotropoulos (S'88–M'94–SM'06) was born in Kavala, Greece, in 1965. He received the Diploma degree with honors in electrical engineering in 1988 and the Ph.D. degree in electrical and computer engineering in 1993, both from the Aristotle University of Thessaloniki, Thessaloniki, Greece.

He is currently an Associate Professor in the Department of Informaticse, Aristotle University of Thessaloniki. From 1989 to 1993, he was a Research and Teaching Assistant in the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki. In 1995, he joined the Department of Informatics, Aristotle University of Thessaloniki, as a Senior Researcher and served then as a Lecturer from 1997 to 2001 and as an Assistant Professor from 2002 to 2007. He has also conducted research in the Signal Processing Laboratory, Tampere University of Technology, Finland, during the summer of 1993. He has coauthored 37 journal papers, 145 conference papers, and contributed six chapters to edited books in his areas of expertise. He is coeditor of the book *Nonlinear Model-Based Image/Video Processing and Analysis* (Wiley , 2001). His current research interests include audio, speech, and language processing; signal processing; pattern recognition; multimedia information retrieval; biometric authentication techniques; and human-centered multimodal computer interaction.

Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a member of EURASIP, IAPR, ISCA, and the Technical Chamber of Greece.