AUTOMATIC PHONEMIC SEGMENTATION USING THE BAYESIAN INFORMATION CRITERION WITH GENERALISED GAMMA PRIORS

George Almpanidis and Constantine Kotropoulos
Aristotle University of Thessaloniki, Dept. of Informatics
Box 451, Thessaloniki 541 24, GREECE. Email: {galba, costas}@aiia.csd.auth.gr

ABSTRACT

Speech segmentation at a phone level imposes high resolution requirements in the short-time analysis of the audio signal. In this work, we employ the Bayesian information criterion corrected for small samples and model 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 attest that the proposed adjustments yield significant performance improvement in noisy environments.

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. It can benefit segment-based speech recognition, concatenative speech synthesis and automatic transcription systems.

Many recent works in the phonemic segmentation employing statistical methods exist in the literature. [1] introduces a novel approach for text-independent speech segmentation, where preprocessing is based on critical-band perceptual analysis. It obtains 74% segmentation accuracy, while limiting over-segmentation to a minimum. [2] examines a maximum a-posteriori (MAP) decoding strategy for segment-based speech recognition where landmarks are modelled in addition to phonemic acoustic units. In [3], a two-step hidden Markov model (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%. [4] also deals with phoneme recognition using a hierarchical structure of multilayer perceptrons, where a block of spectral vectors is split into several blocks processed separately. An overview of machine learning techniques exploited for phone segmentation is given in [5]. Evaluating HMMs, artificial neural networks (ANN), dynamic time warping (DTW), Gaussian mixture models (GMM), and pronunciation modelling, it is concluded that they yield 85-90% detection accuracy, when training data are available and a 20ms tolerance is assumed. [6] also uses 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 DTW is prone to gross labelling errors, while HMM modelling exhibits a systematic bias of 15ms [7].

In this paper, we propose an unsupervised automatic acoustic change detection algorithm that identifies phone boundaries in speech, using the Bayesian information criterion (BIC) for statistical inference. Avoiding the need for linguistic constraints and training data, the algorithm is suitable for speech enhancement in telecommunications, speech transcription in computer-aided systems as well as multilingual speech recognition and synthesis applications. In this paper, the representation power of generalised gamma distribution (GΓD) is exploited, instead that of Gaussian distribution (GD), in order to model the noisy speech signal efficiently while, at the same time, the limited availability of information in frame-based speech processing is accounted for thanks to small-sample approximations of BIC for statistical model comparisons.

2. PHONEMIC SEGMENTATION USING THE BAYESIAN INFORMATION CRITERION

Speech may be roughly considered as the result of sequential linking of *phones*. In short-time analysis, 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.

Common statistical methodology in speech segmentation embraces binary decision-making strategies. When statistical parameters are estimated from random samples, they are also 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). The Bayesian approach to model selection is based on posterior model probabilities. Given a model selection problem, in which we have to choose between two models M_0 and M_1 parameterised by parameter vectors $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, on the basis of a data vector $\boldsymbol{x}$, we can choose the

model with the higher posterior probability, using Bayes' theorem, by calculating their marginal likelihood ratio

$$\frac{P(M_0|\boldsymbol{x})}{P(M_1|\boldsymbol{x})} = \frac{P(\boldsymbol{x}|M_0)P(M_0)}{P(\boldsymbol{x}|M_1)P(M_1)} = BF \times prior\ odds, \qquad (1)$$

where BF is the *Bayes factor*, i.e. the ratio of the integrated likelihoods for the two competing models

$$BF = \frac{P(\boldsymbol{x}|M_0)}{P(\boldsymbol{x}|M_1)} = \frac{\int_{\Theta_0} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0, M_0)\, p(\boldsymbol{\theta}_0\,|\,M_0) d\boldsymbol{\theta}_0}{\int_{\Theta_1} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1, M_1)\, p(\boldsymbol{\theta}_1\,|\,M_1) d\boldsymbol{\theta}_1}. \qquad (2)$$

(2) is similar to a likelihood-ratio test (LRT), but instead of maximising the likelihood, we average it over the parameters. 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.

A large-sample approximation of BF is BIC. It is an asymptotically optimal method for estimating the best model using only sample estimates [8]. BIC is viewed as a penalised maximum likelihood (ML) technique, because it imposes a penalty for model complexity in order to battle over fitting. It is defined as

$$BIC(M_k) = -2L(\hat{\boldsymbol{\theta}}_k) + d_k \ln(n) \qquad (3)$$

where $\hat{\boldsymbol{\theta}}_k$ are the ML estimation (MLE) parameters, $L(\hat{\boldsymbol{\theta}}_k) = \ln P(\hat{\boldsymbol{\theta}}_k\,|\,M_k)$ is the maximised log-likelihood function under model M_k, $d_k = \dim(\boldsymbol{\theta}_k)$ is the dimension of the parameter space for M_k, and n is the sample size. 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 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. BIC 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. It can be regarded as a prior distribution that contains the same amount of information as a single observation.

An acoustic change detection system based on BIC is DISTBIC, 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 [9]. First, assuming that the audio signal is Gaussian, distances of adjacent fixed-length windows are computed in order to identify possible candidates for a change point. Different criteria such as the Kullback-Leibler distance, the generalised LRT or BIC can be applied. When the window is sufficiently small, it can be assumed as a homogenous segment. The offset of the sliding window defines the resolution of the system. Next, a plot of distances is created and significant local peaks are selected as candidate change points by applying heuristic rules. In the last step, a sliding window moves over the signal, making statistical decisions at each candidate point t_i. The boundaries of this window are determined by candidate points t_{i-1} and t_{i+1}. The adjacent signal sub-windows 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 GD or whether a two-segment representation describes it more accurately. The decision problem is undertaken by using BIC as a model selection criterion. This step can be iterated in order to validate or discard the candidates determined in the first step. Let $\boldsymbol{x_i}$ be Q-dimensional feature vectors in a transformed domain, i.e. Mel Frequency Cepstral Coefficients (MFCC) representation, $\boldsymbol{\Sigma_Z}$ 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 $\boldsymbol{Y}$, while m and n-m are the number of feature vectors for each subset.

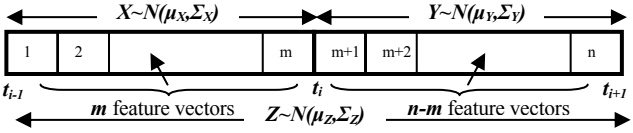

Figure 1 - Models for two adjacent speech segments

For the purpose of phonemic segmentation, we would need to evaluate the following statistical hypotheses at the time instant t_i:

- H_0: $(\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_n}) \sim N(\boldsymbol{\mu_Z},\boldsymbol{\Sigma_Z})$: the data sequence comes from one source $\boldsymbol{Z}$ (i.e., noisy speech/silence, the same phone)
- H_1: $(\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_m}) \sim N(\boldsymbol{\mu_X},\boldsymbol{\Sigma_X})$, $(\boldsymbol{x_{m+1}},\boldsymbol{x_{m+2}},...,\boldsymbol{x_n}) \sim N(\boldsymbol{\mu_Y},\boldsymbol{\Sigma_Y})$: the data sequence comes from two sources $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e. there is a transition from speech utterance to silence, or a transition between two different phones or vice versa

DISTBIC follows a variable window scheme by using relatively small window sizes in areas where boundaries are very likely to occur, while increasing the window size more generously when boundaries are unlikely to occur. Incorporating long-term speech information (i.e. more observations) to the decision rule benefits the speech/pause discrimination and phone transition detection. DISTBIC has been found efficient in detecting acoustic changes that are relatively close one another, but at the price of many falsely detected changes. 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. MODIFICATION OF BIC FOR SMALL SAMPLES

A central problem in statistical inference is dealing with situations, where little information from data is provided. This is clearly the case with phone segmentation where the duration of a single phone can be as small as a few milliseconds. BIC is a *dimension consistent* information criterion that attempts to consistently estimate the dimension of the true model. Assuming that the true model exists and it is in the set of candidate models, BIC will select it asymptotically with probability 1 as sample size increases. This consistency imposes large sample size requirements in order to achieve efficient statistical inference. Due to this limitation, the application of the BIC measure to target domains containing an insufficiently small number of samples requires caution.

A method widely used in probabilistic modelling to approximate the value of marginal likelihoods in model

comparison is Laplace's method. The integrals in (2) can be efficiently approximated using this method provided that the posterior density is highly peaked. Using this approximation 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. Subsequently, it is possible to derive modified BIC criteria that perform better than ordinary BIC for model selection and for different sample sizes. Bollen's approximation ABF2 is derived by considering a specific implicit unit information prior that is more flexible than BIC [10]. By using Laplace's approximation for likelihood, choosing a specially scaled unit information prior

$$P(\boldsymbol{\theta}_k \mid M_k)_{ABF2} \approx N\left(\hat{\boldsymbol{\theta}}_k, \left[w\frac{I_0(\hat{\boldsymbol{\theta}}_k)}{n} \right]^{-1}\right) \tag{4}$$

where I_0 is the observed Fisher information matrix, and setting the optimum value of scale w so that it maximises $P(\boldsymbol{x} \mid M_k)$, we obtain

$$ABF2 = \begin{cases} -2L(\hat{\boldsymbol{\theta}}_k) + d_k\left(1+\ln\dfrac{d_k}{\hat{\boldsymbol{\theta}}_k{}^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k}\right), \\ \qquad\qquad \text{if } d_k > \hat{\boldsymbol{\theta}}_k{}^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k \\ -2L(\hat{\boldsymbol{\theta}}_k) - \hat{\boldsymbol{\theta}}_k{}^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k, \quad \text{otherwise} \end{cases} \tag{5}$$

where $\overline{I_E}$ is the expected information matrix per observation. It has been attested that the performance gain using ABF2 instead of BIC in small samples (n<60) is significant [10]. BICC (*BIC corrected for small samples*) is another approximation that performs better than BIC, both in terms of mean squared error of the parameter estimates and in terms of prediction error [11]. It is defined as:

$$BICC = -2L(\hat{\boldsymbol{\theta}}_k) + \frac{d_k n \ln(n)}{n - d_k - 1}. \tag{6}$$

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.

4. PHONEMIC SEGMENTATION USING GΓD

A critical parameter that affects the performance of statistical speech segmentation methods is the choice of distribution for modeling 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. Nevertheless, many works have demonstrated that Laplacian (LD) and Gamma (ΓD) distributions are more suitable than GD for approximating active voice segments for many frame sizes [12]. Recently, it has been asserted that the generalised Gamma, GΓD, fits the voiced speech signal even better, especially in small frame sizes, and consequently it offers great perspectives for very short-time speech analysis and phone boundary detection [13]. [13] used a two-sided, 3-parameter version of GΓD, defined as

$$f_x(x) = \frac{cb^a}{2\Gamma(a)} \mid x \mid^{ac-1} e^{-b|x|^c} \tag{7}$$

where b, a, and c are positive real values corresponding to scale (b) and shape (a, c) parameters, respectively. GD is a special case of (7) for c=2 and a=0.5. For c=1 and a=1, (7) yields the LD, while for c=1 and a=0.5, it represents the common ΓD. This special property, allows us to model both clean speech and noise/silence with a GΓD

In this paper, we introduce an improved version of the DISTBIC algorithm, DISTBIC-Γ, where we modify the pre-segmentation and refinement steps by assuming a GΓD distribution model for our signal in the analysis windows instead of GD. Zero inputs are ignored, because the GΓD in (7) cannot be specified exactly, when the argument equals zero. 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 BF associated with the hypothesis test. The distribution parameters in (7) are estimated using a computationally efficient on-line algorithm based the gradient ascent algorithm that has been introduced in [13]. Starting from an initial value, the shape parameter c is numerically determined with the gradient ascend algorithm according to the MLE principle. Using a learning factor, we can then re-estimate the value of c that locally maximizes the log-likelihood function L, until L convergences, by iteratively updating it over the sample data. Using this value and the data samples, we can determine the scale (b) and shape (a) parameters. Once the parameters for each of the Q components have been estimated, the likelihood ratio can be calculated. Here, we are making the assumption that the noisy speech signal has uncorrelated components in the MFCC domain. Depending on the window size, this gives a reasonable approximation for the multivariate probability densities with the marginal probability distribution functions. Since the multivariate equivalents of likelihood are simple products over the Q components, the average BF can be easily calculated. 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 [9]. In the second step, using the BIC test as a merging criterion, we compute BIC values for each segmentation point candidate in order to validate the results of the first step. In order to further improve performance we also propose alternative versions of the algorithm, DISTBICC-Γ and DISTABF2-Γ, where the classical BIC is replaced by BICC and ABF2 respectively. Let $\psi_0()$ and $\psi_1()$ denote the digamma, and trigamma functions. The model complexity penalty for ABF2, assuming GΓD modelling is given by (5) where

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k{}^T \overline{I_E}(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k = {} & \hat{a}\psi_0(\hat{a})^2 - 2(\hat{a}\ln\hat{b} - 1)\psi_0(\hat{a}) + \\ & + (\hat{a}^2 + \hat{a})\psi_1(\hat{a}) + \hat{a}\ln\hat{b}^2 - \\ & \hat{a} - 2\ln\hat{b} + 3 \end{aligned} \tag{8}$$

and $\hat{\boldsymbol{\theta}}_k = (\hat{a}, \hat{b}, \hat{c})^T$ is the parameter vector for model M_k at MLE. The four algorithms, DISTBIC, DISTBIC-Γ, DISTBICC-Γ, and DISTABF2-Γ are tested in the setting of phonemic segmentation in the next section.

5. EVALUATION

The performance of the proposed methods is evaluated 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 [14]. 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. 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 ±10ms from a hand-labelled segmentation point, which implies a 20ms tolerance. In the second set of experiments, we used 192 utterances from the Core Set Test of the TIMIT dataset [15] totalling 583 seconds of speech time. The segmentation performance was evaluated against the pre-existing phoneme labelling. For both experiments we used the same set of parameter values and features (50ms initial window, 20ms shift of analysis window, first 12 MFCCs excluding the energy component). White and babble noise from the NOISEX-92 database [16] was added to the clean speech samples at various signal to noise ratios (SNR) levels ranging from 20 to 5 dB.

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

$$PRC = \frac{CFC}{DET}100\% \qquad RCL = \frac{CFC}{ACP}100\% \tag{9}$$

where *CFC* is the number of correctly found changes, *DET* is the number of changes detected by the system, and *ACP* is the number of actual change points. The overall objective effectiveness of the system is assessed by the F_1-measure.

$$F_1 = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL} \tag{10}$$

Fig. 2 and 3 depict the overall performance of the 4 algorithms. The Recall-Precision and F_1-measure results of our tests are also illustrated in Tables 1 and 2. For each case, we calculate the average rates over all recordings. The improved results over the baseline algorithm DISTBIC demonstrate the higher representation power of the GΓD despite making independence assumptions regarding the distribution components. Furthermore, we deduce that small-sample corrected versions of BIC allow additional improvement in detection accuracy. DISTABF2-Γ performs best, followed by DISTBICC-Γ. The performance improvement is most notable at low SNRs.

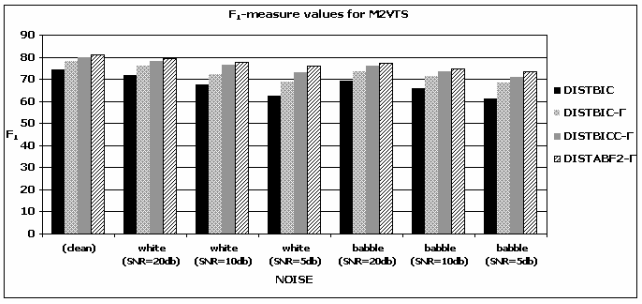

Figure 2 - Overall system evaluation using the F_1 measure for the M2VTS dataset

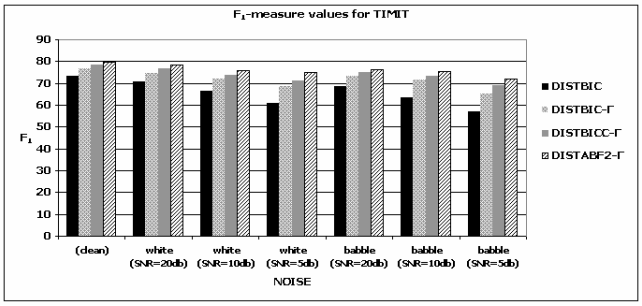

Figure 3 - Overall system evaluation using the F_1 measure for the TIMIT dataset

Table 1 - 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)	-	69.4	79.9	74.2	74.6	82.1	78.1	76.1	83.6	79.6	78.1	84.6	81.2
white	20	67.2	76.8	71.6	72.2	80.5	76.1	74.6	81.9	78.1	76.3	82.9	79.4
white	10	63.6	72.2	67.6	68.3	76.1	72.0	72.8	80.0	76.2	74.5	81.2	77.7
white	5	58.2	67.6	62.5	64.9	73.4	68.9	69.2	77.4	73.0	73.1	79.5	76.1
babble	20	64.7	74.2	69.1	69.1	78.2	73.3	71.9	80.8	76.0	73.7	81.2	77.3
babble	10	61.9	70.7	66.0	67.6	75.9	71.5	69.3	77.9	73.4	71.5	78.7	74.9
babble	5	57.2	66.1	61.2	64.3	73.0	68.4	67.3	75.0	70.9	69.9	77.1	73.3

Table 2 - Recall, Precision, and F_1 measure in TIMIT

		DISTBIC			DISTBIC-ΓD			DISTBICC-ΓD			DISTABF2-ΓD		
Noise	**SNR (dB)**	**PRC**	**RCL**	**F1**	**PRC**	**RCL**	**F1**	**PRC**	**RCL**	**F1**	**PRC**	**RCL**	**F1**
(clean)	-	67.9	77.7	73.3	72.3	82.1	76.8	74.4	82.8	78.4	76.6	83.4	79.8
white	20	65.4	76.3	70.7	70.2	79.4	74.4	72.2	81.7	76.6	74.8	82.7	78.5
white	10	60.9	71.4	66.6	67.3	77.5	72.0	69.6	78.8	73.9	72.7	79.8	76.0
white	5	54.5	66.0	60.7	63.3	74.6	68.5	67.1	76.2	71.3	70.9	79.3	74.9
babble	20	62.5	73.2	68.5	69.6	77.5	73.3	72.0	78.5	75.1	72.5	80.7	76.4
babble	10	58.4	66.5	63.5	67.7	75.8	71.5	69.1	77.8	73.1	71.4	79.6	75.3
babble	5	51.1	62.1	57.0	60.6	71.0	65.3	64.9	73.3	68.8	68.8	75.9	72.1

6. CONCLUSIONS

We have demonstrated that by representing noisy speech with a GΓD in the MFCC domain, we are able to yield improved results compared to GD for an offline two-pass phone segmentation algorithm. 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. Moreover, by considering small-sample corrections to BIC, we were able to further improve phone segmentation accuracy. Evaluation 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 more flexible implicit unit information prior than BIC.

Acknowledgement: This work has been partially supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation, and Learning (FP6-507 752).

REFERENCES

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

[2] J. Glass, "A Probabilistic Framework for Segment-Based Speech Recognition", *Computer Speech and Language*, vol. 17, pp. 137-152, 2003.

[3] 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. Information and Systems*, E89-D(3), pp. 1082-1091, 2006.

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

[5] J. Adell and A. Bonafonte, "Towards phone segmentation for concatenative speech synthesis", in Proc. *5th ISCA Speech Synthesis Workshop (SSW5)*, pp. 139-144, June 2004.

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

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

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

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

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

[11] M. Tremblay and D. Wallach, "Comparison of parameter estimation methods for crop models", *Agronomie*, vol. 24, pp. 351-365, 2004.

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

[13] J. W. Shin and J. -H. Chang, "Statistical modeling of speech signals based on generalized gamma distribution", *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258-261, Mar. 2005.

[14] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (Release 1.00)", in Proc. *1st Int. Conf. Audio- and Video-based Biometric Person Authentication*, pp. 403-409, 1997.

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

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