Time-Scale Information Measures for Text-Independent Phone Segmentation

A. S. Cherniz[† ‡ *] M. E. Torres[† § *] H. L. Rufiner[‡ § *] A. Esposito[♭]

†Laboratorio de Señales y Dinámicas no Lineales , ‡Laboratorio de Cibernética
Facultad de Ingeniería, Universidad Nacional de Entre Ríos,
C.C. 47 Suc. 3 - 3100 Paraná (E.R.), Argentina
metorres@santafe-conicet.gov.ar
§Centro de I+D en Señales, Sistemas e Inteligencia Computacional
Fac. de Ing. y Cs. Hídricas, Univ. Nac. del Litoral, Santa Fe, Argentina
Consejo Nacional de Investigaciones Científicas y Técnicas
♭ Department of Psychology and IIASS
Second University of Naples, Caserta, Italy

***Abstract*— In this work, speech parameterization based on the continuous multiresolution divergence is used to modify a text-independent phone segmentation algorithm. This encoding is employed as input and also replaces an stage of the segmentation procedure responsible for the estimation of the intensity of changes in signal features. The segmentation performance of this representation has been compared with the original algorithm using as input a classical Melbank parameterization and speech representation based on the continuous multiresolution divergence. The results indicate that the modification here proposed increases the ability of the algorithm to perform the segmentation task. This suggests that continuous multiresolution divergence provides valuable information related to acoustic features that take into account phoneme transitions. Moreover, this parameterization gives enough information for its direct use without further processing.**

***Keywords*— Information measures, Divergence, Multiresolution analysis, Automatic speech segmentation.**

1. INTRODUCTION

Segmentation and labeling of speech material according to phonetic or similar linguistic rules is a fundamental task in applications like automatic speech recognition, coding, text-to-speech synthesis and corpora annotation, among others [4, 5]. The aim of speech segmentation is that the sequence of speech frames resulting from short-term analysis could be organized into homogeneous segments associated with a set of symbols representing phones, words, syllables, or other specific acoustic units.

Automatic speech segmentation methods has been faced through several strategies [7]. Some of them incorporate linguistic knowledge, such as the hidden Markov model (HMM) approach [6]. In the case of text-independent segmentation methods, no prior knowledge of the linguistic content contained in the waveform is needed [1]. These procedures can be useful to perform the segmentation when a phonetic transcription is unavailable or inaccurate, or in applications like speaker or language identification systems, concatenative speech synthesis, among others [4, 1].

The text-independent phone segmentation algorithm proposed in [4] is a novel method that accomplishes the phonetic segmentation based on the detection of spectral instability in multiple frequency bands.The algorithm works on an arbitrary number of time-varying features, obtained through a short-term analysis of the speech signal and consists of three stages. The first one corresponds to a jump function computation section, used to estimate the intensity of abrupt changes. The second stage is a relative thresholding step. Finally, a fitting procedure is performed, which takes care of combining different sharp transition events, detected by distinct features, into a single indication of phone boundary. The authors have been proved that this algorithm gives better performance than other methods of the same class [4].

Information about the changes in the dynamics of speech signal can be hidden in different time-varying features. Tools based on entropy notions have been used to characterize the complexity degree of physiological signals. Their use has been extended over different time–scale distributions. The multiresolution entropy gives account of the temporal evolution

[*] This work is supported by Universidad Nacional de Entre Ros, Universidad Nacional del Litoral and CONICET, under Projects PID 6111 and 6106, PAE 37122 and PAE-PICT-2007-00052.

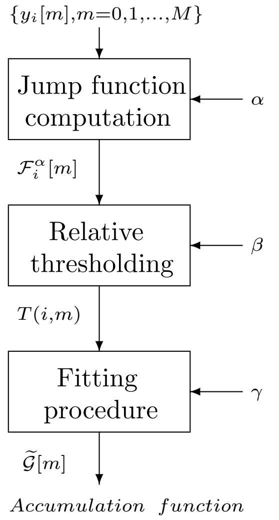
$\{y_i[m], m=0,1,...,M\}$
Jump function computation
α
$\mathcal{F}_i^\alpha[m]$
Relative thresholding
β
$T(i,m)$
Fitting procedure
γ
$\widetilde{\mathcal{G}}[m]$
Accumulation function

Figure 1: Scheme of the segmentation algorithm stages, described in Sec. 2.1.

of Shannon or Tsallis entropies computed over the wavelets coefficients. Continuous multiresolution entropy (CME) [8] has shown to be robust to additive white noise, in the detection of slight changes in the underlying nonlinear dynamics corresponding to physiological signals [2]. Recently, speech parameterization based on CME and continuous multiresolution divergence (CMD), using the Shannon entropy and the Kullback-Leibler distance respectively, have been used to perform text-independent phone segmentation, giving promising results [3].

In this paper we utilize the CMD, computed with the Kullback-Leibler distance, to characterize the speech signal. These time-varying features are used as input in the automatic speech segmentation procedure proposed in [4], but directly in the second stage. This means that the first step of the algorithm, the jump function computation, is skipped. The results of the segmentation obtained with the modification here introduced are compared with those obtained using the original algorithm with a classical Melbank encoding and the CMD-based parametrization.

2. METHODS

In this section we introduce the main characteristics of the speech encodings based on divergence, the text-independent algorithm used and the experiments performed.

2. 1. Time-independent speech segmentation algorithm

The speech segmentation algorithm, proposed by Esposito and Aversano [4], performs the segmentation task working on the time-varying features obtained through the short-term analysis of the speech signal. This is regulated by three operational parameters: α, β and γ. Figure 1 depicts its stages. Parameter α identifies how many consecutive frames, in the values of the speech features, are used to estimate the height (or the intensity) of an abrupt change. Thus, given the set of i time-sequence speech features $\vec{y}_i = \{y_i[m], m = 0, 1, ..., M\}$, for $i = 1, \ldots, \mathcal{J}$, obtained for example using a Melbank parameterization, the function $\mathcal{F}_i^\alpha$ is computed as:

$$\mathcal{F}_i^\alpha[m] = \left| \sum_{\mu=m-\alpha}^{m-1} \frac{y_i[\mu]}{\alpha} - \sum_{\mu=m+1}^{m+\alpha} \frac{y_i[\mu]}{\alpha} \right|. \tag{1}$$

A relative thresholding procedure, with parameter β, is used to identify the frame m^* where a possible transition from one phoneme to another is localized. The relative thresholding procedure is accomplished as follows. Given an interval $[u, v] \subset [\alpha, M - \alpha]$, where $\mathcal{F}_i^\alpha[u]$ and $\mathcal{F}_i^\alpha[v]$ are two valleys of function $\mathcal{F}_i^\alpha$, the frame $m^* \in [u, v]$ is selected so that $\mathcal{F}_i^\alpha[m^*]$ has its maximum value in this interval. That is:

$$\mathcal{F}_i^\alpha[m^*] > \mathcal{F}_i^\alpha[m^* - 1] > ... > \mathcal{F}_i^\alpha[u], \tag{2}$$

with $\mathcal{F}_i^\alpha[u] < \mathcal{F}_i^\alpha[u - 1]$ and

$$\mathcal{F}_i^\alpha[m^*] \geq \mathcal{F}_i^\alpha[m^* + 1] \geq ... \geq \mathcal{F}_i^\alpha[v], \tag{3}$$

where $\mathcal{F}_i^\alpha[v] < \mathcal{F}_i^\alpha[v + 1]$.

A relative height η is then computed as:

$$\eta = \min \left[\mathcal{F}_i^\alpha[m^*] - \mathcal{F}_i^\alpha[u], \mathcal{F}_i^\alpha[m^*] - \mathcal{F}_i^\alpha[v] \right]. \tag{4}$$

The frame m^*, corresponding to a peak of equation (1), is considered as a possible phone transition and stored in the binary matrix $\mathbf{T} = \{T(i, m)\}$, when η exceeds the threshold β. Here, $T(i, m)$ is equal to 1 if a valid transition has been detected for the time-sequence i at the frame m, and 0 otherwise.

It has been observed that sharp transitions do not occur simultaneously for each component of the speech features, even though they take place in a close time interval. A fitting procedure takes care of combining the different transition events stored in the matrix $\mathbf{T}$ into a single indication of phone boundary. In this way the transitions detected in the neighboring of frame m^* are combined into a unique indication of phone boundary. The parameter γ is used to identify the width of the neighborhood where this barycenter is individuated. This is carried out in the undermentioned form: for every $m = 1, ..., M - \gamma + 1$ an interval $V = [m, m + 1, ..., m + \gamma - 1]$ is considered, where the following function is computed:

$$\mathcal{G}[c] = \sum_{\mu=c}^{c+\gamma-1} \sum_{i=1}^{\mathcal{J}} T(i, \mu) \left| \mu - c \right|, \quad c \in V. \tag{5}$$

The possible barycenter of interval V is the frame $\widetilde{c}$ where:

$$\mathcal{G}[\widetilde{c}] = \min_{c \in V} \mathcal{G}[c], \quad m \leq \widetilde{c} \leq m + \gamma - 1. \tag{6}$$

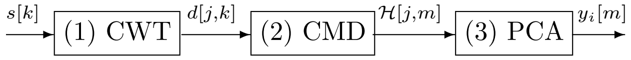

Figure 2: Scheme of the stages of the CMD processing.

For each frame m, the value $\widetilde{\mathcal{G}}[m]$ indicates how many barycenters $\widetilde{c}$ have been found on it. This leads to a new function where the peaks correspond to the indication of a possible phone boundary.

In [4] various standard multi-band representations of speech signals have been tested, using different number of parameters. The authors have proved that the Melbank encoding with 8 coefficients, provides the best results. The Melbank parameterization is an standard short-term processing of speech signal, which is based on a bank-of-filters model [6].

2. 2. Parameterization based on Continuous Multiresolution Divergence

The stages of speech signal parameterization based on CMD are outlined in Fig. 2.

Given a discrete signal $\vec{s} = \{s[k]\}$ of length K, a discretized decomposition $\{d[j,k]\} \in \mathbb{R}^{J\times K}$ in the time–scale plane is obtained applying the quasi–continuous wavelet transform. This decomposition is performed by making $d[j,k] = \Psi_s(a = j\delta, b = k)$, where

$$\Psi_s(a,b) \triangleq \int_{-\infty}^{\infty} |a|^{-1/2} s(t)\ \bar{\psi}\left(\frac{t-b}{a}\right) dt, \qquad (7)$$

is the continuous wavelet transform (CWT) of signal $s(t)$. Observe that $\psi_{a,b}(t) = |a|^{-1/2}\psi\left((t-b)/a\right)$ are dilated and translated versions of $\psi(t)$, with dilation and translation factors a and b respectively. As in practice, $s(t)$ is a discretized signal, i.e. $s(t) = s[k]$ if $t \in [k, k+1]$ for $k = 1,\ldots,K$, to numerically compute the CWT defined by (7) a piecewise constant interpolation is used. This leads to a discretized version, $\Psi_s(j,k) = \Psi_s(a = j\ \delta, b = k)$ with scales $j = 1,\ldots,J$, $J \in \mathbb{Z}$, $\delta \in \mathbb{R}^+$ and time sample $k \in \mathbb{Z}$, the "quasi-continuous" wavelet transform.

For the sake of notational simplicity, for a fixed scale j the CWT coefficient's temporal evolution will be named as $\{d_j[k]\}$ in what follows, with $d_j[k] = d[j,k]$.

Let us consider now a set of rectangular sliding windows $\mathcal{W}^j = \left\{W^j(m,L,\Delta), m = 0,1,2,...,M\right\}$, with $W^j(m,L,\Delta) = \{d_j[k], k = l + m\Delta,\ l = 1,...,L\}$, which determine the frames of analysis. They depend on two parameters, width $L \in \mathbb{N}$ and shift $\Delta \in \mathbb{N}$, which are chosen such that $L \leq K$ (the signal length) and $(K-L)/\Delta = M \in \mathbb{Z}$. Their selection is accomplished in agreement with the windowing performed in order to obtain the Melbank parameterization of the speech signal used in [4] (see Sec. 2. 1)..

An equipartition $d_j^0 = \min_k\{d_j[k]\} < d_j^1 < \cdots < d_j^N = \max_k\{d_j[k]\}$ over the range of values of each window $W^j(m,L,\Delta)$ is considered. This provides a subset $I_n^j = \left\{\left[d_j^{n-1}, d_j^n\right), n = 1,...,N\right\}$ of N disjoint subintervals, such that $W^j(m,L,\Delta) = \bigcup_{n=1}^{N} I_n^j$.

Let us denote with $p_m^j(I_n^j)$ the probability that a given $d_j[k] \in W^j(m,L,\Delta)$ belongs to the interval I_n^j. Therefore, for each window $W^j(m,L,\Delta)$ a set $P^j[m]$ of N probabilities $p_m^j(I_n^j)$ is obtained by means of the regular histogram. Observe that here m represents the time–evolution at the considered scale j.

Having in mind $P^j[m]$, we now consider a second set, $R^j[m] = \{r_m^j(I_n^j),\ n = 1,\ldots,N\}$, corresponding to the next window $W^j(m+1,L,\Delta)$. Thus, the Kullback-Leibler divergence over each set of consecutive windows is:

$$\mathcal{D}_{\mathbf{d}}[j,m] = \sum_{n=1}^{N} p_m^j(I_n^j) \ln\left(\frac{p_m^j(I_n^j)}{r_m^j(I_n^j)}\right). \qquad (8)$$

Observe that here $\mathcal{D}_{\mathbf{d}}$ stands for $\mathcal{D}_{\mathbf{d}}(P,R)$. The probability reference has been skipped in order to make the notation more readable. At each fixed scale j and for each fixed m, the divergence value corresponding to the wavelet coefficients on the window $W^j(m,L,\Delta)$ is computed. Therefore, $\{\mathcal{D}_{\mathbf{d}}[j,m],\ m = 0,1,\ldots,M\}$ represents the Kullback-Leibler divergence evolution at the time–control m. This procedure, when accomplished for all the scales, gives the matrix $\{\mathcal{D}_{\mathbf{d}}[j,m],\ j = 1,\ldots,J,\ m = 0,\ldots,M\}$, denoted as **CMD**, where $CMD(j,m) = \mathcal{D}_{\mathbf{d}}[j,m]$, corresponds to the continuous multiresolution divergence.

The PCA is used to extract the time-varying features that will compose the CMD-based parameterization. The matrix of principal components is computed as:

$$\mathbf{Y} = \mathbf{Q}^T\mathbf{CMD}^*. \qquad (9)$$

where $\mathbf{Q}$ is the eigenvector matrix of $\sigma_{\mathbf{CMD}} = \mathbf{U}\mathbf{U}^T$, and $\mathbf{U} = \mathbf{CMD}^*$ is the statistical normalized matrix associated to **CMD**.

The rows of $\mathbf{Y}$ are the projected **CMD** data in the sense of maximum variability. The principal component of $\mathbf{Y}$ is the row $\vec{y}_1$ that corresponds to the maximum value of $\mathbf{\Lambda}$, the diagonal matrix of eigenvalues, which evolves with the time–control m.

The CMD-based parameterization is obtained using the first eight rows of $\mathbf{Y}$, associated with the eight larger values of $\mathbf{\Lambda}$: $\vec{y}_i = \{y_i[m]\}$, with $i = 1,..,\mathcal{J}$ ($\mathcal{J} = 8$ in this case). The elements $y_i[m]$ of the components $\vec{y}_i$ are now the new features that represent the frame m.

The chosen $\mathcal{J} = 8$ components for the new parameterization is in agreement with the amount of features of the Melbank encoding scheme used to perform the experiments. Moreover, from the point of view of PCA, the first eight components have more than 95% of the total variability of the divergence of the wavelet transform of the speech signal.

For more details on CMD computation, see [3].

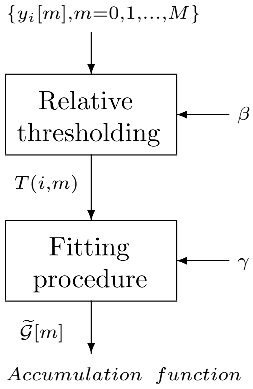

Figure 3: Scheme of the modified segmentation algorithm. The original version was described in Sec. 2. 1.

2. 3. Phone segmentation experiments

The speech signal encoding based on CMD (Sec. 2.2)) has been used as input for the text independent speech segmentation algorithm described in Sec. 2.1, with good results [3]. In the present work, this parameterization is used not only as input for the algorithm but also to replace its first stage. Figure 3 depicts this new approach, that will be referred as the modified algorithm, to differentiate it from the original algorithm described in Sec. 2. 1.

As an example, we show in Fig. 4 the time evolution behavior of one feature of Melbank speech parameterization and CMD-based encoding. In Fig. 4(a) a part of the labeled speech signal of the sentence: " ¿Cómo se llama el mar que baña Valencia?" (What is the name of the sea that border Valencia?) is shown. Fig. 4(b) shows the time evolution of the jump function, $\mathcal{F}_1^{\alpha}$, computed over the Melbank feature $\vec{y}_1$. The segmentation obtained with this parameterization is indicated with the inferior dotted lines with star markers. Fig. 4(c) depict the feature $\vec{y}_1$ obtained through the CMD-based encoding described in Sec. 2. 2. The inferior dotted lines with star markers showed in 4(c) correspond to the segmentation obtained using the modified algorithm with the parameterization based on CMD. Upper dotted lines with circle indicate the database reference labeling in the three figures.

It can be observed from Fig. 4 that the tool based on CMD detect phone boundaries that are ignored by jump function, for example, the third inferior dotted line indication in 4(c)). Also the modified algorithm detects some points that are missing when jump computation is used, such as the sixth an seventh inferior dotted line indication of 4(c). These findings motivate the use of the CMD based parameterization as a part of the algorithm.

The segmentation performance of the modification here proposed was compared with the original algorithm (2.1). using as inputs a Melbank parameterization and the CMD-based encoding described in 2. 2.

2. 4. Signals and Database

A subset of the Albayzin speech corpus, consisting of 600 sentences, 200 words vocabulary, related to Spanish geography, was used [3]. Speech utterances had 3.55 sec. mean phrase duration, and they were spoken by 6 males and 6 females from the central area of Spain (average age 31.8 years). The labeled speech files, consisting in a hand-corrected HMM forced alignment segmentation, recorded the position of the phone boundaries expressed in milliseconds and were used as the reference segmentation.

Each phrase in the corpus has been normalized in mean, pre-emphasized and Hamming windowed in segments of 20 ms length, shifted 10 ms. An 8 coefficients encoding were used for the three evaluated conditions.

2. 5. Indexes of segmentation performance evaluation

The percentage of correctly detected phone boundaries (PC) and the percentage of erroneously inserted points (PI) were computed in order to evaluate the obtained segmentation.

The PC index relates the number of correctly detected boundaries, B_C, with the overall number of phone boundaries contained in the database, B_T, using a tolerance of ±20 ms:

$$PC = 100\left(\frac{B_C}{B_T}\right). \qquad (10)$$

The PI index relates the number of phone boundaries erroneously detected $B_I = B_D - B_C$ (B_D is the whole number of segmentation points detected by the algorithm), with the total number of frames F_T in the signal:

$$PI = 100\left(\frac{B_I}{F_T}\right). \qquad (11)$$

Another indexes derived from the theory of signal detectability, similar to those defined above but mathematically more accurate, have been assessed: the false alarm rate P_{fa} and the missed detection rate P_{md}. They are defined as:

$$P_{fa} = \frac{B_I}{F_T - B_T}100 \qquad (12)$$

$$P_{md} = \frac{B_T - B_C}{B_T}100. \qquad (13)$$

With the values of P_{fa} and P_{md} we have constructed receiver operating characteristic (ROC) curves for each of the encoding schemes evaluated.

3. RESULTS AND DISCUSSION

We now present and discuss the results of the phone segmentation task obtained with the speech encoding based on CMD (Sec. 2.2) into the algorithm of Sec. 2.1. We also compare the modification proposed in this paper with the original algorithm using both, Melbank and CMD-based parameterizations.

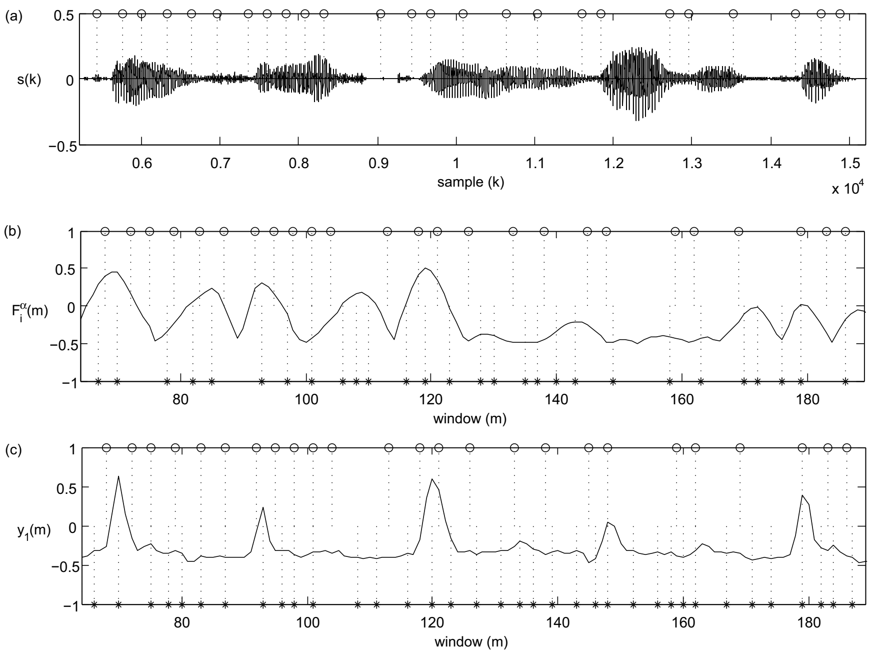

Figure 4: (a) Speech signal with Albayzin labeling (upper dotted lines with circle markers). (b) Jump function computation $\mathcal{F}_i^\alpha[m]$ corresponding to the Melbank feature $y_1[m]$ of signal (a); upper dotted lines with circle markers indicate the database labeling and inferior dotted lines with star markers the segmentation obtained using the Melbank representation as input for the original algorithm. (c) Feature $y_1[m]$ obtained using the CMD-based encoding of the signal displayed in (a); upper dotted lines with circle markers indicate the database labeling and inferior dotted lines with star markers the segmentation obtained using the modified algorithm. Operational parameters used for the automatic segmentation procedures are: $\alpha = 6$, $\beta = 0.05$ and $\gamma = 3$.

Table 1 shows the PC and the PI indexes of segmentation obtained using the modified algorithm and the original algorithm with Melbank and CMD-based parameterizations inputs. The operational parameters used have been: $\gamma = 2$, 3 and 4, $\beta = 0.01$, 0.05, 0.1 and $\alpha = 6$ when original algorithm was used. Notice that there is no need to set parameter α in the modified algorithm case. It is worthwhile to notice that the optimal PC and PI indexes are 100% and 0% respectively. Other results concerning original algorithm with this parameterization, using different values for the operational parameters, can be seen in [3]. It can be observed that for almost all the set of operational parameters, the modified algorithm gives better results when $\gamma \leq 3$. The proposed method increases the correctly detected bounds decreasing also the erroneously inserted points. This can be appreciated by comparing similar PC and PI indexes without having in mind the operational parameters used. For example, the modified algorithm gives PC=97.81% and PI=14.10% (second row) and the original algorithm using Melbank parameterization gives PC=97.33% and PI=17.50% (first row). Also, the modified algorithm give PC=94.62% and PI=11.22% (fifth row) and the original algorithm using now the CMD-based encoding give PC=94.10% and PI=15.57% (fourth row).

In Fig. 5 we show the ROC curves, constructed with the P_{fa} and P_{md} indexes in order to compare the performance of the modified algorithm and the original one using the two encoding schemes considered in this work. In this figure we have used $\gamma = 3$, and β varying between 0 and 1 (with step of 0.01) and $\alpha = 6$ for the original algorithm. High values of P_{fa} occur because many over-segmentations have been obtained. High values of P_{md} indicate that the algorithm has under–segmented the signal. As can be seen from the figure, when P_{fa} is low, P_{md} is high and viceversa. The closer the curve is to the bottom and to the left axes, the more accurate is the detection. We can see that, since the modified algorithm allows to reduce the P_{fa}, reducing also the P_{md}, it gives better results than the original one using either Melbank or CMD-based parameterizations.

These results indicate that the CMD-based parameterization provides information related with abrupt changes in the speech signal, improving the phone segmentation algorithm performance. This approach also reduces the need of further processing the speech parametrization in order to perform the segmentation.

4. CONCLUSIONS

In this work we have proposed a modification to the algorithm introduced in [4], using the continuous mul-

Table 1: Percentage of correctly detected phone boundaries (PC) and percentage of erroneously inserted points (PI) obtained for the modified algorithm and the original algorithm using as inputs the Melbank and CMD-based parameterization. Different operational parameters were tested. Parameter $\alpha = 6$ when original algorithm was used.

Parameters		Modified algorithm		Original algorithm			
				Melbank		CMD	
γ	β	PC	PI	PC	PI	PC	PI
2	0.01	99.54	22.88	97.33	17.50	93.07	16.21
	0.05	97.81	14.10	95.44	16.10	91.27	9.26
	0.1	95.06	11.45	92.06	11.06	87.13	7.71
3	0.01	93.91	18.44	93.21	15.22	94.10	15.57
	0.05	94.62	11.22	90.44	12.01	91.47	9.83
	0.1	92.69	9.13	87.40	6.85	88.15	8.64
4	0.01	95.29	15.52	86.92	13.84	95.44	14.98
	0.05	94.17	9.25	83.55	9.21	93.05	9.98
	0.1	91.47	7.38	79.44	5.59	89.43	8.91

tiresolution divergence, based on Kullback-Leibler distance. This encoding scheme has been used as input for the phone segmentation algorithm, replacing its first stage, corresponding to the jump function computation. An interesting side effect of this modification is the reduction in the number of tunable parameters of the algorithm. The performance of this approach has been compared with the original algorithm using as input the classical Melbank representation and a CMD-based parameterization. The results indicate that the modification proposed increases the algorithm ability to perform the segmentation task. The number of correctly detected boundaries increases and the amount of erroneously inserted points decreases.

This demonstrates that CMD based measures provide valuable information related to acoustic features that take into account transitions from one phoneme to another. Furthermore, this information needs no additional refinement to be used to perform the segmentation.

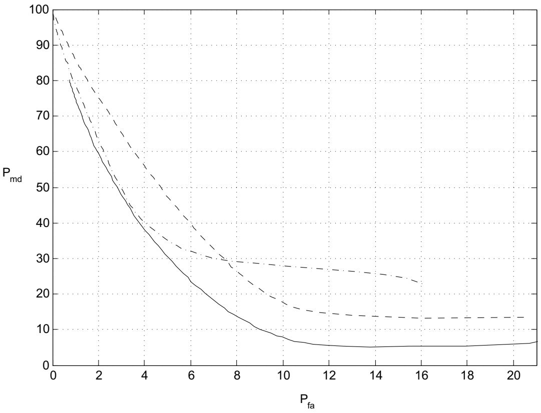

Figure 5: ROC curves for the phone segmentation algorithm using the modified algorithm (solid line) and the original algorithm using Melbank (dashed line) and CMD-based parameterization (dash-dot line). Operational parameters are $\alpha = 6$, for $\gamma = 3$, and β vary form 0 to 1 with increments of 0.01.

REFERENCES

[1] G. Almpanidis and C. Kotropoulos. Phoneme segmentation using the generalized gamma distribution and small sample bayesian information criterion. *Speech Communication*, 50(1):38–55, 2008.

[2] M. M. Añino, M. E. Torres, and G. Schlotthauer. Slight parameter changes detection in biological models: A multiresolution approach. *Physica A*, 324(3–4):645–664, 2003.

[3] A. Cherniz, M. E. Torres, H. Rufiner, and A. Esposito. Multiresolution analysis applied to text-independent phone segmentation. *Journal of Physics Conference Series*, 90, 2007.

[4] A. Esposito and G. Aversano. Text independent methods for speech segmentation. In G. Chollet et al., editor, *Nonlinear Speech Modeling And Applications: Advanced Lectures and Revised Selected Papers*, pages 261–290. Springer, Berlin, Germany, 2005.

[5] H. Kawai and T. Toda. An evaluation of automatic phone segmentation for concatenative speech synthesis. In *Proc. of the Inter. Conf. on Acoustics, Speech, and Signal Processing, ICASSP '04*, volume I, pages 677–680, Montreal, Canada, 2004. IEEE.

[6] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[7] V. Stouten, K. Demuynck, and H. Van hamme. Automatically learning the units of speech by non-negative matrix factorisation. In *8th Annual Conf. of the Inter. Speech Communication Association, INTERSPEECH 2007*, pages 1937–1940, Antwerp, Bélgica, 2007. ISCA.

[8] M. E. Torres, L. Gamero, P. Flandrin, and P. Abry. On a multiresolution entropy measure. In A. Aldroubi et al., editor, *SPIE'97, Wavelet Applications in Signal and Image Processing V*, volume 3169, pages 400–407, Washington, USA, 1997. SPIE Int. Soc. for Optical Engineering.