Multiresolution Analysis applied to Text-Independent Phone Segmentation

Analía S Cherniz[12], María E Torres[13], Hugo L Rufiner[23] and Anna Esposito[4]

[12]Laboratorio de Señales y Dinámicas no Lineales and Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, C.C. 47 Suc. 3 - 3100 Paraná (E.R.), Argentina
[3]Laboratorio de Señales e Inteligencia Computacional, Universidad Nacional del Litoral, Santa Fe, Argentina
[4]Department of Psychology and IIASS, Second University of Naples, Caserta, Italy
Type the author addresses here

E-mail: metorres@ceride.gov.ar

Abstract. Automatic speech segmentation is of fundamental importance in different speech applications. The most common implementations are based on hidden Markov models. They use a statistical modelling of the phonetic units to align the data along a known transcription. This is an expensive and time-consuming process, because of the huge amount of data needed to train the system. Text-independent speech segmentation procedures have been developed to overcome some of these problems. These methods detect transitions in the evolution of the time-varying features that represent the speech signal. Speech representation plays a central role is the segmentation task. In this work, two new speech parameterizations based on the continuous multiresolution entropy, using Shannon entropy, and the continuous multiresolution divergence, using Kullback-Leibler distance, are proposed. These approaches have been compared with the classical Melbank parameterization. The proposed encodings increase significantly the segmentation performance. Parameterization based on the continuous multiresolution divergence shows the best results, increasing the number of correctly detected boundaries and decreasing the amount of erroneously inserted points. This suggests that the parameterization based on multiresolution information measures provide information related to acoustic features that take into account phonemic transitions.

1. Introduction

Segmentation and labelling of speech material according to phonetic or similar linguistic rules is a fundamental task in different speech applications. The aim is that the sequence of speech frames, resulting from short-term analysis, could be organized into homogeneous segments, associated with phones, words, syllables or other specific acoustic units. Traditionally, this task was accomplished manually by a trained phonetician, using listening and visual cues. However, this procedure can be tedious, time-consuming, subjective and error prone, especially for spontaneous speech recordings [1].

The most common automatic speech segmentation methods perform a statistical modelling of transitions between phonetic units using hidden Markov models (HMM) [2]. The HMMs are trained based on the given speech database with the corresponding transcripts and then they are used to align the data along a known phonetic transcriptions. This led to an expensive and time-consuming process because of the huge amount of data needed to train the system.

Different text-independent speech segmentation procedures have been suggested to overcome some of these problems [3, 4, 5]. In [3] the authors proposed an algorithm that works on an arbitrary number of time-varying features, obtained through a short-term analysis of the speech signal. This procedure tries to detect transitions in the speech frames where the value of the parameters changes significantly and quickly.

Entropy notions have been used to characterize the complexity degree of speech and other physiological signals [6]. Spectral entropy has been used in different ways for word/sentence segmentation and silence detection tasks [7, 8]. The continuous multiresolution entropy (CME) gives account of the temporal evolution of Shannon or Tsallis entropy computed over the wavelets coefficients of the continuous wavelet transform (CWT) [9]. Recently, speech representations using CME have been included in an automatic speech recognition system, improving its performance [10].

In this paper we present new speech parameterizations based on the CME and the continuous multiresolution divergence (CMD), using Shannon entropy and the Kullback-Leibler distance, respectively. These new time-varying features are used as inputs for the automatic speech segmentation procedure proposed in [3]. The results of the segmentation obtained with the speech encoding here proposed are compared with those obtained using the classical Melbank parameterization.

2. Materials and methods

2.1. Melbank parameterization

A mel-frequency bank of filters (Melbank) parameterization is a standard short-term processing of speech signal that gives a measurement of the signal energy in a given frequency band [11]. These filters are spanned using a not uniform scale which tries to reproduce the particular spectral resolution of the human ear:

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right). \tag{1}$$

In [3] various standard parameterizations have been tested, using different number of parameters, and the 8 coefficients Melbank encoding provided the best results.

2.2. Parameterization based on CME

Given the discrete speech signal $\mathbf{s} = \{s[k], k = 1,...,K\}$, of length *K*, its CME is obtained by first computing its quasi–continuous wavelet transform $\mathbf{\Psi}_{\mathbf{s}}(a,b)$. This led to a discretized decomposition in the time–scale plane $\{d[j,k]\} = \{\mathbf{\Psi}_{\mathbf{s}}(a = j\delta, b = k)\}$, where $j = 1,...,J \in \mathbb{N}$, $\delta \in \mathbb{R}^{+}$ and $b = k$ is the time–control. In what follows, for each fixed j the CWT coefficient's temporal evolution will be denoted as $\mathbf{d}_j = \{d_j[k]\}$.

We consider a set $W^j(m,L,\Delta) = \{d_j[k], k = l + m\Delta, l = 1,...,L\}$ of rectangular sliding windows for $m = 0,1,...,M$, where $L \in \mathbb{N}$ is the width and $\Delta \in \mathbb{N}$ the shift of the windows. These parameters are chosen such that $L \leq K$ and $(K-L)/\Delta = M \in \mathbb{N}$. These parameters selection will be directly related with maximum speed of significant vocal tract morphology modification [12].

Each window $W^j(m,L,\Delta)$ is divided in a subset of *N* disjoint subintervals I_n and we denote with $p_m^j(I_n)$ the probability that a given $d_j[k] \in W^j(m,L,\Delta)$ belongs to one of such subintervals. Thus, a set of probabilities is obtained for each window:

$$\left\{P^j[m]\right\} = \left\{p_m^j(I_n), n = 1,\dots,N\right\} \tag{2}$$

Using equation (2) the Shannon entropy is computed in the following form:

$$\mathcal{H}_{\mathbf{d}}[j,m] = -\sum_{n=1}^{N} p_m^j(I_n^j)\ln\left(p_m^j(I_n^j)\right). \tag{3}$$

The matrix $\mathbf{CME} = \left\{\mathcal{H}_{\mathbf{d}}[j,m]\right\}$ is the continuous multiresolution entropy of **s**, for each scale j and frame m.

Principal component analysis (PCA) is applied to extract the temporal components of higher variance, in order to obtain the new speech parameterization. The correlation matrix $\sigma = \mathbf{U}\mathbf{U}^T$ is computed from $\mathbf{U} = \mathbf{CME}^*$ (the statistical normalized matrix **CME**). The eigenvectors matrix **Q** and its corresponding eigenvalues matrix **Λ** are obtained, such as $\sigma_{\mathbf{CME}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$. The matrix of principal components is computed as:

$$\mathbf{Y} = \mathbf{Q}^T\mathbf{U}. \tag{4}$$

The principal component of **Y** is the row $\mathbf{y}_1 = \left\{y_1[m], m = 0,1,\dots,M\right\}$, corresponding to the maximum element of **Λ**.

A new parameterization is proposed using the rows of **Y** associated to the $\mathcal{J}$ larger elements of **Λ**: $\mathbf{y}_i$, $i = 1,\dots,\mathcal{J}$, with $\mathcal{J}$ =8. This number of components is chosen in agreement with the Melbank parameterization used to compare the results. Furthermore, these components accumulate more than 95% of the signal variability.

2.3. Parameterization based on CMD

Having in mind the probability set mentioned in equation (2), for the window $W^j(m, L, \Delta)$, we consider now a second set $\left\{R^j[m]\right\} = \left\{r_m^j(I_n^j), n = 1,\dots,N\right\}$, corresponding to the next window $W^j(m+1, L, \Delta)$. The Kullback–Leiber divergence of these consecutive windows can be computed as:

$$\mathcal{D}_{\mathbf{d}}[j,m] = \sum_{n=1}^{N} p_m^j(I_n^j)\ln\left(\frac{p_m^j(I_n^j)}{r_m^j(I_n^j)}\right). \tag{5}$$

This procedure accomplished for all windows and scales gives the continuous multiresolution divergence matrix $\mathbf{CMD} = \left\{\mathcal{D}_{\mathbf{d}}[j,m]\right\}$.

We obtain the CMD-based parameterization applying PCA. This is performed in the same way as for the CME-based encoding. Again, a set of $\mathcal{J}$ =8 time-varying features is obtained for each frame m: $\mathbf{y}_i = \left\{y_i[m], m = 0,1,\dots,M\right\}$, $i = 1,\dots,\mathcal{J}$.

2.4. Speech segmentation algorithm

The speech segmentation algorithm proposed in [3], performs the segmentation using the speech encodings mentioned above. This algorithm is regulated by three operational parameters: α, β and γ.

Parameter α identifies how many consecutive frames are needed to estimate the intensity of an abrupt change. Thus, given $\left\{y_i[m]\right\}$, the following function is computed:

$$\mathcal{F}_i^{\alpha}[m] = \left|\sum_{\mu=m-\alpha}^{m-1}\frac{y_i[\mu]}{\alpha} - \sum_{\mu=m+1}^{m+\alpha}\frac{y_i[\mu]}{\alpha}\right|. \tag{6}$$

A relative thresholding procedure is used to identify the frame m^* where a peak, related to a possible transition from one phoneme to another, is detected according to parameter β. $\mathcal{F}_i^{\alpha}[u]$ and

$\mathcal{F}_i^{\alpha}[v]$ are two valleys of the function given by equation (6), for $u, v \in [\alpha, M-\alpha]$, *u*<*v*. The frame m^* is selected so that $\mathcal{F}_i^{\alpha}[m^*]$ is a local maximum in that interval. A relative height η is computed as:

$$\eta = \min\left[\mathcal{F}_i^{\alpha}[m^*] - \mathcal{F}_i^{\alpha}[u], \mathcal{F}_i^{\alpha}[m^*] - \mathcal{F}_i^{\alpha}[v]\right] \tag{7}$$

A matrix $\mathbf{T} = \{T[i,m]\}$ is constructed, where $T[i,m]=1$ if $\eta \geq \beta$ for the time-sequence *i* at the frame *m*, and 0 otherwise.

The transitions detected by distinct features *i* do not occur simultaneously, even though they occur in a close time interval. The segmentation algorithm uses a fitting procedure to combine into a barycentre the events of each group of quasi-simultaneous sharp transitions. The parameter γ is used to identify the width of the neighborhood where the barycentre is individuated. An interval $V = [m, m+\gamma-1] \in \Box$ is considered for every $m = 1, \ldots, M-\gamma+1$ and the following function is computed:

$$G[c] = \sum_{\mu=c}^{c+\gamma-1} \sum_{i=1}^{\mathcal{I}} T(i,\mu)\left|\mu - c\right|, c \in V. \tag{8}$$

The possible barycentre of interval *V* is the frame $\tilde{c}$ where $G[\tilde{c}] = \min_{V} G[c]$. The value $\tilde{G}[m]$ indicates how many barycenters $\tilde{c}$ have been found on each frame *m*. This leads to a new function where the peaks correspond to the indication of a possible phone boundary.

2.5. Signals and Database

A subset of the Albayzin speech corpus, consisting of 600 sentences, 200 words vocabulary, related to Spanish geography, was used. Speech utterances had 3.55 secs. 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 user-assisted segmentation, recorded the position of the phone boundaries expressed in milliseconds and were used as the exemplar 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 [12]. An 8 coefficients encoding was used for the three evaluated parameterizations.

2.6. 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:

$$PC = 100\left(\frac{B_C}{B_T}\right). \tag{9}$$

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), and the total number of frames F_T in the signal:

$$PI = 100\left(\frac{B_I}{F_T}\right). \tag{9}$$

In order to evaluate the statistical significance of the obtained results, we have estimated the probability that the proposed encodings were better than the Melbank parameterization ($\Pr(\varepsilon < \varepsilon_{ref})$). To perform this test we assumed the statistical independence of the detection errors and the binomial distribution of the errors has been approximated by means of a Gaussian distribution.

3. Results

Table 1 shows the *PC* and the *PI* indexes for the proposed encoding schemes (presented in Secs. 2.2 and 2.3), which are compared with the Melbank parameterization, using the following operational parameters for the segmentation algorithm: γ=3, α=2, 4, 6 and β=0.01, 0.05, 0.1. Bold numbers indicate the best *PC* obtained.

Table 1: *PC* and *PI* obtained for the three encoding schemes evaluated in the phone segmentation algorithm using different operational parameters. Bold numbers indicate the best results for each parameterization.

Parameters			Encoding					
			Melbank		CME		CMD	
γ	α	β	*PC*	*PI*	*PC*	*PI*	*PC*	*PI*
3	2	0.01	**86.52**	16.63	81.64	16.79	84.65	15.12
		0.05	83.97	15.37	84.40	17.44	81.80	08.36
		0.1	79.97	10.49	85.48	17.51	77.20	06.96
	4	0.01	81.11	13.56	83.79	14.90	83.09	12.59
		0.05	78.78	11.96	83.65	15.00	79.31	07.01
		0.1	75.92	07.24	83.39	14.83	74.98	06.08
	6	0.01	75.41	13.31	86.25	17.43	**86.75**	13.87
		0.05	72.02	08.77	**86.91**	17.16	83.51	09.04
		0.1	68.05	05.22	86.09	16.61	79.94	08.05

As can be seen from the indexes in bold, the CME-based encoding provided the best *PC*, but its corresponding *PI* index was high. On the other hand, the CMD-based encoding offered a suitable *PC*, with a *PI* lower than Melbank parameterization. It is worthwhile to notice that the optimal *PC* and *PI* indexes are 100% and 0% respectively.

The statistical significance evaluation of these results for the *PC* index shows $\Pr(\varepsilon < \varepsilon_{ref}) >$ 93.02% for the CME-based parameterization and $\Pr(\varepsilon < \varepsilon_{ref}) >$ 80.57% for the CMD-based encoding. The *PI* index for the CME-based encoding performed worse than the *PI* of Melbank, with $\Pr(\varepsilon > \varepsilon_{ref}) >$ 96.11%. The *PI* corresponding to the CMD-based parameterization was significantly lower than both encoding schemes, with $\Pr(\varepsilon < \varepsilon_{ref}) >$ 99.99% with respect to the Melbank.

It can be observed that, in general, the *PC* index obtained for the parameterization based on the CME and CMD are higher than those achieved with Melbank. The CMD-based parameterization gives the lower values of *PI* index, indicating a better performance and suggesting the use of this encoding scheme for the segmentation algorithm.

In Fig. 1 we compare the performance of the segmentation algorithm using the proposed encodings and Melbank, for α=2, 3, 4, 5, 6, with γ=3 and β=0.01. Fig. 1(a) shows the *PC* index obtained for the different values of α. The performance of the segmentation is better when the line is closer to the upper limit. It can be observed that the proposed encoding schemes present a better performance when α increases, while Melbank decreases. In Fig. 1(b) the *PI* for these experiments is displayed. The lower dash-dotted line indicates the best results. Here, the CMD-based encoding shows the best performance. Although Melbank performs better when α=5 and 6, their *PI* values are comparable to those obtained with the parameterization based on the CMD.

The results of the segmentation using the encodings here proposed seem to be more stable than those obtained with Melbank when parameter α changes. This could be because while the changes on the Melbank parameters evolve in a smooth way and appear in many frames of the speech signal, the evolution of the CME and CMD coefficients is sharper and more concentrated. Therefore, the number of frames to perform the detection, determined by α, affects mainly the Melbank parameterization.

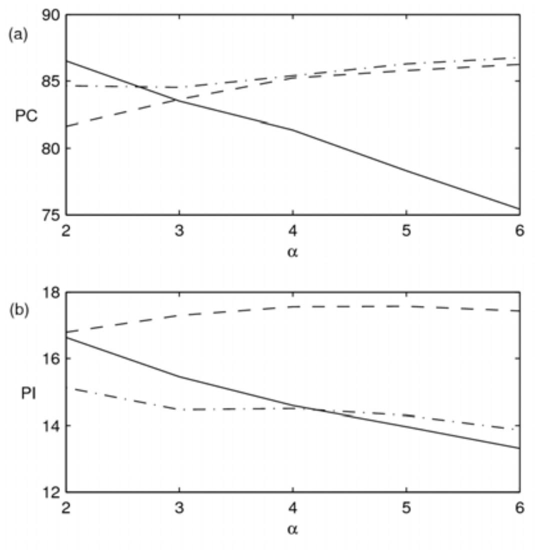

Figure 1. Percentages of (a) correctly detected phone boundaries PC and (b) erroneously inserted points PI, obtained for the phone segmentation algorithm using the classical Melbank parameterization (solid line), the CME-based parameterization (dashed line) and the CMD-based parameterization (dash-dotted line), for α=2, 3, 4, 5, 6, β=0.01 and γ=3.

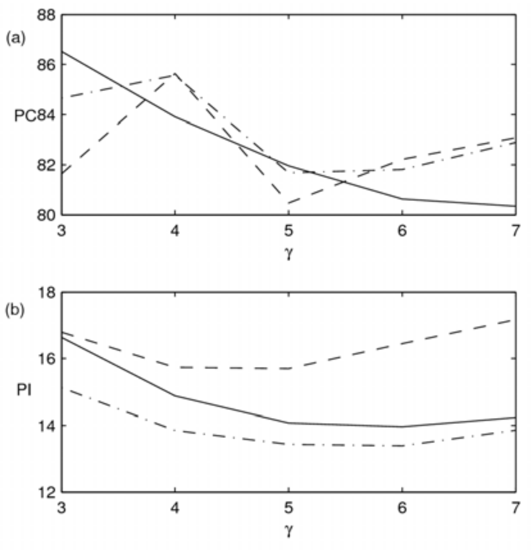

Figure 2. Percentages of (a) correctly detected phone boundaries PC and (b) erroneously inserted points PI, obtained for the phone segmentation algorithm using the classical Melbank parameterization (solid line), the CME-based parameterization (dashed line) and the CMD-based parameterization (dash-dotted line), for γ=3, 4, 5, 6, 7, α=2 and β=0.01.

The property of CME mentioned above, which allows the detection of changes in the parameters in a concentrated way, is the responsible of the high *PI* values obtained with the encoding based on it. The reason is that CME enhances the small changes that appear frame by frame, producing many false positives. Instead, the CMD is more robust because it only takes into account the differences between frames.

Fig. 2 shows the indexes *PC* in (a) and *PI* in (b) obtained for the three evaluated encoding schemes, using γ=3, 4, 5, 6, 7, α=2 and β=0.01. The CMD-based parameterization shows the best performance (high *PC* values) especially for γ =4, 6 and 7, and its *PI* line is the lowest for all γ values.

It can be observed that lines in Fig. 2(b) have a concave shape, with minima around γ=5. This parameter determines the neighborhood of frames used to perform the fitting procedure. Low γ values produce narrow neighborhoods, which could not cover the whole range where the phonetic transition is present. This may cause that one phonetic change could be considered as two transitions, increasing the *PI* index. On the other hand, when γ is high, two transitions appearing in close frames can be processed as only one detection point, producing a worse segmentation performance. This can be observed also in the *PC*, especially for Melbank.

4. Conclusion

In this paper we have presented two parameterizations based on Shannon entropy and Kullback-Leibler distance computed in time–scale plane, which have been used as input for a phone segmentation algorithm. The results indicate that the two proposed parameterizations increase the ability of the algorithm to perform the segmentation task. In particular, the parameterization based on the CMD shows the best performance, since it not only increases the number of correctly detected boundaries but also decreases the amount of erroneously inserted points. This suggests that measures based on the CME and the CMD supply valuable information related to acoustic features, taking into account transitions from one phoneme to another. The information measures used here could give knowledge about changes on the dynamics of the vocal tract, providing an important tool to perform the segmentation.

Acknowledgments

This work is supported by A.N.P.C.yT., under Project PICT Nro 11-12700, Argentina.

References

[1] Binnenpoorte D, Goddijn S and Cucchiarini C 2003 How to improve human and machine transcriptions of spontaneous speech *Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition* (*Tokyo, Japan*) pp 147–50.

[2] Torre D, Hernández L and Villarrubia L 2003 Automatic phonetic segmentation *IEEE Transations on Speech and Audio Processing* **11** pp 617–25.

[3] Esposito A and Aversano G 2005 Text independent methods for speech segmentation *Nonlinear Speech Modeling And Applications: Advanced Lectures and Revised Selected Papers* ed G Chollet (Berlin:Springer) pp 261–90

[4] Gómez J and Castro M 2002 Automatic segmentation of speech at the phonetic level *Proc. of the Joint IAPR Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition* (London:Springer-Verlag) p 672

[5] Sharma M and Mammone R 1996 Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge *Proc. of 4th Int. Conf. on Spoken Language Processing* (*Philadelphia, USA*) pp 1237–40

[6] Rufiner H, Torres M, Gamero L and Milone D 2004 Introducing complexity measures in nonlinear physiological signal: application to robust speech recognition *Physica A* **332** pp 496–508

[7] Wu B and Wang K 2006 Noise spectrum estimation with entropy-based VAD in non-stationary environments *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* **E89**A pp 479–85

[8] Weaver K, Waheed K and Salem F 2003 An entropy based robust speech boundary detection algorithm for realistic noisy environments *Proc. of the Int. Joint Conf. on Neural Networks* pp 680–85

[9] Torres M, Gamero L, Flandrin P and Abry P 1997 On a multiresolution entropy measure *SPIE'97 Wavelet Applications in Signal and Image Processing V* (*Washington, USA*) pp 400–7

[10] Torres M, Rufiner H, Milone D and Cherniz A 2006 Comparison between temporal and time-scale information measures applied to speech recognition *WSEAS Transactions on signal Processing* **9** pp 1153–59

[11] Rabiner L and Juang B 1993 *Fundamentals of Speech Recognition*. (New Jersey:Prentice-Hall)

[12] Deller J, Proakis J, and Hansen J 1993 *Discrete Time Processing of Speech Signals* (New York:Macmillan Publishing)