Volume 385, issue 1 1 November 2007 ISSN 0378-4371

ELSEVIER

PHYSICA A

STATISTICAL MECHANICS AND ITS APPLICATIONS

Recognized by the European Physical Society

Available online at
ScienceDirect
www.sciencedirect.com

http://www.elsevier.com/locate/physa

Available online at www.sciencedirect.com
ScienceDirect
Physica A 385 (2007) 319–332
PHYSICA A
www.elsevier.com/locate/physa

Multiresolution information measures applied to speech recognition[☆]

María E. Torres[a,c,*], Hugo L. Rufiner[b,c], Diego H. Milone[c], Analía S. Cherniz[a,b]

[a]Laboratorio de Señales y Dinámicas no Lineales, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, C.C. 47 Suc. 3- 3100 Paraná (E.R.), Argentina
[b]Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, C.C. 47 Suc. 3- 3100 Paraná (E.R.), Argentina
[c]Laboratorio de Señales e Inteligencia Computacional, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Argentina

Received 11 June 2006
Available online 3 July 2007

Abstract

Considerable advances in automatic speech recognition have been made in the last decades, thanks specially to the use of hidden Markov models. In the field of speech signal analysis, different techniques have been developed. However, deterioration in the performance of the speech recognizers has been observed when they are trained with clean signal and tested with noisy signals. This is still an open problem in this field. Continuous multiresolution entropy has been shown to be robust to additive noise in applications to different physiological signals. In previous works we have included Shannon and Tsallis entropies, and their corresponding divergences, in different speech analysis and recognition systems. In this paper we present an extension of the continuous multiresolution entropy to different divergences and we propose them as new dimensions for the pre-processing stage of a speech recognition system. This approach takes into account information about changes in the dynamics of speech signal at different scales. The methods proposed here are tested with speech signals corrupted with babble and white noise. Their performance is compared with classical mel cepstral parametrization. The results suggest that these continuous multiresolution entropy related measures provide valuable information to the speech recognition system and that they could be considered to be included as an extra component in the pre-processing stage.

PACS: 43.72.+q; 05.45.−a; 05.90.+m; 43.50.+y

Keywords: Complexity measures; Entropy; Divergence; Multiresolution analysis; Robust speech recognition

1. Introduction

Automatic speech recognition (ASR) has been an active research field in the past two decades, in which three wide areas can be mentioned: speech signal analysis, acoustic and language modeling. Statistical methods

[☆] This work is supported by A.N.P.C.yT., under Project PICT No. 11-12700.

[*] Corresponding author.
E-mail address: metorres@ceride.gov.ar (M.E. Torres).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.06.031

of Markov source or hidden Markov models (HMM) [1] have lead to high performance levels in ASR. In the field of speech signal analysis different techniques have been applied. In those experiments such as linear predictive coding (LPC) [2], cepstral and mel frequency cepstral coefficients (MFCC) [3], where production and perception features of human speech were taken into account, the best results have been obtained [4].

When the ASR system is trained with clean speech signals and then it is tested with added noise, an important performance deterioration is observed, which can lead to increase in the recognition errors over 80% [5,6]. In a similar way, when ASR system is trained with speech signal recorded with high quality audio systems and then tested with speech registered with a simple home microphone, errors can grow up to 50% [5,6]. This is the scope of ‘‘robust’’ speech recognition, which has became an important research field in the last years. The aim of robust speech recognition is to obtain ASR systems that can be used in real environments, with noise, reverberation, loss in the transmission channels, home quality audio systems, etc. Research is oriented to two main areas: (1) techniques based on transformation of the speech signal in the feature space (pre-processing) and (2) adaptation of models to noise or particular environmental conditions [7].

In terms of speech perception, it is also important to compare human and machine performance. Human listeners are significantly more robust than artificial ASR systems when performing similar tasks [8]. This leads to insights about similarities and differences, and promotes the selective introduction of new methods into the ASR domain [9]. In this context, there is some research focused on improving ASR systems by using human auditory processing models as front-ends [3]. In addition, noise reduction technology for digital hearing aids is being transferred for its use in ASR. New noise reduction systems, based on physiological and psychoacoustical knowledge, are being designed. Some techniques used in ASR have been useful for the design of hearing aids both in the estimation of speech intelligibility and in its quality [10,11].

There are several pre-processing methods to improve the ASR system's performance [5]. Often it is supposed that both signal and noise are generated by linear systems and that noise has special features allowing one to be easily modeled. In practice none of them is a real assumption: often noise consists in other voices in a conversation and the signal corresponds to a nonlinear system. Thus, the robustness problem of ASR systems is still an ‘‘open’’ research topic, specially for low signal-to-noise ratio (SNR).

Entropy notions have been used to characterize the complexity degree of different physiological signals [12–14]. The application of these quantitative measures provides information about the dynamics of the underlying nonlinear systems, and helps to gain a better understanding of them. Their use has been extended over different time-scale distributions [15]. In speech processing, spectral entropy has been used for relatively simple tasks like segmentation and silence detection [16,17]. The multiresolution entropy, proposed by Torres et al. in Ref. [18], is a tool based on the wavelet transform which gives account of the temporal evolution of the wavelets coefficients' Shannon entropy. Later it has been used with Tsallis entropy [19]. Combined with the continuous wavelet transform (CWT) [20], continuous multiresolution entropy (CME) has shown to be robust to additive white noise, in the detection of slight changes in the underlying nonlinear dynamics corresponding to physiological signals [21,22]. Good results have also been obtained in applications to speech signals corrupted with additive noise in experiments of self-organizing map clustering [23]. All these motivate us to explore this tool over the pre-processing front-end in an ASR system. Recently, Shannon and Tsallis entropies and their corresponding divergences have been included in an ASR system, providing information of the temporal evolution of the complexity degree of speech signals, improving its performance [24]. These information measures and other ones, such as Jensen–Shannon divergence, have been used in different speech analysis applications [25–29].

In this paper we present the extension of the CME to different divergences and we propose to compare the results obtained while included as new parameters in an ASR system. For that purpose, a new approach will be introduced here to the front-end stage of an ASR system: new dimensions taking into account information about the changes in the dynamics of speech signal for different scales, provided by different multiresolution entropies or divergences, will be concatenated to a MFCC classic parametrization. The paper is organized as follows. In Section 2 the basic concepts used in this paper are described. In Section 3 the materials and methods are detailed. In Section 4 we analyze and discuss the obtained results. Finally, in Section 5 conclusions are presented.

2. Basic concepts

2.1. Continuous wavelet transform

Given $s(t)$, a continuous signal, its complexity measures are obtained in the time-scale plane by performing first its CWT, defined as [30]

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

where $\bar{\psi}(t)$ is the complex conjugate of the mother wavelet ψ, which can be any oscillatory function with zero mean and whose Fourier transform $\hat{\psi}(\omega)$ satisfies

$$C_\psi = 2\pi \int_{-\infty}^{\infty} |w|^{-1} |\hat{\psi}(w)|^2 \, \mathrm{d}w < \infty. \tag{2}$$

Observe that $\psi_{a,b}(t) = |a|^{-1/2}\psi((t-b)/a)$ are dilated and translated versions of $\psi(t)$. Thus, the CWT allows to analyze the signal at different size's structures.

Since 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 in (1) a piecewise constant interpolation is used by [31], which translates it as

$$\Psi_s(a,b) = \sum_k \int_k^{k+1} s[k] \bar{\psi}_{a,b}(t) \, \mathrm{d}t. \tag{3}$$

This leads to a discretized version of $\Psi_s(a,b)$ when $a = j\,\delta$, $j = 1, \ldots, J$, with $J \in \mathbb{Z}$ and $\delta \in \mathbb{R}^+$, and $b = k$, which corresponds to the so named "quasi-continuous" wavelet transform.

2.2. Information measures

In this work we will consider Shannon and Tsallis entropies, with their corresponding relative entropies, and Jensen–Shannon divergence.

Shannon entropy is a measure related with the information needed to localize a system in a given state. Given a discrete random variable (rv) $\mathbf{x} \in I = [x_{min}, x_{max}]$, its Shannon entropy $\mathscr{H}_\mathbf{x}$ is computed as [32]

$$\mathscr{H}_\mathbf{x}(P) \triangleq -\sum_{n=1}^{N} p_n \ln(p_n), \tag{4}$$

where $P = \{p_n, n = 1, \ldots, N\}$, p_n is the probability that $\mathbf{x}$ belongs to a considered interval, N is the number of partitions in which the range I is uniformly divided, and with the agreement that $p_n \ln(p_n) = 0$ if $p_n = 0$.

Tsallis entropy depends on a real parameter $q \neq 1$ and it is defined as [33]

$$\mathscr{H}_\mathbf{x}^q(P) \triangleq (q-1)^{-1} \sum_{n=1}^{N} (p_n - (p_n)^q). \tag{5}$$

It has been applied in the analysis of complex biological signals for slight parameter changes detection in the context of nonlinear dynamical systems [22,34].

In the case of Shannon entropy, the associated relative entropy between two probability distributions P and R, corresponding to two rv's $\mathbf{x}$ and $\mathbf{y}$, is expressed as [35]

$$D_{\mathbf{x},\mathbf{y}}(P|R) \triangleq \sum_{n=1}^{N} p_n \ln\left(\frac{p_n}{r_n}\right). \tag{6}$$

$D_{\mathbf{x},\mathbf{y}}(P|R)$ is also called Kullback–Leiber distance, where $p_n \ln(p_n/r_n) = 0$ if $p_n = 0$. In our case, P and R will correspond to two different consecutive segments of the same signal and for this reason we will notate this divergence as $D_\mathbf{x}(P|R)$.

In a similar way, the divergence corresponding to the q-entropy [24,36] is here given by

$$D_{\mathbf{x}}^{q}(P|R) \triangleq \frac{1}{1-q}\sum_{n=1}^{N} p_n \left[1-\left(\frac{p_n}{r_n}\right)^{q-1}\right]. \tag{7}$$

Finally, we consider the Jensen–Shannon divergence [37], which shares similar properties than the above mentioned ones, and is here defined as

$$D_{\mathbf{x}}^{JS}(P|R) \triangleq \mathscr{H}_{\mathbf{x}}(\pi_P P + \pi_R R) - (\pi_P \mathscr{H}_{\mathbf{x}}(P) + \pi_R \mathscr{H}_{\mathbf{x}}(R)), \tag{8}$$

where $\mathscr{H}_{\mathbf{x}}(\cdot)$ is the Shannon entropy and π represents the weight assigned to each distribution.

2.3. Principal component analysis

Principal component analysis (PCA) is a statistical method used for data analysis, feature extraction and compression [38]. Given the data $\mathbf{H} = \{h[j,m]\} \in \mathbb{R}^{J\times M}$, let us suppose that it is produced by a statistical model $\mathbf{\Phi}^{-1}$, with $\mathbf{\Phi} \in \mathbb{R}^{J\times J}$, from a linear combination of noncorrelated hidden sources $\mathbf{Y} = \{y[j,m]\} \in \mathbb{R}^{J\times M}$:

$$\mathbf{H} = \mathbf{\Phi}^{-1}\mathbf{Y}. \tag{9}$$

The goal of PCA is to obtain the components $y[j,m]$ that better explain the data $h[j,m]$ in the sense of the maximum variance directions. In such a way, PCA allows to reduce the dimension of the available data, revealing its most relevant components.

For a fixed row j, the rv $\mathbf{y}_j = \{y[j,1],\ldots,y[j,M]\}$, obtained as $\mathbf{y}_j = \phi_j\mathbf{H}$, with $\phi_j = \{\phi[j,1],\ldots,\phi[j,J]\}$, is called the jth principal component of $\mathbf{H}$ if

$$\tilde{\phi}_j = \arg \underset{i}{max}\, \mathscr{E}[(\phi_i\mathbf{H})^2], \tag{10}$$

where $\mathscr{E}[\cdot]$ stands for the expectation value of the corresponding variable.

In practice, this is solved choosing the rows of $\mathbf{\Phi}$ as the eigenvectors of $\mathbf{H}\mathbf{H}^{\mathrm{T}}$, assuming Gaussian distribution for $\mathbf{Y}$. This diagonalizes the covariance matrix of $\mathbf{Y}$,

$$\tilde{\boldsymbol{\sigma}}_{\mathbf{Y}} = \frac{1}{M-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}, \tag{11}$$

and entails a redundancy reduction, ensuring the independence of sources $\mathbf{Y}$ for second order statistics. The term $(M-1)^{-1}$ is a normalization constant.

The covariance matrix of $\mathbf{H}$ can be computed in the same way as (11) and satisfies $\boldsymbol{\sigma}_{\mathbf{H}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q}$ is its eigenvector matrix, $\mathbf{Q}^{\mathrm{T}} \equiv \mathbf{\Phi}$ and $\mathbf{\Lambda} = diag\{\lambda_1,\ldots,\lambda_J\}$ is the diagonal matrix of eigenvalues associated to $\mathbf{Q}$, with $\lambda_1 > \cdots > \lambda_J$. Then, (9) can be written as

$$\mathbf{Y} = \mathbf{Q}^{\mathrm{T}}\mathbf{H}^{*}, \tag{12}$$

where $\mathbf{H}^{*}$ is the normalized (zero mean) version of $\mathbf{H}$. Thus, in order to solve (9) two procedures are performed: mean subtraction of $\mathbf{H}$ and eigenvector computation of $\mathbf{H}\mathbf{H}^{\mathrm{T}}$. The first principal component corresponds to the one that has the maximum eigenvalue and it is called "the principal component" (PC).

3. Materials and methods

In this section we introduce the main characteristics of the ASR baseline system and the multiresolution entropy approach to be used in this paper. The basic concepts of the different tools will be presented.

A modification to the classical speech signal pre-processing stage will be here outlined. The multiresolution entropy will be concatenated as a new dimension to the mel frequency cepstral coefficients parametrization of speech signal.

The block diagram showed in Fig. 1 depicts each stage of the algorithm proposed in this paper, providing the reader a guide of what follows. Starting from the sampled speech signal, two branches are open. In one of them the CME is computed (Fig. 1: Steps 1–2). In the other branch the classic MFCC parametrization is

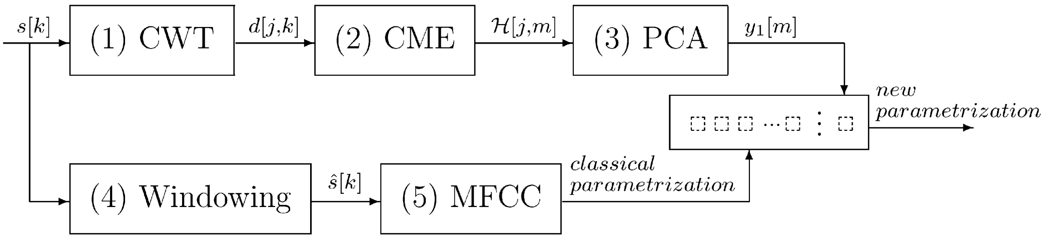

Fig. 1. Schemes of the stages of the proposed method, for one PC, which are explained in the text.

obtained (Fig. 1: Steps 4–5). For major details on MFCC implementation see appendix in Ref. [24]. In order to calculate the CME, for each scale of CWT matrix, information measures are evaluated. We will consider different entropies and divergences: Shannon and Tsallis entropies and their corresponding relative entropies and the Jensen–Shannon divergence. The CME and the continuous multiresolution divergence (CMD) details are presented in Sections 3.1 and 3.2. PCA is applied (Fig. 1: Step 3) to extract the temporal components of higher variance. The obtained values are concatenated as new dimensions in the MFCC parametrization (Section 3.3). The performance of this new approach will be compared to the classical front-end in different noisy conditions.

3.1. Continuous multiresolution entropy

In this section we recall the basic concepts concerning the CME [20].

Given a discrete signal $s[k], k = 1, \ldots, K$, and applying the quasi-CWT, a discretized decomposition $\{d[j,k]\} \in \mathbb{R}^{J \times K}$ in the time-scale plane is obtained, with $d[j,k] = \Psi_s(a = j\delta, b = k)$ (see Section 2.1). For each fixed scale j the CWT coefficient's temporal evolution will be named as $d_j[k]$ in what follows. Let us consider a set of rectangular sliding windows $\mathscr{W}^j = \{W^j(m, L, \Delta), m = 0, 1, 2, \ldots, M\}$, with

$$W^j(m, L, \Delta) = \{d_j[k], k = l + m\Delta, l = 1, \ldots, L\}, \quad m = 0, 1, 2, \ldots, M, \tag{13}$$

which depend on two parameters, width $L \in \mathbb{N}$ and shift $\Delta \in \mathbb{N}$, and where L and Δ are chosen such that $L \leqslant K$ (the signal length) and $(K - L)/\Delta = M \in \mathbb{Z}$. The selection of these values is accomplished in agreement with the windowing performed to obtain the MFCC parametrization of speech signal (see Fig. 1). In this case, the windows length is directly related with maximum speed of significant vocal tract morphology modification [3].

Over each window $W^j(m, L, \Delta)$ an equipartition $d_j^0 = min_k\{d_j[k]\} < d_j^1 < \cdots < d_j^{N-1} < d_j^N = max_k\{d_j[k]\}$ is considered, providing a subset of N disjoint subintervals:

$$I_n^j = \{[d_j^{n-1}, d_j^n), n = 1, \ldots, N\}, \tag{14}$$

such that $W^j(m, L, \Delta) = \overline{\bigcup_{n=1}^N I_n^j}$.

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:

$$P^j[m] = \{p_m^j(I_n^j), n = 1, \ldots, N\}. \tag{15}$$

Observe that here m represents the time-evolution at the considered scale j.

Following the seminal ideas of multiresolution entropies in Refs. [18,20], we are in a condition to compute the information measures over each window $W^j(m, L, \Delta)$, in order to obtain the corresponding continuous multiresolution measures. The Shannon entropy (4) can now be written as

$$\mathscr{H}_{\mathbf{d}}[j,m] = -\sum_{n=1}^{N} p_m^j(I_n^j) \ln(p_m^j(I_n^j)), \quad m = 0, 1, \ldots, M. \tag{16}$$

Observe that here $\mathscr{H}_{\mathbf{d}}$ stands for $\mathscr{H}_{\mathbf{d}}(P)$, and in what follows we will skip the probability reference in all the information measures in order to make the notation more readable.

At each fixed scale j and for each fixed m, the entropy value corresponding to the wavelet coefficients on the window $W^j(m, L, \Delta)$ is computed. Observe that $\{\mathscr{H}_{\mathbf{d}}[j,m], m = 0, 1, \ldots, M\}$ represents the Shannon entropy evolution at the time-control m. In this way, $\{\mathscr{H}_{\mathbf{d}}[j,m], j = 1, \ldots, J, m = 0, \ldots, M\}$ is a matrix that will be denoted as **CME**, where $CME(a = j\delta, m) = \mathscr{H}_{\mathbf{d}}[j,m]$, and named as the continuous multiresolution entropy.

Under the same considerations, the evolution of q-entropy of $d_j[k]$ is computed over each window $W^j(m, L, \Delta)$ and it is obtained as

$$\mathscr{H}^q_{\mathbf{d}}[j,m] = (q-1)^{-1} \sum_{n=1}^{N} (p^j_m(I^j_n) - \left(p^j_m(I^j_n)\right)^q) \tag{17}$$

and the corresponding continuous multiresolution q-entropy matrix $\mathbf{CME}_q$ is obtained, with $CME_q(a = j\delta, m) = \mathscr{H}^q_{\mathbf{d}}[j,m]$.

3.2. Continuous multiresolution divergence

In this section, we extend the ideas of multiresolution entropy to the relative information measures, using now the Kullback–Leiber distance, the q-divergence and the Jensen–Shannon divergence.

Having in mind the probability set $P^j[m]$ mentioned above (15), corresponding to one window $W^j(m, L, \Delta)$, we consider now also a second set $R^j[m] = \{r^j_m(I^j_n), n = 1, \ldots, N\}$, where $r^j_m(I^j_n)$ is the probability that a given $d_j[k]$ corresponding at the next window $W^j(m+1, L, \Delta)$ belongs to the interval I^j_n. In this way, the Kullback–Leiber divergence (6) of two consecutive windows can be computed as

$$D_{\mathbf{d}}[j,m] = \sum_{n=1}^{N} p^j_m(I^j_n) \ln\left(\frac{p^j_m(I^j_n)}{r^j_m(I^j_n)}\right). \tag{18}$$

This procedure accomplished for all scales gives the corresponding continuous multiresolution divergence **CMD**, where $CMD(a = j\delta, m) = D_d[j,m]$.

In a similar way the relative q-entropy (7) is computed as

$$D^q_{\mathbf{d}}[j,m] = \frac{1}{1-q} \sum_{n=1}^{N} p^j_m(I^j_n) \left[1 - \left(\frac{p^j_m(I^j_n)}{r^j_m(I^j_n)}\right)^{q-1}\right] \tag{19}$$

and the continuous multiresolution q-divergence $\mathbf{CMD}_q$ is obtained, being $CMD_q(a = j\delta, m) = D^q_d[j,m]$.

Finally, the Jensen–Shannon divergence reads as

$$D^{JS}_{\mathbf{d}}[j,m] = \mathscr{H}_{\mathbf{d}}(\pi_P\, P^j[m] + \pi_R\, R^j[m]) - (\pi_P \mathscr{H}_{\mathbf{d}}(P^j[m]) + \pi_R \mathscr{H}_{\mathbf{d}}(R^j[m])), \tag{20}$$

computing the $\mathbf{CMD}_{JS}$ as above. Weights π_P and π_R where chosen according to the lengths of $P^j[m]$ and $R^j[m]$ [37].

As an example, we show in Fig. 2 the behavior of one of this multiresolution divergences while applied to a speech signal with and without noise. In Fig. 2(a) a part of the labeled speech signal of sentence: ‘‘Cómo se llama el mar que baña Valencia?’’ (*What is the name of the sea that border Valencia?*) is shown. Fig. 2(b) shows the scalogram ($|d[j,k]|^2$) corresponding to the signal showed in (a), obtained with the Daubechies wavelet of order 16. In Fig. 2(c) the corresponding $\mathbf{CMD}_q$ is shown, for $q = 0.2$. Figs. 2(d)–(f) show results obtained for the same signal but corrupted with additive background conversation noise at 10 dB SNR. It can be observed in Figs. 2(c) and (f) that in this case $\mathbf{CMD}_q$ has higher values in those points labeled as transitions from one phoneme to another, both in the clean signal and in the corrupted one. This result suggests that an appropriate inclusion in the model of the information provided by this tool could improve the ASR system performance in the presence of noise, making it more robust.

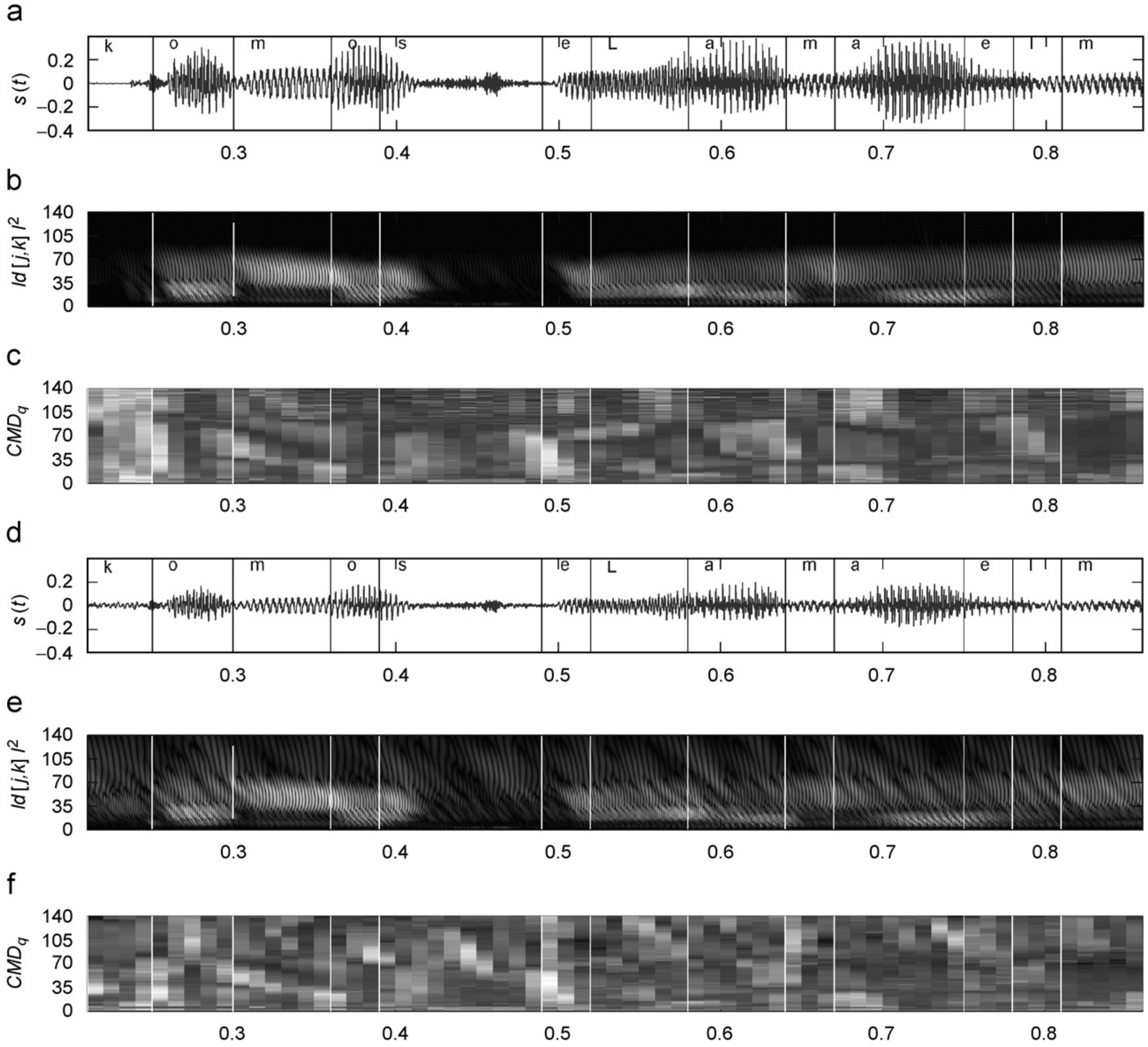

Fig. 2. (a) Labeled speech signal. (b) Scalogram corresponding to the signal displayed in (a). (c) $\mathbf{CMD}_q$ ($q = 0.2$) of scalogram showed in (b). (d) The same signal shown in (a) with additive babble noise (10 dB SNR). (e) Scalogram corresponding to the signal displayed in (d). (f) $\mathbf{CMD}_q$ ($q = 0.2$) of scalogram displayed in (e).

3.3. *Different CME-based parametrization approaches*

While working with HMMs it is important to limit the number of free parameters to be estimated, for this reason, once the multiresolution information measures are obtained, a PCA is performed in order to keep a relative low dimension for the final coefficient vector.

The PCA is here used in three different ways and, in what follows, they are described in detail for the **CME** and can be easily extended to the other multiresolution information measures presented above.

3.3.1. *Method 1: first PC (PC_1)*

Given **CME**, we denote as $\mathbf{U} = \mathbf{CME}^*$ the statistical normalized matrix associated to it. Normalization is obtained scaling to get zero mean and unit variance over each row. Defining the correlation matrix of data $\boldsymbol{\sigma}_{\mathbf{CME}} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$ we find its eigenvector matrix $\mathbf{Q}$ and the diagonal matrix of eigenvalues Λ associated to $\mathbf{Q}$, such as $\boldsymbol{\sigma}_{\mathbf{CME}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}}$. Thus, according to (12), the matrix of PC is computed as

$$\mathbf{Y} = \mathbf{Q}^{\mathrm{T}}\mathbf{CME}^{*}. \tag{21}$$

The row vector of $\mathbf{Y}$ corresponding to the maximum value of $\boldsymbol{\Lambda}$ is the PC and we denote it as $\mathbf{y}_1$. The component $\mathbf{y}_1[m]$ evolves with the time-control m and it will be concatenated to the classical MFCC to obtain our new parametrization.

3.3.2. Method 2: first and second PC (PC_{12})

In method 1 we obtained vector $\mathbf{y}_1$ from (21). Here we also obtain the second component of $\mathbf{Y}$, associated to the second larger value of $\boldsymbol{\Lambda}$, $\mathbf{y}_2$. Thus, both elements $\mathbf{y}_1[m]$ and $\mathbf{y}_2[m]$ are concatenated to the MFCC to generate the new parametrization vector.

The motivation of this method arises from the characteristics of noisy CMD showed in Fig. 2(b). Comparing it with (2(a)), corresponding to clean signal, new structures can be observed, which are related with noise. These structures appear almost separately from the ones corresponding to the clean signal. Therefore, it could be supposed that taking into account two PCs would provide information about both components, signal and noise, increasing the information provided to the model.

3.3.3. Method 3: scale dependent PC (PC_{SD})

For this case, we consider two submatrices from **CME** to apply PCA:

$$\mathbf{U}^{(1)} = (CME(j\delta, m))^{*} \quad \text{where } 1 \leqslant j \leqslant J/2 \text{ and}$$

$$\mathbf{U}^{(2)} = (CME(j\delta, m))^{*} \quad \text{with } J/2 < j \leqslant J,$$

and we compute both correction matrices: $\sigma^{(i)}_{\mathbf{CME}} = \mathbf{U}^{(i)}(\mathbf{U}^{(i)})^{\mathrm{T}}$ for $i = 1, 2$. Thus, the columns of $\mathbf{Q}^{(i)}$ will contain the eigenvector of $\boldsymbol{\sigma}^{(i)}_{\mathbf{CME}}$ and the diagonal matrix $\boldsymbol{\Lambda}^{(i)}$ will provide the associated eigenvalues, for $i = 1, 2$. Two expressions equivalent to (21) are obtained, but over each halves of subdivided matrix **CME**. In this way, we have

$$\mathbf{Y}^{(i)} = (\mathbf{Q}^{(i)})^{\mathrm{T}}\mathbf{U}^{(i)} \quad \text{for } i = 1, 2. \tag{22}$$

From each one of these equations the corresponding PCs $\mathbf{y}_1^{(1)}$ and $\mathbf{y}_1^{(2)}$ are obtained. Both PCs are accordingly concatenated to the classic MFCC parametrization.

Under the same ideas considered for method 2, it is observed from 2(b) that structures related with noise appear mainly at high scales, while the corresponding to signal are in lower scales, suggesting that the two components obtained as above would provide information about signal and noise in a relatively separate way.

These three procedures are accomplished in a similar fashion over the other multiresolution measures. These new coefficients incorporate information about dynamic changes of speech signal in the time-scale plane.

3.4. Automatic speech recognition experiments

In this section we describe the ASR system, the speech database and the cross validation tests used in the experiments.

3.4.1. Automatic speech recognition system

In order to compare the classical parametrization with the alternative one proposed here, we build a state of the art ASR system for a Spanish speech corpus. In what follows we briefly describe its characteristics and the procedure to obtain such reference system. For further details, read Ref. [39].

Three state semi-continuous HMMs (SCHMMs) have been used for context-independent phonemes and silences [2]. Probability density functions for observations have been modeled with Gaussian mixtures. A complete model was built for all the phrases and four reestimations have been accomplished using the Baum–Welch algorithm [40]. Parameter tying was accomplished using a pool of 200 Gaussians for each model state. In the same phoneme model tied mixtures reduce the total effective amount of parameters from 855 000 to 26 200. This stage is necessary in order to improve the estimation robustness because of the reduced training set used [41]. Finally the remaining reestimations have been computed in order to complete a total of 16. For language modeling, backing-off smoothed bigrammars [40] have been estimated with transcriptions of the training database.

For the reference system, each phrase has been normalized in mean, preemphasized and Hamming windowed in segments of 25 ms length, shifted 10 ms. Each segment have been parameterized with 28 coefficients: 13 MFCC, 1 energy coefficient (E) and their temporal derivatives ($\Delta MFCC$ and ΔE) [3].

For each of the methods explained in Section 3.3 new vectors were concatenated to the classical front-end, which incorporate information about dynamic changes of speech signal in the time-scale plane. Therefore, the following parameterizations were considered:

- PC_1 method: 12 MFCC, which allows to maintain the number of coefficients of the reference front-end, one energy coefficient and the coefficient obtained from PCA over each segment, $\mathbf{y}_1$ and its respective temporal derivatives. Thus, our approximation stays: $[MFCC|E|\mathbf{y}_1|\Delta MFCC|\Delta E|\Delta\mathbf{y}_1]$.
- PC_{12} method: 11 MFCC, the energy coefficient and both information measure coefficients, with its corresponding derivatives. This is: $[MFCC|E|\mathbf{y}_1|\mathbf{y}_2|\Delta MFCC|\Delta E|\Delta\mathbf{y}_1|\Delta\mathbf{y}_2]$.
- PC_{SD} method: 11 MFCC, energy coefficient, the two values corresponding to information measures at both low and high scales and their temporal derivatives: $[MFCC|E|\mathbf{y}_1^{(1)}|\mathbf{y}_1^{(2)}|\Delta MFCC|\Delta E|\Delta\mathbf{y}_1^{(1)}|\Delta\mathbf{y}_1^{(2)}]$.

Notice that the bar '|' is used to indicate the concatenation of the different vectors whose elements are used to form the new patterns.

3.4.2. Signals and database

The signals used for training and testing of the ASR system have been obtained from a subset of Spanish corpus Albayzin [42]. This subset consists of 600 sentences, with a vocabulary of 200 words, related to Spanish geography. Its speech utterances have a phrase duration average of 3.35 s. They were spoken by six males and six females from the central area of Spain, with an average age of 31.8 years. The speech signal was registered in a recording study and has been re-sampled at 8 kHz with 16 bits of resolution.

Speech signals corrupted with noise were used to test robustness of the ASR system. White and babble noise from the database NOISEX-92 [43] have been used. White noise has been digitalized from a high-quality analog noise generator. The source of babble noise was background conversation of 100 persons talking in a bar. Both noises have been re-sampled at 8 kHz and have been mixed additively with data at different SNR levels.

3.4.3. Cross validation test

In order to test the methods proposed in this paper we have applied a leave-k-out cross validation [44]. Ten models have been build and trained. For each model, different training/test partitions of signals have been randomly selected. For system training, each partition has been constructed using 80% of the Albayzin subset, and the remaining 20% has been used for testing.

The recognition has been evaluated computing the word error rate (WER), considering as errors the word deletion and substitutions [39]. In order to highlight the differences of our proposal, compared with the reference system, the percentage improvement of relative error has been computed as

$$\Delta\varepsilon_{\%} = \frac{\varepsilon_{ref} - \varepsilon}{\varepsilon_{ref}}, \tag{23}$$

where ε is the WER value and ε_{ref} is the reference WER.

4. Results and discussion

As explained in the previous section, the methods proposed in this work have been included in an ASR system trained with clean speech and tested with speech signals corrupted with both babble and white noise. We present and discuss here the results obtained while comparing the recognition with the one obtained with a classical front-end.

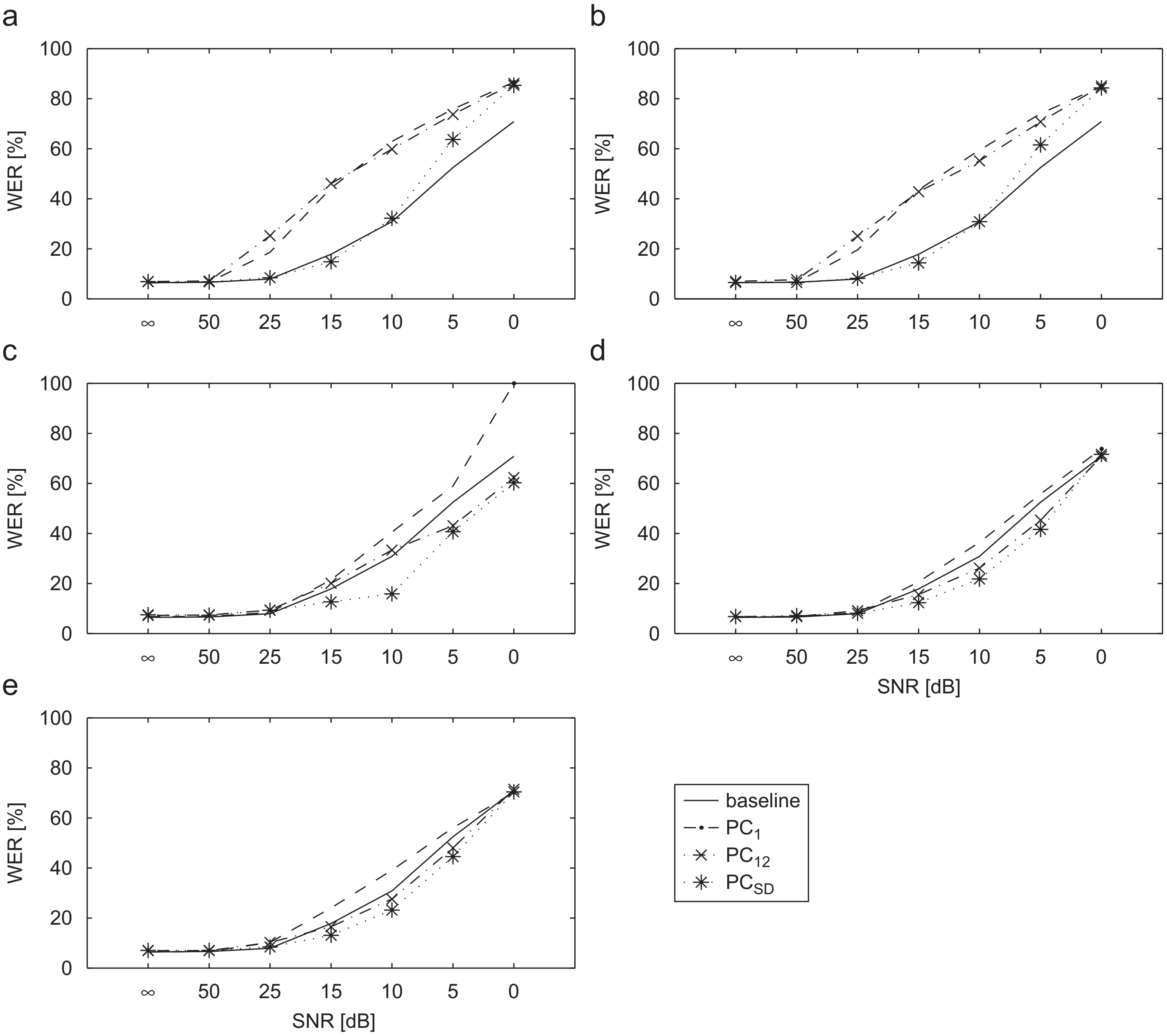

Fig. 3. Word error rate of ASR system *vs.* SNR using signals corrupted with babble additive noise. Comparison between the classical pre-processing (solid line) and the proposed methods: PC$_1$, PC$_{12}$, PC$_{SD}$, computed with (a) Shannon entropy, (b) q-entropy, with $q = 0.2$, (c) Kullback–Leiber distance, (d) q-divergence, with $q = 0.2$, and (e) Jensen–Shannon divergence.

In Fig. 3 we compare the WER obtained with classical parametrization and with the methods proposed here for different SNR and babble noise. Fig. 3(a) shows the WER obtained with the methods PC$_1$, PC$_{12}$ and PC$_{SD}$, when Shannon entropy based CME is concatenated to the MFCC vector. In Fig. 3(b) q-entropy is used. A previous work [24] suggests $q = 0.2$ as an optimal value for this type of experiments. Figs. 3(c)–(e) show the WER obtained with Kullback–Leiber distance, q-divergence and Jensen–Shannon divergence, respectively. These results suggest that the methods PC$_{12}$ and PC$_{SD}$ have a better performance than baseline in the cases (c), (d) and (e). In the case of Kullback–Leiber distance (c) we can observe that, when the parametrization of method PC$_{SD}$ is applied, its word recognition error is under the one of baseline for SNR equal and less than 15 dB.

Fig. 4 shows the WER obtained with classical parametrization *vs.* the methods proposed in this work for white noise at different SNRs, in similar way as the previous figure. In this case, the method PC$_{SD}$ displays an error rate lower than the one obtained for the classical parametrization, in particular for 5, 10 and 15 dB SNRs. We can appreciate that Kullback–Leiber distance (c) displays the best performance, especially for low

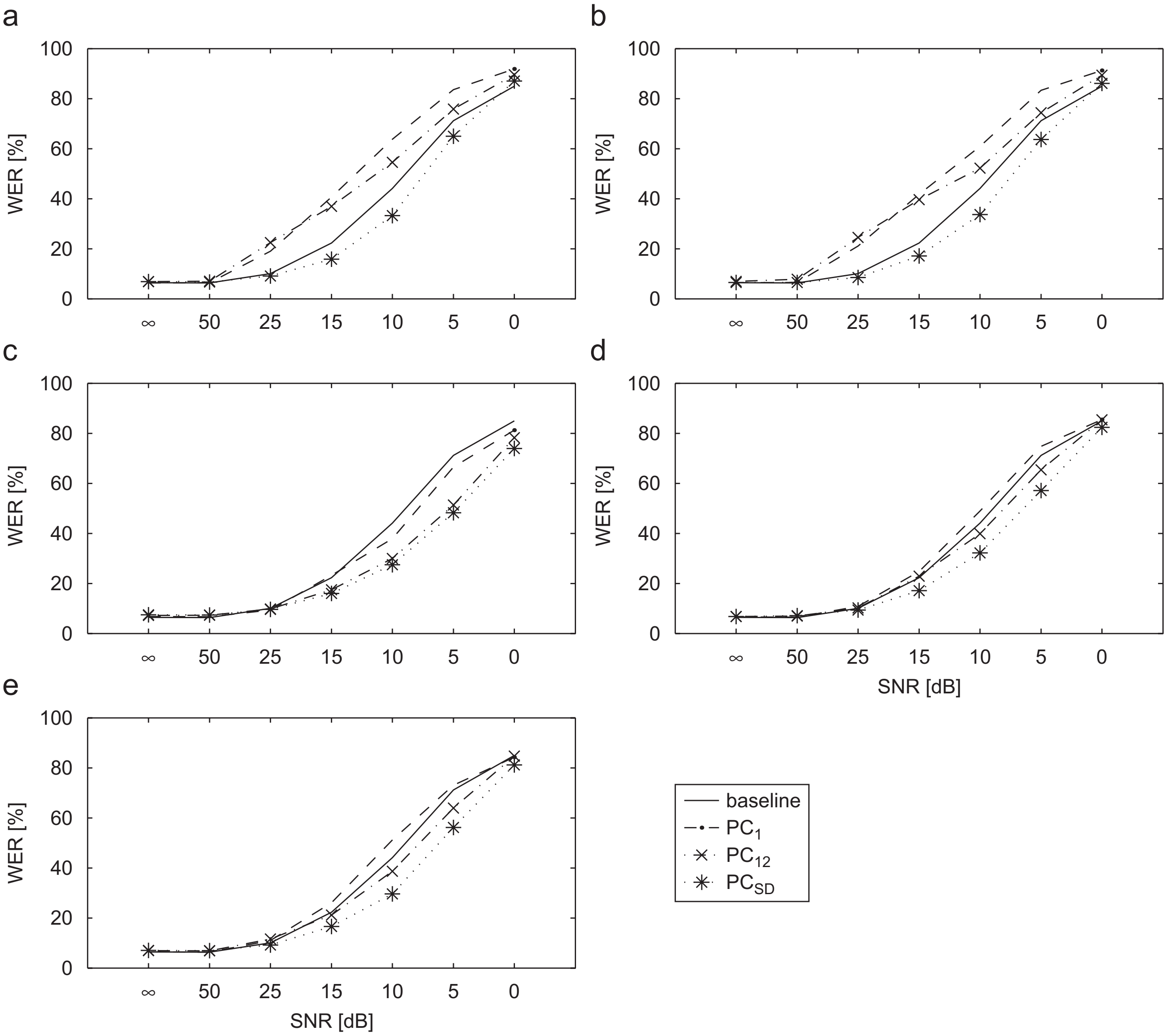

Fig. 4. Word error rate of ASR system *vs.* SNR using signals corrupted with white additive noise. Comparison between the classical pre-processing (solid line) and the proposed methods: PC_1, PC_{12}, PC_{SD}, computed with (a) Shannon entropy, (b) q-entropy, with $q = 0.2$, (c) Kullback–Leiber distance, (d) q-divergence, with $q = 0.2$, and (e) Jensen–Shannon divergence.

SNRs (less than 10 dB for method PC_1 and less than 15 dB for the other methods). For high SNRs the recognition rates are near the baseline.

Comparing Figs. 3(d,e) and 4(d,e), we observe that the systems have similar performance with any of the three methods proposed here, for both kinds of noise. Nevertheless, given that babble noise is less stationary than white noise, and recalling that we are computing the relative entropies between consecutive temporal windows, it is not surprising that the relative measures behave better than the Shannon and Tsallis entropies when babble noise is added to the signal.

Previous results suggest that method PC_{SD} offers the best performance in combination with relative information measures. This could be related to the characteristics of CMD already observed in Fig. 2, where the structures belonging to the clean speech signal are mainly at the lowest scales. In the presence of noise the structures at the higher scales are highly modified, suggesting that babble noise information is more concentrated at these scales.

Table 1
Relative error improvement ($\Delta\varepsilon_{\%}$) of the different measures compared with the classical front-end for speech signal corrupted with *babble* noise

	SNR_{dB}	∞	50	25	15	10	5	0
CME	PC_1	1.00	−2.31	−57.24	−59.67	−51.04	−30.82	−18.19
	PC_{12}	−7.14	−7.29	−68.55	−61.25	−48.52	−28.74	−17.74
	PC_{SD}	−6.28	−3.57	−5.81	**20.63**	−4.46	−17.60	−16.97
$\text{CME}_{q=0.2}$	PC_1	−0.62	−2.09	−59.42	−59.10	−48.23	−29.11	−16.52
	PC_{12}	−7.46	−13.59	−68.37	−58.34	−44.13	−25.87	−16.72
	PC_{SD}	−2.20	0.78	−2.28	**24.27**	−0.23	−14.65	−15.97
CMD	PC_1	−5.71	−2.71	−6.01	−16.62	−24.08	−10.99	−29.16
	PC_{12}	−10.77	−11.10	−15.07	−10.83	−7.50	**21.77**	**13.74**
	PC_{SD}	−14.67	−9.99	−14.78	**40.86**	**94.52**	**29.07**	**17.60**
$\text{CMD}_{q=0.2}$	PC_1	−2.70	−1.87	−4.50	−13.53	−15.40	−5.85	−3.90
	PC_{12}	−5.43	−2.58	−12.52	**14.63**	**18.01**	**15.76**	0.16
	PC_{SD}	−5.43	−7.76	−1.11	**44.75**	**41.39**	**26.18**	−1.02
CMD_J	PC_1	−0.06	−0.65	−22.61	−25.60	−21.22	−6.51	0.97
	PC_{12}	−7.59	−5.10	−22.12	8.12	**11.67**	**9.35**	−0.83
	PC_{SD}	−9.21	−6.87	−5.65	**36.23**	**33.19**	**17.93**	0.62

Bold numbers indicate $\Pr(\varepsilon < \varepsilon_{ref}) > 99.99\%$.

From the point of view of PCA, in method PC_1, when we only take into account one global PC, the raw signal information and the noise information are simultaneously included, and provided in the vector of coefficients so the system cannot discriminate between them. In method PC_{12}, when first and second PCs are used, we could expect that the information not provided by the first PC could appear in the second one, giving additional information, but it is still not well established which one corresponds to the speech signal. This ambiguity appears to be solved by the third method proposed here.

In order to evaluate the statistical significance of these results, we have estimated the probability that a given recognizer is better than the reference system ($\Pr(\varepsilon < \varepsilon_{ref})$). To perform this test we have assumed the statistical independence of the recognition errors for each word and we have approximated the binomial distribution of the errors by means of a Gaussian distribution. This is possible because we have a large number of words (11 077 words, if we take into account all the test partitions).

Table 1 shows the relative errors $\Delta\varepsilon_{\%}$ obtained with each method using signal corrupted with babble noise at different SNRs (corresponds to Fig. 3). In Table 2 the results corresponding to additive white noise are presented. A positive value means that the results provide lower error than the baseline. We have marked in bold fonts those results with higher statistical significance than 99.99%. It can be observed that for babble noise, method PC_{SD} with the Kullback and the parametric divergence provides $\Pr(\varepsilon < \varepsilon_{ref}) > 99.99\%$, for SNR between 5 and 15 dB. For white noise it is obtained for 0–15 dB, for the same methods.

Results obtained in previous work showed relative error improvements of 21.25% for white noise and 24.31% for babble noise in the case of 15 dB SNR [24]. Comparing these results with the methods proposed here, it can be seen that method PC_{SD} provides great percentage of relative error improvement.

5. Conclusions

In this work we have introduced information measures, computed in time-scale plane, in the front-end of an ASR system. The three methods proposed here have been tested with speech signals corrupted with babble and white noise. Performance of these approaches has been compared with classical MFCC parametrization. The results indicate that the method PC_{SD} provides a significant increase in recognition rates over the baseline. This behavior was observed both in babble and white noise, specially for 15, 10 and 5 dB SNRs and for the relative information measures.

Table 2
Relative error improvement ($\Delta\varepsilon_{\%}$) of the different measures compared with the classical front-end for speech signal corrupted with *white* noise

	SNR_{dB}	∞	50	25	15	10	5	0
CME	PC_1	1.00	−1.26	−47.16	−45.17	−30.91	−14.89	−7.44
	PC_{12}	−7.14	−9.78	−55.35	−39.67	−19.17	−6.18	−5.10
	PC_{SD}	−6.28	−5.15	10.61	**40.68**	**32.51**	**9.55**	−2.39
$\mathrm{CME}_{q=0.2}$	PC_1	−0.62	−1.91	−52.48	−46.62	−27.69	−14.65	−6.91
	PC_{12}	−7.46	−18.01	−59.04	−43.70	−15.51	−4.31	−4.87
	PC_{SD}	−2.20	−1.68	**18.02**	**30.21**	**31.00**	**11.77**	−1.25
CMD	PC_1	−5.71	−4.30	7.58	−4.23	**16.31**	**6.94**	**4.62**
	PC_{12}	−10.77	−12.37	2.29	**27.57**	**47.07**	**38.50**	**8.58**
	PC_{SD}	−14.67	−13.65	3.92	**39.16**	**60.27**	**47.60**	**14.92**
$\mathrm{CMD}_{q=0.2}$	PC_1	−2.70	−1.85	−7.56	−9.95	−9.98	−4.87	−0.60
	PC_{12}	−5.43	−8.66	−1.69	−2.68	**10.55**	**8.78**	−0.47
	PC_{SD}	−5.43	−10.65	7.22	**30.02**	**36.76**	**24.51**	**3.13**
CMD_J	PC_1	−0.06	0.14	−8.88	−14.47	−13.90	−2.65	0.81
	PC_{12}	−7.59	−7.41	−13.61	5.10	**13.89**	**11.31**	0.33
	PC_{SD}	−9.21	−7.84	8.87	**33.98**	**49.03**	**26.62**	**4.66**

Bold numbers indicate $\Pr(\varepsilon < \varepsilon_{ref}) > 99.99\%$.

Results obtained not only overcome the baseline but also those reached in our previous work [24], where we used similar information measures, but in the time domain. This demonstrates that CME and CMD related measures provide valuable information to the ASR system in order to perform the recognition, even in the presence of additive noise. This could be related to the fact that the detection of the dynamical changes of the vocal tract is an important cue for speech recognition. Moreover, in order to decode the message carried by a speech signal, human auditory system simultaneously use information from different temporal scales. The wavelet based analysis tool presented here resembles these biological features, providing a new approach to include this information in the pre-processing stage ASR system.

References

[1] L.R. Rabiner, in: Proceedings of the IEEE, vol. 2, IEEE Press, USA, 1989, pp. 257–286.
[2] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[3] J. Deller, J. Proakis, J. Hansen, Discrete Time Processing of Speech Signals, Macmillan, New York, 1993.
[4] M. Cooke, D.P. Ellis, Speech Commun. 35 (2001) 141.
[5] J.C. Junqua, J.P. Haton, Robustness in Automatic Speech Recognition: Fundamentals and Applications, Kluwer Academic Publishers, Boston, 1996.
[6] O. Viiki (Ed.), Noise robust ASR, Speech communication 34 (2001) 1–2 (Special issue).
[7] G.M. Davis (Ed.), Noise Reduction in Speech Applications, CRC Press, USA, 2002.
[8] R.P. Lippmann, Speech Commun. 22 (1997).
[9] J. Tchorz, M. Kleinschmidt, B. Kollmeier, in: Proceedings of Neural Information Processing Systems (NIPS'2000), MIT Press, Denver, USA, 2001, pp. 821–827.
[10] W. Frank, in: R. Stuckless (Ed.), Lovejoy Symposium on Applications of Automatic Speech Recognition with Deaf and Hard of Hearing People, Rochester Institute of Technology, Rochester, New York, 1997.
[11] M. Kleinschmidt, Acta Acustica United Acustica 88 (2002) 416.
[12] P. Saparin, A. Witt, J. Kurths, B. Anishchenko, J. Chaos Solitons and Fractals 4 (1994) 1907.
[13] J.S. Richman, J.R. Moorman, Am. J. Physiol. Heart Circ. Physiol. 278 (2000) 2039.
[14] X.-S. Zhang, R.J. Roy, E.W. Jensen, IEEE Trans. Biomed. Eng. 48 (2001) 1424.
[15] M. Costa, A.L. Goldberger, C.-K. Peng, Phys. Rev. E 71 (2005) 1.
[16] L.S. Huang, C.H. Yang, in: Proceedings of the 2000 International Conference on Acoustics, Speech, and Signal Processing, vol. 3, IEEE, Istanbul, Turkey, 2000, pp. 1751–1754.

[17] P. Renevey, A. Drygajlo, in: Proceedings of Seventh European Conference on Speech Communication and Technology, International Speech Communication Association, Aalborg, Denmark, 2001, pp. 1887–1890.

[18] M. Torres, L. Gamero, E. D'Attellis, Latin Am. Appl. Res. 53 (1995) 53.

[19] L.G. Gamero, A. Plastino, M.E. Torres, Physica A 246 (1997) 487.

[20] M.E. Torres, L. Gamero, P. Flandrin, P. Abry, in: A.F.L. Akram Aldroubi, M. Unser (Eds.), SPIE'97 Wavelet Applications in Signal and Image Processing V, vol. 3169, SPIE International Society for Optical Engineering, Washington, 1997, pp. 400–407.

[21] M.M. Añino, M.E. Torres, G. Schlottahauer, Physica A 324 (2003) 645.

[22] M.E. Torres, M.M. Añino, L.G. Gamero, M.A. Gemignani, Int. J. Bifurcations Chaos. 11 (2001) 967.

[23] H.M. Torres, J.A. Gurlekian, H.L. Rufiner, M.E. Torres, Physica A 361 (2006) 337.

[24] H.L. Rufiner, M.E. Torres, L. Gamero, D.H. Milone, Physica A 332 (2004) 496.

[25] R.D.M. Brigitte Bigi, Y. Huang, in: International Conference on Spoken Language Processing (ICSLP), International Speech Communication Association, ISCA, Jeju Island, South Korea, 2004.

[26] L. Lee, in: Artificial Intelligence and Statistics 2001, Morgan Kaufmann, Key West, Florida, 2001, pp. 65–72.

[27] L. Lee, F. Pereira, in: 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Morgan Kaufmann, University of Maryland, USA, 1999, pp. 33–40.

[28] M. Torres, et al., in: Proceedings of the 25th Annual International Conference of the IEEE–EMBS, IEEE Press, Cancun, Mexico, 2003.

[29] J. Weeds, D. Weir, D. McCarthy, in: Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004, Morgan Kaufmann, Geneva, Switzerland, 2004, pp. 1015–1021.

[30] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, New York, USA, 1999.

[31] M. Misiti, Y. Misiti, G. Oppenheim, J.-M. Poggi, Wavelet Toolbox for Use with MATLAB, first ed., The MathWorks, Inc., ⟨http://www.mathworks.com⟩, 1996. User's Guide.

[32] C. Shannon, Bell Syst. Tech. J. 27 (1948) 379.

[33] C. Tsallis, Chaos Solitons Fractals 6 (1995) 539.

[34] M.E. Torres, L.G. Gamero, Physica A 286 (2000) 457.

[35] T.M. Cover, J.A. Thomas, Information Theory, Wiley, NY, 1991.

[36] M.E. Torres, Ph.D. Thesis, Universidad Nacional de Rosario—Argentina, 1999 (Math. D. Thesis).

[37] I. Grosse, et al., Phys. Rev. E 65 (2002) 1.

[38] M. Akay, Detection and Estimation Methods for Biomedical Signals, Academic Press, San Diego, CA, 1996.

[39] S. Young, et al., The HTK Book (for HTK Version 3.1), Cambridge University Engineering Department, Cambridge, ⟨http://htk.eng.cam.ac.uk⟩, 2002.

[40] F. Jelinek, Statistical Methods for Speech Recognition, The MIT Press, Cambridge, MA, 1999.

[41] S. Young, in: IEEE Workshop on Speech Recognition, IEEE Press, Snowbird, Utah, 1995.

[42] J.E.D. Verdejo, et al., in: Proceedings of the First International Conference on Language Resources and Evaluation, vol. 1, European Language Resources Association, Granada, 1998, pp. 497–502.

[43] A. Varga, H. Steeneken, Speech Commun. 12 (1993) 247.

[44] D. Michie, D.J. Spiegelhalter, C.C. Taylor, Machine Learning, Neural and Statistical Classification, Ellis Horwood, University College, London, 1994.