A Multiresolution Information Measure approach to Speech Recognition*

MARÍA E. TORRES, HUGO L. RUFINER, DIEGO H. MILONE, ANALÍA S. CHERNIZ
Applied Research Group on Signal Processing and Pattern Recognition
Universidad Nacional de Entre Ríos, Facultad de Ingeniería
Universidad Nacional del Litoral, Facultad de Ingeniería y Ciencias Hídricas
CC 47 Suc 3 (3100) Paraná
ARGENTINA
metorres@ceride.gov.ar,{lrufiner, acherniz,}@bioingenieria.edu.ar, dmilone@fich.unl.edu.ar

Abstract: When automatic speech recognition systems trained with clean signals are tested with noisy signals, deterioration in their performance have been observed. Continuous multiresolution entropy have shown to be robust to additive noise in applications to different physiological signals and, in particular, in some speech signal contexts. In this paper we present its extension to different divergences and we propose them as new dimensions at the pre–processing stage of a speech recognizer. Methods proposed here are tested with speech signals corrupted with babble and white noise. Their performance are compared with the classical mel cepstra parametrization. Results suggest that continuous multiresolution entropy related measures provide valuable information that could be considered as an extra component in a pre–processing stage.

Key–Words: Entropy, Divergence, Automatic speech recognition, Continuous multiresolution entropy

1 Introduction

Over the last two decades considerable efforts have been made concerning Automatic Speech Recognition (ASR). However, important performance deteriorations are observed when ASR systems trained with clean speech signals recorded with high quality audio systems are tested with signals registered with simple home microphones or with added noise. This is the scope of "robust" speech recognition, which aim is to obtain ASR systems that can be used in real enviroments, with noise, reverberation, home quality audio systems, etc [1].

There are several pre-processing methods to improve ASR system's performance, in which it is often supposed that both signal and noise are generated by linear systems and noise can be easily modeled. In practice none of them is a real assumption and the robustness problem of ASR systems is still "open", specially for low signal-to-noise ratio (SNR).

Entropy notions have been used to characterize the complexity degree of different physiological signals. The application of these quantitative measures provides information about the underlying dynamics of non linear systems and helps to gain a better understanding of them. Recently, Shannon and Tsallis entropies and their corresponding divergences have been included in the pre–processing stage of an ASR system, providing information of the temporal evolution of the complexity degree of speech signals, improving its performance [2].

The multiresolution entropy, proposed by Torres et al. in [3], is a tool based on the wavelet transform which gives account of the temporal evolution of the wavelets coefficients' Shannon entropy. Combined with the continuous wavelet transform (CWT) [4], the tool known as continuous multiresolution entropy (CME) has shown to be robust to additive white noise in the detection of slight changes in the underlying nonlinear dynamics of physiological signals [5]. In applications to speech signals corrupted with additive noise, good results have been obtained in experiments of self-organizing map clustering [6]. This motivates us to explore this tool over the parametrization of an ASR system in order to test its robustness.

In this paper we present the extension of the continuous multiresolution entropy to different divergences and we propose to compare the results obtained when these information measures are introduced as new dimensions to the front–end stage of an ASR system. These new parameters, taking into account information about the changes in the dynamics

[1]This work is supported by A.N.P.C.yT., under Project PICT #11-12700 and PICT2004 #25984. CAID 12-72. CONICET.

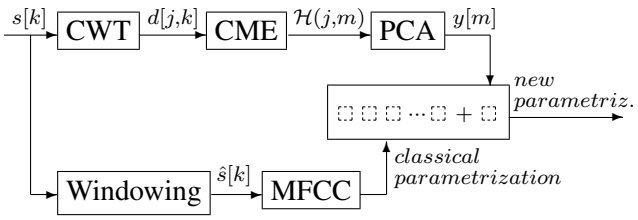

Figure 1: Schemes of the stages of the proposed method, which are explained in the text.

of speech signal for different scales, will be concatenated to a MFCC classic parametrization.

2 Materials and Methods

A modification to the classical speech signal preprocessing stage will be here outlined. The block diagram of Fig. 1 depicts each stage of the algorithm proposed in this paper, providing a guide of what follows to the reader.

2.1 Continuous Multiresolution Entropy

Given a discretized signal $s[k]$, $k = 1, \ldots, K$, its complexity measures in the time–scale plane are obtained by performing first its 'quasi-continuous" wavelet transform, $\Psi_s(a, b)$. This leads to a discretized distribution $\{d[j,k]\} \in \mathbb{R}^{J\times K}$, where $\{d[j,k]\} = \Psi_s(a = j\delta, b)$, $j = 1, \ldots, J \in \mathbb{Z}$, $\delta \in \mathbb{R}^+$ and $b = k$ is the time–control. For each fixed j the CWT coefficient's temporal evolution will be denoted as $d_j[k]$ in what follows.

We consider a set of rectangular sliding windows $W_j(m, L, \Delta) = \{d_j[k], k = l + m\Delta, l = 1, ..., L\}$, with $m = 0, 1, 2, ..., M$, which depends on two parameters width $L \in \mathbb{N}$ and shift $\Delta \in \mathbb{N}$, that are chosen such that $L \leq K$ (the signal length) and $(K - L)/\Delta = M \in \mathbb{Z}$. This 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 [7].

Over each window $W_j(m, L, \Delta)$ we consider a subset of N disjoint subintervals I_n and we denote with $p_{j,m}(I_n)$ the probability that a given $d_j[k] \in W_j(m, L, \Delta)$ belongs to one of such subintervals. Thus, for each window, a set $P[j, m]$ of N probabilities $p_{j,m}(I_n)$ is obtained:

$$P[j, m] = \{p_{j,m}(I_n),\ n = 1, \ldots, N\}, \qquad (1)$$

where m represents the time–evolution at the considered scale j.

We can compute the information measures over each window $W_j(m, L, \Delta)$ following the seminal ideas of multiresolution entropies in [3, 4]. The Shannon entropy [8] can be writen as:

$$\mathcal{H}_d[j, m] = -\sum_{n=1}^{N} p_{j,m}(I_n) \ln\left(p_{j,m}(I_n)\right),$$

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. The Shannon entropy evolution at the time–control m is a matrix $CME(a = j\delta, m) = \mathcal{H}_d[j, m]$, denoted as the continuous multiresolution entropy.

The evolution of Tsallis entropy or q–entropy [9], $q \neq 1$, computed over each window of $d_j[k]$ is:

$$\mathcal{H}^q{}_d[j, m] = (q - 1)^{-1} \sum_{n=1}^{N} (p_{j,m}(I_n) - (p_{j,m}(I_n))^q).$$

$CME_q(a = j\delta, m) = \mathcal{H}^q{}_d[j, m]$ is the corresponding continuous multiresolution q-entropy matrix.

2.2 Continuous Multiresolution Divergence

In this section, we extend the ideas of multiresolution entropy to the relative information measures. We use the Kullback-Leiber distance [10], the relative entropy associated with Shannon entropy, the q–divergence [2, 11], related with q–entropy, and the Jensen–Shannon divergence [12], which shares similar properties than the above mentioned ones.

Having in mind the probability set $P[j, m]$ mentioned above (1), corresponding to one window $W_j(m, L, \Delta)$, we consider now also a second set $R[j, m] = \{r_{j,m}(I_n),\ n = 1, \ldots, N\}$, where $r_{j,m}(I_n)$ corresponds to the probability at the consecutive window $W_j(m + 1, L, \Delta)$. In this way, the Kullback–Leiber divergence corresponding to two consecutive windows can be computed as:

$$D_d[j, m] = \sum_{n=1}^{N} p_{j,m}(I_n) \ln\left(\frac{p_{j,m}(I_n)}{r_{j,m}(I_n)}\right).$$

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

In a similar way the relative q–entropy is:

$$D^q{}_d[j, m] = \frac{1}{1 - q} \sum_{n=1}^{N} p_{j,m}(I_n) \left[1 - \left(\frac{p_{j,m}(I_n)}{r_{j,m}(I_n)}\right)^{q-1}\right]$$

and the Continuous Multiresolution q–Divergence is $CMD_q[a = j\delta, m] = D^q{}_d[j, m]$.

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
M. E. Torres, H. L. Rufiner, D. H. Milone & A. S. Cherniz, "A Multiresolution Information Measure approach to Speech Recognition"
E-Proceedings of 6th WSEAS International Conference on Signal, Speech and Image Processing (SSIP '06). pp. 187--192, Sep, 2006.

For the Jensen–Shannon divergence we have:

$$D^{JS}{}_d[j,m] = \mathcal{H}_d(\pi_P\, P[j,m] + \pi_R\, R[j,m]) - \Big(\pi_P\, \mathcal{H}_d(P[j,m]) + \pi_R\, \mathcal{H}_d(R[j,m])\Big),$$

where $\mathcal{H}_d(\cdot)$ is the Shannon entropy and π represents the weight assigned to each distribution. We obtain the Continuous Multiresolution Jensen–Shannon divergence, $\mathbf{CMD}_{JS}$ as above.

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 (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), 2(e) and 2(f) show results obtained for the same signal but corrupted with additive background conversation noise at 10dB SNR. It can be observed at 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 of this tool to the model could improve the ASR system performance in the presence of noise, making it more robust.

2.3 Different CME–based parametrization approaches

Once the multiresolution information measures are obtained, principal component analysis (PCA) is performed in order to keep a relative low dimension for the final coefficients vector. It is used here in three different ways, described in what follows for the $\mathbf{CME}$. This can be easily extended to the other multiresolution information measures.

Method 1. First PC ($\mathbf{PC}_1$): Given $\mathbf{CME}$, we obtain the matrix of principal component as:

$$\mathbf{Y} = \mathbf{Q}^T\, (\mathbf{CME})^*, \qquad (2)$$

where $(\mathbf{CME})^*$ is the statistical normalized matrix and $\mathbf{Q}$ contain the eigenvector of its correlation matrix.

The element of $\mathbf{Y}$ corresponding to the maximum eigenvalue of $\sigma_{\mathbf{CME}}$ is the principal component and we denote it as $y_1[m]$. This vector evolve with the time–control m and it will be concatenated to the classical MFCC to obtain our new parametrization.

Method 2. First and second PC ($\mathbf{PC}_{12}$): In method 1 we obtained vector $y_1[m]$ from (2). Here we also obtain the second component of $\mathbf{Y}$, $y_2[m]$, which is associated with the major eigenvalue that follow to the biggest one. Both elements, $y_1[m]$ and $y_2[m]$, are concatenated to the MFCC to generate the new parametrization vector.

Method 3. Scale dependent PC ($\mathbf{PC}_{SD}$): For this case, we consider two submatrices from $\mathbf{CME}$ to apply PCA: $\mathbf{U}^{(1)} = \{\mathbf{CME}^*[j\delta, m],\, = 1, ..., J/2\}$ and $\mathbf{U}^{(2)} = \{\mathbf{CME}^*[j\delta, m],\, j = 1 + J/2, ..., J\}$. From both we have two correlation matrices on which compute its corresponding eigenvalues, the columns of $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$, respectively. Applying (2) with each halves of subdivided matrix $\mathbf{CME}(a,m)$ we compute $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$, the corresponding principal component matrices. From each one of them we obtain $y_1^{(1)}[m]$ and $y_1^{(2)}[m]$, which are concatenated to the classic MFCC parametrization.

2.4 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 Spanish speech corpus [13].

A 3 state semi-continuous HMMs (SCHMMs) have been used for context–independent phonemes and silences. Observations probability density functions 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.Parameters tying was accomplished using a pool of 200 Gaussians for each model state. Finally the remaining reestimations have been computed in order to complete the total of sixteen. For language modeling, backing-off smoothed bigrammars have been estimated with transcriptions of the training database.

For the reference system, each phrase has been normalized in mean, pre-emphasized 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 (Δ MFCC+ΔE) [7]. For each of the methods proposed the following parametrizations were considered:

$\mathbf{PC}_1$ method: MFCC+E+y_1+ΔMFCC+ΔE+Δy_1.

$\mathbf{PC}_{12}$ method:MFCC+E+y_1+y_2+ΔMFCC+ΔE+Δy_1 +Δy_2.

$\mathbf{PC}_{SD}$ method:MFCC+E+$y_1^{(1)}$+$y_1^{(2)}$+ΔMFCC+ΔE+ +$\Delta y_1^{(1)}$+$\Delta y_1^{(2)}$.

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
M. E. Torres, H. L. Rufiner, D. H. Milone & A. S. Cherniz, "A Multiresolution Information Measure approach to Speech Recognition"
E-Proceedings of 6th WSEAS International Conference on Signal, Speech and Image Processing (SSIP '06). pp. 187--192, Sep, 2006.

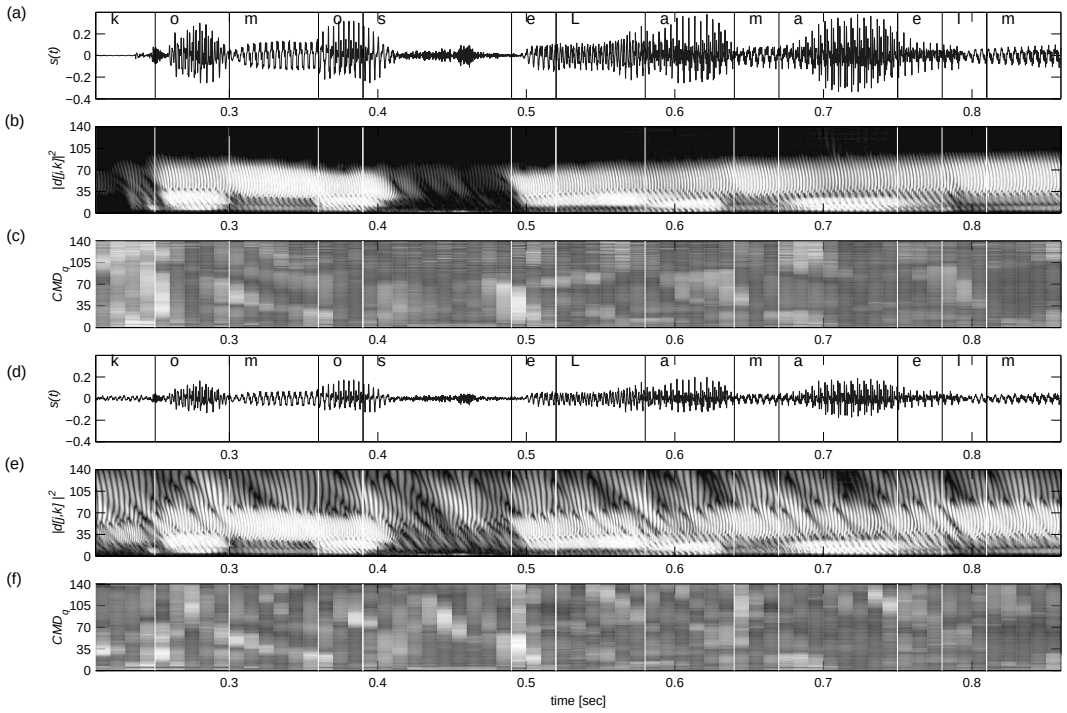

Figure 2: (a) Labeled speech signal. (b) Scalogram corresponding to the signal displayed in (a). (c) CMD_q ($q = 0.2$) of scalogram showed in (b). (d) The same signal shown in (a) with additive babble noise (10 dB SNR). (e) Scalogram corresponding to the signal displayed in (d). (f) CMD_q ($q = 0.2$) of scalogram displayed in (e).

2.5 Database and cross validation tests

A subset of the Albayzin speech corpus [14], consisting of 600 sentences, 200 words vocabulary, related to Spanish geography, was used. Speech utterances, registered in a recording study, had 3.55 secs. phrase duration average, and they were spoken by 6 males and 6 females from the central area of Spain (average age 31.8 years). In order to test robustness of the ASR system, speech signals corrupted with white and babble noise from the NOISEX-92 database were used [15]. Both noises have been mixed additively with data at different SNR levels. Data was re–sampled at 8 kHz and 16 bits of resolution.

Tests were accomplished using the *leave-k-out* cross validation method. Ten models were built and trained with different partitions on the same subset of data. For each partition, 80% of sentences have been randomly selected for system training and the remaining 20% has been used for testing. Recognition has been evaluated computing the word error rate (WER), considering as errors the word deletion and substitutions [13]. The percentage of relative error improvement has been computed as:

$$\Delta\varepsilon_{\%} = (\varepsilon_{ref} - \varepsilon)\,/\varepsilon_{ref},$$

where ε is the WER value of the different methods and ε_{ref} is the reference WER of baseline front–end.

3 Results and Discussion

We present and discuss here the results obtained while comparing the recognition obtained with methods proposed in this work and classical front–end.

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 percentage obtained with the methods PC_1, PC_{12} and PC_{SD}, when Shannon entropy is concatenated to the MFCC vector. In Fig. 3(b) q–entropy is used. A previous work [2] suggests $q = 0.2$ as an optimal value for this type of experiments. Figs. 3(c), 3(d) and 3(e) show the WER obtained with Kullback-Leiber distance, q–divergence and Jensen–Shannon divergence respectively. It suggests that the methods PC_{12} and PC_{SD} had 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

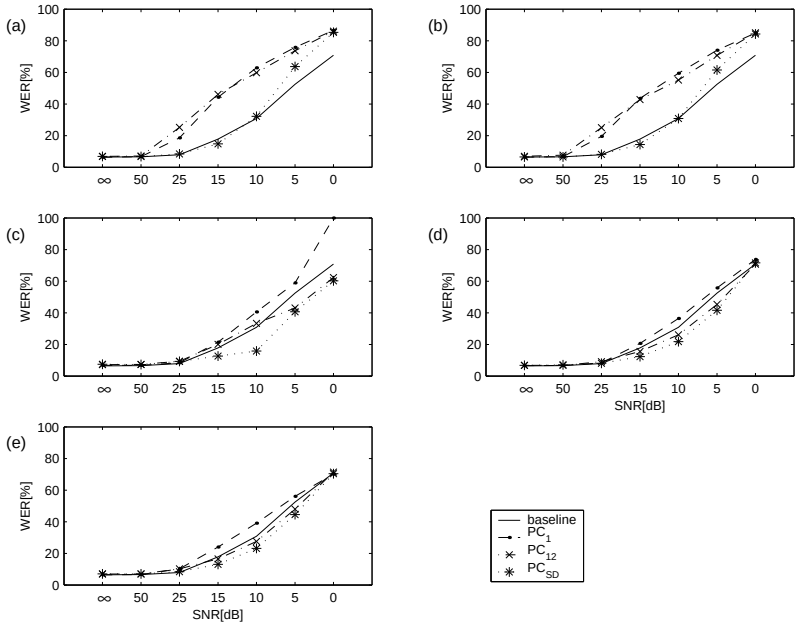

Figure 3: Word error rate of ASR system *vs.* SNR using signals corrupted with babble additive noise. Comparison between the classical pre–processing (solid line) and the proposed methods: PC_1, PC_{12}, PC_{SD}, computed with (a) Shannon entropy, (b) q–entropy, with $q = 0.2$, (c) Kullback-Leiber distance, (d) q–divergence, with $q = 0.2$, and (e) Jensen–Shannon divergence.

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 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 and 4, (d) and(e), we observe that the system's performance was similar for the three proposed methods. 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_q already observed in Fig. 2, where the structures belonging to speech signal are mainly at the lowest scales. In presence of noise the structures at the higher scales are highly modified, suggesting that babble noise information is more concentrated at these scales.

From the point of view of PCA, in method PC_1, when we only take into account one global principal component, the raw signal information and the noise information are simultaneously included, and provided in the vector of coefficients and the system cannot discriminate between them. In method PC_{12}, when first and second principal components 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 here proposed.

4 Conclusions

In this this work we have introduced information measures, computed in time–scale plane, in the parametrization of an ASR system. Methods proposed here were tested with speech signals corrupted with babble and white noise. Performance of these approaches was compared with classical MFCC parametrization. Method PC_{SD} provided an increase on 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.

The results obtained not only overcome the baseline but also those reached in a previous work of some of the authors [2], where similar information measures were used, but only in time domain.

These results suggest that these CME related measures provide valuable information to the ASR system in order to perform the recognition. Because

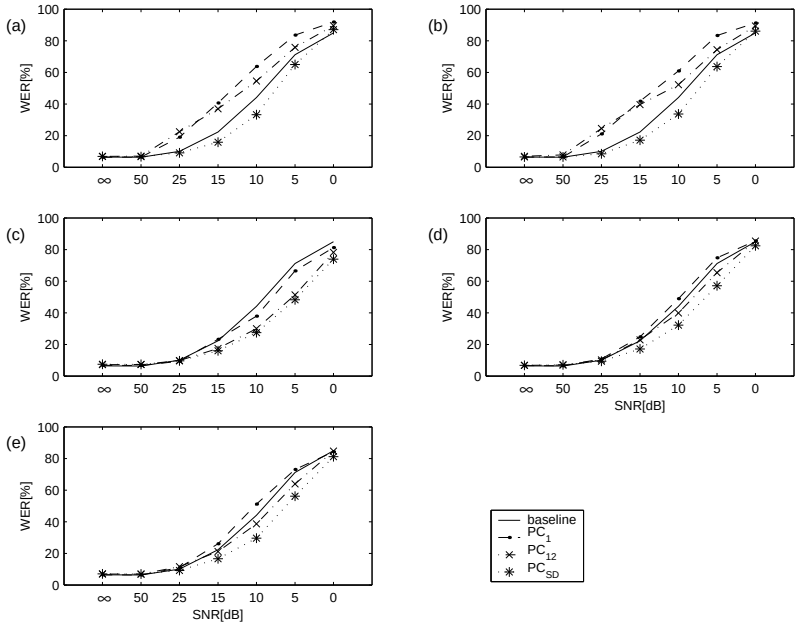

Figure 4: Word error rate of ASR system *vs.* SNR using signals corrupted with white additive noise. Comparison between the classical pre–processing (solid line) and the proposed methods: PC_1, PC_{12}, PC_{SD}, computed with (a) Shannon entropy, (b) q–entropy, with $q = 0.2$, (c) Kullback-Leiber distance,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

of that they could be consided to be included as an extra component in a pre–processing stage.

References:

[1] O. Viiki, editor, Noise Robust ASR, Speech communication, Special Issue, 34, 1–2, 2001.

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

[3] M. E. Torres, L. Gamero, and E. D'Attellis, Latin American Applied Research **53**, 53 (1995).

[4] M. E. Torres, L. Gamero, P. Flandrin, and P. Abry, in *SPIE'97 Wavelet Applications in Signal and Image Processing V*, edited by A. F. L. Akram Aldroubi and M. Unser (SPIE Int. Soc. for Optical Engineering, Washington, 1997), Vol. 3169, pp. 400–407.

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

[6] H.M.Torres, J.A.Gurlekian, H.L.Rufiner, and M.E.Torres, Physica A **337**, (2005).

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

[8] C. Shannon, The Bell System Technical Journal **27**, 379 (1948).

[9] C. Tsallis, Chaos, Solitons and Fractals **6**, 539 (1995).

[10] T. M. Cover and J. A. Thomas, *Information Theory* (John Wiley and Sons, NY, 1991).

[11] M. E. Torres, Ph.D. thesis, Universidad Nacional de Rosario - Argentina, 1999, (Math. D. Thesis).

[12] I. Grosse *et al.*, Physical Review E **65**, 1 (2002).

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

[14] J. E. D. Verdejo *et al.*, in *Proceedings of the First International Conference on Language Resources and Evaluation* (European Language Resources Association, Granada, 1998), Vol. 1, pp. 497–502.

[15] A. Varga and H. Steeneken, Speech Communication **12**, 247 (1993).