# Multiresolution Information Measures applied to Speech Recognition

María E. Torres[1], Hugo L. Rufiner[2,3], Diego H. Milone[3], and Analía S. Cherniz[1,2]

[1] Laboratorio de Señales y Dinámicas no Lineales
Universidad Nacional de Entre Ríos, Facultad de Ingeniería,
C.C. 47 Suc. 3 - 3100 Paraná, Entre Ríos, Argentina
metorres@ceride.gov.ar, acherniz@bioingenieria.edu.ar
[2] Laboratorio de Cibernética
Universidad Nacional de Entre Ríos, Facultad de Ingeniería,
C.C. 47 Suc. 3 - 3100 Paraná, Entre Ríos, Argentina
lrufiner@bioingenieria.edu.ar
[3] Laboratorio de Señales e Inteligencia Computacional
Universidad Nacional del Litoral, Santa Fe, Argentina
d.milone@ieee.org

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

**Abstract.** Considerable advances in automatic speech recognition have been made through application of hidden Markov models and development of different speech signal parametrization techniques. However, deterioration in the speech recognizers performance have been observed when systems trained with clean signals are tested with noisy signals. In this paper we present the extension of the continuous multiresolution entropy to different divergences and we introduce these information measures as new dimensions to the front–end stage of an ASR system. These new parameters take into account information about the changes in the dynamics of speech signal for different scales and is concatenated to a MFCC classic parametrization. Proposed methods are tested with speech signals corrupted with babble and white noise and compared with a classical mel cepstra parametrization. Results suggest that multiresolution information measures provide valuable information to the speech recognition system and could be included as an extra component in the pre–processing stage.

## 1 Introduction

Automatic speech recognition (ASR) has been an active field in the passed two decades, where Hidden Markov (HMM) have lead to high performance levels in the area of acoustic and language modeling and different techniques have been applied in the field of speech signal analysis [1]. In those experiments where production and perception features of human speech, such as linear predictive

[1] This work is supported by A.N.P.C.yT., under Project PICT N° 11-12700.

coding (LPC) [1], cepstral and mel frequency cepstral coefficients (MFCC) [2] were taken into account in the model, the best results have been obtained [3].

An important performance deterioration is observed when the ASR system is trained with clean speech signals or speech recorded with high quality audio systems and then it is tested with signals registered with simple home microphones or with added noise [4, 5]. 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, lost in the transmition channels, home quality audio system, etc. Robust speech recognition investigation is oriented in two main areas: techniques based on transformation of speech signal in the feature space (pre-processing) and adaptation of models to noise or particular enviroment conditions [6].

There are several pre-processing methods to improve the ASR system's performance [4], where often it is supposed that both signal and noise are generated by linear systems and that noise have special features allowing to 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 [7–9]. The application of these quantitative measures provides information about the dynamics of the underlying 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 an ASR system, providing information of the temporal evolution of the complexity degree of speech signals, improving its performance [10]. Their use has been extended over different time–scale distributions [11] and spectral entropy has been used in speech processing for relatively simple tasks like segmentation and silence detection [12, 13].

The multiresolution entropy, proposed by Torres et al. in [14], 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 [15]. Combined with the continuous wavelet transform (CWT) [16], 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 corresponding to physiological signals [17–19]. Good results have also been obtained in applications to speech signals corrupted with additive noise in experiments of self-organizing map clustering [20]. 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 of speech signal for different scales, provided by different multiresolution entropies or divergences, will be concatenated to a MFCC classic parametrization.

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

## 2 Materials and methods

In this section we introduce the main characteristics of both, the ASR baseline system and multiresolution entropy, used in this paper. A modification to the classical speech signal pre-processing stage will be here outlined.

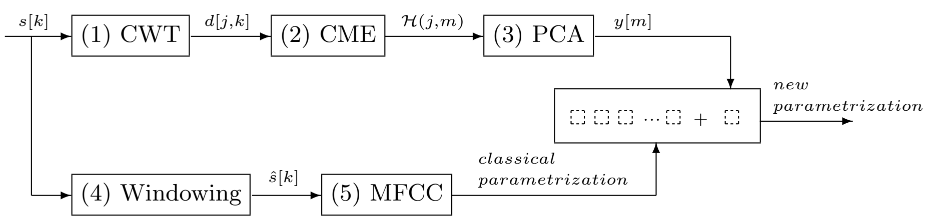


**Fig. 1.** Schemes of the stages of the proposed method, which are explained in the text.

The block diagram showed in Fig. 1 depicts each stage of the algorithm here proposed, providing a guide of what follows to the reader. Starting from the sampled speech signal two branches are open. In one of them the classic MFCC parametrization are obtained. In the other branch the continuous multiresolution entropy is computed. 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 divergences (or relative entropies) and the Jensen-Shannon divergence. PCA is applied to extract the temporal components of higher variance. The obtained values are concatenated as new dimensions in the MFCC parametrization. The performance of this new approach will be compared to the classical front–end in the presence of different noisy conditions.

### 2.1 Continuous Multiresolution Entropy

Given $s(t)$, a continuous signal, its complexity measures are obtained in the time–scale plane by performing first its continuous wavelet transform [21]. 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 a piecewise constant interpolation is used [22], which translates it as:

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

We denote as $\{d[j, k]\} \in \mathbb{R}^{J \times K}$ the discretized distribution in the time–scale plane, obtained with $d[j, k] = \Psi_s(a = j\delta, t = k)$, with $j = 1, \ldots, J$, $J \in \mathbb{Z}$, $\delta \in \mathbb{R}^+$ and $b = k$. That corresponds to the so named "quasi-continuous" wavelet

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

transform. For each fixed $j$ the CWT coefficient's temporal evolution will be named as $d_j[k]$ in what follows.

Let us consider $\mathcal{W} = \{W(m, L, \varDelta), m = 0, 1, 2, ..., M\}$, a set of rectangular sliding windows which depends on two parameters, width $L \in \mathbb{N}$ and shift $\varDelta \in \mathbb{N}$, and with:

$$W_j(m, L, \varDelta) = \{d_j[k], k = l + m\varDelta,\ l = 1, ..., L\}\,,\quad m = 0, 1, 2, ..., M, \qquad (2)$$

where $L$ and $\varDelta$ are chosen such that $L \leq K$ (the signal length) and $(K-L)/\varDelta = M \in \mathbb{Z}$. 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 [2].

An equipartition $d_j^0 = \min_k\{d_j[k]\} < d^1 < \cdots < d^{N-1} < d^N = \max_k\{d_j[k]\}$ is considered over each window $W_j(m, L, \varDelta)$, providing a subset of $N$ disjoint subintervals: $I_n = \left\{\left[d^{n-1}, d^n\right), n = 1, ..., N\right\}$, such that $W_j(m, L, \varDelta) = \overline{\bigcup_{n=1}^{N} I_n}$.

Let us denote with $p_{j,m}(I_n)$ the probability that a given $d_j[k] \in W_j(m, L, \varDelta)$ belongs to the interval $I_n$. Therefore, for each window $W_j(m, L, \varDelta)$ a set $P[j, m]$ of $N$ probabilities $p_{j,m}(I_n)$ is obtained:

$$P[j, m] = \{p_{j,m}(I_n),\ n = 1, \ldots, N\}. \qquad (3)$$

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

We can compute the information measures over each window $W_j(m, L, \varDelta)$ following the seminal ideas of multiresolution entropies in [14, 16]. The Shannon entropy [23] can be writen as:

$$\mathcal{H}_d[j, m] = -\sum_{n=1}^{N} p_{j,m}(I_n) \ln\left(p_{j,m}(I_n)\right) \qquad m = 0, 1, 2, \ldots, M. \qquad (4)$$

At each fixed scale $j$ and for each fixed $m$, the entropy value corresponding to the wavelet coefficients on the window $W_j(m, L, \varDelta)$ is computed. Observe that $\{\mathcal{H}_d[j, m],\ m = 0, 1, 2, \ldots, M\}$ represents the Shannon entropy evolution at the time–control $m$. This is a matrix that will be denoted as **CME**, where $CME[a = j\delta, m] = \mathcal{H}_d[j, m]$, named as the continuous multiresolution entropy.

The Tsallis entropy or $q$–entropy depends on a real parameter $q \neq 1$ [24]. The evolution of this entropy for $d_j[k]$, computed over each window $W_j(m, L, \varDelta)$, is:

$$\mathcal{H}^q{}_d[j, m] = (q-1)^{-1} \sum_{n=1}^{N} (p_{j,m}(I_n) - (p_{j,m}(I_n))^q), \qquad (5)$$

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

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

### 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 [25], the relative entropy associated with Shannon entropy, the $q$–divergence [10, 26], related with $q$–entropy, and the Jensen–Shannon divergence [27], which shares similar properties than the above mentioned ones.

Having in mind the probability set $P[j,m]$ above mentioned (3), 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 next window $W_j(m+1, L, \Delta)$. In this way, the Kullback–Leiber divergence corresponding to two consecutive windows can be computed as:

$$D_d[j,m] = \sum_{n=1}^{N} p_{j,m}(I_n) \ln \left( \frac{p_{j,m}(I_n)}{r_{j,m}(I_n)} \right) . \tag{6}$$

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

In similar way the relative $q$–entropy is computed as:

$$D^q{}_d[j,m] = \frac{1}{1-q} \sum_{n=1}^{N} p_{j,m}(I_n) \left[ 1 - \left( \frac{p_{j,m}(I_n)}{r_{j,m}(I_n)} \right)^{q-1} \right] \tag{7}$$

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 corresponding Jensen–Shannon divergence reads as:

$$D^{JS}{}_d[j,m] = \mathcal{H}_d(\pi_P\, P[j,m] + \pi_R\, R[j,m]) - (\pi_P\, \mathcal{H}_d(P[j,m]) + \pi_R\, \mathcal{H}_d(R[j,m])) \,, \tag{8}$$

where $\mathcal{H}_d(\cdot)$ is the Shannon entropy and $\pi$ represents the weight assigned to each distribution. We obtain the $\mathbf{CMD}_{JS}$ 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 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), 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.

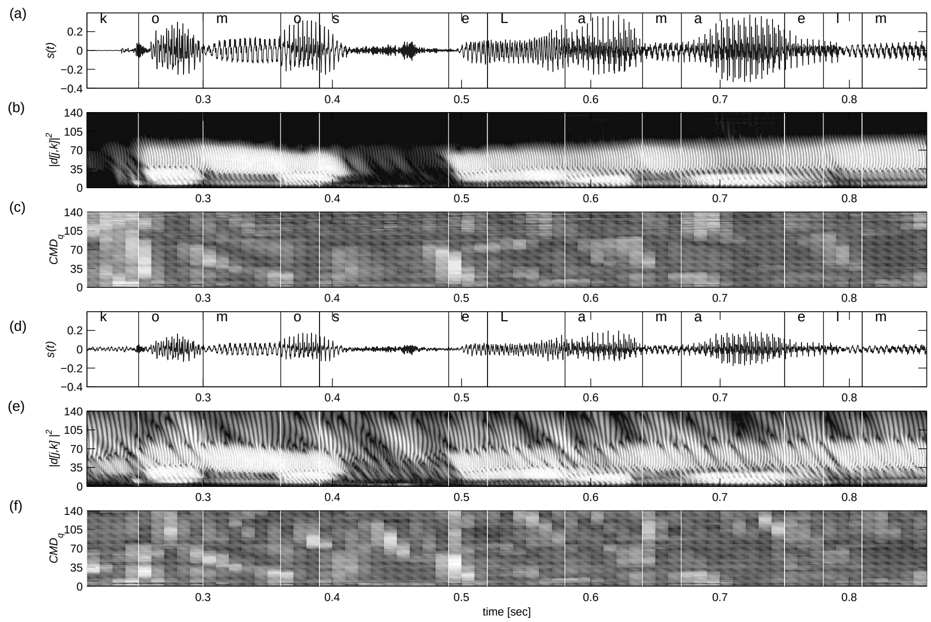


**Fig. 2.** (a) Labeled speech signal. (b) Scalogram corresponding to the signal displayed in (a). (c) $CMD_q$ ($q = 0.2$) of scalogram showed in (b). (d) The same signal shown in (a) with additive babble noise (10 dB SNR). (e) Scalogram corresponding to the signal displayed in (d). (f) $CMD_q$ ($q =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

### 2.3 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, principal component analysis (PCA) is performed in order to keep a relative low dimension for the final coefficients vector. PCA is a statistical method used for data analysis, feature extraction and compression [28]. It is used here in three different ways, described in what follows for the **CME**. This can be easily extended to the other multiresolution information measures.

*Method 1. First PC ($PC_1$):* Given **CME**, we obtain the matrix of principal component as:

$$\mathbf{Y} = \mathbf{Q}^T \mathbf{CME}^*, \tag{9}$$

where $\mathbf{CME}^*$ is the statistical normalized matrix. $\mathbf{Q}$ is the eigenvector matrix of the correlation matrix of data $\sigma_{\mathbf{CME}} = \mathbf{U}\mathbf{U}^T$, with $\mathbf{U} = \mathbf{CME}^*$.

The row 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 value evolves with the time–control $m$ and it will be concatenated to the classical MFCC to obtain our new parametrization.

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

*Method 2. First and second PC ($PC_{12}$):* In method 1 we obtained vector $y_1[m]$ from (9). Here we also obtain the second component of $\mathbf{Y}$, $y_2[m]$, which is associated with the major eigenvalue that follows to the biggest one. Both elements, $y_1[m]$ and $y_2[m]$, 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 this with the corresponding to clean signal (Fig. 2(a)) can be observed new structures, which are related with noise. These structures appear almost separately from the corresponding to 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.

*Method 3. Scale dependent PC ($PC_{SD}$):* For this case, we consider two submatrices from **CME** to apply PCA: $\mathbf{U}^{(1)} = \{CME^*[j\delta, m],\ j = 1, ..., J/2\}$ and $\mathbf{U}^{(2)} = \{CME^*[j\delta, m],\ j = 1 + J/2, ..., J\}$. From both we have two correlation matrices on which we compute their corresponding eigenvalues, the columns of $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$, respectively. Applying (9) with each halves of subdivided matrix **CME** we compute $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$, the corresponding principal component matrices. From each of them we obtain $y_1^{(1)}[m]$ and $y_1^{(2)}[m]$, which are concatenated to the classic MFCC parametrization.

Under the same assumption made in Method 2, it is observed from 2(b) that the structures related with noise appear mainly at high scales, while the corresponding to signal are in lower scales. Therefore, this two components would give us information about signal and noise relatively in a separated way. These three procedures are accomplished alike over the other information measures.

### 2.4 Automatic speech recognition experiments

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

*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 [29]. In what follows we briefly describe its characteristics and the procedure to obtain such reference system.

A 3 state semi-continuous HMMs (SCHMMs) have been used for context–independent phonemes and silences [1]. 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 [30]. Parameters tying was accomplished using a pool of 200 Gaussians for each model state. Tied mixtures reduces the total effective amount of parameters from 855000 to 26200. This stage is necessary in order to improve the estimation robustness because of the reduced training set used [31]. Finally, the remaining reestimations have been computed in order to complete

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

the total of sixteen. For language modeling, backing-off smoothed bigrams [30] 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 ($\Delta$ MFCC+$\Delta$E) [2].

For each of the methods explained in Section 2.3 new coefficients were concatenated to the classical front–end, which incorporates information about dynamic changes of speech signal in the time–scale plane. Therefore, the following parametrizations were considered:

- $\mathrm{PC}_1$ method: 12 MFCC, in order to maintain the number of coefficients of the reference front–ends, 1 energy coefficient and the coefficient obtained from PCA over each segment, $y_1$, and its respective temporal derivatives. Thus, our approximation stay: MFCC+E+$y_1$ + $\Delta$MFCC+$\Delta$E+$\Delta y_1$.
- $\mathrm{PC}_{12}$ method: 11 MFCC, the energy coefficient and both information measure coefficients, with its corresponding derivatives. This is: MFCC+E+$y_1$ + $y_2$ + $\Delta$MFCC+$\Delta$E+$\Delta y_1$ + $\Delta y_2$.
- $\mathrm{PC}_{SD}$ method: 11 MFCC, energy coefficient, the two values corresponding to information measures at both, low and high scales and their temporal derivatives: MFCC+E+$y_1^{(1)} + y_1^{(2)}$ + $\Delta$MFCC+$\Delta$E+$\Delta y_1^{(1)} + \Delta y_1^{(2)}$.

*Database and cross validation tests:* A subset of the Albayzin speech corpus [32] was used for the experiments. This subset consists of 600 sentences, with a vocabulary of 200 words, concerning Spanish geography. 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). Data was re–sampled at 8 kHz and 16 bits of resolution.

In order to test robustness of the ASR system, speech signals corrupted with white and babble noise from the NOISEX-92 database were used [33]. White noise has been digitalized from an 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.

Tests were accomplished using the *leave-k-out* cross validation method [34]. Ten models have been build and trained, using different partitions on the same subset of speech 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 [29]. The percentage of relative error improvement of the different measures compared with the baseline front–end has been computed as:

$$\Delta\varepsilon_{\%} = \frac{\varepsilon_{ref} - \varepsilon}{\varepsilon_{ref}}, \tag{10}$$

where $\varepsilon$ is the WER value and $\varepsilon_{ref}$ is the reference WER.

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

## 3 Results and Discussion

As explained in 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.

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 $\mathrm{PC}_1$, $\mathrm{PC}_{12}$ and $\mathrm{PC}_{SD}$, when Shannon entropy is concatenated to the MFCC vector. In Fig. 3(b) $q$–entropy is used. A previous work [10] 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 $\mathrm{PC}_{12}$ and $\mathrm{PC}_{SD}$ 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 $\mathrm{PC}_{SD}$ is applied, its WER 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

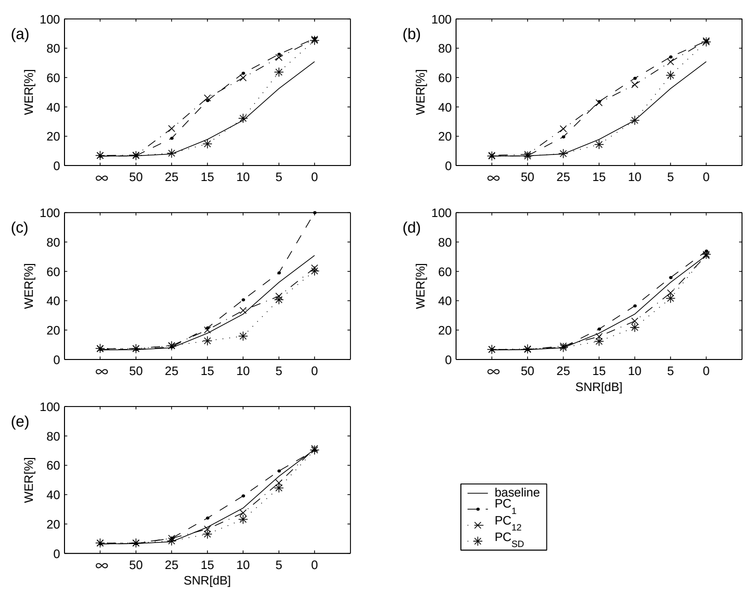


**Fig. 3.** Word error rate of ASR system *vs.* SNR using signals corrupted with babble additive noise. Comparison between the classical pre–processing (solid line) and the proposed methods: $\mathrm{PC}_1$, $\mathrm{PC}_{12}$, $\mathrm{PC}_{SD}$, computed with (a) Shannon entropy, (b) $q$–entropy, with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

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

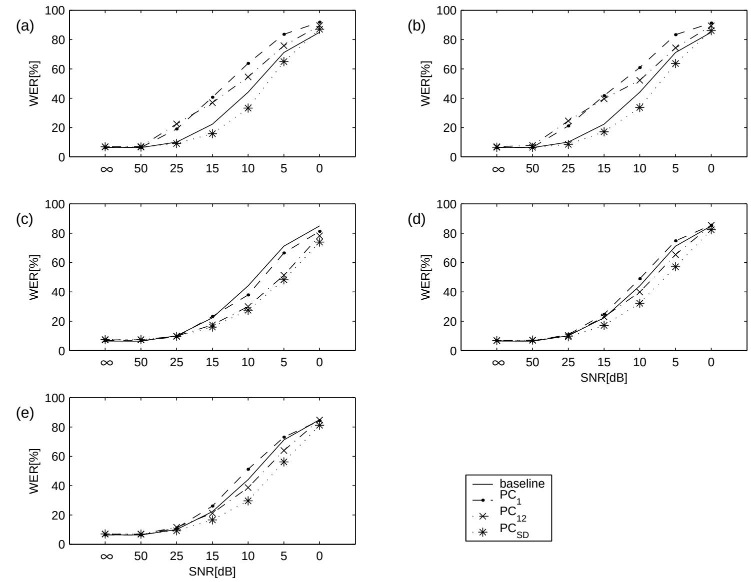


**Fig. 4.** Word error rate of ASR system *vs.* SNR using signals corrupted with white additive noise. Comparison between the classical pre–processing (solid line) and the proposed methods: $\mathrm{PC}_1$, $\mathrm{PC}_{12}$, $\mathrm{PC}_{SD}$, computed with (a) Shannon entropy, (b) $q$–entropy, with $q = 0.2$, (c) Kullback-Leiber distance, (d) $q$–divergence, with $q = 0.2$, and (e) Jensen–Shannon divergence.

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 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.

These results suggest that method $\mathrm{PC}_{SD}$ 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 $\mathrm{PC}_1$, when we only take into account one global principal component, the raw signal information and the

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

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 informations.

The results obtained not only overcome the baseline but also those reached in our previous work [10], where similar information measures were used, but only in time domain.

These results suggest that this CME related measures give valuable information to the ASR system in order to perform the recognition. Because of that they would be consided to be included as an extra component in the pre–processing stage.

## References

1. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, New Jersey, 1993).
2. J. Deller, J. Proakis, and J. Hansen, *Discrete Time Processing of Speech Signals* (Macmillan Publishing, New York, 1993).
3. M. Cooke and D. P. Ellis, Speech Comunication **35**, 141 (2001).
4. J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications* (Kluwer Academic Publishers, Boston, 1996).
5. O. Viiki, editor, Noise Robust ASR, Speech communication, Special Issue, 34, 1–2, 2001.
6. *Noise Reduction in Speech Applications*, edited by G. M. Davis (CRC Press, USA, 2002).
7. P. Saparin, A. Witt, J. Kurths, and B. Anishchenko, J. Chaos, Solitons and Fractals **4**, 1907 (1994).
8. J. S. Richman and J. R. Moorman, Am J Physiol Heart Circ Physiol **278**, 2039– (2000).
9. X.-S. Zhang, R. J. Roy, and E. W. Jensen, IEEE Transactions on Biomedical Engineering **48**, 1424 (2001).
10. H. L. Rufiner, M. E. Torres, L. Gamero, and D. H. Milone, Physica A **332**, 496 (2004).

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.

11. M. Costa, A. L. Goldberger, and C.-K. Peng, Physical Review E **71**, 1 (2005).
12. L. S. Huang and C. H. Yang, in *Proceedings of the 2000 International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Istanbul, Turkey, 2000), Vol. 3, pp. 1751–1754.
13. P. Renevey and A. Drygajlo, in *Proceedings of 7th European Conference on Speech Communication and Technology* (International Speech Communication Association, Aalborg, Denmark, 2001), pp. 1887–1890.
14. M. Torres, L. Gamero, and E. D'Attellis, Latin American Applied Research **53**, 53 (1995).
15. L. G. Gamero, A. Plastino, and M. E. Torres, Physica A **246**, 487 (1997).
16. 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.
17. M. M. Añino, M. E. Torres, and G. Schlottahauer, Physica A **324**, 645 (2003).
18. M. E. Torres, M. M. Añino, L. G. Gamero, and M. A. Gemignani, Int. J. of Bifurcations and Chaos. **11**, 967 (2001).
19. M. E. Torres and L. G. Gamero, Physica A **286**, 457 (2000).
20. H.M.Torres, J.A.Gurlekian, H.L.Rufiner, and M.E.Torres, Physica A **361**, 337 (2005).
21. S. Mallat, *A Wavelet Tour of Signal Processing* (Academic Press, New York, USA, 1999).
22. M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, *Wavelet Toolbox for Use with MATLAB*, 1 ed., The MathWorks, Inc., http://www.mathworks.com, 1996, user's Guide.
23. C. Shannon, The Bell System Technical Journal **27**, 379 (1948).
24. C. Tsallis, Chaos, Solitons and Fractals **6**, 539 (1995).
25. T. M. Cover and J. A. Thomas, *Information Theory* (John Wiley and Sons, NY, 1991).
26. M. E. Torres, Ph.D. thesis, Universidad Nacional de Rosario - Argentina, 1999, (Math. D. Thesis).
27. I. Grosse *et al.*, Physical Review E **65**, 1 (2002).
28. M. Akay, *Detection and Estimation Methods for Biomedical Signals* (Academic Press, San Diego, CA, 1996).
29. S. Young *et al.*, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk, 2002.
30. F. Jelinek, *Statistical Methods for Speech Recognition* (The MIT Press, Cambrige, Masachussets, 1999).
31. S. Young, in *IEEE Workshop on Speech Recogniton* (IEEE Press, Snowbird, Utah, 1995).
32. 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.
33. A. Varga and H. Steeneken, Speech Communication **12**, 247 (1993).
34. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural and Statistical Classification* (Ellis Horwood, University College, London, 1994).

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; "Multiresolution Information Measures applied to Speech Recognition"
7th Argentine Symposium on Computing Technology. 2006.