Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Physica A ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮
PHYSICA A
www.elsevier.com/locate/physa

Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition☆

Hugo L. Rufiner*, María E. Torres, Lucas Gamero, Diego H. Milone

Universidad Nacional de Entre Ríos, Facultad de Ingeniería, C.C. 47 Suc. 3, Paraná (E.R.) 3100, Argentina

Received 2 May 2003; received in revised form 1 July 2003

Abstract

Information measures have been used in the context of nonlinear systems presenting abrupt complexity changes and related to nonlinear time series analysis. In this study, complexity measures such as Shannon entropy, q-entropy and their associated divergences have been added to a robust speech recognizer front-end. The method proposed here is tested on continuous speech and compared with a classical mel-cepstral analysis. The recognition degradation has been evaluated in both systems in presence of white and babble noise. The results suggest that complexity measures provide additional valuable information for speech recognition in noisy conditions.

PACS: 43.72.+q; 02.50.−r; 05.90.+m; 87.80.−y

Keywords: Complexity measures; Entropy; Cepstral analysis; Robust speech recognition

1. Introduction

In the last two decades the analysis of time series obtained from experimental data, where the underlying dynamics is not well known, has been a field of active research.

☆ This work is supported by U.N.E.R. under Projects PID No. 6036, 6060 and 6062.

* Corresponding author.
E-mail address: lrufiner@fi.uner.edu.ar (H.L. Rufiner).

0378-4371/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.physa.2003.09.050

A rich variety of self-oscillating regimes that involve either regular or complex behavior are present [1]. Several notions of entropy have been used to characterize the complexity degree in nonlinear physiological signals. The application of quantitative measures for the analysis of such time series has helped to gain a better understanding of the system dynamics [2].

Linear autoregressive models have been used in the first attempts to obtain a speech synthesis system. In most of the physical models proposed to represent human speech a stationary behavior is assumed on those segments corresponding to voiced phonemes, as vowel and nasal consonants. A suitable model for unvoiced sounds demands the inclusion of random components. In other studies, voiced speech was well characterized by nonlinear and low dimensional behavior [3]. Compared with the pronunciation of isolated and well-pronounced words, signal complexity is increased in natural speech. This is due to the way in which transitions appear between different phonemes in continuous speech, for example, during the transition from a nasal consonant to a vowel due to nasal co-articulation [4]. Some nonlinearities like those produced by radiation at the lips level or turbulence during constrictions [5] are not able to be assessed by a linear model, although this type of model is broadly used.

Considerable progress has been made in the last two decades in *automatic speech recognition* (ASR). The incorporation of *hidden Markov models* (HMM) led to very high levels of performance, mainly thanks to the way this technique enables speech time variability modelization. Research using various HMM paradigms have resulted in the incorporation of a number of features that attempt to model human speech perception and production. In the field of acoustic modeling several speech parameterization techniques were applied, and important advances have been made, such as *linear predictive coding* [6], *cepstral* and *mel frequency cepstral coefficients* (MFCC) with delta and acceleration coefficients [7]. Complexity measures have been previously used for easier tasks than ASR, such as end-point detection or segmentation. In these works, entropy has been usually computed in frequency domain and referred to as *spectral entropy* [8,9].

An important performance deterioration of ASR systems is observed when trained with clean speech and tested with noisy speech [10]. In this case the reduction of their capacities can lead to increased error up to 80%. Similarly, when the ASR system is trained with speech registered by means of a high-quality audio system and is then tested with a simple home microphone, the errors can grow up to 50%. This is the area of “robust” speech recognition research. Thus, the main objective is to obtain ASR systems that can be used in real environments, with noise, reverberation, losses in the transmission channel, home quality audio systems, etc. There are two main approaches that guide this research: techniques based on the transformation of speech into the feature space and techniques based on speech model adaptation to noise. In many cases the noise possesses very special characteristics that allow it to be modeled easily and to significantly improve the ASR system’s performance. Nevertheless, often the noise consists of other voices in a conversation and, specially for low *signal-to-noise ratio* (SNR), the problem is still open. For example, Fig. 1(a) shows a part of a labeled speech signal of sentence: “Cómo se llama el mar que baña Valencia?” (What is the name of the sea that border Valencia?). In Fig. 1(b) the corresponding spectrogram

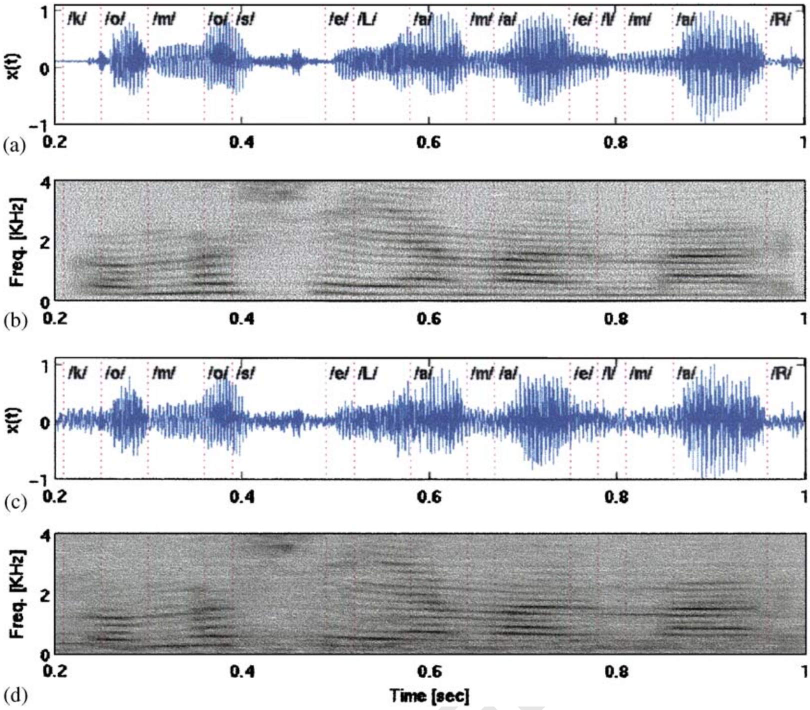

Fig. 1. (a) Labeled speech signal. (b) $|H(f,t)|$ corresponding to the signal displayed in (a). (c) The same signal shown in (a) with additive white noise (20 dB). (d) $|H(f,t)|$ corresponding to the signal displayed in (c).

analysis of this clean speech signal can be observed. Figs. 1(c) and (d) show the same signal but contaminated with a pub conversation at 20 dB SNR. The clean spectrogram shows better spectral contrast that allows an easier indentification of important acoustic clues for phoneme recognition.

In this work a new approach to the front-end stage of an ASR system is introduced. One dimension, that takes into account the dynamical changes of speech signal, is added to a classical MFCC parameterization. This dimension is computed by means of the evaluation of a temporal complexity measure. The Shannon entropy and the more general Harvda–Charvat–Daróvczy–Tsallis [11–13] (q-entropies) and their corresponding relative informations are considered. We contrast the performance of a classical ASR system front-end and the ones obtained including these measures in the presence of noise.

This paper is organized as follows. In Section 2 the materials and methods are introduced. The results obtained are discussed in Section 3 and the conclusions are presented in Section 4. Computational aspects concerning the information measures are considered here and also some basic speech parametrization elements are discussed in Appendix A.

2. Materials and methods

In this section we briefly review the complexity measures used in this paper and the ASR system used in this study. Technical details and description of the data used to perform the experiments are given.

2.1. Complexity measures

2.1.1. Entropies

Given a signal x, its Shannon entropy is defined as

$$\mathscr{H} = -\sum_{i=1}^{M} p_i \ln(p_i) , \tag{1}$$

where p_i is the probability that the signal belongs to a considered interval, with the understanding that $p\ln(p) = 0$ if $p = 0$, and M is the number of partitions [14]. The entropy $\mathscr{H}$ is a measure of the information needed to locate a system in a certain state, meaning that $\mathscr{H}$ is a measure of our ignorance about the system.

The q-entropy [11–13], that depends on a single real parameter $q \neq 1$, reads as:

$$\mathscr{H}_q = (q-1)^{-1} \sum_{i=1}^{M} (p_i - p_i^q) . \tag{2}$$

Applications of q-entropy in the context of nonlinear dynamical systems have been recently introduced to the detection of slight changes on the system's parameter by the analysis of the complex biological signal [15,16].

2.1.2. Relative entropies

The relative entropy (or Kullback–Leiber distance) $D(f|g)$ between two probability densities f and g is defined by

$$D(f|g) = \int_X f(x) \ln(f(x)/g(x))\, \mathrm{d}x \tag{3}$$

with the understanding that $y\ln(y) = 0$ if $y = 0$ [17]. In the q-entropies case for $q \in \mathbb{R} - \{1\}$, the corresponding relative q-entropies are given by [18]

$$D_q(f|g) = (1-q)^{-1} \int_X f(x)(1 - (f(x)/g(x))^{q-1})\, \mathrm{d}x . \tag{4}$$

In our case, given two probabilities p_i and r_i, corresponding to different frames of the same signal, the discrete versions read as

$$D(p|r) = \sum_{i=1}^{M} p_i \ln\left(\frac{p_i}{r_i}\right) \tag{5}$$

and

$$D_q(p|r) = \frac{1}{1-q} \sum_{i=1}^{M} p_i \left[1 - \left(\frac{p_i}{r_i}\right)^{q-1}\right] . \tag{6}$$

The presented complexity measures $\mathscr{H}$, D, $\mathscr{H}_q$ and D_q, with different values of q, and have been explored in this work inside an ASR framework. For computational details we refer the reader to Appendix A. For more comprehensive discussions, see e.g. [11,12,15,19]. In particular, in Ref. [15], these measures have been applied to the detection of dynamical changes in complex biological signals.

2.2. Automatic speech recognition

A "state of the art" multi-speaker ASR system for continuous speech in Spanish was built, based on HMM. It is our baseline reference for comparisons with the alternative proposed here. We briefly present here the most relevant aspects of this baseline system used in the experiments. For computational details we refer the reader to Appendix A. For further details, and due to the extent of this topic we remit readers not familiarized with ASR to [20].

The "front-end" (or speech parametrization) used here was based on classical MFCC analysis. A three state semi-continuous HMMs (SCHMMs) have been used for context-independent phonemes and silences [6]. 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 [21]. Parameters tying was accomplished using a pool of 200 Gaussians for each model state. Tied mixtures reduces 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 [22]. Finally the remaining reestimations have been computed in order to complete the total of 16. For language modeling, *backing-off* smoothed bigrammars [21] 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 has been parameterized with 28 coefficients: 13 MFCC, 1 energy coefficient (E) and their temporal derivatives ($\Delta\mathrm{MFCC} + \Delta E$) [7]. For the alternative front-end proposed here, we considered 12 MFCC, 1 energy coefficient and their temporal derivatives. One coefficient related to a given complexity measure of the speech segment or "frame", and its temporal derivative have been added to obtain the same feature vector dimension of the reference system (see Appendix A for more details). These two new coefficients incorporate information about the dynamical changes and complexity of the speech signal.

2.3. Database and cross validation tests

A subset of the Albayzin speech corpus [23] was used for the experiments. This subset consists of 600 sentences concerning Spanish geography, and a vocabulary size of 200 words. The speech utterances from this corpus, registered in a recording study, had 3.55 s phrase duration average, and they were spoken by six males and six females from the central area of Spain (average age 31.8 years). The data was digitized at 8 KHz, 16 bits and a μ-law sampling has been used.

In order to determine the system robustness, once trained with clean data, it has been tested with speech contaminated with different kinds of noise. White and babble noise of the NOISEX-92 database were used [24]. The white noise data was digitized from a high-quality analog noise generator. The source of background conversation noise data (babble) was 100 persons speaking in a pub. The noise has been re-sampled to 8 KHz and has been mixed with speech data at different SNRs.

Tests were accomplished using the *leave-k-out* cross-validation method [25] with 10 different partitions on the same subset of speech data. For each partition, the 20% of sentences not used during the training stage, are randomly selected to play as the test set. For recognition error measurement, the *word error rate* (WER) was evaluated, considering as errors the words deletions and substitutions [26]. The percentage of relative error improvement of the different measures compared with the baseline front-end has also been computed

$$\Delta\varepsilon_{\%} = \frac{\varepsilon_{ref} - \varepsilon}{\varepsilon_{ref}} \times 100 \,,$$

where ε is the WER value.

3. Results and discussion

In this section we present and discuss the most relevant results obtained. Fig. 2(a) shows the same speech signal of Fig. 1(a). The D_q evolution is displayed in Fig. 2(b). We can appreciate that there is a correspondence between the divergence variations and the phonemic changes. In Fig. 2(c) the same signal is shown but now corrupted with additive white noise (20 dB) and Fig. 2(d) is the corresponding D_q evolution. The peaks are located in similar positions with respect to the ones obtained for the clean signal. This suggests a certain robustness of this measure, at least for this SNR level. Similar results have been reported in Ref. [8], using Shannon's spectral entropy in end-point speech detection. In other biological signals robustness has been improved by means of multiresolution analysis [27].

The WER for the ASR systems described below has been computed. Figs. 3 and 4 present the results obtained for signals corrupted with white and babble additive noise, respectively, vs. SNR. The subplots compare the classical front-end (dotted lines) with the alternative ones (full lines) obtained using the Shannon entropy for (a), the q-entropy (q=0.5) for (b), the Kullback–Leiber divergence for (c) and the q-divergence (q=0.1) for (d). The reference value for the classical front-end in the clean signal case was 6.74%. As can be seen in Fig. 3, it is clear that in the white noise case, the best results are obtained for D_q. In particular for SNR between 20 and 50 dB the error rate is reduced (Fig. 3(d)). In the babble noise case (Fig. 4) the D_q advantage is not so clear. In most of the considered measures the WER values obtained for the white noise case are higher than the ones obtained for babble noise. This last result is consistent with the fact that white noise is more destructive of the signal spectral properties and the information measures provide additional information for the recognition.

The percentage of relative error improvement of the different measures compared with the classical front-end was computed. The results are summarized in Table 1 for

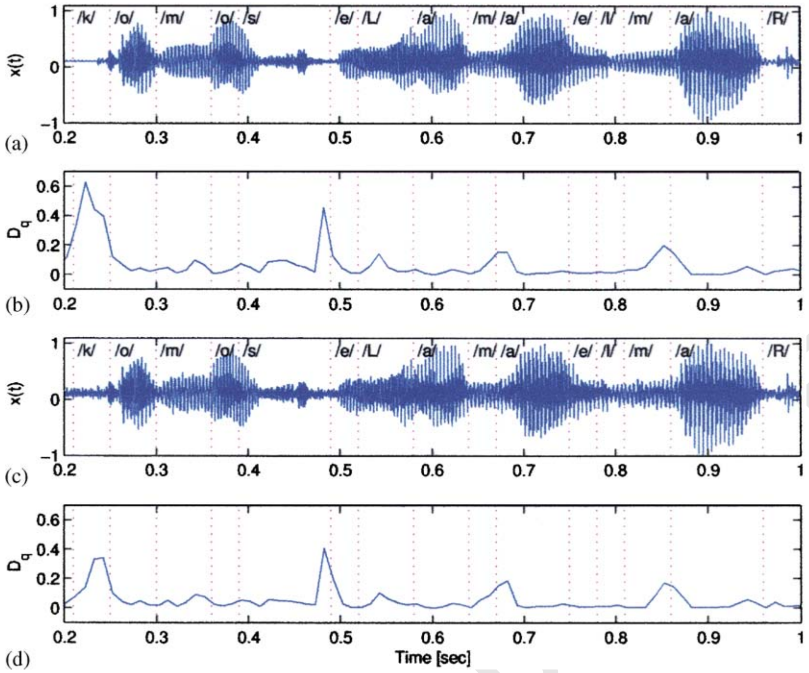

Fig. 2. (a) Labeled speech signal. (b) D_q evolution corresponding to the signal displayed in (a). (c) The same signal shown in (a) with additive white noise (20 dB). (d) D_q evolution corresponding to the signal displayed in (c).

Table 1
Percentage of relative error improvement ($\Delta\varepsilon\%$) of the different measures compared with the classical front-end for speech signal corrupted with white noise

SNR_{dB}	$\mathscr{H}$	$\mathscr{H}_{q=0.1}$	$\mathscr{H}_{q=0.5}$	D	$D_{q=0.1}$	$D_{q=0.5}$
∞	0.89	3.55	2.55	0.96	**8.04**	3.89
100	−1.88	0.09	1.48	−2.01	**7.23**	1.42
50	1.20	−8.57	7.19	5.90	**12.83**	2.38
30	8.60	6.39	7.30	−7.43	**21.25**	2.51
20	13.02	15.79	12.29	8.99	**18.38**	−1.14
10	2.25	**3.20**	3.13	−5.45	−1.91	−8.53
0	−1.26	**0.31**	−0.75	−4.07	−2.44	−3.17

Bold numbers indicate the best performance for each SNR value and they remark the results observed in Fig. 3.

different q values and SNRs under white noise. The same analysis using babble noise are given in Table 2. Bold numbers indicate the best performance for each SNR value. The results displayed in Table 1 remark those observed in Fig. 3. From Tables 1 and 2 it can be concluded that with white and babble noise, $D_{q=0.1}$ is the measure that provides better results, in particular for SNRs higher than 20 dB.

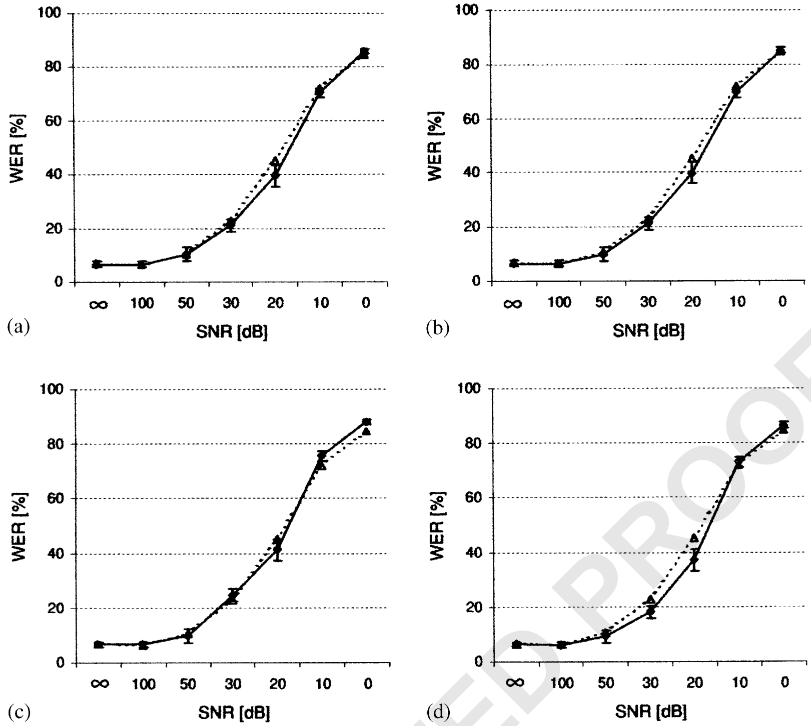

Fig. 3. Recognition WER for the ASR systems vs. SNR, for speech signals corrupted with white additive noise. Comparison of the classical front-end (dotted lines) with the alternative ones obtained using: (a) Shannon entropy; (b) q-entropy with $q=0.5$; (c) Kullback–Leiber divergence; and (d) q-divergence with $q=0.1$.

Table 2
Percentage of relative error improvement ($\Delta\varepsilon\%$) of the different measures compared with the classical front-end for speech signal corrupted with babble noise

SNR$_{dB}$	$\mathscr{H}$	$\mathscr{H}_{q=0.1}$	$\mathscr{H}_{q=0.5}$	D	$D_{q=0.1}$	$D_{q=0.5}$
∞	0.89	3.55	2.55	0.96	**8.04**	3.89
100	1.48	1.68	3.79	1.66	**10.71**	3.92
50	6.46	−1.34	6.84	4.31	**14.11**	1.06
30	17.98	14.50	20.75	17.19	**24.31**	10.13
20	**14.83**	10.42	13.46	4.13	8.16	0.64
10	1.52	**2.67**	2.18	−6.73	−4.53	−11.36
0	−4.35	**−2.00**	−2.77	−8.56	−4.96	−7.25

Bold numbers indicate the best performance for each SNR value. They allow us to assert that in the babble noise case it is again D_q measure that provides better results, for $q = 0.1$, in particular for SNR between 30 and 50 dB.

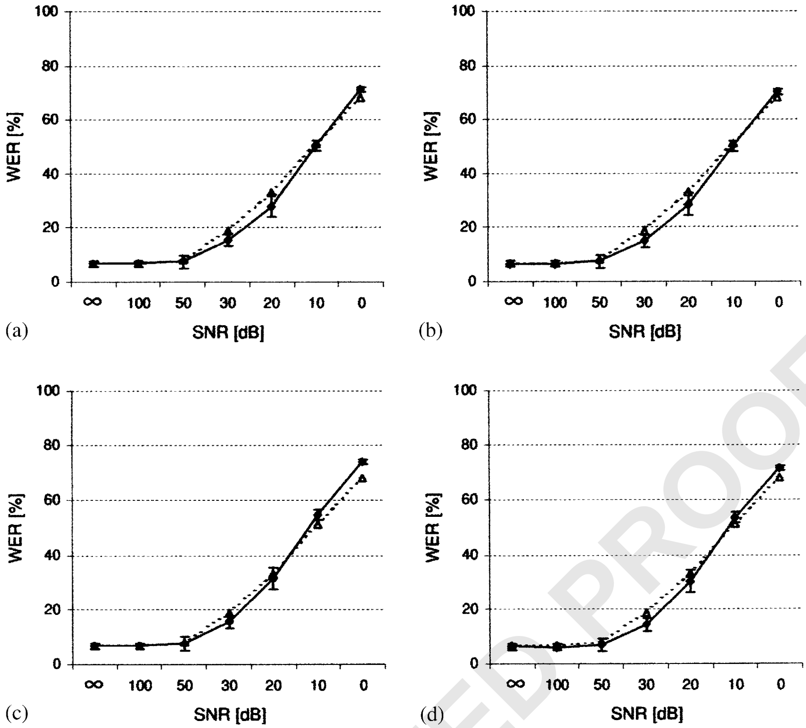

Fig. 4. Recognition WER for the ASR systems vs. SNR, for speech signals corrupted with babble additive noise. Comparison of the classical front-end (dotted lines) with the alternative ones obtained using: (a) Shannon entropy; (b) q-entropy with $q = 0.5$; (c) Kullback–Leiber divergence; and (d) q-divergence with $q = 0.1$.

We have also evaluated the statistical significance of these results by computing the probability that a given recognizer is better than the reference or baseline system ($\Pr(\varepsilon_{ref} > \varepsilon)$). In order to perform this test we assumed the statistical independence of the recognition errors for each word and we approached the errors' Binomial distribution by means of a Gaussian distribution. This is possible because we have a sufficiently high number of words (11 077 words if we take into account all the partitions). In this way, for $q = 0.1$ and SNR between 20 and 50 dB of both noise types we have that $\Pr(\varepsilon_{ref} > \varepsilon) > 99.999\%$.

4. Conclusions

In this work we have introduced the use of complexity measures in the front-end of an ASR system. The original motivation for this approach was the relation observed between complexity measures and speech segmentation. In previous works [15,27,28] we have shown that information measures as the ones used in this paper are well suited both for detecting relative complexity changes and slight parameters variations in nonlinear dynamical systems. These aspects and the fact that speech signals show a nonlinear behavior allowed us to expect that the inclusion of these measures in an ASR system would provide a more reliable information about the speech underlying dynamics, improving recognition capabilities in adverse conditions.

The proposed method has been tested on noisy speech and compared with a classical MFCC parametrization. The expansion of the feature vector through the addition of the $D_{q=0.1}$ measure, demonstrated a significant reduction of the WER in noisy conditions. For both white and babble noise, the best improvements have been obtained for 20 and 30 dB SNRs. Although this would not be considered extremely noisy conditions, these preliminary results suggest that, as expected, complexity measures could provide valuable information derived from vocal tract dynamical changes in speech recognition. Therefore, complexity measures could be considered as part of the information provided to an ASR system.

Appendix A. Computational aspects

A.1. Information measures

Let $x(t)$ be the temporal evolution of a given signal and let us consider its discrete evolution, obtained by a regular sampling, given by

$$S = \{x[k],\ k = 1, 2, \ldots, K\}\ ,$$

where $x[k]$ stands for $x(t_k)$, the signal value at the time $t_k = k\,\Delta t$, where Δt is the sampling period.

The central idea of the analysis proposed here is that of being in a position to generate a suitable probability set from a given signal. Once the probabilities are determined, the information measures are computed according to the pertinent definitions. In order to calculate the different entropies, we have to evaluate the probability distributions of the corresponding sampled signal $x[k]$.

Fixed a desired number of partitions M, we may define on set S a sliding window depending on the two parameters, width $L \in \mathbb{N}$ and sliding factor $\Delta \in \mathbb{N}$:

$$W_R(m; L, \Delta) = \{x[k],\ k = 1 + m\Delta, \ldots, L + m\Delta\}\ ,$$
$$m = 0, 1, 2, \ldots, m_{\max}\ , \qquad \text{(A.1)}$$

with Δ and L selected such that $L \leqslant K$ and $(K - L)/\Delta = m_{\max} \in \mathbb{N}$. Observe that $m_{\max}$ has to be fixed in such a way that as m evolves from 0 to $m_{\max}$, all the data set S is "viewed" through the *rectangular* windows. In this way $W_R(m; L, \Delta)$ is called a "frame".

In the case of speech signal the window's width has direct relation with the maximum speed of significant vocal tract morphology modification. A suitable selection allows the study of the signal within each frame under a stationary hypothesis [7]. In our experiments, $M = 10$, $\Delta = 200$ and $L = 80$ were used, with $\Delta t = 125$ µs.

The center of the frame (A.1) is $x[L/2 + m\Delta]$. The integer m is a time-control and, for each $W_R(m; L, \Delta)$, the equipartition

$$x^0 = x^{\min}[m] < x^1 < x^2 < \cdots < x^M = x^{\max}[m] \, , \tag{A.2}$$

is to be considered, where

$$\begin{aligned} x^{\min}[m] &\mathrel{\hat{=}} \min[W_R(m; L, \Delta)] \\ &= \min_k \{x[k],\ k = 1 + m\Delta, \ldots, L + m\Delta\} \, , \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} x^{\max}[m] &\mathrel{\hat{=}} \max[W(m; L, \Delta)] \\ &= \max_k \{x[k],\ k = 1 + m\Delta, \ldots, L + m\Delta\} \, . \end{aligned} \tag{A.4}$$

Attention is now focused upon the set $\{I^l = [x^{l-1}, x^l[, l = 1, \ldots, M\}$ of disjoint intervals such that

$$[x^0, x^M] = \overline{\bigcup_{l=1}^{M} I^l} \, . \tag{A.5}$$

Let us indicate $p^m(I^l)$ the probability that the element $x[k] \in W_R(m; L, \Delta)$ belongs to the interval I^l. This probability is estimated as the ratio between the number of elements of $W_R(m; L, \Delta)$ in I^l and the total number of elements in this window. In this way the corresponding q-entropy can be computed as

$$H_{q,x}[m] = (q-1)^{-1} \sum_{l=1}^{M} [p^m(I^l) - (p^m(I^l))^q] \, , \tag{A.6}$$

for $m = 0, 1, \ldots, M$.

In conclusion, H_q allows us to follow the corresponding entropy temporal evolution (time-control m) of the given signal. The other information measures that involve pdf functions are computed in a similar way.

A.2. Basic speech parametrization

For basic speech parametrization it is more convenient to use sliding windows other than rectangular windows. In the Hamming window case we modify (A.1)

$$\begin{aligned} W_H(m; L, \Delta) &= \{x_H[k, m],\ k = 1 + m\Delta, \ldots, L + m\Delta\} \, , \\ & \qquad m = 0, 1, 2, \ldots, m_{\max} \, , \end{aligned} \tag{A.7}$$

where $x_H[k, m]$ is the scalar product of the original speech signal and a shifted Hamming window:

$$x_H[k, m] = x[k]\omega_H[k, m; L] \, , \tag{A.8}$$

with Δ and L as above, and $\omega_H[k, m; L]$ the Hamming window

$$\omega_H[k, m; L] = \frac{27}{50} - \frac{23}{50} \cos(2\pi(k - m\Delta)/L) \, . \tag{A.9}$$

If $\mathscr{T}_C[n]$ is the operator for the domain transformation corresponding to the *real cepstral coefficients in mel scale* (MFCC), the windowed signal parametrization process of $x_H[k,m]$ is carried out according to

$$c[n,m] = \mathscr{T}_C[n]\{x_H[k,m]\}, \quad 0 < n \leqslant N \ .$$

That is to say that the application of $\mathscr{T}_C[n]$ to the windowed speech signal generates the MFCC parametrization mentioned in Section 2.2.

As the first step for the definition of $\mathscr{T}_C[n]$ the output of a filter bank $u[j,m]$ for each time-control index m is computed, starting from the weighted summation of the log-spectral coefficients derived from $x_H[k,m]$:

$$u[j,m] = 2 \sum_{i=B(j-1)}^{i=B(j)} \omega_B[i - B(j-1); B(j+1) - B(j)] \log |\hat{x}_H[i,m]|, \quad 0 < j \leqslant J \ ,$$

where $\hat{x}_H[i,m]$ corresponds to the *discrete Fourier transform* of $x_H[k,m]$, and $B(j)$ is a vector that establishes the initial and ending samples for each summation range for each one of the filters, and ω_B is the Bartlett (or triangular) window

$$\omega_B[i;L] = \begin{cases} 2i/L & \text{if } 0 < i \leqslant L/2 \ , \\ 2 - 2i/L & \text{if } L/2 < i \leqslant L \ . \end{cases} \tag{A.10}$$

In order to compute the values of $B(j)$ an expression is used that allows one to locate the central frequencies of filter windows ω_B according to the so-called *mel scale* [7]

$$F_{mel}(f_{Hz}) = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \ .$$

This scale has been derived from studies of the human perception of audio pure tones that allowed to approach the relationship between the perceived frequency and the real one.

Now, $\mathscr{T}_C[n]$ can be defined from the *discrete cosine transform* of $u[j,m]$:

$$c[n,m] = \sqrt{\frac{2}{J}} \sum_{j=1}^{J} u[j,m] \cos\left(\frac{\pi n}{J}(j - 0.5)\right) \ .$$

The experimental results had widely favored this combination. It should be mentioned that, as it is common practice in ASR, only the first $c[n,m]$ were used (12 or 13). These coefficients hold information with respect to the spectral envelope of the vocal tract. In our experiments $J = 24$ filter bands were used.

A.3. Energy and delta coefficients

When the speech feature vector is assembled, it is common practice to add a measure of local energy (see Section 2.2) that is simply defined as [26]

$$E[m] = \log \sum_{k=m\Delta+1}^{m\Delta+L} x_H[k,m]^2 \ . \tag{A.11}$$

An estimate of the temporal derivative was also added for all the computed elements. For a generic feature vector $y[n,m]$ the *delta coefficients* are obtained by means of the regression [26]

$$\Delta y[n,m] = \frac{\sum_{j=1}^{N} j(y[n,m+j] - y[n,m-j])}{2\sum_{j=1}^{N} j^2} ,$$

where N is used to smooth the estimate through the different frames. In our case $N=2$ was used.

References

[1] R.C. Conant, IEEE Trans. Syst. Man Cybern. 2 (1972) 550.
[2] P. Saparin, A. Witt, J. Kurths, B. Anishchenko, J. Chaos, Solitons and Fractals 4 (1994) 1907.
[3] M. Banbrook, S. McLaughlin, I. Mann, IEEE Trans. Speech Audio Process. 7 (1999) 1.
[4] L.S. Su, K.P. Li, K.S. Fu, J. Acoust. Soc. Am. 56 (1974) 1876.
[5] K.N. Stevens, Acoustic Phonetics, MIT Press, Cambridge, MA, 1998.
[6] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[7] J. Deller, J. Proakis, J. Hansen, Discrete Time Processing of Speech Signals, Macmillan, New York, 1993.
[8] 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.
[9] 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.
[10] J.C. Junqua, J.P. Haton, Robustness in Automatic Speech Recognition: Fundamentals and Applications, Kluwer Academic Publishers, Boston, 1996.
[11] J. Havrda, F. Charvat, Kybernetica 3 (1967) 30.
[12] Z. Daróvczy, Inf. Control 16 (1970) 36.
[13] C. Tsallis, Chaos, Solitons and Fractals 6 (1995) 539, and references therein.
[14] C. Shannon, Bell Syst. Tech. J. 27 (1948) 379.
[15] M.E. Torres, L.G. Gamero, Physica A 286 (2000) 457.
[16] M.E. Torres, M.M. Añino, G. Schlotthauer, in: Proceedings of the 2001 Workshop on Nonlinear Signal and Image Processing (NSIP'2001) IEEE-EURASIP, Baltimore, Maryland, USA, 2001, pp. 1–5, Paper No. 1049.
[17] T.M. Cover, J.A. Thomas, Information Theory, Wiley, New York, 1991.
[18] M.E. Torres, Ph.D. Thesis, Universidad Nacional de Rosario—Argentine, 1999, (Math. D. Thesis).
[19] 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.
[20] L.R. Rabiner, B.H. Juang, IEEE Acoust. Speech Signal Process. Mag. 3 (1986) 4.
[21] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambridge, MA, 1999.
[22] S. Young, in: IEEE Workshop on Speech Recognition, IEEE Press, Snowbird, Utah, 1995.
[23] J.E. Diaz, et al., in: Proceedings of the Second European Conference of Speech Communication and Technology, International Speech Communication Association, Berlin, 1993.
[24] A. Varga, H. Steeneken, Speech Commun. 12 (1993) 247.
[25] D. Michie, D.J. Spiegelhalter, C.C. Taylor, Machine Learning, Neural and Statistical Classification, Ellis Horwood, University College, London, 1994.
[26] S. Young, et al., HMM Toolkit, Cambridge University, http://htk.eng.cam.ac.uk, 2000.
[27] M.M. Añino, M.E. Torres, G. Schlottahauer, Physica A 324 (2003) 645.
[28] M.E. Torres, M.M. Añino, L.G. Gamero, M.A. Gemignani, Int. J. Bifurcations Chaos 11 (2001) 967.