Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 26 (2012) 336–348

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Bioinspired sparse spectro-temporal representation of speech for robust classification☆

C. Martínez[a,c,*], J. Goddard[b], D. Milone[a,d], H. Rufiner[a,c,d]

[a] *Centro de I+D en Señales, Sistemas e INteligencia Computacional (SINC(i)), Dpto. Informática, Facultad de Ingeniería, Universidad Nacional del Litoral, CC217, Ciudad Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina*

[b] *Dpto. de Ingeniería Eléctrica, UAM-Iztapalapa, Mexico*

[c] *Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina*

[d] *CONICET, Argentina*

Received 13 April 2010; received in revised form 28 September 2011; accepted 1 February 2012
Available online 7 March 2012

Abstract

In this work, a first approach to a robust phoneme recognition task by means of a biologically inspired feature extraction method is presented. The proposed technique provides an approximation to the speech signal representation at the auditory cortical level. It is based on an optimal dictionary of atoms, estimated from auditory spectrograms, and the Matching Pursuit algorithm to approximate the cortical activations. This provides a sparse coding with intrinsic noise robustness, which can be therefore exploited when using the system in adverse environments. The recognition task consisted in the classification of a set of 5 easily confused English phonemes, in both clean and noisy conditions. Multilayer perceptrons were trained as classifiers and the performance was compared to other classic and robust parameterizations: the auditory spectrogram, a probabilistic optimum filtering on Mel frequency cepstral coefficients and the perceptual linear prediction coefficients. Results showed a significant improvement in the recognition rate of clean and noisy phonemes by the cortical representation over these other parameterizations.

Keywords: Robust phoneme recognition; Approximated auditory cortical representation; Sparse coding

1. Introduction

In previous years, the classic techniques of signal analysis have been applied to automatic speech recognition with relatively good results in controlled conditions. At present, however, there is an increasing need to deal with more complex and real situations, for example robust speech recognition in noisy environments. The ability to solve this and other challenging problems could be improved by the development of new speech representation techniques.

An early stage in the speech recognition process consists in the acoustic modeling of phonemes. In the last years, efforts have been made to provide robustness to this stage by the proposal of different approaches in the speech

☆ This paper has been recommended for acceptance by 'Louis ten Bosch'.

* Corresponding author at: Centro de I+D en Señales, Sistemas e INteligencia Computacional (SINC(i)), Dpto. Informática, Facultad de Ingeniería - Universidad Nacional del Litoral, CC217, Ciudad Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina. Tel.: +54 342 4575233; fax: +54 342 4575224.

E-mail address: cmartinez@fich.unl.edu.ar (C. Martínez).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.02.002

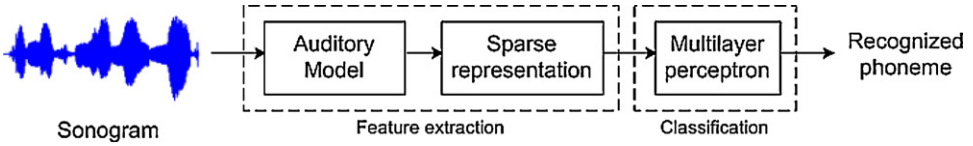

Fig. 1. Block diagram of the proposed phoneme classification system based on the sparse representation of speech.

representation. In Yousafzai et al. (2008), Ager et al. (2008), and Yousafzai et al. (2009), authors use the central segment of the acoustic waveforms of the phonemes. They showed that the mismatch between clean and noisy conditions is better managed by the raw acoustic data than perceptual linear prediction (PLP) coefficients, specially under severe degradation. Recently, a noise compensation technique was proposed to suppress the effect of additive noise with an estimation of the noise envelope (Ganapathy et al., 2010). This work was carried out by processing the speech signal in their time-frequency representation, in a similar way as the approach proposed here, showing better performance at low signal-to-noise ratios (SNRs) than classic speech representations.

The use of biologically inspired, feature extraction methods has improved the performance of artificial systems that try to emulate some aspect of human communication, such as emphasizing the discourse cues. Based on the biological time-frequency analysis the inner ear carries out, auditory representations of speech beyond the cochlea have been widely studied. Different mathematical and computational models have been developed that allow for the estimation of the *auditory spectrograms* (Delgutte, 1996). These investigations enabled modeling the discharge patterns of the auditory nerve.

Moreover, given a speech utterance, a pattern of activations can be found at the primary auditory cortex that encodes a series of meaningful cues contained in the signal. This behavior of the cortical neurons could be emulated using the notion of *spectro-temporal receptive fields* (STRF). The STRF are defined as the optimal linear filter that convert a time-varying stimulus into the firing rate of an auditory cortical neuron, so that it responds with the largest possible activation (Theunissen et al., 2000). Using two-bidimensional discrete dictionaries, an approximated cortical representation can be established by means of techniques related to *independent component analysis* (ICA) and *sparse representations* (Klein et al., 2003; Oja and Hyvärinen, 2000; Rubinstein et al., 2010). Here we used the term *approximated cortical representation* with the meaning of the set of activations that contribute to form a particular pattern from an estimation of the STRF. This estimation intends to model the global statistical characteristics of the discharge patterns in the auditory cortex, in a phenomenological rather than a physiological way. This concept of cortical representation is slightly different from the one applied in neuroscience, where studies about brain activity involves analysis of the cortical areas that are mainly stimulated by viewing images or listening words (Mitchell et al., 2008).

In this work, using the time-frequency representations of the auditory spectrograms of phoneme speech signals, a dictionary of two-dimensional optimal atoms is estimated. Based on this STRF dictionary, a sparse representation that emulates the cortical activation is computed. This representation is then applied to a phoneme classification task in both clean and noisy conditions, designed to evaluate the advantages and robustness of the representation. Fig. 1 resumes the main steps in the operation of the proposed system.

The organization of the paper is as follows. Section 2 presents the method for the speech signal representation used in this work. Section 3 gives the information about the speech data and the noise corpus used in the experiments, along with details of the cortical representation. Section 4 presents the results obtained in the preliminary tuning of the method and the phoneme classification task, which is then compared with other robust parameterizations widely used in this field. Finally, Section 5 summarizes the contributions of this paper and outlines future research.

2. Sparse representations

2.1. Representations based on discrete dictionaries

There are different ways of representing a signal using general discrete and finite dictionaries. For the case where the dictionary forms a basis, in particular for the orthonormal or unitary cases, the techniques are quite simple. This is because, among other aspects, the representation is unique. However, in the general case, a signal can have many different representations for the same dictionary. In these cases, it is possible to find a suitable representation if additional criteria are imposed. For our problem, these criteria can be motivated by obtaining a representation with characteristics

such as sparseness and independence. Furthermore, it is possible to find an optimal dictionary that resembles biological properties of sensorial systems, such as in the primary visual cortex. These visual neurons exhibit a spatially localized, oriented bandpass behavior, similar to the basis functions of a wavelet transform (Olshausen and Field, 1996).

A sparse code is one which represents the information in terms of a small number of descriptors taken from a large set. This means that a small fraction of the elements from the code are used actively to represent a typical pattern. In numerical terms, this means that the majority of the elements are zero, or 'almost' zero, most of the time (Hyvärinen, 1998).

It is possible to define measures or norms that allow us to quantify how sparse a representation is; one way is using either the ℓ_0 or the ℓ_1 norms. An alternative criteria for optimization is to use an *a priori* probability distribution with a large positive kurtosis. This results in a distribution with a large thin peak at the origin and long tails on either side. One such distribution is the Laplacian. In the statistical context it is relatively simple to include aspects related to the independence of the coefficients, like factorial probability distributions, which connects this approach with ICA (Oja and Hyvärinen, 2000).

In the following section a formal description of a statistical method for estimation is given. This method estimates an optimal dictionary and the corresponding sparse representation of the input data.[1]

2.2. Optimal sparse and factorial representations

Let $\vec{x} \in \mathbb{R}^N$ be a signal to represent in terms of a *dictionary* $\vec{\Phi}$, with size $N \times M$, and a set of coefficients $\vec{a} \in \mathbb{R}^M$. In this way, the signal is described as

$$\vec{x} = \sum_{1 \le i \le M} \vec{\phi}_i a_i + \vec{\varepsilon} = \vec{\Phi}\vec{a} + \vec{\varepsilon}, \tag{1}$$

where $\vec{\varepsilon} \in \mathbb{R}^N$ is the term for additive noise and $M \geq N$. The dictionary $\vec{\Phi}$ is composed of a collection of waveforms or parameterized functions ($\vec{\phi}_i$), where each waveform $\vec{\phi}_i$ is an *atom* of the representation.

In the context of this work, $\vec{x}$ corresponds to the reconstruction of the time-frequency representation of the speech at the auditory cortex. The atoms in $\vec{\Phi}$ will further be the representation of the important features found at the cortex for each input stimuli. Finally, an estimation of the coefficients $\vec{a}$ will be the output of the feature extraction stage proposed.

Although (1) appears very simple, the main problem is that for the most general case $\vec{\Phi}$, $\vec{a}$ and $\vec{\varepsilon}$ are unknown, thus there can be an infinite number of possible solutions. Even in the noiseless case (when $\vec{\varepsilon} = \vec{0}$) and given $\vec{\Phi}$, if there are more atoms than the dimension of $\vec{x}$ then multiple representations of the signal are possible. Therefore, an approach that allows us to select one of these representations has to be found. For the complete and noiseless case the relationship between the data and the coefficients is linear and it is given by $\vec{\Phi}^{-1}$. For classical transformations, such as the discrete Fourier transform, this inverse is simplified because $\vec{\Phi}^{-1} = \vec{\Phi}^*$ (with $\vec{\Phi} \in \mathbb{C}^{N \times N}$ and $\Phi^*(i, j) = \overline{\Phi(j, i)}$). In our case – although this is a linear system – the coefficients chosen to be part of the solution generally have a non-linear relationship with the data $\vec{x}$ (Chen et al., 2001).

When $\vec{\Phi}$ and $\vec{x}$ are known, an interesting way to choose the set of coefficients $\vec{a}$ from among all the possible representations, consists of finding those a_i which make the representation as sparse and independent as possible. In order to obtain a sparse representation, a distribution with positive kurtosis can be assumed for each coefficient a_i. Further, assuming the statistical independence of the a_i, the imposed joint *a priori* distribution satisfies

$$P(\vec{a}) = \prod_i P(a_i). \tag{2}$$

The system (1) can also be seen as a generative model. Following the terminology used in the ICA field, this means that signal $\vec{x} \in \mathbb{R}^N$ is generated from a set of sources a_i (in the form of a state vector $\vec{a} \in \mathbb{R}^M$) using a mixing matrix $\vec{\Phi}$, and including an additive noise term $\vec{\varepsilon}$ (Gaussian, in most cases).

[1] Although in our proposed method and experiments two-dimensional patterns are used, for clearness we only describe the one-dimensional case in this section.

The state vector $\vec{a}$ can be estimated from the *posterior* distribution (Lewicki and Sejnowski, 1998)

$$P(\vec{a}|\vec{\Phi},\vec{x}) = \frac{P(\vec{x}|\vec{\Phi},\vec{a})P(\vec{a})}{P(\vec{x}|\vec{\Phi})}. \tag{3}$$

Thus, a *maximum a posteriori* estimation of $\vec{a}$ would be

$$\hat{\vec{a}} = \underset{\vec{a}}{arg\max}[\log\ P(\vec{x}|\vec{\Phi},\vec{a}) + \log\ P(\vec{a})]. \tag{4}$$

When $P(\vec{a}|\vec{\Phi},\vec{x})$ is sufficiently smooth, the maximum can be found by the method of gradient ascent. The solution depends on the functional forms assigned to the distributions for the noise and the coefficients, giving rise to different methods for finding the coefficients. Lewicki and Olshausen (1999) proposed the use of a Laplacian *a priori* distribution with parameter β_i

$$P(a_i) = \alpha\ \exp(-\beta_i|a_i|), \tag{5}$$

where α is a normalization constant. This distribution, with the assumption of Gaussian additive noise $\vec{\varepsilon}$, results in the following updating rule for $\vec{a}$

$$\Delta\vec{a} = \vec{\Phi}^T\vec{\Lambda}_{\vec{\varepsilon}}\vec{\varepsilon} - \vec{\beta}^T|\vec{a}|, \tag{6}$$

where $\vec{\Lambda}_{\vec{\varepsilon}}$ is the inverse of the noise covariance matrix $\mathcal{E}[\vec{\varepsilon}^T\vec{\varepsilon}]$, with $\mathcal{E}[\cdot]$ denoting the expected value. This provides a gradient-based search for the solution of (4).

To estimate the value of $\vec{\Phi}$, the following objective function can be maximized (Lewicki and Olshausen, 1999)

$$\hat{\vec{\Phi}} = \underset{\vec{\Phi}}{arg\max}[\mathcal{L}(\vec{x},\vec{\Phi})], \tag{7}$$

where $\mathcal{L} = \mathcal{E}[\log\ P(\vec{x}|\vec{\Phi})]_{P(\vec{x})}$ is the likelihood of the data. This likelihood can be found by marginalizing the following product of the conditional distribution of the data, given the dictionary and the *a priori* distribution of the coefficients

$$P(\vec{x}|\vec{\Phi}) = \int_{\mathbb{R}^M} P(\vec{x}|\vec{\Phi},\vec{a})P(\vec{a})\,d\vec{a}, \tag{8}$$

where the integral is over the M-dimensional state space of $\vec{a}$.

The objective function in (7) can be maximized using gradient ascent with the following update rule for the matrix $\vec{\Phi}$

$$\Delta\vec{\Phi} = \eta\vec{\Lambda}_{\varepsilon}\mathcal{E}[\vec{\varepsilon}\vec{a}^T]_{P(\vec{a}|\vec{\Phi},\vec{x})}, \tag{9}$$

where η, in the range (0, 1), is the learning rate.

The practical implementation is carried out through an iterative estimation of (6) and (9) allow the calculation of the dictionary $\vec{\Phi}$ and the coefficients $\vec{a}$. In this work, the initialization of both is at random and 1000 iterations are then performed. The parameter η is kept high to get a rough approach to the solution up to 500 iterations, then it is gradually decreased as a function of the number of cycles. This election was guided by preliminary experiments where the evolution of $\mathcal{L}$ was surveyed.

2.3. Matching Pursuit

The computational cost in the estimation of $\hat{\vec{a}}$ is truly expensive. The Matching Pursuit (MP) algorithm is another method to approximate the solution of the sparse representation problem, once the dictionary is provided or estimated (Mallat and Zhang, 1993).

Sparsity is enforced by choosing an appropriate number of terms. Given an initial approximation $\vec{x}^{(0)} = \vec{0}$ and an initial residue $\vec{R}^{(0)} = \vec{x}$, a sequence of approximations is iteratively constructed. At step k the parameter $\gamma = \hat{\gamma}$ is selected, such that the atom $\vec{\phi}_{\hat{\gamma}}^{(k)}$ best correlates with the residue $\vec{R}^{(k)}$, and a multiple of this atom is added to the approximation at step $k-1$, obtaining

$$\vec{x}^{(k)} = \vec{x}^{(k-1)} + a_{\hat{\gamma}}^{(k)}\vec{\phi}_{\hat{\gamma}}^{(k)}, \tag{10}$$

where $a_{\hat{\gamma}}^{(k)} = \langle \vec{R}^{(k-1)}, \vec{\phi}_{\hat{\gamma}}^{(k)} \rangle$, and $\vec{R}^{(k)} = \vec{x} - \vec{x}^{(k)}$. After m steps an approximation to (1) is obtained, with residue $\vec{R} = \vec{R}^{(m)}$. It is said that MP constitutes a greedy solution to the sparse representation problem[2]; therefore it shares the same advantages and disadvantages of this type of optimization methods (fast but generally not optimal methods). Nevertheless, there are investigations that establish that under appropriate conditions of dictionary coherence and sparsity of the state vector $\vec{\hat{a}}$, these algorithms obtain the globally optimal solution (Donoho and Elad, 2003; Donoho et al., 2006).

3. Approximated auditory cortical representation

In neuroscience, it has been established the principle that the brain of an animal adapt its properties (internal configuration) to best describe the statistics of stimuli perceived through its senses (Barlow, 2001). If a simple model of these stimuli is assumed, as the one outlined in (1), it is possible to estimate their properties from the statistical approach presented in the previous section.

The early auditory system codes important cues for phonetic discrimination, such as the ones found in the auditory spectrograms (AS) (Delgutte, 1996). Shamma et al. proposed a model of the processing of the sound carried out in the auditory system based on psychoacoustic facts found in physiological experiments in mammals. The main idea behind the model is to move forward in the representation of the sound from the initial spectral analysis by decomposing this spectrogram in its spectrotemporal modulation content (Chi et al., 2005). While the complete model of Shamma consists of two stages, in this work only the early stage was used. It obtains the *auditory spectrogram*, an internal cochlear representation of the pattern of vibrations along the basilar membrane. This part of the model is composed of a bank of 128 cochlear filters that process the temporal signal $s(t)$ and obtain the outputs

$$y_{\mathrm{ch}} = s(t) \times h_{\mathrm{ch}}(t, f), \tag{11}$$

where h_{ch} is the impulse response of each filter. These outputs are transduced into auditory-nerve patterns

$$y_{\mathrm{an}} = g_{\mathrm{hc}}(\partial_t y_{\mathrm{ch}}(t, f)) \times \mu_{\mathrm{hc}}(t), \tag{12}$$

where ∂_t represents the fluid-cilia coupling (highpass filter), g_{hc} the nonlinear compression in the ionic channels and μ_{hc} the hair-cell membrane leakage (lowpass filter). Finally, the lateral inhibitory network is approximated by a half-wave rectified first-order derivative with respect to the frequency axis as

$$y_{\mathrm{lin}}(t, f) = \max(\partial_f y_{\mathrm{an}}(t, f), 0) \tag{13}$$

and the final output consists in a integration of this signal over short windows.

In these representations – of a higher level in the auditory path – some aspects of the acoustic signal that arrives at the eardrum have been reduced or eliminated. Among these superfluous aspects are the temporal variability of the signal and the relative phase of acoustic waveforms. Hence, following this biological simile, the representation becomes a good starting point to attain more complex ones.

Obtaining a dictionary of two-dimensional atoms $\vec{\Phi}$ using (7), corresponding to time-frequency features estimated from the AS of $\vec{x}$, is equivalent to the STRF of a group of cortical neurons (Klein et al., 2003). Therefore, the activation level of each neuron can be associated with the set of coefficients $\vec{a}$ in (1). Instead of using this vector as the representation, due to the high computational cost in computing it from (4), the preferred approach here is to obtain an approximated solution by the Matching Pursuit algorithm by means of (10). Thus, the feature extraction scheme here proposed obtains the coefficients modeling the activations at the primary auditory cortex in reponse to input stimuli. These features are named *approximated auditory cortical representation* (AACR).

Fig. 2 shows a schematic diagram of the method adopted for estimating the AACR, once the dictionary has been trained following the process previously described. The acoustic signal corresponding to a complete utterance, $s(t)$, is processed by the ear model. It obtains the spectrogram at the early auditory level, $y(t, f)$ from (13). Finally, from these time-frequency representations, the feature extraction scheme obtains the coefficients $\vec{\hat{a}}$ of the AACR as a subproduct of (10) in the Matching Pursuit algorithm.

[2] MP is a greedy algorithm that minimizes $\|\vec{x} - \vec{\Phi}\vec{a}\|_2$.

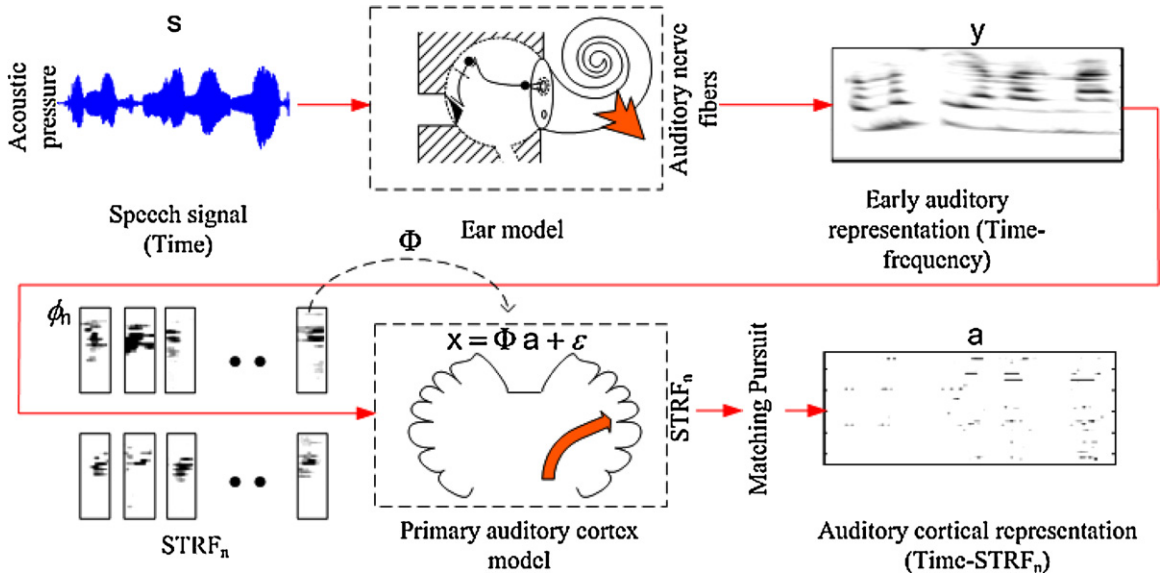

Fig. 2. Schematic diagram of the method used for estimating the approximated auditory cortical representation (AACR).

Table 1
Distribution of patterns per class for training and test data in region DR1 of the TIMIT corpus.

Phoneme	Train, # (%)	Test, # (%)
/b/	211 (3.26)	66 (3.43)
/d/	417 (6.45)	108 (5.62)
/jh/	489 (7.56)	116 (6.04)
/eh/	2753 (42.58)	799 (41.63)
/ih/	2594 (40.13)	830 (43.25)
Total	6464 (100.00)	1919 (100.00)

4. Materials and methods

The feasibility of building a robust classification system based on the described scheme was studied for an initial simpler task of phoneme classification. The classifiers were trained with approximated auditory cortical patterns calculated from clean speech and then tested with patterns obtained from noisy speech, where controlled amounts of white noise were added. The task consisted in the classification of the set of five easily confused phonemes in English: */b/*, */d/*, */jh/*, */eh/*, */ih/*, in a context-independent approach (Stevens, 2000).

4.1. The signals

The clean speech data were extracted from the TIMIT corpus, which contains a total of 6300 sentences recorded from 630 speakers (10 sentences each) (Garofolo et al., 1993). In this work, the training (38 speakers) and test (11 speakers) data corresponding to region DR1 were used. The number of AACR patterns calculated from the TIMIT data is showed in Table 1. It can be observed that there is a noticeable imbalance in the distribution, which could be counterproductive for the generalization capabilities of the classifiers. Thus, the training and test sets were balanced by selecting the same number of patterns for each phoneme in each set (211 and 66 patterns, respectively).

For the estimation of the dictionaries, an AS from the original clean signals sampled at 16 kHz was obtained by means of an early auditory model (Yang et al., 1992). In order to process less data, the frequency resolution was downsampled by half. Thus, AS with 64 frequency coefficients per frame of 32 ms were obtained. After that, a sliding window of one frame in length at intervals of 8 ms, was applied to obtain the set of spectro-temporal patterns. Fig. 3

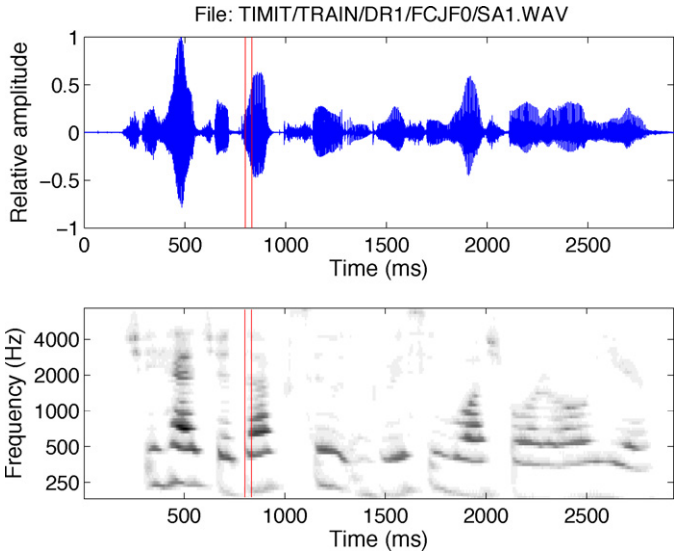

Fig. 3. Principal signals in the process of obtaining the spectro-temporal patterns: sonogram (top) and auditory spectrogram (bottom). A section corresponding to the sliding window, from which each spectro-temporal pattern is generated, has been marked with vertical lines.

shows the principal signals of this process, while in Fig. 4 there is an excerpt of the clean signal and its corresponding low-resolution AS for the five phonemes used in the experiments. Here, the phonemes */b/* and */d/* are shorter than the 32 ms required to calculate the spectro-temporal patterns, so the signals are first zero-padded at the beginning and the end, maintaining the phoneme in the central portion. The spectrograms show in red (light gray) the highest amplitudes, highlighting the high frequency features of the shortest phonemes and the voiced features of the longest ones.

In a previous work using clean speech, we obtained different dictionaries of two-dimensional atoms through the Basis Pursuit algorithm and trained a number of neural networks with the spectro-temporal patterns using (9), with exhaustive tests for both the complete and overcomplete cases (Rufiner et al., 2007). The classification experiments were carried out by means of an artificial neural network, namely a *multi-layer perceptron* (MLP). The best performance was obtained with a dictionary size of 256 atoms. This corresponds to the complete case given that each atom has a

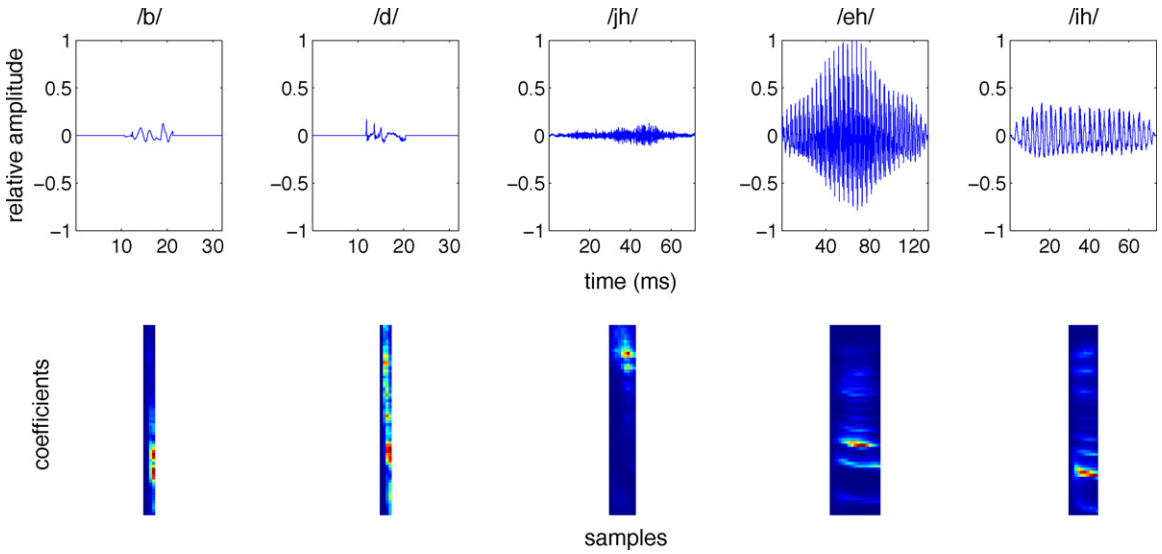

Fig. 4. Examples of the five phonemes used in the experiments showing the sonogram (top) and their respective auditory spectrogram (bottom). The spectrograms have 64 frequency coefficients in height from 0 to 8 kHz and a minimum of 4 coefficients in length, as it can be seen in the shortest phonemes.

dimension of 64 frequency coefficients by 4 frames. In spite of there being no evidence that it would be the best choice for noisy speech as well, it is the configuration used in this work.

In order to obtain the patterns that are the inputs to the classifiers, speech utterances corresponding to the phonemes are processed by the auditory model and their AS are obtained. Then, using the dictionary previously computed, the activation coefficients are calculated. This operation is carried out using the MP algorithm explained in Section 2.3, giving patterns with 256 coefficients (recall that only a subset of them are different from zero).

The noisy version of the corpus was obtained by mixing the clean data with white noise taken from the NOISEX-92 database (Varga and Steeneken, 1993). The noise was first conveniently resampled at 16 kHz with a resolution of 16 bits in order to match the characteristics of the clean signals. Finally, both signals were additively mixed at different signal-to-noise ratios.

4.2. The features

The suitability of the AACR approach for robust phoneme recognition was evaluated by comparing the performance in classification against different parameterizations used in this area: the Mel frequency cepstral coefficients (MFCC) (Deller et al., 1993), the auditory spectrogram, the PLP coefficients, the *Relative Spectral Transform* applied to the previous ones (RASTA-PLP) (Hermansky, 1990) and the *Probabilistic Optimum Filtering* (POF) applied to the MFCC coefficients. The POF analysis consists of a mapping between a pair of acoustic spaces: the clean and noisy speech features. The mapping tries to estimate the clean feature vectors by means of a probabilistic piece-wise linear transformation from the noisy features (Neumeyer and Weintraub, 1994).

4.3. The classifier

This paper focuses on assessing the advantages of the proposed AACR method over the other parameterizations in the representation of isolated phonemes. Thus, a static classifier was used because in this stage of the investigation there is no need to incorporate either a language model or temporal dynamics, in spite of the fact that the patterns are of variable length.

The MLP was used as classifier. It has a fixed number of input units that receives one vector of 256 activations at each time, corresponding to a STRF in $\mathbb{R}^{64\times4}$. Longer phonemes contribute with more patterns due to the sliding window used to extract the STRF. The architecture of the MLPs consisted of one input layer, where the number of input units depended on the dimension of the patterns; one hidden layer with a variable number of units and one output layer of 5 units. The training of the networks was conducted with the standard backpropagation algorithm with momentum term (Haykin, 1999).

5. Results and discussion

5.1. Dictionary for the sparse representation

Fig. 5 shows some of the STRFs corresponding to the complete estimated dictionary $\vec{\Phi} \in \mathbb{R}^{256x256}$. The STRFs are in $\mathbb{R}^{64\times4}$, with frequency content from 0 to 4 kHz and 32 ms in length.

The obtained STRFs present some characteristics of typical behaviors. It can be observed that they act like detectors of diverse significant features present in the spectrogram: unique frequencies, stable speech formant patterns, changes in the speech formants, unvoiced or fricative components (e.g. atom located in the dictionary at row/column 2/2 in the figure) and well-located patterns in time and/or frequency (e.g. the dark marks in atoms located at 1/2 and 1/6). The general similarities with the STRF patterns found in mammals are also revealed by comparing a pair of them taken from sound signals in animals with patterns here estimated, as can be seen to the left of Fig. 5 (Kording et al., 2002).

5.2. Initial tuning of the method

The first series of experiments, using clean speech, was devoted to find the optimum number of coefficients in the Matching Pursuit feature extraction scheme.

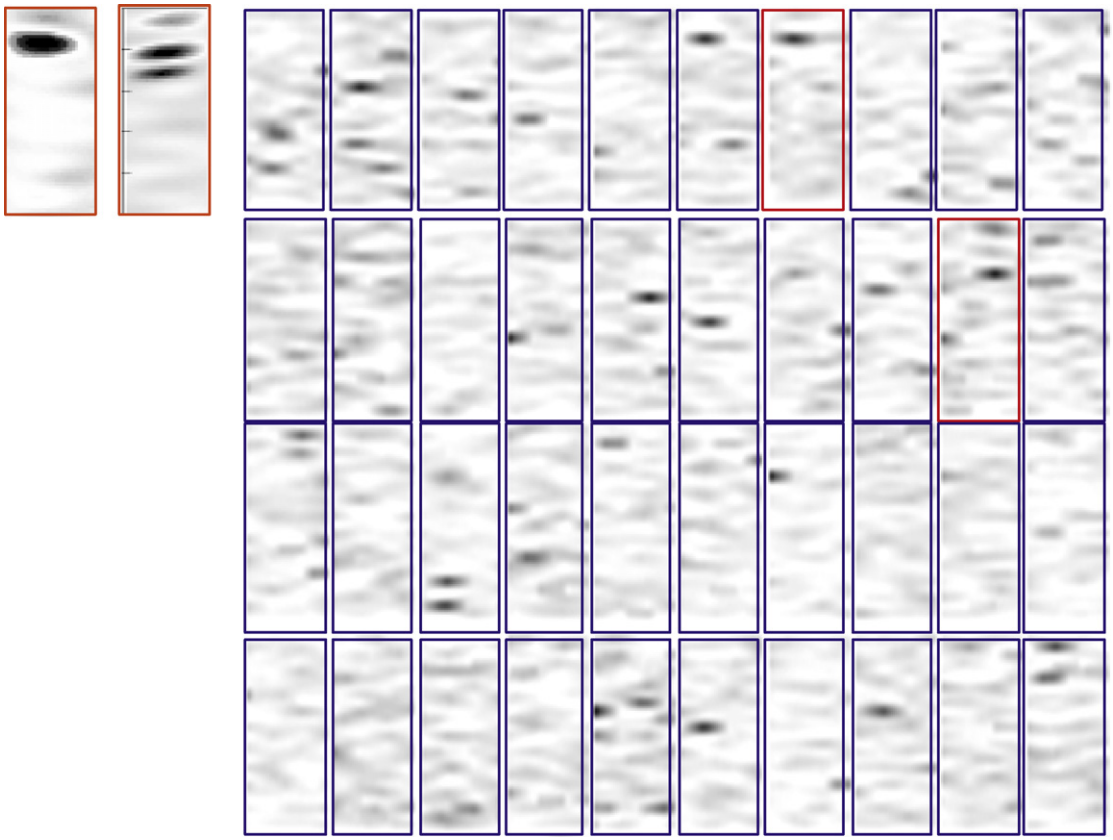

Fig. 5. Example of spectro-temporal receptive fields (STRF) in $\mathbb{R}^{64\times4}$ calculated from the early auditory representation of phoneme utterances. Two examples of biological STRF as found in animals are shown to the left and compared (in red) with similar patterns as estimated in the discrete dictionary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

From the TIMIT DR1 training set, we constructed two subsets for carrying out these experiments. In order to avoid bias in the results due to the class imbalance (showed in Table 1), the data consisted of 100 patterns of each phoneme for the training subset and 25 patterns for the test subset. The exploration was carried out with 4, 8, 16, 32, 64 and 128 selected coefficients of the complete vector in $\mathbb{R}^{256}$. Also, the best network architecture was found by varying the number of hidden units with a powers-of-two law, from 4 up to 512 units. Each experiment consisted of 3 runs with different initial weights at random, reporting the mean value obtained on the test subset.

The results of this initial tuning are presented in Fig. 6, where a similar behavior for all the curves was observed in general. They showed a lower performance when the size of the hidden layer was reduced, because of the limited capability of the MLPs to learn the key aspects of the patterns. Also, the performance achieved a maximum and then flattened as the size of the hidden layer was increased, due to the greater number of weights to adapt. Regarding the differences found when varying the number of selected coefficients, the best performances were obtained with few selected coefficients, showing the curves a general dropping when this parameter increased. This situation may arise owing to the fact that the patterns contain more non-relevant information to the classification.

From these curves, it can be seen that the best recognition rate is achieved by retaining only 8 coefficients in the MP algorithm. This is the situation expected, since the representation obtained is truly sparse (few active atoms for each analyzed pattern). In these conditions, the important cues of each pattern would be encoded in about 3% from the total of atoms in the dictionary. Also, this representation is better processed by an MLP with a low dimension in the hidden layer, in this case 32 nodes. This makes clear the generalization capabilities of the networks, given that the patterns carry only the most important information and therefore fewer weights are required. Therefore, this configuration of the MP algorithm and MLP architecture is set for the following experiments.

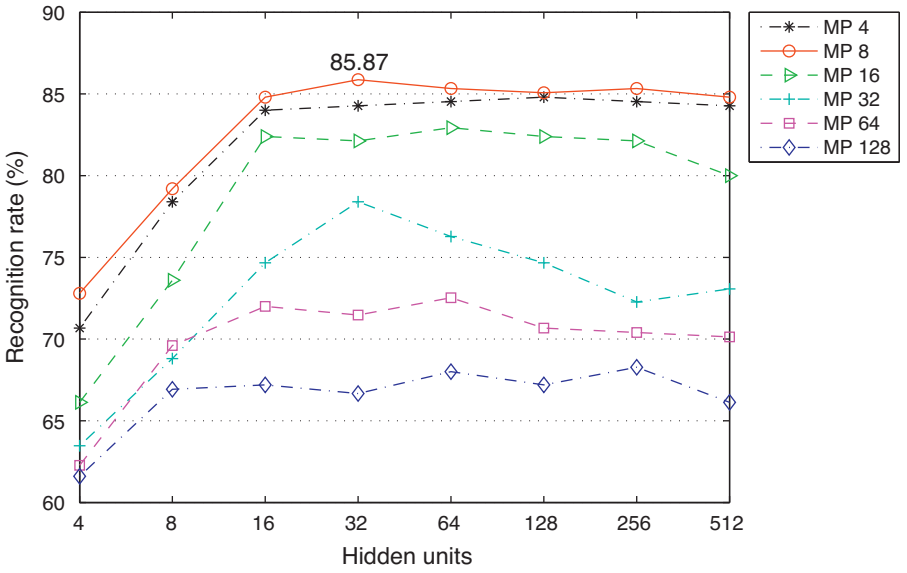

Fig. 6. Initial tuning of the number of selected coefficients in the algorithm and hidden units in the neural networks. The best performance is obtained for 8 selected coefficients and 32 nodes in the hidden layer of the MLP.

5.3. Robust phoneme classification

With the aim of evaluating the performance of the cortical representation in the presence of noise and to compare its robustness with other parameterizations, the next series of experiments consisted in training the MLPs with clean speech and testing them in different noisy conditions.

The feature extraction for MFCC, PLP and RASTA-PLP was fixed to 12 coefficients with frame energy and delta coefficients added, resulting in patterns of 26 coefficients. The patterns obtained from the AS have 256 coefficients. For all these networks, the number of hidden units was fixed to the same number of input units, as it was found to be the optimal configuration in preliminary experiments.

In each experiment, a different SNR was fixed from clean speech up to 0 dB (equal energy levels of noise and speech). Then, for each parameterization, a series of 10 runs with different initial network weights chosen at random was conducted. This initialization method seems to be good enough for our purposes, given that the generalization peak of the MLPs is reached at roughly 10–20 iterations of the backpropagation algorithm in a 200 iterations cycle. Moreover, the variance in the results remains between 1% and 3% in the case of clean test signals.

The obtained results are shown in Fig. 7. The curves show the general behavior of artificial systems in the presence of noise: they achieve a good recognition rate with clean speech, with performance falling as the noise content in the signal increases. The performance of the less robust parameterization, the MFCC, quickly drops in severe noisy conditions (SNR near to 0 dB). All the other parameterizations obtain higher rates in these conditions, as can be seen for 5 and 0 dB. In Neumeyer and Weintraub (1995) authors showed better performance of POF mapping than MFCC at higher noise levels also, but those results were obtained on continuous speech and using hidden Markov models as classifiers. These experimental conditions are very different from our configuration of isolated phoneme recognition by a static classifier.

The AACR approach here proposed always achieves the highest classification rates with respect to the other parameterizations, for all the SNRs evaluated including clean speech. This result would be given by the intrinsic robustness of the AACR, where only the more important activations are retained by the algorithm. Thus, the selected coefficients are acting as phonetic clues that capture the particularities of each phoneme and enable their characterization.

A more in depth analysis of the results is presented by the confusion matrices showed in Table 2. They show the mean recognition rate in percent for each phoneme using the best configuration found with the proposed AACR approach: 8 coefficients in the MP algorithm and the MLP with 32 units in the hidden layer. The rates correspond to the mean test values for the 10 initializations, and two different conditions are evaluated: with clean signals and with noise added

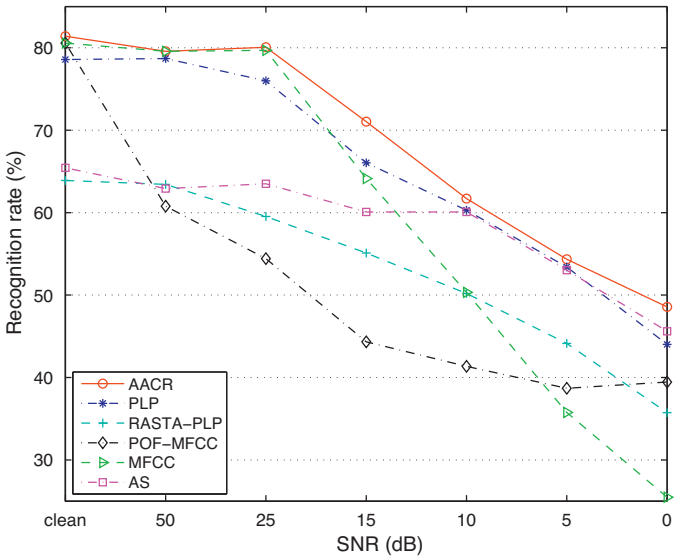

Fig. 7. Recognition rates in percent for the classification of the 5 phonemes in the presence noise at different SNR, from clean speech up to same energy levels of speech and noise (SNR = 0 dB).

Table 2
Confusion matrices showing the classification percentages for the AACR approach with clean and noisy speech at SNR = 15 dB. In rows: teaching output, in columns: classification. Mean recognition rate: 83% (clean speech) and 71% (noisy speech).

Actual phoneme	Clean speech					Noisy speech at SNR = 15 dB				
	/b/	*/d/*	*/jh/*	*/eh/*	*/ih/*	*/b/*	*/d/*	*/jh/*	*/eh/*	*/ih/*
/b/	84	16				33	67			
/d/	16	70	10		4	5	94	1		
/jh/		1	99				12	88		
/eh/	5			93	2	5	1		92	2
/ih/	2	2	10	26	60	3	17	15	18	46

at SNR = 15 dB. For each phoneme in the first column (teaching output of the MLPs), the matrices show in rows the percentages of each phoneme given by the networks.

Results showed that, in the clean condition, the MLP is able to carry out an adequate classification of almost all the classes except the phoneme */ih/*, which is spread in the rest of classes (last row). In the noisy condition, it can be seen that the networks classify very well the phonemes */d/*, */jh/* and */eh/*, whereas phoneme */ih/* is mainly assigned to the other classes. The case for phoneme */b/* is also interesting to analyze. In the clean case a good performance is obtained (84%), with a minor confusion with phoneme */d/*. With the introduction of noise, even in moderate amounts, this confusion is increased: 67% of */b/* are recognized as */d/*. This behaviour could be explained by the fact that in clean speech these plosive, voiced consonants show a high energy content at low frequencies in the AS, but phoneme */d/* also presents more energy at higher frequencies unlike phoneme */b/* (see Fig. 4). When white noise is added, the AS of */b/* resembles more to that of */d/*, giving rise to the misclassification found. A study in line with this idea was presented in Mesgarani et al. (2008), where authors showed that these phonemes are very confusable given their high acoustic similarity (Euclidean distance between their average auditory spectrograms). Similarly, due to the energy content of phoneme */d/*, when noise is added the patterns become more similar to the learned examples. Therefore, the initial confusion of phoneme */d/* with others is reduced given that the noisy phoneme is very different from the clean */b/*, */jh/* and */ih/*.

The statistical significance of these results was evaluated considering the probability that the classification error of a given classifier ϵ is smaller than the one of the reference system ϵ_{ref}. In order to make this estimation, the statistical independence of the errors for each frame was assumed, and the binomial distribution of the errors was

modeled by means of a Gaussian distribution (this is possible because a sufficiently large number of test patterns are given). Therefore, comparing our approach against the second best result (auditory spectrogram) for the worst case, SNR = 0 dB, a $Pr(\epsilon_{ref} > \epsilon) > 96.54\,\%$ was obtained. The standard deviation for the AACR ranges from 0.88 (clean speech) up to 2.31 (SNR = 0 dB), whereas for the PLP coefficients the same parameter has a higher variation: from 0.87 up to 10.71, respectively.

The use of additive white noise is probably the most studied and straightforward way to simulate the operation of an artificial system in adverse conditions. In speech applications, perhaps the babble noise would be the more difficult one to deal with, due to their concentration of energy in the same range as speech formants, masking important features for the classification. A previous study in robust speech recognition support this hypothesis (Rufiner et al., 2003). The performance of phoneme recognition systems with this and other types of noise is a topic to further explore.

6. Conclusions

In this paper, a biologically inspired sparse method for speech parameterization was presented. From the auditory spectrograms, the technique calculates an optimal dictionary of atoms. The extracted feature vectors consist of the activation coefficients obtained with the Matching Pursuit algorithm, which selects the more representative ones. Using a dictionary of 256 atoms, the optimal sparse representation for each speech segment is obtained by selecting only 8 atoms. Thus, in adverse environments, a sort of thresholding of noisy components is carried out, while the most important cues are preserved.

The feasibility of building a robust phoneme recognizer using this representation was evaluated in the classification of five highly confusing English phonemes. The performance of our approach was compared against a number of standard and robust parameterizations, namely the PLP and MFCC among others. The recognition experiments were carried out using multilayer perceptrons. Results showed that the approximated cortical representation always improves the recognition rate obtained by the rest of parameterizations, for both the clean case and in the presence of additive white noise at different signal to noise ratios (from 50 dB up to 0 dB).

Future direction in this investigation would be devoted to optimize the denoising of the speech activation patterns, explore a discriminative learning of the dictionaries and to explore this feature extraction scheme in the major problem of large vocabulary continuous speech recognition.

Acknowledgements

The authors wish to thank: the *Agencia Nacional de Promoción Científica y Tecnológica*, the *Universidad Nacional de Litoral* (with CAI+D 012-72, PAE 37122, PAE-PICT-2007-00052), the *Universidad Nacional de Entre Ríos* (with PID 61111-2 and 6106), the *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina, for their support and the anonymous reviewers for their thoughtful suggestions.

References

Ager, M., Cvetković, Z., Sollich, P., Yu, B., 2008. Towards robust phoneme classification: augmentation of PLP models with acoustic waveforms. In: Proceedings of EUSIPCO.

Barlow, H., 2001. Redundancy reduction revisited. Network: Computation in Neural Systems 12, 241–253.

Chen, S., Donoho, D., Saunders, M., 2001. Atomic decomposition by basis pursuit. SIAM Review 43 (1), 129–159.

Chi, T., Ru, P., Shamma, S., 2005. Multiresolution spectrotemporal analysis of complex sounds. Journal of the Acoustic Society of America 118 (2), 887–906.

Delgutte, B., 1996. Physiological models for basic auditory percepts. In: Hawkins, H.H., McMullen, T.A., Popper, A.N., Fay, R.R. (Eds.), Auditory Computation. Springer, New York.

Deller, J., Proakis, J., Hansen, J., 1993. Discrete Time Processing of Speech Signals. Macmillan Publishing, New York.

Donoho, D., Elad, M., 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proceedings of the National Academy of Sciences 100 (5), 2197–2202.

Donoho, D., Elad, M., Temlyakov, V., 2006. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Transactions on Information Theory 52 (1), 6–18.

Ganapathy, S., Thomas, S., Hermansky, H., 2010. Temporal envelope compensation for robust phoneme recognition using modulation spectrum. Journal of the Acoustic Society of America 128 (6), 3769–3780.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. DARPA TIMIT Acoustic-phonetic continuous speech corpus documentation. Technical report. National Institute of Standards and Technology.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Pearson Education.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustic Society of America 87 (4), 1738–1752.

Hyvärinen, A., 1998. Sparse code shrinkage: denoising of nongaussian data by maximum-likelihood estimation. Technical Report, Helsinki University of Technology.

Klein, D., Konig, P., Kording, K., 2003. Sparse spectrotemporal coding of sounds. EURASIP Journal on Applied Signal Processing 2003 (7), 659–667.

Kording, K., Konig, P., Klein, D., 2002. Learning of sparse auditory receptive fields. In: Proceedings of the International Joint Conference on Neural Networks, IJCNN'02, vol. 2, Honolulu, HI, United States, pp. 1103–1108.

Lewicki, M., Sejnowski, T.,1998. Learning overcomplete representations. In: Proceedings of Advances in Neural Information Processing 10, NIPS'97. MIT Press, pp. 556–562.

Lewicki, M., Olshausen, B., 1999. A probabilistic framework for the adaptation and comparison of image codes. Journal of the Optical Society of America 16 (7), 1587–1601.

Mallat, S.G., Zhang, Z., 1993. Matching pursuit with time-frequency dictionaries. IEEE Transactions on Signal Proccesing 41, 3397–3415.

Mesgarani, N., David, S., Britz, J., Shamma, S., 2008. Phoneme representation and classification in primary auditory cortex. Journal of the Acoustic Society of America 123 (2), 899–909.

Mitchell, T., Shinkareva, S., Carlon, A., Chang, K-M., Malave, V., Mason, R., Just, M., 2008. Predicting human brain activity associated with the meanings of nouns. Science 320, 1191–1195.

Neumeyer, L., Weintraub, M., 1994. Probabilistic optimum filtering for robust speech recognition. In: Proceedings of ICASSP, vol. 1, pp. 17–20.

Neumeyer, L., Weintraub, M., 1995. Robust speech recognition in noise using adaptation and mapping techniques. In: Proceedings of ICASSP, vol. 1, pp. 141–144.

Oja, E., Hyvärinen, A., 2000. Independent component analysis: a tutorial. Neural Networks 13 (4–5).

Olshausen, B., Field, D., 1996. Emergence of simple cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609.

Rubinstein, R., Zibulevsky, M., Elad, M., 2010. Double sparsity: learning sparse dictionaries for sparse signal approximation. IEEE Transactions on Signal Processing 58 (3), 1553–1564.

Rufiner, H.L., Torres, M.E., Gamero, L., Milone, D.H., 2003. Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition. Physica A: Statistical Mechanics and its Applications 332, 496–508.

Rufiner, H., Martínez, C., Milone, D., Goddard, J., 2007. Auditory cortical representations of speech signals for phoneme classification. In: MICAI 2007: Advances in Artificial Intelligence, vol. 4827 of Lecture Notes in Computer Science. Springer-Verlag, pp. 1004–1014.

Stevens, K.N., 2000. Acoustic Phonetics. MIT Press.

Theunissen, F., Sen, K., Doupe, A., 2000. Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. Journal of Neuroscience 20, 2315–2331.

Varga, A., Steeneken, H., 1993. Assessment for automatic speech recognition. II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12 (3), 247–251.

Yang, X., Wang, K., Shamma, S., 1992. Auditory representations of acoustic signals. IEEE Transactions on Information Theory 38, 824–839, Special Issue on Wavelet Transforms and Multiresolution Signal Analysis.

Yousafzai, J., Cvetković, Z., Sollich, P., Ager, M., 2008. Robustness of phoneme classification using support vector machines: a comparison between PLP and acoustic waveform representations. In: Proceedings of ICASSP.

Yousafzai, J., Cvetković, Z., Sollich, P., 2009. Tuning support vector machines for robust phoneme classification with acoustic waveforms. In: Proceedings of INTERSPEECH.