Auditory Cortical Representations of Speech Signals for Phoneme Classification

Hugo L. Rufiner[1,2], César E. Martínez[1,2], Diego H. Milone[2], and John Goddard[3]

[1] Laboratorio de Señales e INteligencia Computacional (SINC), Depto Informática
Facultad de Ingeniería y Cs Hídricas - Universidad Nacional del Litoral
CC 217, Ciudad Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina
Tel.: +54 (342) 457-5233 x 148
lrufiner@fich.unl.edu.ar
[2] Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
[3] Dpto. de Ingeniería Eléctrica - UAM-Iztapalapa, México

Abstract. The use of biologically inspired, feature extraction methods has improved the performance of artificial systems that try to emulate some aspect of human communication. Recent techniques, such as independent component analysis and sparse representations, have made it possible to undertake speech signal analysis using features similar to the ones found experimentally at the primary auditory cortex level. In this work, a new type of speech signal representation, based on the spectro-temporal receptive fields, is presented, and a problem of phoneme classification is tackled for the first time using this representation. The results obtained are compared, and found to greatly improve both an early auditory representation and the classical front-end based on Mel frequency cepstral coefficients.

1 Introduction

The development of new techniques in the signal analysis and representation fields promises to overcome some of the limitations of classical methods for solving real problems with complex signals, such as those related to human speech recognition. Furthermore, novel techniques for signal representation, for example those using overcomplete dictionaries, provide an important new way of thinking about alternative solutions to problems such as denoising or automatic speech recognition. Significant connections have been found between the way the brain processes sensory signals and some of the principles that support these new approaches [1,2].

In the process of human communication, the inner ear –at the cochlea level– carries out a complex time-frequency analysis and encodes a series of meaningful cues in the discharge patterns of the auditory nerve. These early auditory representations, or *auditory spectrograms*, have been widely studied and mathematical and computational models have been developed that allow them to be estimated approximately [3].

A. Gelbukh and A.F. Kuri Morales (Eds.): MICAI 2007, LNAI 4827, pp. 1004–1014, 2007.

In spite of the knowledge available about early auditory representations, the principles that support speech signal representation at higher sensorial levels, as in the primary auditory cortex, are still the object of research [4]. Among these principles two can be singled out, namely, the need for very few active elements in a signal's representation, and the statistical independence between these elements. Using these principles, it is then possible to establish a model for cortical representations that displays important correlations with experimentally obtained characteristics from their physiological counterparts [5,6].

In order to obtain this cortical model, techniques related to *independent component analysis* (ICA) and *sparse representations* (SR) are used [7,8]. These techniques can emulate the behavior of cortical neurons using the notion of *spectro-temporal receptive fields* (STRF) [9]. STRF can be defined as the required optimal stimulus so that an auditory cortical neuron responds with the largest possible activation. Different methods, such as inverse correlation, are used to estimate them from mammal neuronal activity data [10].

In this work, by making use of the time-frequency representations of the auditory spectrograms of 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, designed to evaluate the representation's suitability.

This work is organized as follows: Section 2 presents the method for the speech signal representation that is used in the paper. In particular, 2.3 explains how this representation can include one involving the primary auditory cortex. Section 3 details the data used in the phoneme classification experiments as well as the steps used to obtain the cortical representation patterns. Section 4 presents the experimental results together with a discussion. Finally, Section 5 summarizes the contributions of the present paper and outlines future research.

2 Sparse and Factorial Representations

2.1 Representations Based on Discrete Dictionaries

There are different ways of representing a signal using general discrete dictionaries. For the case where the dictionary forms a unitary or orthogonal transformation, the techniques are particularly simple. This is because, among other aspects, the representation is unique. However, in the general, non-orthogonal case, a signal can have many different representations using a single 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 sensorial characteristics which are sparse and independent [11], as mentioned in the introduction. Furthermore, it is possible to find an optimal dictionary using these criteria [12].

A sparse code is one which represents the information in terms of a small number of descriptors taken from a large set [12]. This means that a small fraction of the elements from the code are used actively to represent a typical

pattern. In numerical terms, this signifies that the majority of the elements are zero, or 'almost' zero, most of the time [13,14].

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 way is to use a probability distribution. In general one uses a 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, which connect this approach with ICA [7].

In the following subsection a formal description is given of a statistical method which estimates an optimal dictionary and the corresponding representation[1].

2.2 Optimal Sparse and Factorial Representations

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

$$\boldsymbol{x} = \sum_{\gamma \in \Gamma} \boldsymbol{\phi}_\gamma a_\gamma + \boldsymbol{\varepsilon} = \boldsymbol{\Phi a} + \boldsymbol{\varepsilon} \ , \tag{1}$$

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

Although (1) appears very simple, the main problem is that for the most general case $\boldsymbol{\Phi}$, $\boldsymbol{a}$ and $\boldsymbol{\varepsilon}$ are unknown, thus there can be an infinite number of possible solutions. Furthermore, in the noiseless case (when $\boldsymbol{\varepsilon} = \boldsymbol{0}$) and given $\boldsymbol{\Phi}$, if there are more atoms than there are samples of $\boldsymbol{x}$, or if the atoms don't form a base, then non-unique representations of the signal are possible. Therefore, an approach that allows us to select one of these representations has to be found. In this case –although this is a linear system– the coefficients chosen to be part of the solution generally form a non-linear relation with the data $\boldsymbol{x}$. For the complete, noiseless case the relationship between the data and the coefficients is linear and it is given by $\boldsymbol{\Phi}^{-1}$. For classical transformations, such as the discrete Fourier transform, this inverse is simplified because $\boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^*$ (with $\boldsymbol{\Phi} \in \mathbb{C}^{N \times M}$ and $\Phi^*(i,j) = \overline{\Phi(j,i)}$).

When $\boldsymbol{\Phi}$ and $\boldsymbol{x}$ are known, an interesting way to choose the set of coefficients $\boldsymbol{a}$ from among all the possible representations, consists in 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, a joint *a priori* distribution satisfies:

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

[1] Although two-dimensional patterns are used, for clearness we only describe the one-dimensional case.

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

The state vector $\boldsymbol{a}$ can be estimated from the *posterior* distribution:

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

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

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} \left[\log P(\boldsymbol{x}|\boldsymbol{\Phi}, \boldsymbol{a}) + \log P(\boldsymbol{a})\right] \quad . \tag{4}$$

When $P(\boldsymbol{a}|\boldsymbol{\Phi}, \boldsymbol{x})$ 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 [15] proposed the use of a Laplacian *a priori* distribution with parameter β_i:

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

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

$$\Delta\boldsymbol{a} = \boldsymbol{\Phi}^T \boldsymbol{\Lambda}_{\boldsymbol{\varepsilon}} \boldsymbol{\varepsilon} - \boldsymbol{\beta}^T \left|\boldsymbol{a}\right| \quad , \tag{6}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\varepsilon}}$ is the inverse of the noise covariance matrix $\mathcal{E}\left[\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}\right]$, with $\mathcal{E}[\cdot]$ denoting the expected value.

To estimate the value of $\boldsymbol{\Phi}$, the following objective function can be maximized [15]:

$$\widehat{\boldsymbol{\Phi}} = \arg\max_{\boldsymbol{\Phi}} \left[\mathcal{L}(\boldsymbol{x}, \boldsymbol{\Phi})\right] \quad , \tag{7}$$

where $\mathcal{L} = \mathcal{E}[\log P(\boldsymbol{x}|\boldsymbol{\Phi})]_{P(\boldsymbol{x})}$ 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 coefficients, together with the coefficients *a priori* distribution:

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

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

The objective function in (7) can be maximized using gradient ascent with the following update rule for the matrix $\boldsymbol{\Phi}$ [16]:

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

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

In this iterative way, the dictionary $\boldsymbol{\Phi}$ and the coefficients $\boldsymbol{a}$ were obtained.

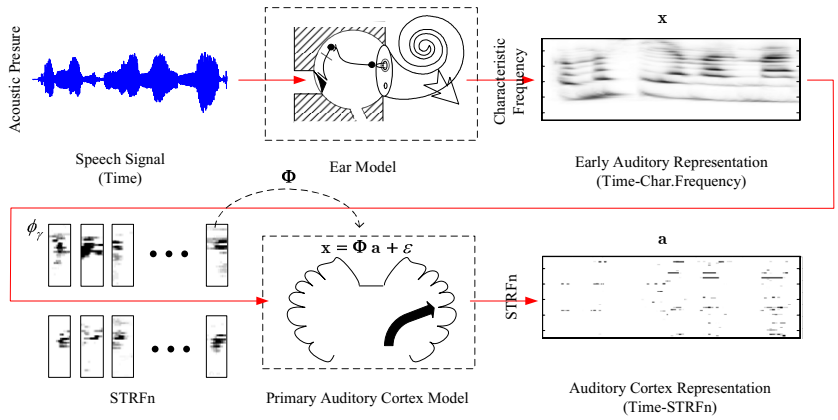

Fig. 1. Schematic diagram of the method used for estimating the auditory cortical representation

2.3 Auditory Cortical Representations

The properties of sensorial systems should coincide with the statistics of their perceived stimuli [17]. 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. In these representations –of a higher level than the acoustic one– some non-relevant aspects of the temporal variability of the sound pressure signal that arrives at the eardrum have been eliminated. Among these superfluous aspects, for example, is the relative phase from some acoustic waveforms [18]. Hence, following this biological simile, this representation forms a good starting point to attain more complex ones.

The obtention of a dictionary of two-dimensional atoms $\boldsymbol{\Phi}$, corresponding to time-frequency features estimated from data $\boldsymbol{x}$ of the auditory spectrogram, is equivalent to the STRF of a group of cortical neurons. Therefore, the activation level of each neuron can be assimilated with the coefficients a_γ in (1). Figure 1 shows a schematic diagram of the method adopted for estimating the cortical representation.

Kording *et al* carried out a qualitative analysis of dictionaries obtained in a similar way, and they found that their properties compared favorably with those of the natural receptive fields [5].

3 Experiments and Data

According to the previous considerations an experiment of phoneme classification was designed to evaluate the performance of a system that uses a cortical representation for this task. The speech data, corresponding to region DR1 of

Table 1. Distribution of patterns per class for training and test data

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)

TIMIT corpus [19] for the set of five highly confusing phonemes */b/*, */d/*, */jh/*, */eh/*, */ih/*, were used (See Table 1).

For each one of the emissions, sampled at 16 KHz, the corresponding auditory spectrogram was calculated from an early auditory model [20]. Then, the frequency resolution of the data was reduced so as to diminish its dimensions. After that, auditory spectrograms with a total of 64 frequency coefficients per time unit were obtained. Finally, by means of a sliding window of 32 ms in length at intervals of 8 ms, the set of spectro-temporal patterns that served as a base for the estimation of the dictionaries were obtained. In Figure 2, the main steps in this process, as well as the corresponding signals are shown.

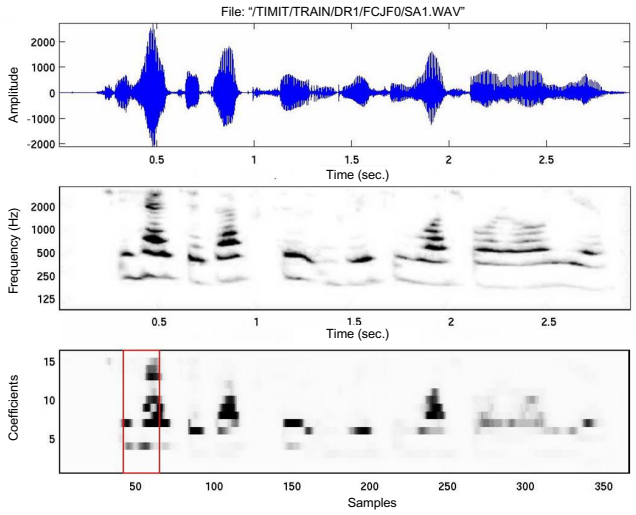

Fig. 2. Main steps in the process used to generate the spectro-temporal patterns that serve as a basis for obtaining the STRF: sonogram (top), original auditory spectrogram (center) and low-resolution spectrogram (bottom). In this last representation, a section corresponding to the sliding window, from which each spectro-temporal pattern is generated, has been marked.

From these spectro-temporal patterns, different dictionaries of two-dimensional atoms were trained using (9) [21]. Several tests for both the complete and overcomplete cases of each configuration were conducted.

Once the STRF were obtained, the activation coefficients were calculated in an iterative form using (6) from the auditory spectrograms. For comparison purposes the *mel frequency cepstral coefficients* (MFCC) with an energy coefficient (MFCC+E) were calculated in the usual way for two consecutive frames, resulting in patterns in $\mathbb{R}^{28}$ [22].

The classification was carried out for each experiment using an artificial neural network, namely a *multi-layer perceptron* (MLP). The network architecture consisted of one input layer, where the number of input units depended on the dimension of the patterns, one hidden layer, and one output layer of 5 units. The number of units in the hidden layer was varied, depending on the experiment.

4 Results and Discussion

An example of some of the STRFs obtained is shown in Figure 3. This case corresponds to the complete dictionary $\boldsymbol{\Phi} \in \mathbb{R}^{256x256}$, using patterns of 64x4. Several typical behaviors can be observed, which are useful for discriminating between the different phonemes used. The relative position of each element in the dictionary is related to its similarity with the other elements in the dictionary (in terms of the ℓ_2 norm of their differences). It is possible to observe that some STRF seem to act like detectors of diverse significant phonetic characteristics,

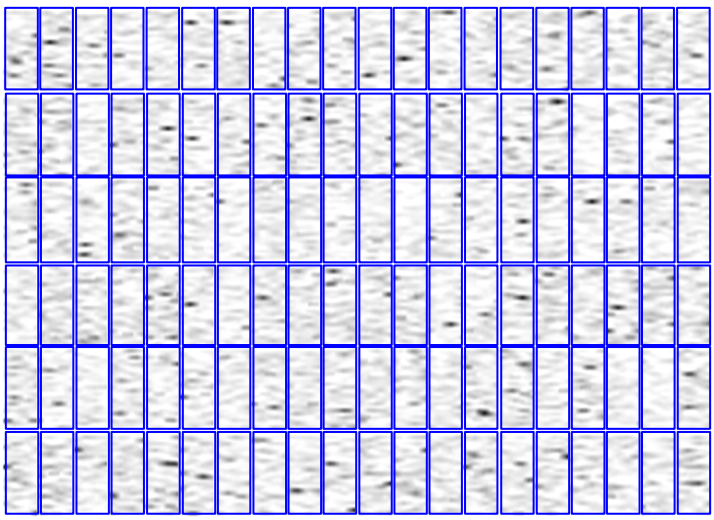

Fig. 3. Some spectro-temporal receptive fields as obtained from the patterns of 64x4 points of the early auditory representations. Each STRF has 4 KHz height and 32 ms width. The speech utterances were taken from five phonemes of TIMIT corpus, region DR1.

Table 2. Recognition rates of phoneme classification experiments with MLP using the representations generated by means of the early auditory models, the cortical representation obtained from the activation of the STRF, and MFCCs (best results in bold)

N°	Experiment		Network	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Auditory	64x4	256/4/5	45.84	44.76	0.00	0.00	6.90	100.00	6.27
2			256/8/5	44.35	43.25	0.00	0.00	4.31	100.00	3.13
3			256/16/5	64.28	65.03	0.00	0.00	9.48	94.99	57.59
4			256/32/5	68.92	69.67	0.00	0.00	100.00	95.87	54.82
5			256/64/5	70.70	72.69	0.00	0.00	83.62	72.34	86.75
6			256/128/5	70.50	72.17	4.55	0.00	62.93	84.73	76.14
7			256/256/5	72.15	73.74	0.00	0.00	97.41	85.23	74.82
8			256/512/5	69.21	71.76	0.00	0.00	100.00	94.49	60.96
9	Cortical	64x4	256/4/5	77.04	75.72	40.91	56.48	97.41	84.86	69.16
10			256/8/5	79.64	**77.64**	46.97	62.96	93.97	84.86	72.77
11			256/16/5	75.60	76.08	65.15	51.85	97.41	89.99	63.73
12			256/32/5	79.72	74.73	65.15	67.59	98.28	79.22	68.80
13			256/64/5	87.27	76.86	74.24	66.67	95.69	88.24	64.82
14			256/128/5	100.00	**78.37**	72.73	70.37	96.55	78.35	77.35
15			256/256/5	98.10	77.07	65.15	71.30	91.38	87.11	67.11
16			256/512/5	99.92	**79.16**	71.21	69.44	92.24	80.35	78.07
17	Cortical	64x4x2	512/4/5	78.65	73.79	48.48	59.26	86.21	85.61	64.58
18			512/8/5	80.62	75.51	63.64	59.26	98.28	85.36	65.90
19			512/16/5	78.65	74.26	54.55	53.70	99.14	82.98	66.63
20			512/32/5	82.58	75.66	62.12	66.67	95.69	85.11	66.02
21			512/64/5	87.27	75.87	54.55	65.74	98.28	83.48	68.43
22			512/128/5	84.72	75.98	65.15	56.48	95.69	84.23	68.67
23			512/256/5	81.37	76.55	65.15	62.96	95.69	86.86	66.63
24			512/512/5	82.64	76.32	65.15	61.11	97.41	77.97	74.70
25	MFCC+E	14+14	28/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40

e.g. unique frequencies, stable speech formant patterns, changes in the speech formants, unvoiced or fricative components, and well-located patterns in time and/or frequency.

The results of the experiments described in the previous section are detailed in Table 2. As can be seen from this table, the results of classification on the training and test data for the cortical representation are better than those obtained when using the direct (or early) auditory representation. For the latter representation, some of the classification results are apparently globally good, however, when the individual phoneme classification rates (exhibited in the rightmost columns of the table) are examined, only two or three phonemes are in fact correctly classified (see experiments N° 1-8). This problem arises because of a local minimum error solution that the cortical representation avoids (see the uneven pattern distribution in Table 1).

Moreover, the results for the cortical representation are better than those obtained using the classical MFCC representation on this task (see experiments N^o 16 and 25 in Table 2). Another important aspect is that the performance is satisfactory for relatively small network architectures in relation to the pattern dimensions. This aspect corroborates the hypothesis that the classes are better separated in this new higher dimensional space, and therefore a simpler classifier can complete the task successfully.

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 frames is given). Therefore, comparing the MFCC with the best result of the cortical representation, the $Pr(\epsilon_{ref} > \epsilon) > 92\%$ was obtained.

Harpur [13] conducted some simple experiments of phoneme classification using low entropy codes, with only positive coefficients generated from a filter bank. However, experiments using more complex models of the auditory receptive fields – such as the ones that appear here – have not been previously reported.

5 Conclusions

In this work, a new approach to speech feature extraction, based on a biological metaphor, has been proposed and applied to a phoneme classification task. This approach first finds an early auditory representation of the speech signal at the auditory nerve level. Then, based on analogies established with neuro-sensorial systems, an optimal dictionary is estimated from the auditory spectrograms of speech data.

The method finds a set of atoms which can be related to the spectro-temporal receptive fields of the auditory cortex, and which prove capable of functioning like detectors of important phonetic characteristics. It is worthwhile mentioning, for example, the detection of events based on highly localized spectro-temporal features, such as relatively stationary segments, different types of formant evolution, and non-harmonic zones.

Using representations provided by the method as the input patterns, multilayer perceptrons were trained as phoneme classifiers. The results obtained improve those of both early auditory and standard MFCC representations. The objective was not to find the best possible classifier, but rather to demonstrate the feasibility of the proposed method. Obviously, further experimentation is called for.

Another interesting issue, that also remains to be explored in future works, is the evaluation of the robustness of this type of representation in the presence of additive noise.

Acknowledgements

The authors wish to thank: the *Universidad Nacional de Litoral* (with UNL-CAID 012-72), the *Agencia Nacional de Promoción Científica y Tecnológica* (with ANPCyT-PICT 12700 & 25984) and the *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina and the *Consejo Nacional de Ciencia y Tecnología* (CONACYT) and the *Secretaría de Educación Pública* (SEP) from Mexico, for their support.

References

1. Greenberg, S.: The ears have it: The auditory basis of speech perception. In: Proceedings of the International Congress of Phonetic Sciences, vol. 3, pp. 34–41 (1995)
2. Rufiner, H., Goddard, J., Rocha, L.F., Torres, M.E.: Statistical method for sparse coding of speech including a linear predictive model. Physica A: Statistical Mechanics and its Applications 367(1), 231–250 (2006)
3. Delgutte, B.: Physiological models for basic auditory percepts. In: Hawkins, H., McMullen, T., Popper, A., Fay, R. (eds.) Auditory Computation, Springer, New York (1996)
4. Simon, J.Z., Depireux, D.A., Klein, D.J., Fritz, J.B., Shamma, S.A.: Temporal symmetry in primary auditory cortex: Implications for cortical connectivity. Neural Computation 19(3), 583–638 (2007)
5. Kording, K.P., Konig, P., Klein, D.J.: Learning of sparse auditory receptive fields. In: Proc. of the International Joint Conference on Neural Networks (IJCNN 2002), Honolulu, HI, United States, vol. 2, pp. 1103–1108 (2002)
6. Klein, D., Konig, P., Kording, K.: Sparse Spectrotemporal Coding of Sounds. EURASIP Journal on Applied Signal Processing 2003(7), 659–667 (2003)
7. Oja, E., Hyvarinen, A.: Independent Component Analysis: A Tutorial. Helsinki University of Technology, Helsinki (2004)
8. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proceedings of the National Academy of Sciences 100(5), 2197–2202 (2003)
9. Theunissen, F., Sen, K., Doupe, A.: Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J. Neuroscience 20, 2315–2331 (2000)
10. deCharms, R., Blake, D., Merzenich, M.: Optimizing sound features for cortical neurons. Science 280, 1439–1443 (1998)
11. Olshausen, B.: Sparse codes and spikes. In: Rao, R.P.N., Olshausen, B.A., Lewicki, M.S. (eds.) Probabilistic Models of the Brain: Perception and Neural Function, MIT Press, Cambridge, 2001 (in Press)
12. Olshausen, B., Field, D.: Emergence of simple cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
13. Harpur, G.F.: Low Entropy Coding with Unsupervised Neural Networks. PhD thesis, Department of Engineering, University of Cambridge, Queens' College (1997)
14. Hyvärinen, A.: Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation. Technical report, Helsinki University of Technology (1998)
15. Lewicki, M., Olshausen, B.: A probabilistic framework for the adaptation and comparison of image codes. Journal of the Optical Society of America 16(7), 1587–1601 (1999)

16. Abdallah, S.A.: Towards music perception by redundancy reduction and unsupervised learning in probabilistic models. PhD thesis, Department of Electronic Engineering, King's College London (2002)
17. Barlow, H.: Redundancy reduction revisited. Network: Computation in Neural Systems (12), 241–253 (2001)
18. Kwon, O.W., Lee, T.W.: Phoneme recognition using ICA-based feature extraction and transformation. Signal Processing 84(6), 1005–1019 (2004)
19. Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren.: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation. Technical report, National Institute of Standards and Technology (1993)
20. Yang, X., Wang, K., Shamma, S.A.: Auditory representations of acoustic signals. IEEE Trans. Inf. Theory. Special Issue on Wavelet Transforms and Multiresolution Signal Analysis 38, 824–839 (1992)
21. Lewicki, M., Sejnowski, T.: Learing overcomplete representations. In: Advances in Neural Information Processing 10 (Proc. NIPS 1997), pp. 556–562. MIT Press, Cambridge (1998)
22. Deller, J., Proakis, J., Hansen, J.: Discrete Time Processing of Speech Signals. Macmillan Publishing, New York (1993)