An approach to robust phoneme classification by modeling the auditory cortical representation of speech

C. Martínez[1,2] J. Goddard[3] D. Milone[1,2,4] H. Rufiner[1,2,4]

[1]*Centro de I+D en Señales, Sistemas e INteligencia Computacional (SINC(i)), Dpto. Informática*
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral
CC 217, Ciudad Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina
Tel: +54 (342) 457-5233 ext 148, cmartinez@fich.unl.edu.ar
[2]*Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos*
[3]*Dpto. de Ingeniería Eléctrica - UAM-Iztapalapa, México*
[4]*CONICET, Argentina*

***Abstract*— In this work, a first approach to a robust speech 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 nerve level. It is based on an optimal dictionary of atoms, estimated from auditory spectrograms, and the Matching Pursuit algorithm to obtain the activations. This provides a sparse coding with some basic degree of noise robustness that can be exploited in the application. The recognition task consisted in the classification of a set of 5 highly confusing English phonemes, on clean and noisy conditions. Multilayer perceptrons were trained as phoneme classifiers and the performance was compared to that obtained by a classic parameterization in this task, the Mel frequency cepstral coefficients. Results showed an improvement in the recognition rate using the approximated auditory cortical representation for both the clean case and in the presence of additive white noise.**

***Keywords*— robust speech recognition, auditory cortical representation, phoneme classification, sparse coding**

1. INTRODUCTION

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

In the speech analysis field, based on the biological time-frequency analysis the inner ear carries out, an auditory representation of the speech at cochlea level has been widely studied. Different mathematical and computational models have been developed that allow to approximately estimate the *auditory spectrograms* [3]. These investigations enabled the modeling of the discharge patterns of the auditory nerve, with significant connections between the brain processing and some of the principles that support these new approaches.

More precisely, given a speech utterance, a pattern of activation 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 is emulated using the notion of *spectro-temporal receptive fields* (STRF), which are defined as the required optimal stimulus so that an auditory cortical neuron responds with the largest possible activation [19]. Using discrete dictionaries, an approximated auditory cortical representation (AACR) can be established by means of techniques related to *independent component analysis* (ICA) and *sparse representations* [15].

A very challenging task in ASR is to build systems that perform well on real environments and conditions, where the speech could be registered with background noise, reverberance, other mixing voices, etc. [7]. During the last years, the sparse coding scheme had been applied to speech analysis and representation (for a review of these techniques and applications see [11]). Among other applications like speech enhancement, speaker recognition and speech separation, regarding the ASR task very few studies were presented. In a recent work, a first exploration to the continuous ASR with a system that obtains a sparse spike train from the spectrograms was presented [18]. On a continuous digits recognition task with only clean speech, it obtained worse results than a baseline system with cepstral coefficients and hidden Markov models as classifiers, showing the actual limitations of the approach.

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 in clean and noisy conditions, designed to evaluate the advantages and robustness of the representation.

The organization of the paper is as follows. Section 2 presents the method for the speech signal representation that is used in this work. Section 3 presents the information about the clean speech data and the noise corpus, along with details about the AACR. Section 4 presents the results obtained in the preliminary tuning of the method and the phoneme classification, compared with a classic parameterization in ASR. Finally, Section 5 summarizes the contributions of this paper and outlines future research.

2. SPARSE REPRESENTATION

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 using these criteria that resembles phisiological properties [16].

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 signifies that the majority of the elements are zero, or 'almost' zero, most of the time [8].

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 [15].

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 $\vec{x} \in \mathbb{R}^N$ 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:

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

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 $(\vec{\phi}_\gamma)_{\gamma \in \Gamma}$, where each waveform $\vec{\phi}_\gamma$ is an *atom* of the representation.

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 samples of $\vec{x}$ 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 have a non-linear relation with the data $\vec{x}$ [2]. 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 $\Phi^*(i,j) = \overline{\Phi(j,i)}$).

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 the sense that a_i be i.i.d. variables). 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 mixture matrix $\vec{\Phi}$, and including an additive noise term $\vec{\varepsilon}$ (Gaussian, in most cases).

The state vector $\vec{a}$ can be estimated from the *posterior* distribution [13]:

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

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

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

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

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 [12] proposed the use of a Laplacian *a priori* distribution with parameter β_i:

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

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

$$\Delta\vec{a} = \vec{\Phi}^T \vec{\Lambda}_{\vec{\varepsilon}} \vec{\varepsilon} - \vec{\beta}^T \left|\vec{a}\right| \quad , \qquad (6)$$

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

To estimate the value of $\vec{\Phi}$, the following objective function can be maximized [12]:

$$\hat{\vec{\Phi}} = \arg\max_{\vec{\Phi}} \left[\mathcal{L}(\vec{x},\vec{\Phi})\right] \quad , \qquad (7)$$

where $\mathcal{L} = \mathcal{E}\left[\log P(\vec{x}|\vec{\Phi})\right]_{P(\vec{x})}$ 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:

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

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 $\vec{\Phi}$:

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

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

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

2. 3. Matching Pursuit

Mallat and Zhang [14] proposed the Matching Pursuit (MP) algorithm. MP is a general method to approximate the solution of the signal representation problem, once the dictionary was provided or estimated. Sparsity is directly included by choosing an appropriate number of terms. Given an initial approximation $\vec{x}^{(0)} = \vec{0}$ and an initial residual $\vec{R}^{(0)} = \vec{x}$, 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 residual $\vec{R}^{(k)}$, and a scalar multiple of this atom is added to the approximation at step $k-1$, obtaining:

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

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; thereof it has the same advantages and disadvantages of this type of optimization methods[2].

2. 4. Approximated auditory cortical representations

The properties of sensorial systems should coincide with the statistics of their perceived stimuli [1]. 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 [3]. 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. Hence, following this biological simile, the representation becomes a good starting point to attain more complex ones.

The obtainment of a dictionary of two-dimensional atoms $\vec{\Phi}$ using (7), corresponding to time-frequency features estimated from the auditory spectrogram of $\vec{x}$, is equivalent to the STRF of a group of cortical neurons [9]. Therefore, the activation level of each neuron can be assimilated with the coefficients a_γ in (1).

3. MATERIALS AND METHODS

The feasibility to build a robust ASR system based on the described scheme was studied for an initial task of phoneme classification. The classifiers were trained with the 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 highly confusing phonemes in English: */b/*, */d/*, */jh/*, */eh/*, */ih/*.

For the estimation of the dictionaries, an auditory spectrogram from the original clean signals sampled at 16 KHz was obtained by means of an early auditory model [21]. In order to work with a simpler version of

[2] Greedily minimizes $\left\|\vec{x} - \vec{\Phi}\vec{a}\right\|_2$

the data, the frequency resolution was reduced. Thus, auditory spectrograms 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.

In a previous work [17] we trained different dictionaries of two-dimensional atoms from the spectro-temporal patterns using (9), with exhaustive tests for both the complete and overcomplete cases. The best performance was obtained with a dictionary size of 256 atoms (complete case), which is the configuration used in this work.

In order to obtain the patterns that would feed the classifiers, a speech utterance is processed by the auditory model and its spectrogram is 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, obtaining patterns of 256 coefficients (recall that only a subset of them is different from zero).

The noisy version of the corpus was obtained by corrupting the clean data with white noise taken from the NOISEX-92 database [20]. Noise was first conveniently re-sampled at 16 kHz with a resolution of 16 bits and then mixed additively at different signal-to-noise levels.

The feasibility to build a robust ASR system based on this scheme was studied by comparing the performance in classification against a standard parameterization used in speech recognition, the *mel frequency cepstral coefficients* (MFCC) [4]. The MFCC was calculated with 12 coefficients plus the energy (MFCC+E), as usual in speech recognition. The first derivative of each coefficient was added in the common way for two consecutive frames, resulting patterns in $\mathbb{R}^{26}$.

The classification experiments were carried out by means of an artificial neural network, namely a *multi-layer perceptron* (MLP). The training of the networks was carried out with the standard backpropagation algorithm with momentum term [6]. 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 variable number of units and one output layer of 5 units.

4. EXPERIMENTS AND RESULTS

The clean speech data was extracted from TIMIT corpus, which contains a total of 6,300 sentences recorded from 630 speakers (10 sentences each) [5]. In this work, the train (38 speakers) and test (11 speakers) data corresponding to region DR1 was used.

Figure 1 shows some of the STRF corresponding to the complete estimated dictionary $\vec{\Phi} \in \mathbb{R}^{256x256}$ using patterns of 64x4 with 4 kHz in height and 32 ms in width. The obtained STRFs present some characteristics of typical behaviors. It can be observed that they act like detectors of diverse significant phonetic clues in the spectrogram: 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 similarities with the STRF patterns found in mammals are also exposed by comparing a pair of them with the estimated patterns, as can be seen to the left of Fig. 1 [10].

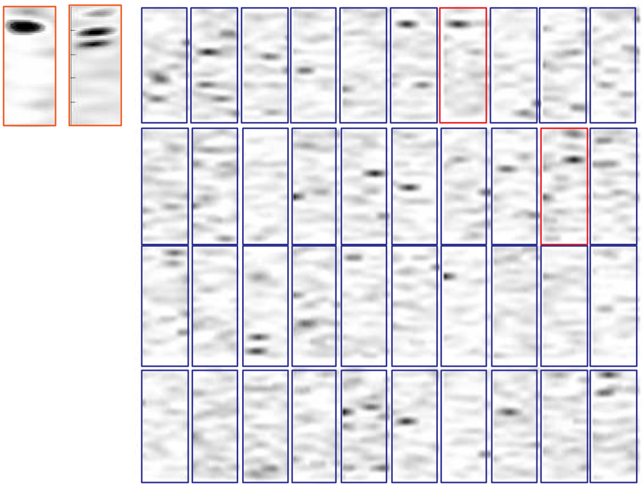

Figure 1: Example of spectro-temporal receptive fields (STRF) calculated from the early auditory representation of phoneme utterances. Each STRF has 64 coefficients with 4 kHz in height and 4 coefficients with 32 ms in width. Two examples of biological STRF as found in animals are shown to the left and compared (in red) with similar patterns as estimated in the discrete dictionary.

The first series of experiments was devoted to find the optimum number of coefficients in the Matching Pursuit feature extraction scheme. Here, the exploration was carried out with 64, 128 and 192 selected coefficients, corresponding to a quarter, a half and three-fourths of the vector dimension (256 coefficients). Also, the best network architecture was found by varying the number of hidden units with a powers-of-two distribution, from 4 up to 1024 units. A subset of the train partition of TIMIT region DR1 was used, where a reduced training set was built from 5 sentences of each speaker and a corresponding validation set was built from other 2 sentences. Each experiment consisted on 3 runs with different initial weights of the networks at random, reporting the mean value obtained on the validation set.

The results of the initial tuning are presented on Fig. 2. It can be observed that the best performance, a recognition rate of 79.73%, is achieved by holding 128 selected coefficients in the MP scheme. In these conditions, from the total of atoms in the dictionary, half of them would be encoding the important cues in the auditory spectrogram. Also, this representation is better processed by the neural network when it has the same dimension in the hidden layer, in this case 256

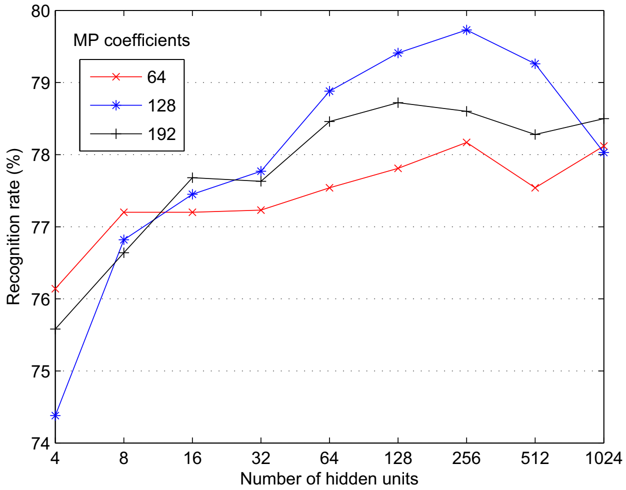

Figure 2: Initial tuning of number of selected coefficients in the Matching Pursuit and number of hidden units in the MLPs, on a reduced training and test data set. The best performance is obtained for 128 coefficients with a MLP of size 256/256/5 (input/hidden/output units).

nodes, than atoms in the complete dictionary. Therefore, this configuration of the MP algorithm and MLP architecture is set for the following experiments.

With the aim to evaluate the performance of the AACR in the presence of noise and to compare its robustness with a standard parameterization, the next series of experiments consisted on the training and evaluation of MLPs using the complete TIMIT region DR1 (with the train and test partitions, respectively). The MFCC feature extraction was fixed to 12 coefficients with frame energy and delta coefficients added, in a vector of dimension 26. The architectures of the MLPs were: 256/256/5 for the cortical patterns and 26/26/5 in the MFCC case.

In each experiment, a series of 10 runs with different initial network weights at random was done and the mean test value was reported. The obtained results are shown in Fig. 3.

As can be seen, the correct classification rates on the test data are better for the AACR than those obtained with the MFCC parameterization. The performance obtained at several SNR reveals an increasing difference between both representations, from 0.71% with clean speech up to 15.64% at 0 dB of SNR. This behavior is given by the intrinsic robustness of the cortical approach, where the more important activations are retained with the Matching Pursuit algorithm.

5. CONCLUSIONS

In this paper, an approach to a robust speech recognition task by means of a biologically-inspired feature extraction method was presented. This technique calculates an optimal dictionary of atoms from the auditory spectrograms. The feature vectors consist of the activation coefficients obtained with the Matching Pursuit algorithm, which selects the more representative ones. In this way, a sort of thresholding of noisy components is applied. The proposed method was applied to the classification of highly confusing phonemes in English.

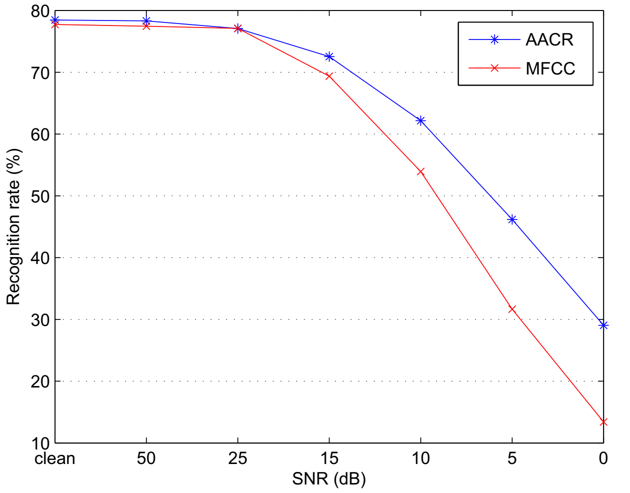

Figure 3: Recognition rate in percent for the set of 5 phonemes in the presence of different signal-to-noise ratios (SNR), from clean speech up to equal energy levels of noise and speech.

The recognition experiments were carried out using multilayer perceptrons. The obtained results showed that the AACR improves the recognition rate over a standard MFCC parameterization, for both the clean case and in the presence of additive white noise.

The feasibility to build a robust speech recognizer based on the cortical representation was explored here using a simple thresholding technique on the parameterization. Future direction in this investigation could be devoted to optimize the denoising of the speech activation patterns and to extend these ideas to the continuous speech recognition task.

6. 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.

References

[1] H. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, (12):241–253, 2001.

[2] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[3] B. Delgutte. Physiological models for basic auditory percepts. In H.H. Hawkins, T.A. McMullen, A.N Popper, and R.R. Fay, editors, *Auditory Computation*. Springer, New York, 1996.

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

[5] Garofolo, Lamel, Fisher, Fiscus, Pallett, and Dahlgren. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation. Technical report, National Institute of Standards and Technology, February 1993.

[6] S. Haykin. *Neural Networks: A Comprenhensive Foundation*. Pearson Education, 1999.

[7] X. Huang, A. Acero, and H-W. Hon. *Spoken Language Processing: a guide to theory, algorithm and system development*. Prentice Hall, 2001.

[8] A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation. Technical report, Helsinki University of Technology, 1998.

[9] D.J. Klein, P. Konig, and K.P. Kording. Sparse Spectrotemporal Coding of Sounds. *EURASIP Journal on Applied Signal Processing*, 2003(7):659–667, 2003.

[10] Konrad P. Kording, Peter Konig, and David J Klein. Learning of sparse auditory receptive fields. In *Proc. of the International Joint Conference on Neural Networks (IJCNN '02)*, volume 2, pages 1103–1108, Honolulu, HI, United States, May 2002.

[11] Te-Won Lee, Gil-Jin Jang, and Oh-Wook Kwon. Sparse representation in speech signal processing. volume 5207, pages 311–320. SPIE, 2003.

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

[13] M.S. Lewicki and T.J. Sejnowski. Learing overcomplete representations. In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 556–562. MIT Press, 1998.

[14] S.G. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. in Signal Proc.*, 41:3397–3415, December 1993.

[15] E. Oja and A. Hyvarinen. Independent Component Analysis: A Tutorial. *Helsinki University of Technology, Helsinki*, 2004.

[16] B.A. Olshausen and D.J. Field. Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[17] H. Rufiner, C. Martínez, D. Milone, and J. Goddard. Auditory Cortical Representations of Speech Signals for Phoneme Classification. In *MICAI 2007: Advances in Artificial Intelligence*, volume 4827 of *Lecture Notes in Computer Science*, pages 1004–1014. Springer-Verlag, 2007.

[18] W. J. Smit and E. Barnard. Continuous speech recognition with sparse coding. *Computer Speech and Language*, 23:200–219, 2009.

[19] F.E. Theunissen, K. Sen, and A.J. Doupe. Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neuroscience*, 20:2315–2331, 2000.

[20] A. Varga and H. Steeneken. 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, 1993.

[21] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Trans. Inf. Theory*, 38:824–839, 1992. Special Issue on Wavelet Transforms and Multiresolution Signal Analysis.