NONLINEAR SPEECH PROCESSING WITH OSCILLATORY NEURAL NETWORKS FOR SPEAKER SEGREGATION

Jean Rouat and Ramin Pichevar
*ERMETIS, DSA, Université du Québec à Chicoutimi, CHICOUTIMI, Québec, CANADA G7H 2B1
GEGI, Université de Sherbrooke, SHERBROOKE, Québec, CANADA J1K 2R1
http://wwwdsa.uqac.ca/ermetis

1 ABSTRACT

Nonlinear masking of space-time representations of speech is a universal technique for speech processing. In the present work we use an AM representation of cochlear filterbank signals in combination with a mask that is derived from a network of oscillatory neurons. The proposed approach does not need any training or learning and the mask takes into account the dependence between points from the auditory derived representation. A potential application is illustrated in the context of speaker segregation.

2 INTRODUCTION

2.1 Space–time representation

Speech enhancement, speaker segregation , speech recognition and coding can be viewed as the result of 3 successive processes.

1. Decomposition of the speech signal into an adequate space–time representation (auditory image representations, spectrograms, wavelet decompositions, etc.);
2. Selection of the relevant information from the chosen representation or masking of the irrelevant information;
3. Synthesis of the speech (enhancement, speaker segregation , coding) or extraction of the parameters from the selected relevant areas of the representation (speech recognition).

In the context of speaker separation the objective is to mask the contribution of the undesirable speakers from the representation and to keep the contribution of the target speaker.

2.2 Nonlinear masking of a time–space representation of speech

It is performed by masking the *irrelevant* information of the representation and by keeping the *interesting* one. The notion of *irrelevant* or *interesting* information is relative to the application. For speech enhancement the objective can be to keep the less corrupted channels (*interesting*) and to throw away the others (*irrelevant*); for speech recognition the objective is to extract the features only from the reliable and discriminant regions (*interesting*) of the representation, etc. There are many ways to perform the masking of time–space representations of speech and for each representation there are also many strategies.

2.3 Examples of masking speech representations

The work by Bahoura and Rouat [2][3] is an example of nonlinear masking by thresholding the wavelet packet representations of speech. It is being used for speech enhancement. Another example of masking being used in speech recognition can be found in the work by Cooke *et al.* [5].

In [2], the evaluation of the mask does not take into consideration the correlation between the scales (yielding discontinuities), while in [5] a preliminary evaluation of the probabilities (training) is required. The idealistic system should instantaneously find the mask (without preliminary training) and take into account the interchannel information (Fig. 1).

We explore the feasibility of the creation of nonlinear masks that do not require any training and that have been designed by taking into account the dependence between points of the space time–representation. We present an exploratory work where the space–time representation is based on signal envelopes obtained from a bank of cochlear filters (Fig. 1), while the mask is obtained via a network of oscillatory neurons.

3 ADAPTIVE NONLINEAR FILTERING AND RECOGNITION WITH NEURONS

The information processing performed by a neuron can be viewed as 1) nonlinear filtering of sequences of spikes via the dendritic tree and the soma, 2) classification by thresholding the nonlinearly filtered information.

The topological organisation of the dendritic tree specifies the pseudo–transfer function of the nonlinear filter (delays, absorptions, summations, products of electrical potentials), while the firing of the neuron will occur only when the filtered electrical potential will be sufficiently high. Schematically, we can say that the neuron fires when it recognises a specific sequence of spikes as input to its dendritic tree [8].

This work has been partially supported by NSERG and Fondation de l'Université du Québec à Chicoutimi. Many thanks to D.L. Wang and his group at OSU for receiving R. Pichevar in August 2001.

As the behavior of the dendrite and the synaptic weights are continuously updated, it is said that plasticity and change of synaptic weights allow adaptive filtering and classification.

What kind of features do the individual properties of neurons carry at the level of group of neurons? In our work, we study one aspect of that question by observing the ability of neurons to dynamically cluster depending on the input speech. For the time being, we fix the topology of the dendrite and we focus on continuous adaption of the synapses. We experiment on a network of chaotic neurons to create adaptive nonlinear filters and clusters of neurons.

4 NONLINEAR MASKING AND NETWORKS OF OSCILLATORY NEURONS

An adequate masking should suppress the undesirable speakers and track the target speaker by finding the streams (c.f. Bregman [4] for definition of streams) that characterise the speakers. To do so, the system should cluster points of the representation according to their relationship with the speakers. Streams are obtained by 1) finding the clusters and 2) classifying the clusters according to their relation to a speaker. It is generally not trivial to establish a relation between the clusters and decide if they belong to the same stream of information (i.e.: come from the same speaker) or not. Usually complex strategies are required to group the clusters into streams of patterns that belong to the same speaker (segregation and fusion of streams).

There is also strong evidence that oscillatory neural networks are able to find clusters of homogeneous information by dynamically nonlinearly filtering the information. For example, it has been shown that an unsupervised neuronal system can perform segmentation of images [18].

The clustering can be performed at the level of group of neurons, based on spike synchrony (see for example the work on "oscillatory correlation" by V.d. Malsburg [7]). It is possible to implement such ability by continuously changing the synaptic weights of spiking neurons.

5 SPEECH STRUCTURE AND MASKING

Speech features are also encoded in the time structure of the signal. In fact, speech has a specific (characteristic) structure that is different from that of most noises and perturbations [11].

According to Bregman [4], the phasic analysis performed by the auditory system, in conjunction with the tonal analysis, is adapted to the perception of speech in adverse environment. The spectral integration (or grouping of sounds) was shown to be partially based on common amplitude modulation characteristics. Furthermore, research work on automatic demodulation of speech can be motivated by the fact that the human brain has neural cells specialized in Amplitude Modulation (AM) and Frequency Modulation (FM). Moreover, simple nonlinear operators can enhance the AM or FM information in a signal [10] [14] and can be used to process the output of a cochlear filterbank in order to obtain AM information characteristics of speech signal and segregate it

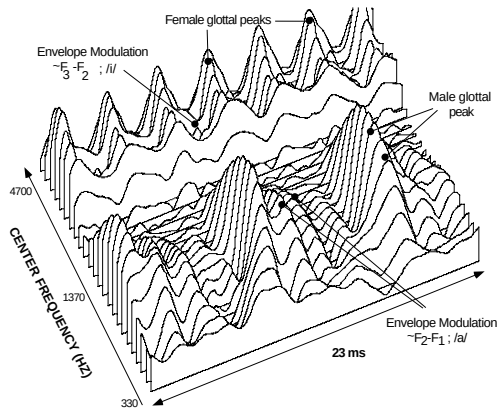


Figure 1: Envelopes $A_i(t)$ for two simultaneous vowels: /i/ by a female, /a/ by a male [11].

from background noise [12]. As we also know the performance of humans in doing ASA (Auditory Scene Analysis) is far better than known sound separation algorithms [4] [1].

6 SEPARATION OF AUDITORY SOURCES USING BIOLOGICAL NEURAL NETWORKS

We applied correlograms of AM envelopes of cochlear filterbank outputs to a network of oscillatory neurons, in order to separate two speakers (or a speaker from a tone). In this approach synchronised regions belong to the same speaker while desynchronised regions with respect to the first speaker's clusters correspond to other speakers (or noise). Our proposed network is composed of chaotic neuronal elements as in [9] but is one dimensional. Our weight adaptation algorithm is a modified version of the rules proposed in [9]. We achieved synchronisation patterns that are different from those in [9]. In fact, we observe periodic and quasi–periodic patterns and believe that it is more biologically plausible. In contrast with other works [15] [17] we do not need any global controller. In fact, our tests on pilot and real data show us that the symmetry breaking is done automatically in this network.

To our knowledge, it is the first time that real speech data are applied to a *one dimensional biologically inspired* neural network. In addition, our network and its associated weight adaptation rule is well suited to multilevel inputs and is not limited to binary data.

6.1 Preprocessing of the speech signal

We tested our network with broadband and narrowband noises. For the broadband noise case, we chose a sound containing two utterances of simultaneous */di:/* and */dae/* . The */di:/* is pronounced by a female speaker and */dae/* by a male speaker. For the narrowband noise case, we used a male speaker sentence contaminated by a tone.

Our preprocessing stage consists of a 24 channel cochlear filterbank that mimics in part the behavior of the human cochlea. The feature extraction algorithm described in [13] has been used and the normalized correlogram is computed for the delays corresponding to the pitch of the target speaker.

In order to find the pitch of the signal we used the pooled correlogram technique [17]. Then the correlograms are quantised to a limited number of levels (4 levels) and applied to our network of chaotic neurons.

6.2 Architecture of the network

The segmentation is based on the "firing" coherence between neurons. The neurons with the same phase belong to the same cluster (note that in the case of chaotic behaviour, output similarity is a measure of synchrony). We used a network of locally connected neurons to segregate sound sources using the "oscillatory correlation" approach [7].

The auditory image is two-dimensional (Fig. 1). One of the dimensions, frequency, is bounded (24 channels in our case), but time is unbounded and can run to infinity. We could segment the auditory image into smaller images that fit in the architecture of a two-dimensional network. By doing so, each point of the auditory image would be applied to the input of a neuron of the network. Therefore, there would be a one-to-one correspondence between frame pixels and neuron inputs [17]. The disadvantage of this technique is that, as soon as a new frame is applied to the network, it forgets previous frames.This is why we decided to use a one dimensional network of neurons. Each neuron corresponds to one channel and the cochlear outputs run freely through the network. We used the chaotic neural network and a relaxation oscillatory network for comparison purposes.

6.2.1 Relaxation Oscillator neurons

We used an array of neurons in a one dimensional network of Locally Excitatory Globally Inhibitory Oscillatory Network (LEGION) [6] [15] . A global controller is used to break the symmetry between different regions. A fixed weighting algorithm as described in [17] is used. The dynamics of each neuron is governed by the modified Van der Pol Equation [16].

6.2.2 Chaotic neurons

An array of chaotic neurons is used to segregate speech. The dynamic of each neuron i is governed by a Chaotic Map [9] :

$$x_i(t+1) = (1-\epsilon)f(x_i(t)) + \frac{\epsilon}{N}\Sigma_{j=1}^{N} f(x_j(t)) \quad (1)$$

$f(x) = ax(1-x)$ is the logistic map, N the number of neurons. We used a modified version of the dynamic neighborhood algorithm described in [9] since we are using a one-dimensional network in contrast to the two dimensional network used in [9] for image segmentation purposes. In addition, our proposed modified weight adaptation rule is able to process non-binary data. The aforementioned proposed algorithm is implemented as follows: Each neuron in the network is connected to other neurons of the network through discrete-time delays (the maximum neighboring distance of connections is set to 10 neurons). In the beginning, each neuron runs freely, that is no synaptic connection is established between neurons. Later, connections are established according to the update formula in Eq. 2. The farther a neuron is from another one, the longer the update delay time is. We chose an exponential delay formula based on the distance between neurons. The delays are constants equal to $d_{i-j} = (2\alpha)^{(i-j)}$, where α is set to 1 in our case and, i and j are neuron indices. The update equations are as follows:

$$w_{ij}(t) = \begin{cases} e^{-5.5*|(x_i(t-d_{i-j}) - x_{i-1}(t-d_{i-j})|} & t - d_{i-j} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

At $t = d_1$, connection strengths between the closest neighbours to all neurons are established. Note that at this instant farther connections are still equal to zero. At $t = d_2$, weights are modified for neurons that are 2 neurons far from neuron i in the network. This continues until connections to all neighbouring neurons are established. In this way, the region of synchrony around a neuron shrinks or grows at fixed time delays according to the defined learning rule.

The mask is generated by using the output of the network. Then, speech is synthesised by weighting the filterbank outputs with that mask. The oscillatory neural network that we use has the advantage of creating a mask that takes into account the mutual information from the cochlear channels and that does not require any training.

6.3 Results

We observed that it is really difficult to achieve synchronization in the multilevel correlogram case for the relaxation oscillator network described earlier as the oscillation frequency is dependent on the magnitude of the input signal. In fact, it is not possible to obtain neuron outputs with equal frequency in the entire network and phase shifts between clusters. As a consequence, "oscillatory correlation" may not be used for sound segregation with our architecture.

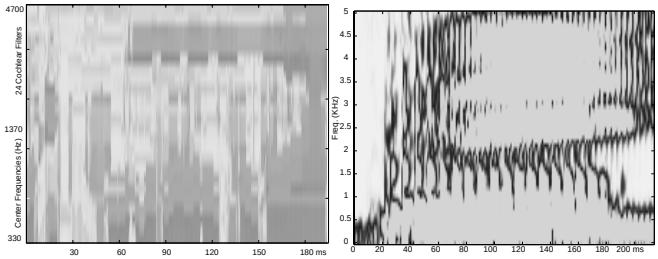
On the other hand, chaotic neurons have much less computational complexity (they require only additions and multiplications), in comparison with the relaxation oscillator network for which third order Van der Pol differential equations [16] must be solved.

Fig. 2 (a) is the output of the chaotic neural network for the two speakers sound segregation case. As we can see, outputs have periodic or quasi-periodic steady state behavior. In fact, steady state chaotic behavior for this one-dimensional network is reached only if the maximum neighborhood distance is chosen very small (4-5).

Subjective hearing tests for the broadband and narrowband noise case showed that in contrast with classical methods (wavelet, inverse filtering, etc.), the output of this approach is smooth and there is no discontinuity or perceptible distortions in the segregated speech. This could be very advantageous when used as a front-end of a speech recogniser.

6.4 Further works

An expansion to more channels and as a consequence to a greater neural network could lead to a finer clustering of regions and to a better audio quality.



(a): Mask (b): Spectrogram

Figure 2: Two simultaneous vowels: /i/ by a female, /a/ by a male. Due to the nonlinear ERB scale in (a), the reader should not try to match the vertical axis with that in (b).

In addition to the analysis already undertaken, more detailed analysis is required to compare relaxation oscillators to chaotic neurons for this application.

More rigorous performance criteria should be found in order to compare results, knowing that the SNR is not a good criteria for sound intelligibility. Although some approaches have good SNR performance, they have quite poor subjective hearing quality.

The update delays in section 6.2 are set empirically on a trial-and-error basis. Further work should be done to find an optimal way to find these parameters. We may include an adaptive learning to adjust updating delays.

7 CONCLUSION

There are limitations to the present approach, one of them, being that the reported experiments have been performed on voiced speech with preliminary estimates of the pitch from the target speaker. Still, the principle of mask generation with such a neural network is interesting and promising. Although the experiments are preliminary, we believe that the approach has a strong potential in relation to nonlinear speech processing.

References

[1] A.J.W. Van der Kouwe, D.L. Wang and G. J. Brown. A comparison of auditory and blind seperation techniques for speech segregation. *IEEE Trans. on Speech and Audio Processing*, 9:189–195, 2001.

[2] M. Bahoura and J. Rouat. A new approach for wavelet speech enhancement. In *proceedings of Eurospeech 2001*, September 2001. Paper nb: 1937.

[3] M. Bahoura and J. Rouat. Wavelet speech enhancement based on the Teager Energy Operator. *IEEE SPL*, 8(1):10–12, Jan 2001.

[4] A. Bregman. *Auditory Scene Analysis*. MIT Press, 1994.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, 2001.

[6] D. Wang and D. Terman. Locally excitatory globally inhibitory oscillator networks. *IEEE Trans. on Neural Networks*, pages 283–286, 1995.

[7] Ch. Von der Malsburg and W. Schneider. A neural cocktail-party processor. *Biol. Cybern.*, pages 29–40, 1986.

[8] S. Deutsch and A. Deutsch. In *Understanding the Nervous System: An Engineering Perspective*. IEEE Press, 1992.

[9] L. Zhao E. Macau. A network of dynamically coupled chaotic map for scene segmentation. *IEEE Trans. on Neural Networks*, pages 1375–1385, 2001.

[10] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. on SP*, 41(10):3024–3051, 10 1993.

[11] J. Rouat. Spatio-temporal pattern recognition with neural networks: Application to speech. In *Artificial Neural Networks-ICANN'97*, Lect. Notes in Comp. Sc. 1327, pages 43–48. Springer, 10 1997. Invited session.

[12] J. Rouat, S. Lemieux, and A. Migneault. A spectro-temporal analysis of speech based on nonlinear operators. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 1629–1632, oct 1992.

[13] J. Rouat, Yong Chun Liu, and D. Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, 21:191–207, 1997.

[14] J. Rouat. Nonlinear operators for speech analysis. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual representations of speech signals*, pages 335–340. J. Wiley and Sons, 1993.

[15] S. N. Wrigley and G. J. Brown . A neural oscillator model of auditory attention. *Lecture Notes in Computer Science*, pages 1163–1170, 2001.

[16] D. Wang. Relaxation oscillators and networks. In *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 396–405. Wiley Sons, 1999.

[17] D. Wang and G. J. Brown. Seperation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10(3):684–697, May 1999.

[18] D.L. Wang and D. Terman. Image segmentation based on oscillatory correlation. *Neural Computation*, 9:805–836, 1997.