A Bio-Inspired Sound Source Separation Technique in Combination with an Enhanced FIR Gammatone Analysis/Synthesis Filterbank

Ramin Pichevar, Jean Rouat
Dept. of Elect. and Computer Eng.
University of Sherbrooke
Sherbrooke, QC, Canada
Ramin.Pichevar@usherbrooke.ca
Jean.Rouat@usherbrooke.ca

Christian Feldbauer, Gernot Kubin
Signal Processing and Speech Comm. Lab.
Graz University of Technology
Graz, Austria
feldbauer@inw.tugraz.at
g.kubin@ieee.org

ABSTRACT

A sound source separation technique based on a two-layered bio-inspired spiking neural network and an enhanced gammatone analysis/synthesis filterbank is proposed. One of the two bio-inspired proposed spectral maps (Cochleotopic / AMtopic or Cochleotopic / Spectrotopic) is used as a front-end to the neural network depending on the nature of the intruding sound. We show that the use of an FIR gammatone filterbank outperforms the previous results obtained by using an IIR gammatone cochlear filterbank, since the FIR implementation has near-perfect reconstruction ability and the cascade of the analysis and synthesis filterbanks is linear-phase.

1 Introduction

The problem of monaural (one-microphone) sound source separation is nowadays a very challenging problem in the speech processing field. Here we propose a bio-inspired solution in the CASA (Computational Auditory Scene Analysis) framework with no prior statistical knowledge of the underlying sources. The processing steps are as follows: analysis filterbank, CAM / CSM generation, auditory stream segregation and integration [1] by the proposed neural network, generation of the mask, and synthesis by the proposed synthesis filterbank. Fig. 1 depicts the block diagram of the analysis-separation-synthesis technique used in this article. In the following sections, we justify each of our choices.

1.1 Motivations for an enhanced gammatone analysis/synthesis filterbank

In a previous work [2] we explored the potential of unsupervised monophonic source separation with no prior knowledge of the underlying sound signals using a bio–inspired solution, in which pseudo auditory images were obtained from two different representations (Cochleotopic/AMtopic Map and Cochleotopic/Spectrotopic Map). Basically, these maps were generated by performing a spectral analysis (the magnitude of the reassigned FFT [3]) on the output of the gammatone filterbank. The approach presented in [2] and the enhanced version presented here are based on finding a mask for the intruding signal using the CAM/CSM maps and then extracting and synthesising the sources by using the output of the cochlear filterbank. In [2] and other works [4, 5] IIR gammatone filterbanks are used as analysis/synthesis filters. We know that the IIR gammatone filterbank introduces nonlinear phase delays and since we use spectral information (no phase information for now) with CAM/CSM maps, to do the segregation, the approach proposed in [2] may introduce phase distortion.

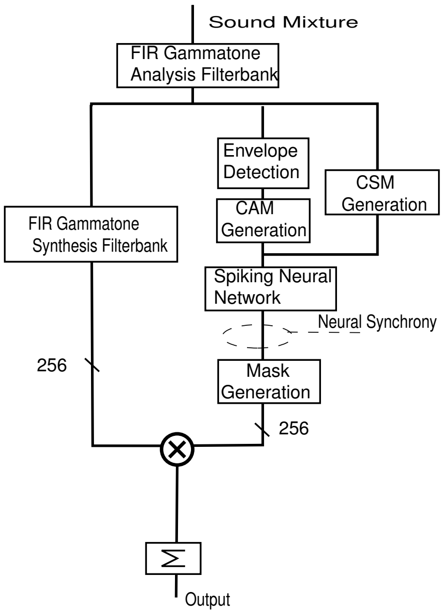

Figure 1: The block diagram for the proposed bio-inspired sound source separation technique

Consequently, a decrease in the quality of the synthesised sounds will be perceived. In addition, the IIR gammatone filterbank is not perfect reconstruction. Therefore, even if someone can find an ideal channel selection strategy for sound segregation, he/she can never achieve the ideal performance due to the distortions caused by the IIR gammatone filterbank during synthesis. On the other hand, as explained in section 1.2, the CAM/CSM generation is based on the magnitude of the FFT and the phase information is ignored. We here propose to use an FIR gammatone analysis/synthesis cochlear filterbank [6] with near perfect reconstruction filters and linear phase.

1.2 Motivations behind the CAM/CSM

We use 2-D bio-inspired maps that we call Cochleotopic/AMtopic (CAM) and Cochleotopic/Spectrotopic (CSM) maps as a front end to our neural architecture. The difference between these two maps is that an AM demodulation is done for the CAM but not for the CSM.

These 2-D representations are used to partially mimic the peripheral auditory system, that adaptively extracts representations suitable for higher auditory nucleus processing. Furthermore, tonotopic maps are observed in the colliculus and specialised cells in the cochlear nucleus. We therefore infer that multiple representations of the same signal are available to the auditory centers and we propose to build a source separation system that can simultaneously use two of these representations.

The selection of either CAM or CSM as a front-end, depending on the nature of the intruding sound, can be explained by recent neurophysiological observations [2].

1.3 Motivations for a spiking neural architecture

Bio-inspired neurons mimic the functional behavior of real biological neurons. In fact, the information in these bio-inspired networks can be coded in the spike phase, in the spike discharge rate, and into the relation between the discharge patterns of the neurons in the network.

Most monophonic source separation systems are based on either expert systems (explicit knowledge), or on statistical approaches (implicit knowledge), or on bio-inspired approaches [7] [4]. Wang and Brown [4] have proposed an original approach that uses features obtained from correlograms (extracted from the outputs of a gammatone IIR filterbank), estimates F0 (the pitch), and uses an oscillatory neural network. Our system neither needs a priori knowledge of the underlying sources, nor does it estimate F0 or compute the computationally expensive correlograms. Our method can then be classified as a non-parametric noise suppression technique (in contrast with techniques that compute speech parameters such as the fundamental frequency). The neural architecture is also designed to handle continuous input signals (even if for now, the CAM and CSM are frame based) and is based on the availability of simultaneous auditory representations of signals. Our proposed architecture has less neurons and less connections for the same task than the architecture proposed in [4], therefore it is computationally less expensive. In addition, by using the analysis/synthesis FIR cochlear filterbank, instead of IIR filters, the segregation and synthesis quality are enhanced.

2 Analysis/Synthesis Filterbank

The proposed method allows to re-synthesise the audio signal of a single sound source from a mixture of sources. Generally speaking, this is achieved using a time-varying filter. The pathway of the audio signal consists of a non-decimated, static analysis filterbank, the time-varying mask, and a static synthesis filterbank.

We use an FIR implementation of the well-known gammatone filterbank [8] as the analysis filterbank. The number of channels is 256 with center frequencies from 100 Hz to 3600 Hz uniformly spaced on an ERB scale. The sampling rate is 8 kHz.

The actual time-varying filtering is done by the mask. Once this mask is obtained by grouping synchronous oscillators of the neural net (see section 4), the output of the synthesis filterbank is multiplied with it. Thus, auditory channels belonging to interfering sound sources are muted and channels belonging to the sound source of interest remain unaffected.

Before the signals of the masked auditory channels are added to form the synthesised signal, they are passed through the synthesis filters, which impulse responses are time-reversed versions of the impulse responses of the corresponding analysis filters. That means that the magnitude of the frequency response of a synthesis filter is the same as of the analysis filter in the same channel. The convolution with the time-reversed impulse responses linearises the phase responses and, if the impulse responses of all filters have same lengths[1] and, therefore, same total group delay in all channels, summation yields a phase-distortion-free result. For a low number of channels, the only distortion of the pair of analysis and synthesis filterbanks would be a minor magnitude ripple in the overall frequency response. But for the high number of channels used in our system, this is absolutely negligible.

This non-decimated FIR analysis/synthesis filterbank was proposed by Irino and Unoki [6] and also used in the perceptual speech coder in [9] (in the latter with 20 channels only).

In an earlier version of our work [2], we used the IIR gammatone filterbank proposed in [10]. We observed phase distortions and an overall reduced signal reconstruction quality. In addition, as stated earlier, since the CAM/CSM takes into account only magnitude information, it cannot guarantee a good separation when nonlinear phase IIR filterbanks are used. The new approach used in the present paper allows us to overcome this problem with a significant increase of reconstruction quality.

3 CAM/CSM Preprocessing

The 2-D bio-inspired maps are generated as follows:

[1] Shorter gammatones of higher-frequency channels need zero padding.

- The sampling rate is 8 kHz, a Butterworth filter of order 10 and cutoff frequency of 3.5kHz is used as an anti-alias filter. The sound is then processed by the analysis filterbank (a 256 channel FIR gammatone filterbank) which frequency range is 100–3600 Hz.

- **For the CSM:** The channels are filtered through the analysis filterbank and the CSM is generated by computing the magnitude of the enhanced FFT [3], so that a 2-D map is generated: one of the dimensions is the cochlear channel number and the other is the frequency bin of the enhanced FFT.

- **For the CAM:** One aspect of the nonlinearities from the hair cells is partially modelised by computing first the Hilbert transform and the envelope of the cochlear channel outputs. Then, the CAM is generated by computing the enhanced FFT as for the CSM.

For the time being, the CAM/CSM selection is done manually depending on the nature of the intruding sound.

4 The neural network

The dynamics of the neurons we use are governed by a modified version of the Van der Pol relaxation oscillator (Wang-Terman oscillators [4]). The state-space equations for this dynamics can be found in [2].

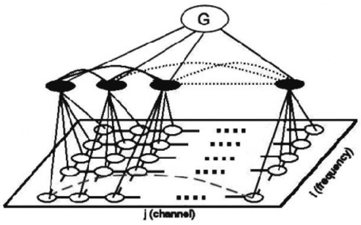

Figure 2: The Two-Layer Neural Network. G: Stands for global controller (the global controller for the first layer is not shown on the figure). One long range connection is shown in the figure.

The first layer is a partially connected network of relaxation oscillators [4]. Each neuron is connected to its four neighbors. The CAM (or the CSM) is applied to the input of the neurons. The first layer is two-dimensional. Our observations have shown that the geometric interpretation of pitch (ray distance criterion) is less clear for the first 24 channels. For this reason, we have also established long-range connections from "clear" (high frequency) zones to "confusion" (low frequency) zones. These connections exist only across the "cochlear channel number" axis of the CAM. This architecture can help the network to better extract harmonic patterns.

A weight normalisation technique described in [2] is used for adapting $w_{i,j,k,m}$ between $neuron_{i,j}$ and $neuron_{k,m}$. The weight adaptation is memoryless and depends only on the actual value of the external inputs to neurons [2]. This same weight adaptation is used for "long range clear to confusion zone" connections in CAM processing case.

The second layer is an array of 256 neurons (one for each channel). Each neuron receives the weighted sum of the outputs of the first layer neurons along the frequency axis of the CAM/CSM. In a modified version of this architecture, we have already shown that multiplicative synapses can further enhance the segregation performance of the network [11].

- For the CAM: Since the geometric (Euclidian) distance between rays (spectral maxima) is a function of the pitch of the dominant source in a given channel, the weighted sum of the outputs of the first layers along the frequency axis tells us about the origin of the signal present in that channel.

- For the CSM: Highly localised energy bursts will be enhanced by that representation.

The selection strategy at the output of the second layer is based on temporal correlation [2]: Neurons belonging to the same source synchronise with the same spiking phase and neurons belonging to other sources desynchronise with a different spiking phase.

5 Results

The utterance "I willingly marry Marilyn" is mixed with a siren noise (taken from Cooke's database [12]). The processing steps are as described in the previous sections. Since the siren is a narrowband noise and that we are looking for energy bursts, CSM is used as front-end. In the case of double-vowel segregation CAM is used (for sound samples see [13]). In fact, for each experiment both the CAM and the CSM are generated and applied to the neural network. The selection between CSM and CAM could be based on the SNR increase between the original signal and the extracted signal. For now, this is manually made. The spectrogram of the original mixture, the extracted siren, and the extracted utterance are shown in Fig. 3 and Fig. 4. The sound file and results for this example and other examples can be found at [13]. Note the difference between results obtained by the IIR gammatone filterbank in Fig. 5 and results obtained by the proposed FIR gammatone filterbank in Fig. 4. The incompleteness in sound separation in the IIR case is due to the fact that sound separation is based on spectral magnitude and not on phase, as explained in 1.1.

6 Conclusion and Further Works

A bio-inspired sound source separation based on spiking neural networks and FIR gammatone filterbank has been proposed. The neural network chooses the channel

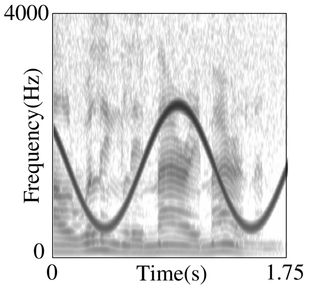

Figure 3: Mixture of a siren and the sentence "I willingly marry Marilyn"

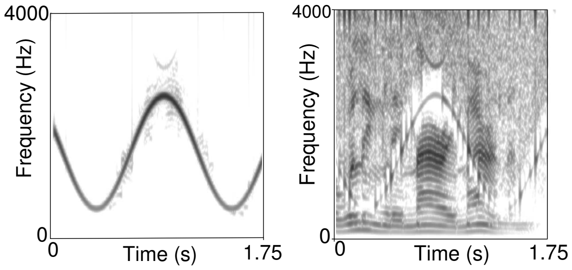

Figure 4: Results with the proposed 256-channel FIR gammatone filterbank . Left: the spectrogram of the extracted siren. Right: the spectrogram of the utterance.

belonging to a source based on the synchronisation between corresponding second-layer neurons giving birth to a cochleo-temporal mask. This mask is then used to synthesise the extracted sound by using a near perfect reconstruction gammatone filterbank. We believe that the qualitative and quantitative results we have obtained from synthesis are very encouraging. More qualitative and quantitative experiments will be performed. Top-down processing can be added to the bottom-up processing proposed in this article by using the Oscillatory Dynamic Link Matching (ODLM) [14]. The computationally very efficient frequency-warped filterbank as described in [15] can also be used as a replacement to the gammatone filterbank. In another experiment, we masked the output of the analysis filterbank before performing the synthesis filtering. The musical noise is greatly reduced, but a pink noise is then observed after synthesis. Non-binary masks with smooth transitions must be used to reduce different types of noise.

Acknowledgments

NSERC, PSIIRI from MDER of gvt du Québec and the University of Sherbrooke financially supported the first two authors. Collaboration between the authors has been initiated and supported within the COST action 277 on Nonlinear Speech Processing.

References

[1] A.S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

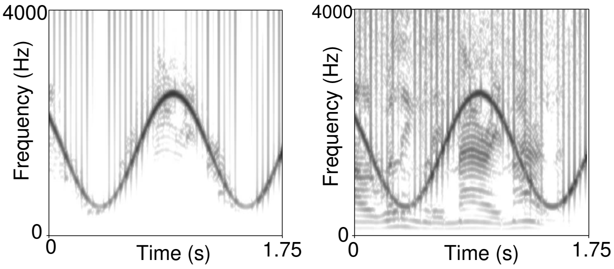

Figure 5: Results with a 256-channel IIR implementation of the gammatone filter. Left: the spectrogram of the extracted siren. Right: the spectrogram of the separated utterance.

[2] R. Pichevar and J. Rouat, "Cochleotopic/AMtopic (CAM) and Cochleotopic/Spectrotopic (CSM) map based sound source separation using relaxation oscillatory neurons," in *IEEE Neural Networks for Signal Processing Workshop, Toulouse, France*, 2003.

[3] F. Plante, G. Meyer, and W. Ainsworth, "Improvement of speech spectrogram accuracy by the method of reassignment," *IEEE Trans. on Speech and Audio Processing*, pp. 282–287, 1998.

[4] D. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[5] G. Brown and M. Cooke, "Computational auditory scene analaysis," *Computer Speech and Language*, pp. 297–336, 1994.

[6] T. Irino and M. Unoki, "A time-varying, analysis/synthesis auditory filterbank using the gammachirp," in *ICASSP 98*, Seattle, Washington, May 1998, vol. 6, pp. 3653–3656.

[7] J. Rouat and R. Pichevar, "Nonlinear speech processing techniques for source segregation," in *EUSIPCO, Toulouse, France*, 2002.

[8] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds., pp. 429–446. Pergamon Press, Oxford, 1992.

[9] Gernot Kubin and W. Bastiaan Kleijn, "On speech coding in a perceptual domain," in *ICASSP 99*, Phoenix, Arizona, Mar. 1999, vol. 1, pp. 205–208.

[10] Malcolm Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Tech. Rep. 35, Apple Computer, Inc, 1993.

[11] R. Pichevar and J. Rouat, "Streaming of audio objects on 2D spectral maps through multiplicative synaptic connection neurons," in *APCAM , Vancouver, Canada*, 2003.

[12] Martin Cooke, ," http://www.dcs.shef.ac.uk/˜martin/.

[13] ," http://www-edu.gel.usherbrooke.ca/pichevar/.

[14] R. Pichevar and J. Rouat, "Oscillatory dynamic link matching for pattern recognition," in *International Workshop on Neural Coding, Aulla, Italy*, 2003.

[15] C. Feldbauer and G. Kubin, "Critically sampled frequency-warped perfect reconstruction filterbank," in *ECCTD, Karkow, Poland*, 2003.