Filterbank optimization for amplitude modulation analysis of audio signals and its relation to auditory filterbanks *

◎ Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono,
Shigeki Sagayama (The University of Tokyo) and Alain de Cheveigné (CNRS, ENS, Paris 5)

1 Introduction

Newborns must learn to structure incoming acoustic information into segments, words, phrases, etc., before they can start to learn language. This process is thought to rely on modulation structure of the speech waveform induced by segmental or prosodic regularities within the speech heard by the infant. Here, we investigate the process by which the initial acoustic processing required by modulation analysis can itself be tuned by exposure to the regularities of speech. Starting from the classic definition of modulation, as applied within channels of the peripheral filter, we formulate a mathematical framework in which the structure of initial spectral filtering is adapted for modulation analysis. Our working hypothesis is that the human ear and brain are adapted to the analysis of modulation, via a data-driven learning process on the scale of development (or possibly evolution). Simulation results are presented and a possible similarity with Gammatone filters is pointed out.

2 Description of the model

2.1 Objective

We are looking for a filterbank which would be adapted as much as possible to extract the modulation present in a signal. Our idea here is that such a filterbank would process the signal in such a way that the sum of the "modulation energy" of its outputs is maximal, where by "modulation energy" we shall denote the energy of the low-passed (under for example say 20Hz) squared signal. In order to avoid trivial solutions, we also need to introduce a constraint on the filterbank.

2.2 Formulation of the objective function

Let us denote by $s(t)$ the input signal, and $F = (f_1, \dots, f_N)$ be a $K \times N$ matrix representing the filter bank to optimize, such that $F_{ij} = f_j(i)$. Each of its columns corresponds to a FIR filter of order K. We suppose that F verifies

$$F^T F = I, \tag{1}$$

that is F lies on the Stiefel manifold $V_N(\mathbb{R}^K)$ of ordered N-tuples of orthonormal vectors of $\mathbb{R}^K$. This means simply that the filters are normalized and mutually orthogonal. This condition is assumed to avoid trivial situations, such as for example all the filters converging to the filter giving the output with highest energy.

We will note $u_j = f_j * s$ and $v_j = u_j^2$. Let $\mathcal{L}_{\omega_c}$ denote a low-pass filter with cut-off frequency ω_c, and

$$w_j = \mathcal{L}_{\omega_c}(v_j) = \mathcal{L}_{\omega_c}\left((f_j * s)^2\right). \tag{2}$$

Our optimization problem can now be stated as the maximization of $\sum_j ||w_j||$ with respect to F under the condition that F lies on the Stiefel manifold. We note $\mathcal{I}(F)$ the objective function to maximize:

$$\mathcal{I}(F) = \sum_j \sqrt{\int \left(\mathcal{L}_{\omega_c}\left((f_j * s)^2\right)\right)^2 (t)\,\mathrm{d}t}. \tag{3}$$

The process leading to the definition of $\mathcal{I}$ is illustrated in Fig. 1. We shall refer to $||w_j||$ as the *modulation energy* of the output of the j-th filter, and to $\mathcal{I}(F)$ as the *total modulation energy*.

2.3 Optimization on Stiefel Manifolds

As it is difficult to obtain an analytical solution here, a gradient method seems to be the only solution, but it suffers from the fact that the updated filterbank is not guaranteed to stay on the Stiefel manifold. An optimization method which would take into account the particular geometrical structure of the constraint space through an update which keeps the filterbank as close to the Stiefel manifold as possible is desirable. The natural gradient method is the natural tool for this kind of task [1], and in the particular case of the Stiefel manifold, the update goes as follows [2]. While the classical gradient method update is

$$F_{(n+1)} = F_{(n)} + G_{(n)}, \tag{4}$$

where

$$G_{(n)} = \mu(n)\frac{\partial \mathcal{I}}{\partial F}(F_{(n)}) \tag{5}$$

is the scaled (Euclidean) gradient of the cost function with respect to F evaluated at $F_{(n)}$, and $\mu(n)$ is a chosen step size sequence, the natural gradient method update can be written

$$F_{(n+1)} = F_{(n)} + G_{(n)}F_{(n)}^T F_{(n)} - F_{(n)}G_{(n)}^T F_{(n)}. \tag{6}$$

Although the natural gradient update can be proven [2] to stay in the constraint space for continuous flows, the discrete-time version presented above is numerically unstable and slowly diverges from the Stiefel manifold, making it impossible to simplify $F_{(n)}^T F_{(n)}$ in (6)). We thus project every few steps on the Stiefel manifold to correct this tendency, through

$$\hat{M} = M(M^T M)^{-\frac{1}{2}}. \tag{7}$$

The derivative of the objective function with respect to F can be obtained as follows:

$$\frac{\partial \mathcal{I}}{\partial F_{i_0 j_0}} = \frac{1}{||w_{j_0}||}\int \left(\mathcal{L}_{\omega_c}(u_{j_0}^2)\right)\left(\mathcal{L}_{\omega_c}(2\mathcal{T}_{i_0}(s)u_{j_0})\right)(t)\,\mathrm{d}t, \tag{8}$$

* 音響信号の振幅変調分析のためのフィルタバンク最適化及び聴覚フィルタバンクとの関係、ルルー・ジョナトン、亀岡弘和、小野順貴、嵯峨山茂樹 (東大情報理工)、ドゥシュベニエ・アラン (CNRS/ENS/Paris 5)

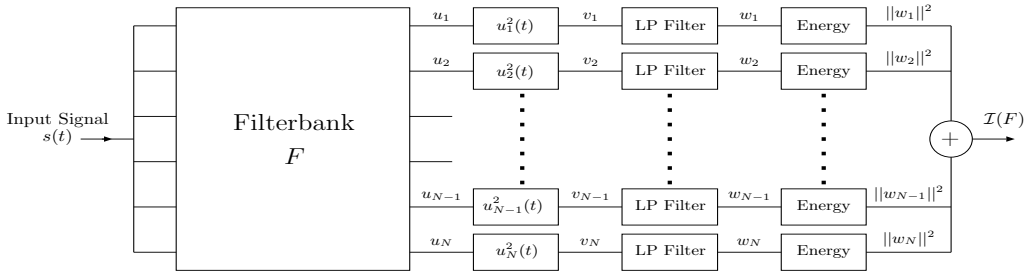

Fig. 1 *Diagram of the model.*

where $\forall t, \mathcal{T}_{i_0}(s)(t) = s(t - i_0)$.

3 Simulations and results

3.1 Experimental Procedure

We performed simulations on speech files with a sampling rate of 16kHz. The low-pass filter's shape was set triangular, and the cut-off frequency was 20Hz. The filterbank was initialised by generating a random matrix with coefficients uniformly distributed on $[-0.5; 0.5)$ and then projecting it back to the Stiefel manifold. The initial value of $\mu(n)$ was set to 0.1, divided by 2 if an update yielded an energy decrease and multiplied by 1.3 after three steps without decrease.

3.2 Results

We show in Fig. 3 the modulation curves for five of the top filters of a filterbank of 30 FIR filters with 250 taps, optimized on a speech sample uttered by a female speaker, normalized such that their L^2 norm is equal to 1. The waveform of the input speech file can be seen in Fig. 2. As a comparison, we show in Fig. 4 the normalized modulation curves computed from five Gammatone filters (with center frequencies of 272Hz, 334Hz, 404Hz, 482Hz, 570Hz) that we selected out of a bank of 30 Gammatone filters ranging from 20Hz to 8000Hz. It is worth mentioning that the modulation curves obtained look surprisingly alike. This fact is particularly interesting, as Gammatone filters are known to be a good model of the cochlear filters.

4 Conclusion

We introduced a framework for data-driven modulation analysis based on the optimization of a filterbank to maximize the modulation energy at its output. We explained how to perform the optimization efficiently, ran simulation experiments to illustrate this procedure, and noticed that one could find some correspondences between the modulation curves obtained through the optimized filters and through Gammatone filters. We shall investigate further in the future the optimization of the filterbanks on large databases of speech sounds, music and environmental sounds, and the degree and reasons of the similarity between optimized filters and Gammatone filters.

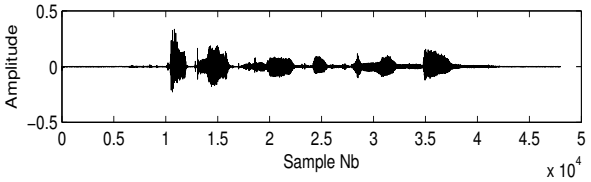

Fig. 2 *Waveform of the input speech file.*

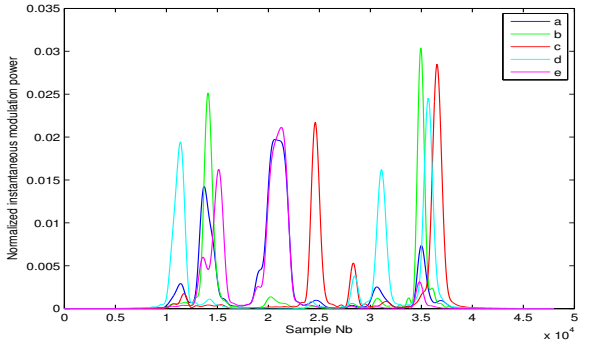

Fig. 3 *Normalized modulation for five optimized filters.*

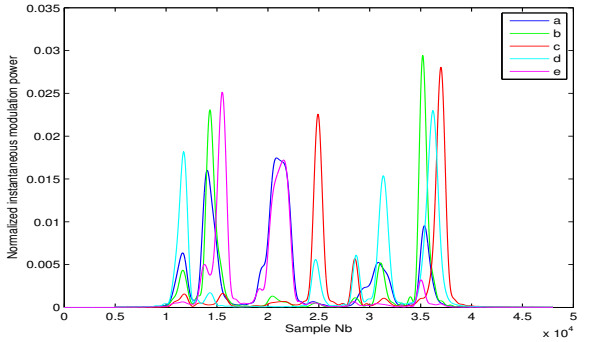

Fig. 4 *Normalized modulation for five Gammatone filters.*

References

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] S. C. Douglas, S.-I. Amari, and S.-Y. Kung. Gradient adaptive paraunitary filter banks for spatio-temporal subspace analysis and multichannel blind deconvolution. *J. VLSI Signal Process. Syst.*, 37(2-3):247–261, 2004.