MODULATION ANALYSIS OF SPEECH THROUGH ORTHOGONAL FIR FILTERBANK OPTIMIZATION

Jonathan Le Roux[1,2], Hirokazu Kameoka[1†], Nobutaka Ono[1], Shigeki Sagayama[1], Alain de Cheveigné[3]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[2]EDITE, Paris, France, [3]CNRS, Université Paris 5, and Ecole Normale Supérieure, Paris, France
{leroux,kameoka,onono,sagayama}@hil.t.u-tokyo.ac.jp, Alain.de.Cheveigne@ens.fr

ABSTRACT

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 comparison with filterbanks classically used in signal processing is done.

Index Terms— modulation spectrum, filter optimization, natural gradient, data-driven analysis.

1. INTRODUCTION

The task of finding words in running speech is difficult because of the lack of obvious acoustic markers at word boundaries, such as onset transients or silent pauses. Adult speakers might conceivably solve the problem by matching templates of words stored in a lexicon to the incoming acoustic stream. Infants do not have this option, as they lack a lexicon. How can one acquire a lexicon without knowledge of how to segment speech into the appropriate chunks to store in this lexicon? Among other hypotheses, it has been suggested [1] that the modulation structure related to prosodic and segmental organization of speech might allow the infant's developing perceptual system to identify initial anchors that facilitate the acquisition of a more complete set of speech part boundaries.

The concepts of *temporal envelope* and *modulation spectrum* are gaining momentum in auditory science (e.g. [2]), speech science (e.g. [3]), automatic speech recognition (e.g. [4]), and evaluation of auditory impairments (e.g. [5]). ”Modulation” can be defined as a part of the temporal structure of the acoustic waveform that is not well captured by standard spectral representations based on the Fourier Transform of the raw waveform. Modulation features extend over wider temporal spans (and thus lower frequencies) than represented in the audio spectrum. They describe the shape of the temporal *amplitude envelope* of the stimulus waveform, rather than the waveform itself. The pitch of a sound and most aspects of its timbre (for example vowel timbre) are usually assumed to reflect the audio spectrum, whereas the perception of roughness, rhythm and long-term temporal structure of speech, are associated with modulation. Processing of temporal envelope structure is assumed to be distinct from that of temporal “fine structure” (e.g. [5]), although there is some overlap in the region of pitch. Both are presumably important for the perception of speech, and a number of studies have attempted to tease apart their respective roles using vocoded speech in which either envelope or fine structure are degraded (e.g. [6, 5]).

Modulation thus seems to play a central role for auditory perception, and if we were to consider the possibility of a tuning of the initial acoustic processing by exposure to the regularities of speech, it would make sense to assume that during the course of development and/or evolution, the human ear and brain adapted for modulation analysis through a data-driven learning process. We shall design a mathematical framework to investigate this hypothesis.

Perception of modulation presumably arises from the analysis of neural activity within each channel from the cochlea. Sensitivity to modulation has been ascribed to the existence of a “modulation filterbank” [2] implemented within the auditory brainstem or midbrain (e.g. [7]) or cortex [8]. We focus our study on the optimization of the combination of peripheral and central filterbanks to best extract the modulation structure of the input data. This is relevant for the hypothesis that such a criterion might in part drive the design of the human auditory system.

Data-driven optimization approaches for the determina-

[†]Now with NTT Communication Science Laboratories

tion of analysis parameters have recently been investigated in speech processing (e.g. [9, 10]), and constitute a trend in Machine Learning to address the arbitrariness of feature choice. The framework we develop might thus also give an insight into new methods for signal analysis.

2. THE TEMPORAL ENVELOPE

Intuitively, the temporal envelope is a smooth function that bounds the oscillations of a signal. We expect the envelope to remain positive and vary slowly, while the carrier or fine structure makes faster positive and negative excursions. It is straightforward to synthesize a waveform based on such a description, but harder to demodulate an existing signal into envelope and fine structure. The task may seem trivial ("draw" a line connecting waveform peaks), but it is hard to perform in full generality. Demodulation usually involves two steps: a non-linearity to produce positive values, and temporal smoothing to give them an "envelope-like" time-course. Popular non-linearities are a full-wave rectifier (absolute value), half-wave rectifier (analogous to cochlear transduction), or square (instantaneous power), possibly followed by a logarithmic transform (dB scale). Smoothing usually involves some form of low-pass filtering. We choose instantaneous power as the non-linearity. For a waveform $s(t)$, we will note

$$v(t) = s(t)^2 \tag{1}$$

and define its "temporal envelope" $w(t)$ as

$$w(t) = \mathcal{L}_{\omega_c}(v)(t) = \mathcal{L}_{\omega_c}\left(s^2\right)(t), \tag{2}$$

where $\mathcal{L}_{\omega_c}$ denotes a low-pass filter with cut-off frequency ω_c. This quantity is not very relevant perceptually if derived directly from the acoustic waveform, as one can argue that the ear has access only to channels filtered by the cochlea. Accordingly, it is common to apply Eq. (2) to the outputs of a filterbank, for example a cochlear model or some other type of spectro-temporal analysis. This produces in effect a *spectro-temporal envelope*, or array of frequency-specific temporal envelopes.

Output $u_j(t)$ of channel j of the initial filterbank is related to the acoustic input $s(t)$ by convolution:

$$u_j(t) = f_j * s(t) = \sum_{k=0}^{K} f_j(k)s(t-k) \tag{3}$$

where $f_j(t)$ is the impulse response of the jth filter (approximated here as a K-tap finite-impulse response filter). The spectro-temporal envelope at time and frequency indices t,j can then be defined as:

$$w_j(t) = \mathcal{L}_{\omega_c}(v_j)(t) = \mathcal{L}_{\omega_c}\left((f_j * s)^2\right)(t). \tag{4}$$

Our goal here is to optimize this initial filterbank under a certain criterion, which we will describe in the next section, such that it is "suited" for modulation analysis.

3. DESCRIPTION OF THE MODEL

3.1. Objective

We are looking for a filterbank which would be adapted to extract the modulation present in a signal. The idea is to maximize the "modulation energy" of the filter outputs, defined as the energy $||w_j||$ of the temporal envelope w_j obtained after rectifying and smoothing (here we low-pass at 20Hz) the output signal u_j. In order to avoid trivial solutions such as several filters converging to the same optimal filter, we also introduce an orthogonality constraint on the filters.

3.2. Formulation of the objective function

Let us denote by $s(t)$ the input signal, and let $F = (f_1, \ldots, f_N)$ 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

$$F^T F = I, \tag{5}$$

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.

Our optimization problem can now be stated as the maximization of the *total modulation energy* $\mathcal{I}(F) = \sum_j ||w_j||$, where w_j is defined as in Eq. (4), with respect to F under the condition that F lies on the Stiefel manifold. The objective function to maximize is thus

$$\mathcal{I}(F) = \sum_j \sqrt{\int \left(\mathcal{L}_{\omega_c}\left((f_j * s)^2\right)\right)^2(t)\,\mathrm{d}t}. \tag{6}$$

3.3. Optimization on Stiefel Manifolds

It is important to find an effective optimization method which is able to take into account the constraint (5). As it is difficult to obtain an analytical solution, a gradient method is indicated but it suffers from the fact that the updated filterbank is not guaranteed to stay on the Stiefel manifold. A first solution to this problem could be to project back to the Stiefel manifold after each update, using the fact that the closest matrix to M on the Stiefel manifold is given by [11]

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

However, it seems slightly risky to leave the Stiefel manifold. An optimization method which would take into account the particular geometrical structure of the constraint space is desirable. The natural gradient method is the natural tool for this kind of task [12], and in the particular case of the Stiefel manifold, the update goes as follows [13]. While the classical gradient method update is

$$F_{(n+1)} = F_{(n)} + G_{(n)}, \tag{8}$$

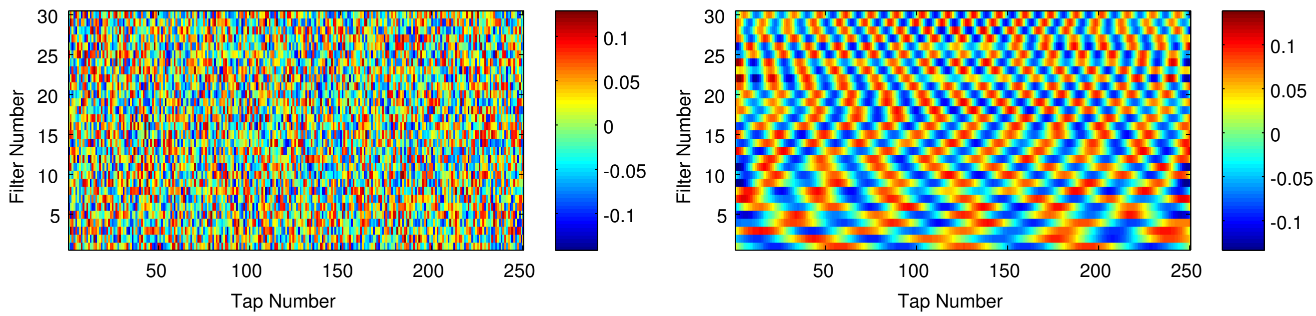

Fig. 1. *Initial orthogonalized random filterbank (left) and optimized filterbank (right).*

where

$$G_{(n)} = \mu(n) \frac{\partial \mathcal{I}}{\partial F}(F_{(n)}) \tag{9}$$

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

$$F_{(n+1)} = F_{(n)} + G_{(n)} F_{(n)}^T F_{(n)} - F_{(n)} G_{(n)}^T F_{(n)}. \tag{10}$$

Although the natural gradient update can be proven [13] 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 (10)). We thus still use the projection (7) every few steps to correct this tendency.

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

$$\frac{\partial \mathcal{I}}{\partial F_{i_0 j_0}} = \frac{1}{||w_{j_0}||} \int \big(\mathcal{L}_{\omega_c}(u_{j_0}^2)\big)\big(\mathcal{L}_{\omega_c}(2\mathcal{T}_{i_0}(s) u_{j_0})\big)(t)\, \mathrm{d}t, \tag{11}$$

where $\mathcal{T}_{i_0}$ is a shift operator, i.e. $\forall t, \mathcal{T}_{i_0}(s)(t) = s(t - i_0)$.

4. SIMULATIONS AND RESULTS

The method is computationally expensive. We present preliminary results on a small sample (12s) of speech uttered by both male and female speakers.

4.1. Experimental Procedure

The sampling rate was 16kHz. The initial low-pass filter applied to the envelope was implemented by convolution with a triangular window, the cutoff frequency being set to 20Hz based on classical speech perception considerations [3]. We chose a low-pass filter with a non-negative impulse response to guarantee that the filtered envelope be non-negative as well. The filterbank consisted of 30 FIR filters with 250 taps, and was initialized by generating a random matrix with coefficients uniformly distributed on $[-0.5; 0.5)$ and then projecting it back to the Stiefel manifold (shown in the left part of Fig. 1). 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.

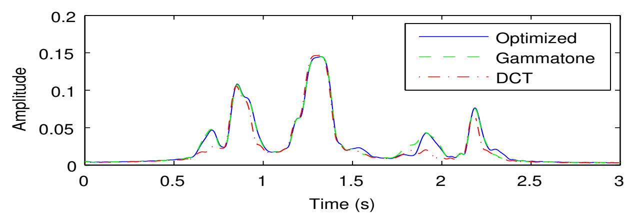

Fig. 2. *Modulation curves for one of the optimized filters.*

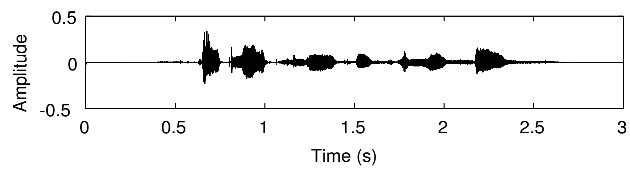

Fig. 3. *Waveform of the input speech file.*

4.2. Results

The optimized filterbank is shown in the right part of Fig. 1. Each horizontal line represents the temporal envelope of one of the filters, which are ordered by decreasing center frequency from top to bottom. We notice that the filters are narrowband, with center frequencies ranging from 230Hz to 1140Hz, and usually go in pairs (in quadrature). Given the small training set, we should not assign too much significance to these values.

As a comparison, we show in Fig. 2 a modulation curve obtained at the output of one of the filters in response to the waveform of Fig. 3 ("I'd like to leave this in your safe" uttered by a female speaker), along with the curves obtained with Gammatone and DCT filters with the same center frequency 287Hz. One can notice that the temporal envelopes extracted are very close. This is confirmed globally by computing the correlations between each modulation curve obtained with an optimized filter and the one obtained with Gammatone and DCT filters, as shown in Fig. 4. Thus our data-driven approach gives results that are quite similar, in particular, to the properties of cochlear filters. The power spectrogram obtained from the optimized filterbank is shown in Fig. 5. We can see by comparing it to a classical FFT-based spectrogram, shown in Fig. 6, that our result is pertinent.

It is too big a step to conclude from this study that the human ear has been developed under such a modulation-based

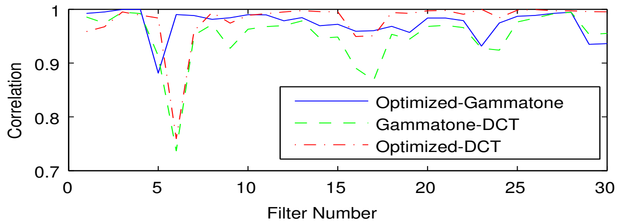

Fig. 4. *Correlations between the modulation curves obtained with three types of filter having the same center frequencies.*

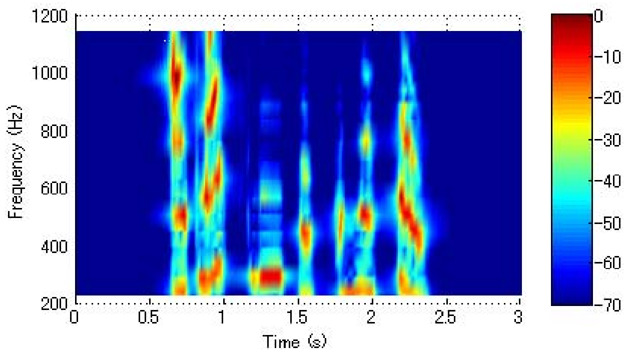

Fig. 5. *Spectrogram obtained using the optimized filterbank.*

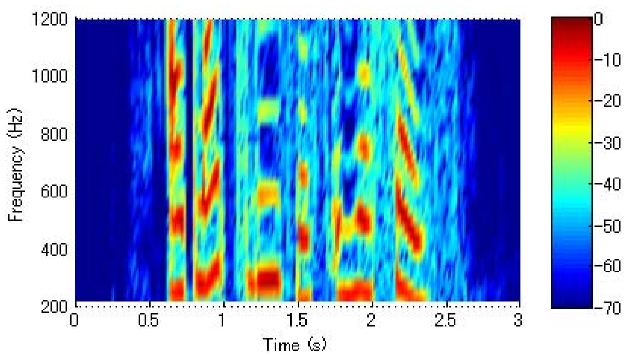

Fig. 6. *FFT spectrogram (window size: 512).*

criterion. However we see it as "proof of concept" that such criteria can be formulated to optimize information processing stages, including initial feature extraction, in a data-driven manner. We have developed a methodology that can be applied to a wider range of optimization criteria and targets to be optimized. An aim for future work will be to derive the number of filter channels and low-pass cut-off frequency from the data, rather than imposing them.

5. CONCLUSION

We introduced a framework for data-driven modulation analysis in which the initial filterbank analysis is optimized based on a criterion of maximum modulation power within the low frequency band, subject to orthogonality constraints between filters. The idea was tested using speech data, and the results obtained were close to classical filterbanks, showing that the hypothesis of a tuning of an auditory system on such a criterion is pertinent. Future work will apply the analysis to large databases of speech and environmental sounds, and perform a comparison of the optimized filterbanks with filtering properties of the auditory periphery and brainstem.

6. REFERENCES

[1] A. Christophe, A. Gout, S. Peperkamp, and J. Morgan, "Discovering words in the continuous speech stream: the role of prosody," *Journal of Phonetics*, vol. 31, pp. 585–598, 2003.

[2] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, vol. 102, pp. 2892–2905, 1997.

[3] S. Greenberg and E.D. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Proc. ICASSP*, 1997, vol. 3, pp. 1647–1650.

[4] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. Interspeech*, 2005.

[5] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B.C.J. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Natl. Acad. Sci.*, vol. 103, no. 49, pp. 18866–18869, 2006.

[6] R.V. Shannon, F.-G. Zeng, and J. Wygonski, "Speech recognition with altered spectral distribution of envelope cues," *J. Acoust. Soc. Am.*, vol. 104, pp. 2467–2476, 1998.

[7] U. Dicke, S.D. Ewert, T. Dau, and B. Kollmeier, "A neural circuit transforming temporal periodicity information into a rate-based representation in the mammalian auditory system," *J. Acoust. Soc. Am.*, vol. 121, pp. 310–326, 2007.

[8] M. Elhilali, J.B. Fritz, J.Z. Simon, and S.A. Shamma, "Dynamics of precise spike timing in primary auditory cortex," *J. Neurosci.*, vol. 24, pp. 1159–1172, 2004.

[9] H. Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech," in *large part published in Proceedings of ASRU-2003*, 2003, number 50, IDIAP-RR 03-50.

[10] P. Paches-Leal, R. C. Rose, and C. Nadeu, "Optimization algorithms for estimating modulation spectrum domain filters," in *Proc. European Conf. Speech Communication and Technology*, Sept. 1999.

[11] D. Luenberger, *Optimization by Vector Space Methods*, Wiley, 1969.

[12] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[13] 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.*, vol. 37, no. 2-3, pp. 247–261, 2004.