Binaural Scene Classification with Time-Frequency Scattering and Deep Convolutional Networks

Vincent Lostanlen
École normale supérieure
45 rue d'Ulm, 75005 Paris, France

***Abstract*—The abstract goes here.**

I. Introduction

Classification of acoustic scenes is only made possible by integrating signal information over a long temporal context. Whereas a few seconds are often sufficient to recognize a speaker, a musical instrument, or a genre, it may require up to 30 seconds to disambiguate closely related acoustic scenes.

II. Time-frequency scattering

Let $\boldsymbol{\psi}[t]$ an analytic band-pass filter of dimensionless frequency ξ and bandwidth ξ/Q. A filter bank of wavelets is built by dilating $\boldsymbol{\psi}$ according to a geometric sequence of scales $2^{-k_1/Q}$, where the log-frequency index k_1 takes integer values in the interval $[\![0;K_1[\![$. We denote by $\boldsymbol{\psi_{k_1}}[t]$ the resulting wavelets. In all subsequent experiments, $\boldsymbol{\psi}$ is designed as a Gammatone wavelet of quality factor $Q=4$, so as to approximate the properties of the human cochlea. The wavelet transform of an audio signal $\boldsymbol{x}[t]$ is obtained by convolution with all wavelets: $\boldsymbol{y_1}[t,k_1] = (\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi_{k_1}})[t]$. Applying pointwise complex modulus to $\boldsymbol{y_1}$ yields the wavelet scalogram $\boldsymbol{x_1}[t,k_1] = |\boldsymbol{y_1}[t,k_1]|$, indexed by time t and log-frequency k_1.

$$\boldsymbol{y_2}[t,k_1,k_2] = \left(\boldsymbol{x_1} \overset{t,k_1}{*} \boldsymbol{\Psi_{k_2}}\right)[t,k_1] \tag{1}$$

where the two-dimensional filter $\boldsymbol{\Psi_{k_2}}[t,k_1]$ is either a temporal low-pass filter $\boldsymbol{\phi}[t]$, a oriented edge detector in the time-frequency plane $(\boldsymbol{\psi_\alpha}[t] \times \boldsymbol{\psi_\beta}[k_1])$, or $(\boldsymbol{\psi_\alpha}[t] \times \boldsymbol{\phi}[k_1])$.

$$\mathbf{W_2}[t,k_1,k_2] = \alpha\psi(\alpha t) \times |\beta|\psi(\beta k_1) \tag{2}$$

III. Deep Convolutional Networks

$$\boldsymbol{x}[t] = r \times \boldsymbol{x}^{\mathsf{L}}[t] + (1-r) \times \boldsymbol{x}^{\mathsf{R}}[t], \tag{3}$$

where r is drawn uniformly at random in the interval $[0,1]$.

IV. Conclusion

The conclusion goes here.

This work is supported by the ERC InvariantClass grant 320959.