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 1 and bandwidth $1/Q$. 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. 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 $\boldsymbol{x_1}[t, k_1] = |\boldsymbol{y_1}[t, k_1]|$, also called constant Q transform (CQT), indexed by time t and log-frequency k_1.

Time-frequency scattering consists in convolving $\boldsymbol{x_2}[t, k_1]$ with a family of two-dimensional wavelets $\boldsymbol{\Psi_{k_2}}[t, k_1]$, where the index k_2 encapsulate two values, i.e. a temporal scale α and a log-frequential scale β. The temporal scale α takes 5 values between $23\,\mathrm{ms}$ and $370\,\mathrm{ms}$ according to a geometric sequence. The log-frequential scale β takes 5 values between $1/4^{\mathrm{th}}$ of an octave and 4 octaves, as well as its 5 "mirror" frequencies. In addition, the edge case $\beta = \infty$ corresponds to a log-frequential moving average along 4 octaves according to all 5 temporal scales α. Finally, the edge case $(\alpha, \beta) = (\infty, 0)$ coresponds to a temporal moving average $\boldsymbol{\phi}[t]$ along $370\,\mathrm{ms}$ without any transformation of the log-frequency axis. In sum, there are $5 \times 5 \times 2 + 5 + 1 = 56$ time-frequency scattering features for every time t and log-frequency k_1.

This work is supported by the ERC InvariantClass grant 320959.

The time-frequency scattering coefficients are thus defined as

$$\begin{aligned}\boldsymbol{y_2}[t, k_1, k_2] &= (\boldsymbol{x_1} \overset{t,k_1}{*} \boldsymbol{\Psi_{k_2}})\,[t, k_1] \\ &= \sum_{\tau, \kappa_1} \boldsymbol{x_1}[t-\tau, k_1-\kappa_1]\boldsymbol{\Psi_{k_2}}[\tau, \kappa_1]. \end{aligned} \tag{1}$$

$$\boldsymbol{x_2}[t, k_1, k_2] = |\boldsymbol{y_2}[t, k_1, k_2]| \overset{t}{*} \boldsymbol{\phi}[t] \tag{2}$$

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.