AUDIO CLASSIFICATION FROM TIME-FREQUENCY TEXTURE

Guoshen Yu

CMAP, Ecole Polytechnique,
91128 Palaiseau Cedex, France

Jean-Jacques Slotine

NSL, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

ABSTRACT

Time-frequency representations of audio signals often resemble texture images. This paper derives a simple audio classification algorithm based on treating sound spectrograms as texture images. The algorithm is inspired by an earlier visual classification scheme particularly efficient at classifying textures. While solely based on time-frequency texture features, the algorithm achieves surprisingly good performance in musical instrument classification experiments.

Index Terms— Audio classification, visual, time-frequency representation, texture.

1. INTRODUCTION

With the increasing use of multimedia data, the need for automatic audio signal classification has become an important issue. Applications such as audio data retrieval and audio file management have grown in importance [2, 18].

Finding appropriate features is at the heart of pattern recognition. For audio classification considerable effort has been dedicated to investigate relevant features of divers types. Temporal features such as temporal centroid, auto-correlation [13, 3], zero-crossing rate characterize the waveforms in the time domain. Spectral features such as spectral centroid, width, skewness, kurtosis, flatness are statistical moments obtained from the spectrum [13, 14]. MFCCs (mel-frequency cepstral coefficients) derived from the cepstrum represent the shape of the spectrum with a few coefficients [15]. Energy descriptors such as total energy, sub-band energy, harmonic energy and noise energy [13, 14] measure various aspects of signal power. Harmonic features including fundamental frequency, noisiness and inharmonicity [5, 13] reveal the harmonic properties of the sounds. Perceptual features such as loudness, shapeness and spread incorporate the human hearing process [22, 12] to describe the sounds. Furthermore, feature combination and selection have been shown useful to improve the classification performance [6].

While most features previously studied have an acoustic motivation, audio signals, in their time-frequency representations, often present interesting patterns in the visual domain. Fig. 2 shows the spectrograms (short-time Fourier representations) of solo phrases of eight musical instruments. Specific patterns can be found repeatedly in the sound spectrogram of a given instrument, reflecting in part the physics of sound generation. By contrast, the spectrograms of different instruments, observed like different textures, can easily be distinguished from one another. One may thus expect to classify audio signals in the visual domain by treating their time-frequency representations as texture images.

In the literature, little attention seems to have been put on audio classification in the visual domain. To our knowledge, the only work of this kind is that of Deshpande and his colleges [4]. To classify music into three categories (rock, classical, jazz) they consider the spectrograms and MFCCs of the sounds as visual patterns. However, the recursive filtering algorithm that they apply seems not to fully capture the texture-like properties of the audio signal time-frequency representation, limiting performance.

In this paper, we investigate an audio classification algorithm purely in the visual domain, with time-frequency representations of audio signals considered as texture images. Inspired by the recent biologically-motivated work on object recognition by Poggio, Serre and their colleagues [16], and more specifically on its variant [21] which has been shown to be particularly efficient for texture classification, we propose a simple feature extraction scheme based on time-frequency block matching (the effectiveness of application of time-frequency blocks in audio processing has been shown in previous work [19, 20]). Despite its simplicity, the proposed algorithm relying only on visual texture features achieves surprisingly good performance in musical instrument classification experiments.

The idea of treating instrument timbres just as one would treat visual textures is consistent with basic results in neuroscience, which emphasize the cortex's anatomical uniformity [11, 8] and its functional plasticity, demonstrated experimentally for the visual and auditory domains in [17]. From that point of view it is not particularly surprising that some common algorithms may be used in both vision and audition, particularly as the cochlea generates a (highly redundant) time-frequency representation of sound.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

2. ALGORITHM DESCRIPTION

The algorithm consists of three steps, as shown in Fig. 1. After transforming the signal in time-frequency representation, feature extraction is performed by matching the time-frequency plane with a number of time-frequency blocks previously learned. The minimum matching energy of the blocks makes a feature vector of the audio signal and is sent to a classifier.

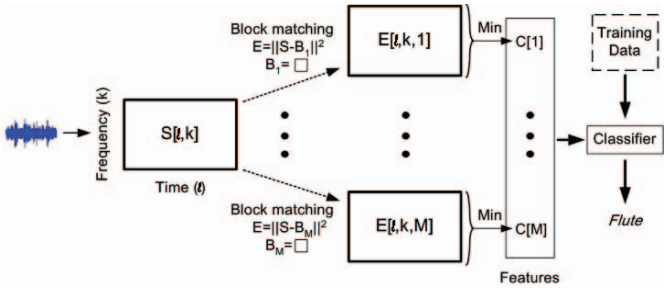

Fig. 1. Algorithm overview. See comments in text.

2.1. Time-Frequency Representation

Let us denote an audio signal $f[n]$, $n = 0, 1, \ldots, N-1$. A time-frequency transform decomposes f over a family of time-frequency atoms $\{g_{l,k}\}_{l,k}$ where l and k are the time and frequency (or scale) localization indices. The resulting coefficients shall be written:

$$F[l,k] = \langle f, g_{l,k} \rangle = \sum_{n=0}^{N-1} f[n]\, g^*_{l,k}[n] \qquad (1)$$

where * denotes the conjugate. Short-time Fourier transform is most commonly used in audio processing and recognition [19, 9]. Short-time Fourier atoms can be written: $g_{l,k}[n] = w[n - lu] \exp\left(\frac{i2\pi kn}{K}\right)$, where $w[n]$ is a Hanning window of support size K, which is shifted with a step $u \leq K$. l and k are respectively the integer time and frequency indices with $0 \leq l < N/u$ and $0 \leq k < K$.

The time-frequency representation provides a good domain for audio *classification* for several reasons. First, of course, as the time-frequency transform is invertible, the time-frequency representation contains complete information of the audio signal. More importantly, the texture-like time-frequency representations usually contain distinctive patterns that capture different characteristics of the audio signals. Let us take the spectrograms of sounds of musical instruments as illustrated in Fig. 2 for example. Trumpet sounds often contain clear onsets and stable harmonics, resulting in clean vertical and horizontal structures in the time-frequency plane. Piano recordings are also rich in clear onsets and stable harmonics, but they contain more chords and the tones tend to transit fluidly, making the vertical and horizontal time-frequency structures denser. Flute pieces are usually soft and smooth. Their time-frequency representations contain hardly any vertical structures, and the horizontal structures include rapid vibrations. Such textural properties can be easily learned without explicit detailed analysis of the corresponding patterns.

As human perception of sound intensity is logarithmic [22], the classification is based on log-spectrogram

$$S[l,k] = \log |F[l,k]|. \qquad (2)$$

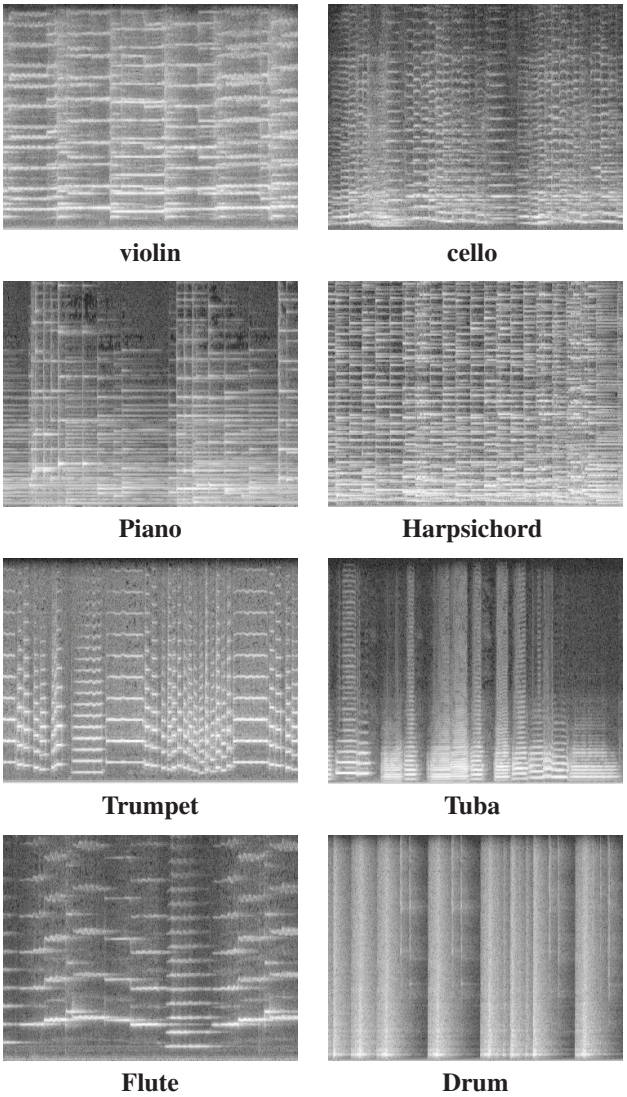

Fig. 2. Log-spectrograms of solo phrases of different musical instruments.

2.2. Feature Extraction

Assume that one has learned M time-frequency blocks B_m of size $W_m \times L_m$, each block containing some time-frequency structures of audio signals of various types. To characterize an

audio signal, the algorithm first matches its log-spectrogram S with the sliding blocks $B_m, \forall m = 1, \ldots, M$,

$$E[l,k,m] = \frac{\sum_{i=1}^{W_m}\sum_{j=1}^{L_m} |\bar{S}[l+i-1, k+j-1] - \bar{B}_m[i,j]|^2}{W_m L_m}, \quad (3)$$

where $\bar{X}$ denotes the normalized block with unity energy $\bar{X} = X/\|X\|$ that induces the loudness invariance. $E[l,k,m]$ measures the degree of resemblance between the patch $\bar{B}_m$ and locally normalized log-spectrogram $\bar{S}$ at position $[l,k]$. A minimum operation is then performed on the map $E[l,k,m]$ to extract the highest degree of resemblance locally between $\bar{S}$ and $\bar{B}_m$:

$$C[m] = \min_{l,k} E[l,k,m]. \quad (4)$$

The coefficients $C[m]$, $m = 1, \ldots, M$, are time-frequency translation invariant. They constitute a feature vector $\{C[m]\}$ of size M of the audio signal. Note that a fast implementation of the block-matching operation (3) can be achieved by using convolution.

The feature coefficient $C[m]$ is expected to be discriminative if the time-frequency block B_m contains some salient time-frequency structures. In this paper, we apply a simple random sampling strategy to learn the blocks as in [16, 21]: each block is extracted at a random position from the log-spectrogram S of a randomly selected training audio sample. Blocks of various sizes are applied to capture time-frequency structures at different orientations and scales [19]. Since audio log-spectrogram representations are rather stationary images and often contain repetitive patterns, the random sampling learning is particularly efficient. Patterns that appear with high probability are likely to be learned.

2.3. Classification

The classification uses the minimum block matching energy C coefficients as features. While various classifiers such as SVMs can be used, a simple and robust nearest neighbor classifier will be applied in the experiments.

3. EXPERIMENTS AND RESULTS

The audio classification scheme is evaluated through musical instrument recognition. Solo phrases of eight instruments from different families, namely flute, trumpet, tuba, violin, cello, harpsichord, piano and drum, were considered. Multiple instruments from the same family, violin and cello for example, were used to avoid over-simplification of the problem.

To prepare the experiments, great effort has been dedicated to collect data from divers sources with enough variation, as few databases are publicly available. Sound samples were mainly excerpted from classical music CD recordings of personal collections. A few were collected from internet. For each instrument at least 822-second sounds were assembled from more than 11 recordings, as summarized in Table 1. All recordings were segmented into non-overlapping excerpts of 5 seconds. 50 excerpts (250 seconds) per instrument are randomly selected to construct respectively the training and test data sets. The training and test data did not contain certainly the same excerpts. In order to avoid bias, *excerpts* from the *same recording* were never included in both the training set and the test set.

	Vio.	Cel.	Pia.	Hps.	Tru.	Tuba	Flu.	Drum
Rec.	27	35	31	68	11	15	12	22
Time	7505	7317	6565	11036	822	1504	2896	2024

Table 1. Sound database. *Rec* and *Time* are the number of recordings and the total time (second). Musical instruments from left to right: violin, cello, piano, harpsichord, trumpet, tuba, flute and drum.

Human sound recognition performance seems not degrade if the signals are sampled at 11000 Hz. Therefore signals were down-sampled to 11025 Hz to limit the computational load. Half overlapping Hanning windows of length 50 ms were applied in the short-time Fourier transform. Time-frequency blocks of seven sizes 16×16, 16×8 and 8×16, 8×8, 8×4 and 4×8 and 4×4 that cover time-frequency areas of size from $640\text{Hz} \times 800\text{ms}$ to $160\text{Hz} \times 200\text{ms}$ were simultaneously used, same number for each, to capture time-frequency structures at different orientations and scales. The classifier was a simple nearest neighbor classification algorithm.

Fig. 3 plots the average accuracy achieved by the algorithm in function of the number of features (which is seven times the number of blocks per block size). The performance rises rapidly to a reasonably good accuracy of 80% when the number of features increases to about 140. The accuracy continues to improve slowly thereafter and becomes stable at about 85%, very satisfactory, after the number of features goes over 350. Although this number of visual features looks much bigger than the number of carefully designed classical acoustic features (about 20) commonly used in literature [7, 6], their computation is uniform and very fast.

The confusion matrix in Table 2 shows the classification details (with 420 features) of each instrument. The highest confusion occurred between the harpsichord and the piano, which can produce very similar sounds. Other pairs of instruments which may produce sounds of similar nature, such as flute and violin, were occasionally confused. Some trumpet excerpts were confused with violin and flute — these excerpts were found to be rather soft and contained mostly harmonics. The most distinct instrument was the drum, with the lowest confusion rate. Overall, the average accuracy was 85.5%.

4. CONCLUSION AND FUTURE WORK

An audio classification algorithm is proposed, with spectrograms of sounds treated as texture images. The algorithm

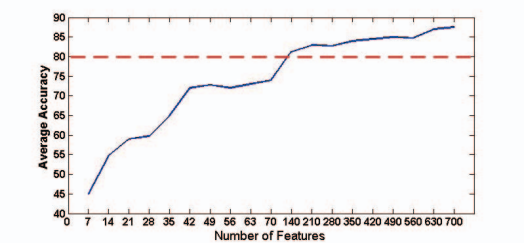

Fig. 3. Average accuracy versus number of features.

	Vio.	Cel.	Pia.	Hps.	Tru.	Tuba	Flu.	Drum
Vio.	94	0	0	0	2	0	4	0
Cel.	0	84	6	10	0	0	0	0
Pia.	0	0	86	8	6	0	0	0
Hps.	0	0	26	74	0	0	0	0
Tru.	8	2	2	0	80	0	8	0
Tuba	2	4	2	0	0	90	0	2
Flu.	6	0	0	0	0	0	94	0
Drum	0	0	0	0	0	2	0	98

Table 2. Confusion matrix. Each entry is the rate at which the row instrument is classified as the column instrument. Musical instruments from top to bottom, left to right: violin, cello, piano, harpsichord, trumpet, tuba, flute and drum.

is inspired by an earlier biologically-motivated visual classification scheme, particularly efficient at classifying textures. In experiments, this simple algorithm relying purely on time-frequency texture features achieves surprisingly good performance at musical instrument classification.

In future work, such image features could be combined with more classical acoustic features. In particular, the still largely unsolved problem of instrument separation in polyphonic music may be simplified using this new tool. In principle, the technique could be similarly applicable to other types of sounds, such as e.g. "natural sounds" in the sense of [10]. It may also be applied to other sensory modalities, e.g. in the context of tactile textures as studied by [1].

Acknowledgements: We are grateful to Emmanuel Bacry, Jean-Baptiste Bellet, Laurent Duvernet, Stéphane Mallat, Sonia Rubinsky and Mingyu Xu for their contribution to the audio data collection.

5. REFERENCES

[1] Adelson, E.H., Personal communication.

[2] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[3] J. Brown, "Musical instrument identification using autocorrelation coefficients", *in Proc. Int. Symp. Musical Acoustics*, 1998, pp. 291-295.

[4] H. Deshpande and R. Singh and U. Nam, "Classification of music signals in the visual domain", *Proc. the COST-G6 Conf. on Digital Audio Effects*, 2001.

[5] B. Doval and X. Rodet, "Fundamental frequency estimation and tracking using maximumlikelihood harmonic matching and HMMs", *Proc. IEEE ICASSP, Minneapolis*, 1993.

[6] S. Essid, G. Richard and B. David, "Musical Instrument Recognition by pairwise classification strategies", *IEEE Transactions on Speech, Audio and Language Processing*, vol.14, no.4, pp 1401-1412, 2006.

[7] E. Zwicker, H. Fastl, "Content-based Audio Classification and Retrieval using SVM Learning", *IEEE Transactions on Neural Networks*, vol.14, no.1, pp 209-215, 2003.

[8] J. Hawkins, S.Blakeslee, *On Intelligence*, Times Books, 2004.

[9] S. Mallat, *"A Wavelet Tour of Signal Processing, 2nd edition"*, New York Academic, 1999.

[10] J.H. McDermott, A.J, Oxenham "Spectral completion of partially masked sounds", *PNAS*, 105 (15), 5939-5944, 2008.

[11] V. Mountcastle, "An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System", The Mindful Brain, MIT Press, 1978.

[12] B.C.J. Moore, B.R. Glasberg and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness", *J. Audio Eng. Soc*, vol.45, no.4, pp.224-240, 1997.

[13] *Information Technology*, *Multimedia Content Description Interface C Part 4: Audio, Int. Standard*, , ISO/IEC FDIS 15938C4:2001(E), Jun. 2001.

[14] P. Viola and M. Jones, "A Large Set of Audio Features for Sound Description(Similarity and Classification)", *in the CUIDADO Project, IRCAM, Paris, France, Tech. Rep.*, 2004.

[15] L. Rabiner and B. Juang, *Fundamentals of Speech Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1993, Prentice-Hall Signal Processing Series.

[16] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms", *IEEE Trans. PAMI*, vol.29, no.3, pp.411-426, 2007.

[17] Von Melchner, L., Pallas, S.L. and Sur, M, "Visual behavior mediated by retinal projections directed to the auditory pathway", *Nature*, 404, 2000.

[18] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio", *IEEE Multimedia Mag.*, vol. 3, pp. 27C36, July 1996.

[19] G. Yu, S. Mallat and E. Bacry, "Audio Denoising by Time-Frequency Block Thresholding", *IEEE Transactions on Signal Processing*, vol.56, no.5, pp.1830-1839, 2008.

[20] G. Yu. E. Bacry, S. Mallat, "Audio Signal Denoinsing with Complex Wavelets and Adaptive Block Attentuation", *Proc. IEEE ICASSP, Hawaii*, 2007.

[21] G. Yu and J.J. Sloine, "Fast Wavelet-based Visual Classification", *Proc. IEEE ICPR, Tampa*, 2008.

[22] E. Zwicker, H. Fastl, *Psychoacoustics: Facts and Models*, Berlin, Springer-Verlag, 1990.