Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

The texture of natural sounds

T. C. Andringa and R. R. Violanda

University of Groningen, Dept. Artificial Intelligence, P.O. Box 407, 9700 AK Groningen, Netherlands
t.andringa@ai.rug.nl

The texture of sound sources is a robust and characteristic perceptual property that listeners use for source recognition. This work introduces a method to determine the presence of sound textures associated with noise, pulses and tonal contributions and applies this to sound source similarity prediction. Local textures are estimated by applying a Kolmogorov-Smirnov test to compare the histogram of locally estimated texture information with the histogram for broadband noise. This method can determine the texture of TF-regions as small as 50 ms by 1 octave. With this measure we analyze cochleograms (i.e. spectrograms based on a cochlea-model) in terms of the presence of three basic textures: broadband noise, tones, and pulses. The relative contributions of the three classes are used as a distance measure between sounds. These distances are compared with the differences between sounds that listeners reported in an experiment by Brian Gygi. Our results lead to an organization of natural sounds similar to the perceptual organization that resulted from Gygi's sound similarity study. We conclude that human sound classification first and foremost is aimed at separating input sounds into broadband noise, pulses and tonal components.

1 Introduction

We investigate the importance of three simple sound textures, corresponding to noisy, tonal and pulse-like contributions, to describe the global perceptual characteristics of sound sources. This paper defines sound textures as spectro-temporal patterns of energy. Typical sound textures are the babble-noise of multiple speakers, the sound of running water, the pulsed sound of a helicopter, or the noisy sound of breaking surfs on a beach. These textures are a useful way to describe sounds because they capture information indicative of certain environmental processes. In fact listeners tend to couple specific textures to specific sources in verbal reports like '*It sounds like running water*'.

Sound texture estimation is also very robust. As our day-to-day experience tells us, the influence of transmission effects and concurrent sounds can become quite severe before source specific textures and the corresponding verbal reports cannot be associated with any measure or reliability. The robustness of human sound texture estimation in combination with the fact that listeners seem to refer to textures in terms of the sources that might produce them, suggest an intimate link between texture estimation and sound source recognition.

In an earlier study [1] we addressed environment classification by estimating *source specific* cues for sounds such as birds, vehicles, speech, and footsteps. These cues were not based on spectral envelope and level, but on texture patterns of tones, pulses and noises that captured source specific patterns. The current article does not rely on *source specific* cues, but on very general cues in the form of the relative importance of tones, pulses, and noises in signals. This choice is based on the results reported by Gygi [2] that show that humans order sounds in the first place in broad classes that Gygi refers to as continuous sounds, harmonic sounds, discrete impact sounds.

Conform Gygi's finding we will develop sound textures detectors for broadband noise, tones, and pulses. This choice is based on the physical aspects of sound production. Tones for example are often produced by sources that exhibit resonances. Pulses are produced by a short and possibly regular/rhythmic impact interaction of hard materials. Noises can be considered as a broadband excitation that can result from uncorrelated mixtures of many tonal and pulse-like contributions. Based on these three basic texture classes, similarities between different sounds are computed and compared with the similarities of the same set of sounds as judged by the human subjects.

We first address a working definition of a sound texture that is both intuitively appealing as well as useful from a perspective of sound source recognition. Using this definition we analyze cochleograms, i.e. spectrograms based on a cochlea-model, in terms of the presence of three basic textures: broadband noise, tones, and pulses. Next, the relative contribution of these textures is computed for the 50 sound pairs in Gygi's study, which are selected to be maximally dissimilar. The relative contributions of the three classes are used as a distance measure between sounds. These distances are compared with the differences between sounds that listeners reported [2, 3]. We will confirm Gygi's finding and conclude that human sound classification is first and foremost aimed at separating input sounds into broadband noise, pulses and tonal components.

2 Defining sound textures

While sound textures are important it is unclear when a sound, or part of it, can be called a sound texture. For example the sound texture that defines a bowling alley is substantially determined by the sounds of bowling balls rolling towards the cones, bowling balls hitting cones, and cones hitting each other and tumbling against the walls and the floor. For a bowling alley with many lanes and a large number of players, the overall sound texture is mainly determined by the statistical properties of these sound-producing events, in combination with the acoustic properties of the hall. In this realistic, but not overly complex, acoustic environment it is not evident what to call a sound texture. Clearly the overall acoustic properties in the hall are to some degree constant over an extended period, but the realized sound texture changes as the pattern of sound producing activities changes.

A way to make the concept of a sound texture more tangible and intuitively acceptable is to couple it, as in the examples above, to (physical) events. A sound texture can then be defined as a spectro-temporal energy pattern associated with some physical process. Since physical processes can be defined on several levels and temporal scales, the associated sound texture must also be defined on different scales. Using this definition the overall sound texture of a bowling alley is defined as a constrained mixture of lower level sound textures associated with bowling balls rolling and cones being hit.

It is important to make a distinction between the texture and the actual resulting spectro-temporal pattern that is an instance of the texture. We treat the sound texture as a topdown, a priori determined, and constant, statistical description in the form of a histogram of values indicative of the process. We will develop a method to estimate for each point in the time-frequency plane if its preceding spectro-temporal environment justifies the point to be classified as either noisy, tonal or pulse-like. We will do this by estimating how likely it is that a certain cochleogram area stems from a statistical process that generates noise.

3 Methods

This section starts with a description of the time-frequency analysis and the computation of a local measure for tonality and pulsality. These values are used in a running histogram where they lead to statistical information about the local sound texture in the form of running histograms. These local running histograms are compared to a reference histogram of noise

A special characteristic of all statistical measures introduced below is the choice to rescale them so that the distribution of broadband noises (such as white or pink noise) has zero mean and unit standard deviation. Since these distributions are approximately normal it is possible to reduce the number of spurious values from the flanks of the probability density function in a predictable way by choosing a higher threshold. Conversely, reducing the threshold will make it possible to select events that resemble noise-like events, but at the cost of an increase in spurious contributions. Because all described measures are frequency channel dependent, rescaling to zero mean and unit standard deviation makes it possible to work with frequency independent thresholds.

3.1 Cochleogram, pulsality, tonality

The first step in the signal processing is the conversion from amplitude in the time domain, to energy in the time-frequency domain. This conversion is performed by filtering the audio signal with a 100-channel gammachirp filter bank. The filter output is squared and leaky-integrated to result in a time–frequency-energy representation, called a cochleogram. The center frequencies of the filter bank channels (also called segments) are chosen conform an ERB scale between 30 and 6000 Hz. Compared to a human cochlea, the time-frequency response is shifted towards frequency specificity and results in an increased group-delay [4]. The cochleogram is sampled every 5 milliseconds. Each second corresponds to 200 frames by 100 segments, which comprises 20,000 TF-points.

To find TF-points dominated by tones or pulses, we convolve the cochleogram with two segment-dependent filters. These filters match either the shape of the tone response or the shape of an impulse response for the associated segment. The segment dependent broadness of the filter-shapes corresponds to a width equivalent to the peak broadness at two standard deviations of white noise below the peak of the response, which, depending on the segment, corresponds to -3 to -6 dB for high frequency and low frequency segments. The convolutions result in two representations: the first indicating tonality, the other indicating the pulsality per TF-point. The tonality and pulsality values for white noise are approximately normal. Both representations are rescaled to zero mean and unit standard deviation for each segment.

Because the tonality and pulsality distribution of noise is almost normal, about 95-98% of the points of the TF-plane of a broadband noise signal complies with the demand that both pulsality and tonality values are in the range [-2, 2] standard deviations. In a similar way tonality corresponds to high tonality values (typically 3 to 5 standard deviations) in combination with pulsality values close to zero (due to limited energy development in the temporal direction). Pulsality corresponds to high pulsality in combination with low tonality.

3.2 Running histogram

While the individual tonality and pulsality values are indicative, they are not conclusive for the local texture since their TF-scope is limited to the broadness of the peak shapes, which correspond to a few frames or segments and is unlikely to include a pattern of multiple peaks. Furthermore, we want a measure indicating whether or not a TF-region of about 100 TF-points represents a pattern that is likely to comply with the expectations for noise. When it does not, we want to know if the local distribution is shifted towards either tonal or pulse-like patterns/textures.

Depending on the texture to be estimated, the scope must be adapted. The tonality is estimated in a range of 11 segments (0.7 octaves) and a temporal scope of 10 periods of the segments best- frequency with a minimum of 10 frames (50 ms). For a 100 Hz segment this entails a scope of 100 ms. Pulsality is estimated from a range of 21 segments (about 1.3 octaves) and 5 local periods with a minimum of 5 frames). In both cases this corresponds to at least 100 TF-points.

The running histogram is computed for both the tonality and the pulsality values by first computing a histogram per segment over the specified range of frames for that segment. Each histogram is defined by 101 bins that span the range of [-5, 5] in steps of 0.1 standard deviations. At each time-step an old value, which is now outside the temporal scope, is removed and a new value is added to the histogram. All histograms are rescaled to have a unit sum and can be treated as probability density functions. At each frame, all segment histograms are updated and the histograms are averaged over the specified segment range so that all histograms at this frame represent information about the specified spectro-temporal scope of at least 100 TF-points.

3.3 Comparing histograms

The histograms for each segment are compared with segment dependent histograms of 1/f-noise using a Kologorov-Smirnov distance [5]. This entails that the maximum distance between the corresponding cumulative probability density functions (or empirical distribution functions) is taken as a correspondence measure between the local histogram and the reference distribution.

This correpondence is computed for all TF-points and yields tonality and pulsality representations that indicate how well the preceding TF-region of each TF-point complies with the expectation based on the reference noise.

For easy thresholding the values are rescaled to ensure that the corresponding reference distribution has zero mean and unit stand deviation for each segment. As an additional statistical descriptor the spread of the distribution is also computed; it is defined as the range between the 5% and the 95% percentile. For the noise reference this corresponds to about 2 standard deviations on each side of the mean.

3.4 Classifying regions

We now have a source of textural information for pulses and tones, based on at least 100 TF-points, to determine whether or not the region preceding each TF-point is noise-like, tonal or pulse-like. The next step is the determination of regions for noise, tonal and pulse-like contributions.

The region for noisy contributions is based on three demands. The Kolmogorov-Smirnov distance (rescaled to unit standard deviation) must be smaller than unity for both tonality and pulsality. The third demand is based on the observation that an increase or decrease in the spread of tonality is balanced by a decrease or increase in the spread of pulsality. This entails that the difference in both measures of spread is smaller than 4 standard deviations for almost all regions dominated by noise. Satisfying all demands leads to inclusion in a noise region.

The regions for tones are defined as the combination of a Kolmogorov-Smirnov distance for tonality larger than unity in combination with a tonality spread larger than 4 standard deviations. Likewise, the region demands for pulsality are identical, but now based on the Kolmogorov-Sminov distance for pulsality.

Almost all TF-points are assigned to one of the three classes. A few points, on the borders of regions, may be assigned to multiple regions.

3.5 Comparing sounds

The results of Gygi [2] indicate that listeners tend to describe sounds in noisy, tonal, and pulse-like contributions. Therefore we described each sound in terms of the fraction of the total log energy represented by each of the contributions. The resulting fractions are depicted in the right panels of figure 1 for two examples of car windscreen wipers.

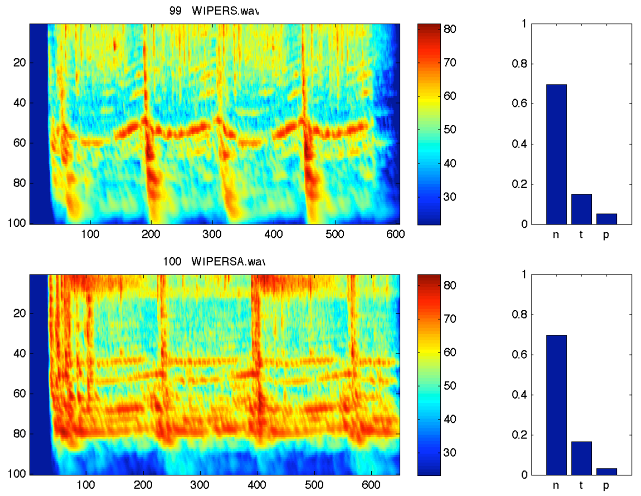

Figure 1 Cochleogram with the fraction of noise (n), tones (t) and pulses (p)

The time-unit of the cochleograms is frame numbers (each 5 ms) and the frequency axis is in segment numbers. Note that, although the cochleograms are quite different, the fraction dominated by the three different contributions is similar in the two files.

The similarity $s_{i,j}$ between sound i and sound j is computed as weighted differences between the different contributions:

$$s_{i,j} = 1 - \frac{1}{3}\left(\left| n_i - n_j \right|^{0.3} + \left| t_i - t_j \right|^{0.3} + \left| p_i - p_j \right|^{0.3} \right) \quad (1)$$

where n_i, t_i and p_i correspond to the noise, tonality and pulsality fraction of sound i. To ensure a range of values between 0 and 1 the similarity is rescaled via

$$s_{i,j} = \left(s_{i,j} - \min(s_{i,j})\right) / \left(1 - \min(s_{i,j})\right) \quad (2)$$

where $\min(s_{i,j})$ corresponds to the minimal value of $s_{i,j}$. All similarities are stored in a matrix $S=(s_{ij})$. S is symmetric with unit diagonal values. All distances in equation [1] are weighted with a power of 0.3. This value is chosen to approach a similar mean of the calculated data and similarity data of the experimental data in matrix *G* provided by Gygi (after rescaling it to [0, 1]). This procedure results in two symmetric matrices S and G with unit diagonal values and a similar average value for the off-diagonal values.

Note that the fractions of the log-energy explained by the three types of textures are dependent in the sense that together they explain all energy. Determining two values fixes the third. As a result the distance measure is based on two-dimensional input.

3.6 MDS analysis

The similarity matrix *S* is obtained from the statistics of the textures for each sound and is compared to the similarity measure obtained from human listening experiment in matrix *G* [2]. To simplify the comparison procedure, the dimensions of the similarity matrices were reduced using the multidimensional scaling (MDS) method [2, 6]. The first three dimensions from the MDS solutions with the highest eigenvalues serve as the axes of a low-dimensional representation. An analysis of the eigenvalues indicates that three dimensions are sufficient to describe the variance.

The correlations between the MDS dimensions for the two similarity matrices *S* and *G* are computed and used as a measure of correspondence. Furthermore, the distribution of the Euclidean distances of the similarities of the individual sound sources projected on the first two MDS axes are calculated and compared. The distribution of the distances between two similar sources (e.g., two bell sounds) are compared to the distribution of the distances for all sounds. A small inter-pair distance, compared to a larger average distance between unrelated sources, indicates that sources with similar textural properties are indeed grouped together.

4 Results

This section addresses three ways to compare the calculated and the experimental similarity measures. The first measure addresses a direct comparison of the two sets. The second measure relies on MDS to form a more efficient

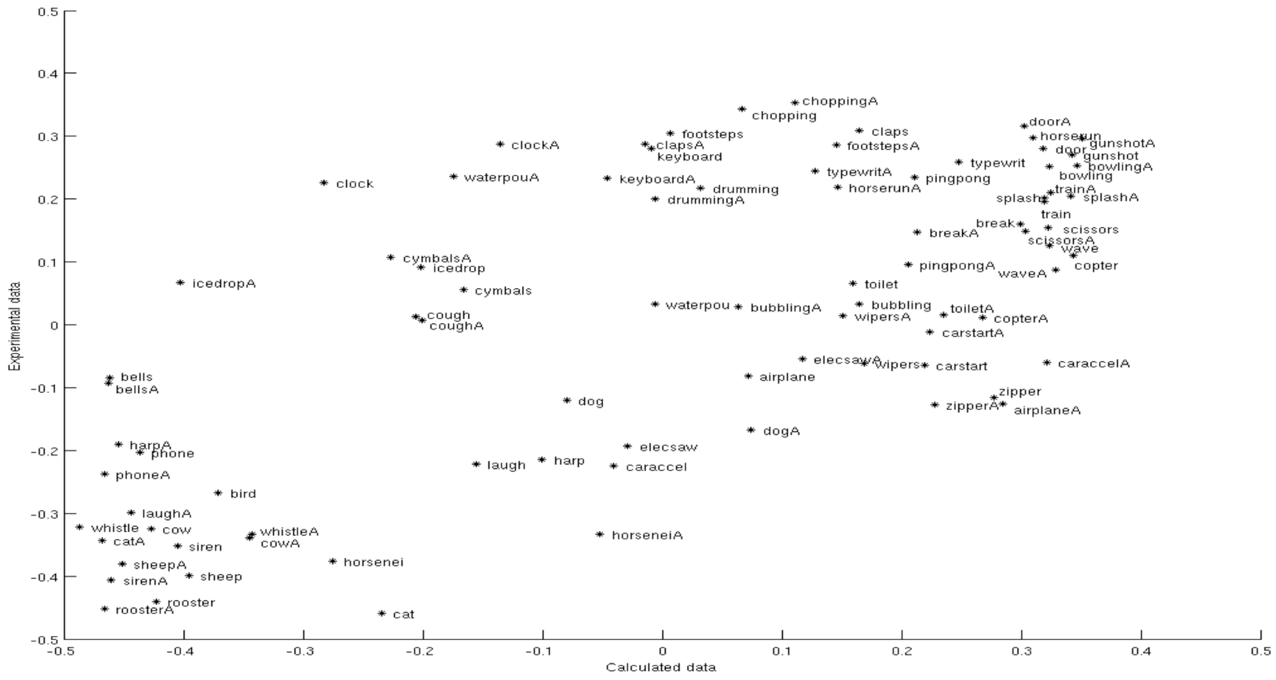

Figure 2 Scatter plot of the relation between the first two MDS dimensions for a selection of the calculated and the experimental data. The associated correlation is 0.71. The second instance of a source pair is denoted by an 'A'

representation of the correlation between the sets. The last measure compares the similarity of two instances of the same source with the similarity of random combinations.

4.1 Distribution of distance values

The distribution of similarity values in the calculated matrix S and the similarity values in G shows a correlation of 0.58 which is equivalent to about 33% explained variance.

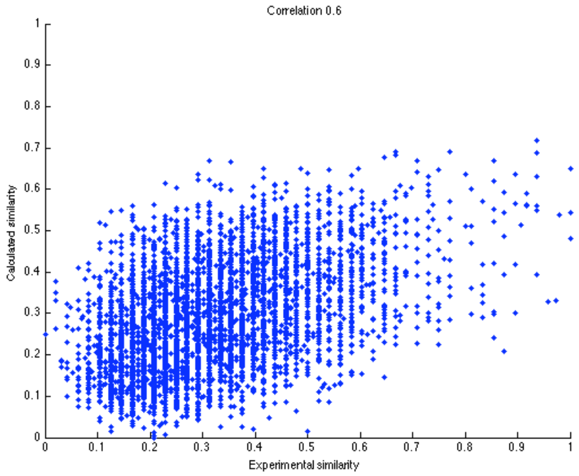

Figure 3 Scatter plot of the distance values for the experimental similarity and the calculated similarity

Given the very simple and low dimensional distance measure this is already an impressive result. It indicates that simple measures of the fractions of tonal, pulses, and noisy contribution can be indicative for a coarse grained classification.

4.2 Correlation between MDS dimensions

Figure 3, at the top of this page, shows the correlation between first MDS dimension of experimental data and first MDS dimension of the calculated data. The correlation on of these first dimensions yields a value of 0.71 which corresponds to 50% of the explained variance. A clustering of the tonal sound sources can be observed in the lower left region of the plot while the pulse-like sound sources are clustered at the top right. The middle part of the plot depicts the clustering of the noisy sound sources. The inter-pair distance is typically small compared to the mean distance between the sounds. This entails a clustering of sounds from similar sources. This fairly high correlation supports the idea that humans tend to identify these very dissimilar sounds based on only a few texture classes.

The correlation between the second MDS dimensions or the calculated and the experimental data shows a correlation of 0.46 which explains another 21% of the variance. Together the simple distance measure based on only two independent values can explain about 70% of the variance of the experimental data. As such it is, given its simplicity, an extremely efficient measure.

4.3 Distances between similar sounds

Plots of the distance distribution calculated from the reduced two-dimensional axes are shown in Fig. 4.

The distribution of the inter-sound distances for the calculated similarity and the listening experiment similarity are shown in the graph as open squares and open circles, respectively. The distance distributions for both similarity measures are similar, which indicates a good agreement between the calculated similarities to those reported by the human subjects. Similarly, a good agreement of the two similarity measures can also be seen in the inter-pair distance distribution (solid square and circle). The comparison of the inter-pair distance with that of the overall distance distribution shows that the experimental and the calculated distances are similar.

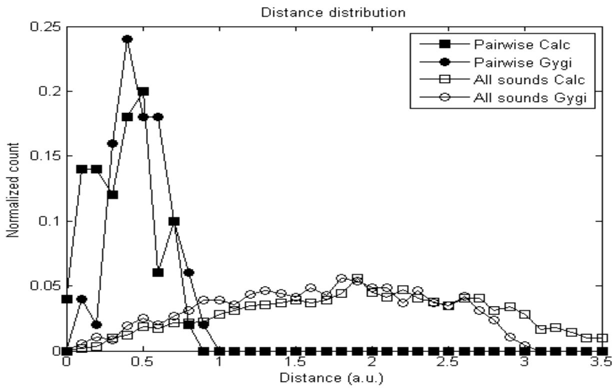

Figure 4 Distance distribution between same source pairs and other possible pairs for the first MDS dimension of Gygi's experimental data and first MDS dimension of the calculated data

This entails that our algorithm, agrees with the human listeners on the similarity of sources with similar physical properties.

5 Discussion

The results show that our procedure leads to inter-sound similarities that correlate to the experimental similarity scores reported in [2]. Considering the simplicity of the distance measure used in the calculation, the MDS dimensions of the experimental results correlate surprisingly strongly. Furthermore the distribution of distances between pairs of the same sound sources and pairs of dissimilar sources are very similar for the experimental and the calculated data. It is important to note that the calculated similarity is only based on two independent values (the third is the remaining energy not assigned to the other two classes). The fact that 70% of the experimental variance can be explained by a distance measure based on only two numbers that describe the individual sounds (which last between 600 ms and 4 seconds) suggests a simple auditory algorithm for source similarity.

Gygi's experiment was designed to investigate sounds with a maximal spread of possible environmental sound sources, which were chosen according to the sound production taxonomy proposed by Gaver [7]. As such the experimental data do not inform us about the fine-grained classifications associated with, for example, speech recognition. The data do however reflect how listeners score perceptual differences between maximally dissimilar environmental sounds. Because these difference estimations do not require fine-grained analyses they are likely to be indicative of the first stages of auditory processing.

The correlation between the calculated and the experimental distances, in combination with calculated distances based on the fraction of the log energy classified as noisy, tonal, or pulse-like suggests that human source similarity estimation might be based on a similarly broad classification. This makes sense from a physical point of view since these regions reflect different physical properties and as such require different algorithmic approaches to convey the information they represent. These results then suggest that texture identification can be combined with separate processing routes for noisy, tonal and pulse-like contributions. These routes allow specific algorithmic approaches that may lead to a maximally informative analysis of the associated texture. Furthermore, the fact that the log-energy weighted contribution of the three components determines the distance between them, suggests that perceptual distances between the wide range of sound classes in this study are determined by the amount of "driving energy" for each of these routes.

Although the previous conclusions are speculative and need more corroboration, one firm conclusion can be drawn. Namely that spectral envelope cues, such as reflected in Mel-scaled Cepstral Coefficients (MFCC) which are commonly used in speech, music, and environment [8] classification, are not necessary to calculate source similarity. In fact the very purpose of this type of descriptors is to code the overall structure of the spectral envelope with as few parameters as possible. As such they are a very efficient way to remove texture cues. The resulting impoverished representations are quite different from what is available in the current study and may therefore not rich enough for general sound source classification.

Acknowledgements

We thank Brian Gygi for providing us with his data and ideas that inspired this research.

References

[1] T.C. Andringa and M. van Grootel, "Predicting listeners' reports of environmental sounds", 19th International Congress on Acoustics, Madrid (2007).

[2] B. Gygi, G. R. Kidd and C. S. Watson, "Similarity and categorization of environmental sounds", *Perception and Psychophysics* (96) 6, 839-855 (2007).

[3] B. Gygi, G. R. Kidd and C. S. Watson, "Spectral-temporal factors in the identification of environmental sounds", *J. Acoust. Soc. Am.*, 115, 1252-1265 (2004).

[4] T.C. Andringa, *Continuity Preserving Signal Processing*, PhD-thesis University of Groningen. (2002)

[5] Chakravarti, Laha, and Roy, *Handbook of Methods of Applied Statistics*, Volume I, John Wiley and Sons, pp. 392-394 (1967)

[6] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika* 29, 1-27 (1964).

[7] W. Gaver, "What in the world do we hear? An ecological approach to auditory event perception", *Ecological Psychology* 5, 1-29 (1993).

[8] Aucouturier, J.-J., Defreville, B., and Pachet, F. "The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music". *Journal of the Acoustical Society of America*, 122(2):881-91, 2007.