Classification of Similar Impact Sounds

Sofia Cavaco and José Rodeia*

CITI, Departamento de Informática
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
scavaco@fct.unl.pt, zrodeia@netcabo.pt

Abstract. Several sound classifiers have been developed throughout the years. The accuracy provided by these classifiers is influenced by the features they use and the classification method implemented. While there are many approaches in sound feature extraction and in sound classification, most have been used to classify sounds with very different characteristics. Here, we propose a similar sound classifier that is able to distinguish sounds with very similar properties, namely sounds produced by objects with similar geometry and that only differ in material. The classifier applies independent component analysis to learn temporal and spectral features of the sounds, which are then used by a 1-nearest neighbor algorithm. We concluded that the features extracted in this way are powerful enough for classifying similar sounds. Finally, a user study shows that the classifier achieves better performance than humans in the classification of the sounds used here.

Keywords: sound classification, feature extraction, natural sounds, acoustic signal processing, independent component analysis.

1 Introduction

While most environmental sound classifiers use sounds with quite different spectral and temporal characteristics, such as door bells, keyboards or whistles, here we focus on the classification of very similar sounds, such as sounds produced by the same event, and by objects with the same geometry and size, that only differ in material. As a consequence these sounds have very similar properties: the temporal envelopes of two sounds produced by the same event (such as impacts on two rods) are more alike than the temporal envelopes of the sound of a door bell and the sound of a dog barking. Also the spectra of two sounds produced by similar objects, such as two metal rods, can be more alike than the spectra of sounds produced by completely different objects.

Sound classifiers are characterized by a stage of sound features extraction and another stage of classification. Many low and high level temporal, spectral and short-time features have been tried [1, 2, 6, 9, 12, 16, 17, 18], but due to the

* Both authors contributed equally to this work.

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 307–314, 2010.
© Springer-Verlag Berlin Heidelberg 2010

difficulty on deciding which are the most appropriate features to characterize the data, many classifiers use a combination of several features to achieve good classification rates. Whereas most sound classifiers use a set of pre-defined features, there are also some classifiers that learn the features using a decomposition method such as matching pursuit and independent component analysis (ICA) [7, 8, 11]. As for the classification stage, some popular techniques in sound classification are the k-nearest neighbor algorithm, neural networks, hidden Markov models and Gaussian mixture models [1, 3, 9, 13, 14, 15].

Even though some of the classification rates obtained with the features (and techniques) mentioned above are very good, we must keep in mind that the sounds used to test the classifiers were produced by very different sources and events. The problem we investigate here is harder due to the similarity of the objects and events used to produce the sounds.

While our sound classifier can use spectral and temporal features, it does not use a set of pre-defined features. Instead, it learns them from the data using Cavaco and Lewicki's method for modeling the intrinsic structures of impact sounds [5]. This method uses ICA to learn the sound features that describe the data more efficiently. The method is able to adapt to the data and learn a small set of features that describe the whole variability in the data and that, at the same time, give enough information to accurately separate samples from different classes. Using these features on a 1-nearest neighbor (1-NN) algorithm, we obtained very high classification rates: 98.33% for spectral features and 100% for temporal features. This allowed us to conclude that the features learned by this method are adequate to classify similar sounds.

2 The Classifier

Instead of extracting pre-defined features of the sounds such as mel frequency cepstral coefficients or other short-time features, our classifier learns them from the data. It starts by representing the sounds with spectrograms (or more precisely, the magnitude of the short-time Fourier transform). It then extracts sound features from these spectrograms (or transposed spectrograms) using Cavaco and Lewicki's method for modeling the intrinsic structures of impact sounds [5]. The method allows us to extract time and frequency-varying functions that we use

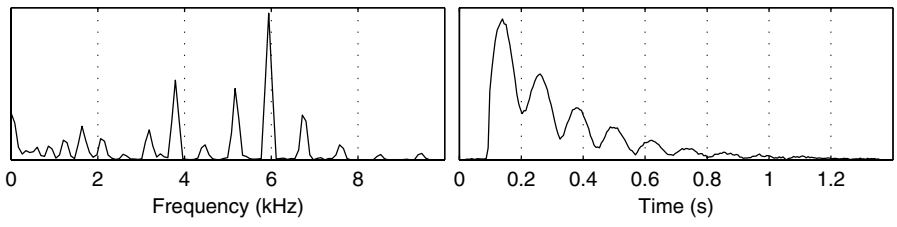

Fig. 1. (a) A spectral basis function. (b) A temporal basis function.

as features for classification in the last stage of the classifier, which consists of a 1-NN algorithm. Here we prove that the features learned in this manner are powerful enough for classifying very similar sounds.

Let us first look into how the spectral features can be learned. In this case, the data is initially represented by spectrograms, one for each sound. Each spectrogram is considered to be a sequence of frames, or in other words, the data runs over frequency. The method performs ICA[1] on the concatenation of these spectrograms, $\left(\mathbf{S}^1, \ldots, \mathbf{S}^{\mathrm{K}}\right)$ [2], where $\mathbf{S}^i$ is the spectrogram of a sound, to learn a set of spectral basis functions $\boldsymbol{\Theta}$ that describe the spectral regularities in the frames in the data set, that is, in $\left(\mathbf{S}^1, \ldots, \mathbf{S}^{\mathrm{K}}\right)$. In this way, the data is now represented in a new space, whose axes are spectral functions (Fig. 1a).

Each frame is represented in this new space of spectral basis functions by a new set of coefficients (one for each basis function). If we consider all the coefficients related to the frames of one spectrogram and one basis function, we have a vector of coefficients. These vectors are commonly called *source signals*, and since here they range over time, we call them *temporal source signals*. The concatenated spectrograms $\left(\mathbf{S}^1, \ldots, \mathbf{S}^{\mathrm{K}}\right)$ can be expressed as the linear combination of these temporal source signals:

$$\left(\mathbf{S}^1, \ldots, \mathbf{S}^{\mathrm{K}}\right) = \boldsymbol{\Theta}\left(\mathbf{P}^1, \ldots, \mathbf{P}^{\mathrm{K}}\right) \quad , \tag{1}$$

where $\boldsymbol{\Theta}$ is a matrix with one spectral basis function per column, and, for every $1 \leqslant k \leqslant \mathrm{K}$, $\mathbf{P}^k$ is a matrix of temporal source signals: there is one such matrix for each spectrogram and the rows of theses matrices contain the source signals. The ith row of every matrix $\mathbf{P}^k$ is associated to the ith basis function in $\boldsymbol{\Theta}$, which corresponds to the ith column of this matrix. Since the method uses ICA to learn $\boldsymbol{\Theta}$ and extract the temporal source signals, these signals are independent. (For more details see [4]).

While the basis functions are the spectral features that will later be used in the classification step, the coefficients in the source signals are the values of those features. Fig. 2 shows four source signals related to the same basis function (from Fig. 1a). Each point in these graphs is the coefficient for one frame. The whole sequence of points is the sequence of coefficients for all the frames in the spectrogram.

Instead of using all these coefficients, the classifier uses only that with the maximum absolute value (the one that corresponds to the highest or deepest peak). These values are marked by circles in Fig. 2. Note how the highest peak in source signals from sounds from the same class have approximate heights

[1] We used the *fastica* software package by Hyvrinen et al. [10]. Because ICA models the variation around the data mean, we used both the spectrogram $\mathbf{S}$ and its negative, i.e., we used the extended matrix $(-\mathbf{S}, \mathbf{S})$, so that the mean would be zero. This was done so that the results from ICA describe the signal rising and falling from zero, rather than the spectrogram mean. The spectrograms of the sounds were computed with a 512-point fast Fourier transform using a 11.6 ms sliding Hanning window. Successive frames overlapped by 5.8 ms.

[2] $(\mathbf{A}, \mathbf{B})$ represents two concatenated matrices.

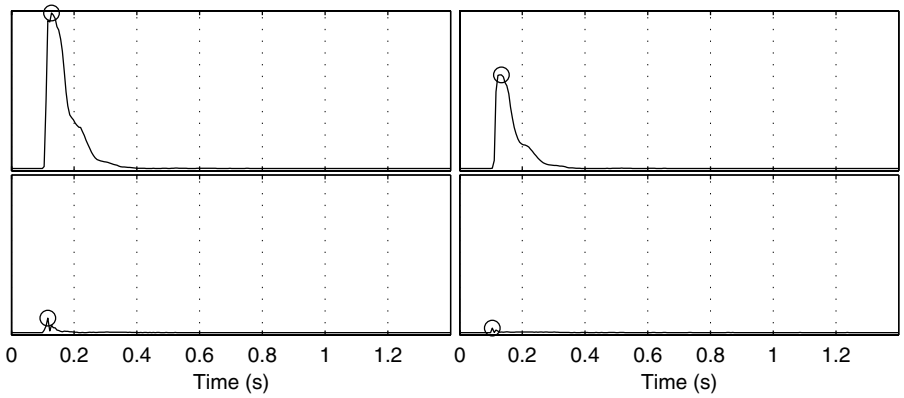

Fig. 2. Temporal source signals from two aluminum sounds (top row) and from two sounds from another rod (bottom row), related to the basis function plotted in figure 1a

(Fig. 2a and b) and how these differ from the heights of the highest peaks in the source signals from the other class (Fig. 2c and d). Also since the maxima from the source signals from Fig. 2a and b are higher than the maxima from the other signals, suggests that the corresponding basis function describes a property of the former sounds, that is, from aluminum.

The method can also be used to learn temporal features. In this case the method uses the transposed spectrograms, which are considered to be a sequence of (transposed) bins, or in other words, the data is initially considered to be running over time. The method performs ICA on the sequence of transposed spectrograms $\left(\left(\mathbf{S}^1\right)^T, \ldots, \left(\mathbf{S}^{\mathrm{K}}\right)^T\right)$ to learn a set of temporal basis functions $\boldsymbol{\Phi}$ that describe the temporal regularities in the bins in the data set (Fig. 1b), and extract a set of independent source signals (i.e. vectors of coefficients). In this way, the data is now represented in a new space, whose axes are temporal functions. If we consider all the coefficients related to the bins of one spectrogram and one basis function, we have a vector of coefficients that runs over frequency. Therefore, we call these signals *spectral source signals*. The data set can be expressed as the linear combination of these spectral source signals:

$$\left(\left(\mathbf{S}^1\right)^T, \ldots, \left(\mathbf{S}^{\mathrm{K}}\right)^T\right) = \boldsymbol{\Phi}\left(\mathbf{C}^1, \ldots, \mathbf{C}^{\mathrm{K}}\right) \quad , \tag{2}$$

where $\boldsymbol{\Phi}$ is a matrix with one temporal basis functions per column, and, for every $1 \leqslant k \leqslant \mathrm{K}$, $\mathbf{C}^k$ is a matrix of temporal source signals (there is one source signal per row and one such matrix per spectrogram). The ith spectral source signal in $\mathbf{C}^k$ is associated to the ith basis function in $\boldsymbol{\Phi}$.

Fig. 3 shows four source signals related to the same basis function (from Fig. 1b). The energy in these signals is a good indicator of the class of the sounds. For example, it can be observed that the spectral source signals in Fig. 3a and b have more energy than those in c and d, which suggests that the basis

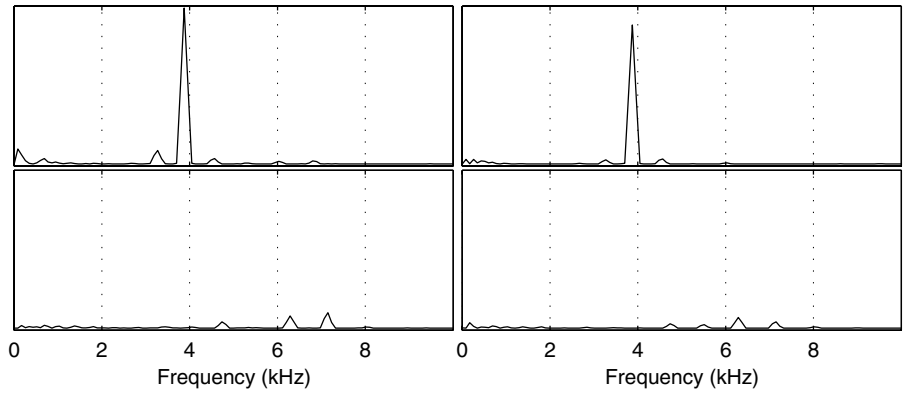

Fig. 3. Spectral source signals from two aluminum sounds (top row) and from two sounds from another rod (bottom row), related to the basis function plotted in figure 1b

function related to these signals is describing a temporal property of sounds from aluminum. Here we used the energy of the spectral source signals to train the classifier.

3 Result Analysis

The data used to test and train the classifier is a set of impacts on rods, which includes 60 samples from four rods with the same length and diameter but different materials (wood, aluminum, steel and zinc plated steel). A wooden rod with the same diameter but much shorter length was used as a mallet. The sounds were all produced by impacts on the same region of the rod (close to the edge), but the location of the impacts and the impact force varied slightly from one instance to the next, since the rods were hit by hand. Even though the sounds differ from each other due to these variations, they are quite similar as they are produced by the same type of event and by objects with the same geometry and size which only differ in material. The sounds were digitized using a sampling frequency of 44100 Hz.

We performed a n-fold cross validation experiment: we used 15 samples from each class (that is, from each rod), organized in 3 sets of 5 samples. Each experiment used 2 sets from each class for training and one set for testing. More explicitly, there were 3 experiments, each having a training set with 10 samples from each rod and a test set with 5 samples from each rod. Therefore, all samples were used for training at two of the experiments and for testing in one of the experiments.

Just like a training sample is an m dimensional vector of the maximum absolute value or the energy in m source signals (associated to m basis functions) from one sound, a test sample also consists of an m dimensional vector computed from m source signals (associated to the same m basis functions). While

the source signals in the training set are learned by Cavaco and Lewicki's method, the source signals in the test set do not need to be learned. Instead, we use the basis functions and the spectrogram of the test sound to determine the source signals with one of the following equations:

$$\mathbf{P}^{\text{test sound}} = \boldsymbol{\Theta}^{-1}\mathbf{S}^{\text{test sound}} \quad , \tag{3}$$

$$\mathbf{C}^{\text{test sound}} = \boldsymbol{\Phi}^{-1}(\mathbf{S}^{\text{test sound}})^T \quad . \tag{4}$$

While the method learns a high number of basis functions, the classifier does not need to use them all. A small subset of m basis functions that completely separates the data set can be chosen. Therefore, the classifier can be trained either with m dimensional vectors of the maximum absolute value of the temporal source signals, or with m dimensional vectors of the energy in the spectral source signals. In the experiments performed m varied from 4 to 11, giving classification rates of 98.33% and 100% for spectral and temporal features, respectively. Such accurate results can be obtained with a low number of features because the method is able to learn basis functions that describe the intrinsic properties present in the sounds of each class. For instance, the method learned features that describe a ringing property from the aluminum sounds. This basis function can be used to separate most aluminum sounds from the rest of the samples.

4 Human Listening Test

A natural question that follows is if the classifier can surpass human ability to classify similar sounds and if the sounds are actually hard (for humans) to distinguish. To answer these questions, we conducted two user studies. In these studies, the subjects were asked to classify the same impact sounds used to train and test the classifier. The studies were performed using headphones. The first user study had 12 participants with ages between 23 and 55. None had hearing problems and two had acoustics knowledge. The second user study had 11 participants with ages between 23 and 50. None had hearing problems and one had a high level of acoustics and music knowledge.

Before the actual studies started, two sounds of each class were presented with indication of the material of the rod that produced the sound. Then, in order to familiarize the subjects with the task, they were asked to classify some sounds (presented in random order). In the first user study, the subjects were presented two sounds from each class and received no feedback on whether their answers were correct or incorrect, while in the second user study the subjects were presented three sounds from each class and were told the correct material right after giving their answers. After this training phase, the actual test began. In the first user study, the subjects heard 41 sounds in random order: 7 of aluminum, 11 of zinc plated steel, 12 of wood and 11 of steel. In the second user study, the subjects heard 32 sounds: 7 of aluminum, 10 of zinc plated steel, 5 of wood and 10 of steel.

The results (Table 1) show that while the wood sounds are easily recognizable, the metal sounds used in these studies are very hard for humans to

Table 1. Results from the user studies. Each row shows the percentage of correct answers obtained for each material. The top percentage in each cell shows the results obtained for the first user study, while the bottom percentage shows the results obtained for the second user study.

	Aluminum	Zinc plated steel	Wood	Steel
Aluminum	22.619%	53.571%	0%	23.81%
	53.247%	23.337%	0%	23.377%
Zinc plated steel	43.939%	24.242%	0%	31.818%
	26.364%	40.909%	0%	32.727%
Wood	1.389%	0%	97.917%	0.694%
	0%	0%	100%	0%
Steel	28.03%	25.758%	0%	46.212%
	27.273%	28.182%	0%	44.545%

distinguish. As can be observed in the diagonal of this table, while more training and receiving feedback during training improves the results, the percentages of right answers for the metal sounds are inferior to 54% in both studies. In the first user study, the percentage of right answers for steel sounds was much higher than the percentage of right answers for the two other metal rods. However, this did not improve in the second user study where users got more training.

The mistaken answers also reveal that the sounds are very similar. For instance, in the second study, when missing the aluminum sounds, we detect the same percentage of wrong answers for steel and zinc plated steel (23.377%): the users assume it can be any of the three metals when confused.

5 Conclusions

A system for classifying very similar sounds has been proposed. While most environmental sound classifiers use sounds with quite different spectral and temporal characteristics, our classifier is able to distinguish sounds with very similar characteristics. The classifier can use temporal and spectral features learned by ICA of the spectrograms (or transposed spectrograms), to train a 1-NN algorithm that can then be used to classify new sounds.

Two user studies allowed us to conclude that while the wood sounds are easily recognizable, the metal sounds used here are very hard for humans to distinguish. These studies also show that the classifier achieves far better results than humans. The classifier missed only one metal sound out of 60 sounds in an n-fold cross validation experiment. The classification rates obtained are, therefore, very high: 98.33% for spectral features and 100% for temporal features.

While we used temporal and spectral features separately, it is possible to combine them in one classifier only. Since our results were so high, we did not try this approach.

References

[1] Berenzweig, A.L., Ellis, D.: Locating singing voice segments within music signals. In: Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio, pp. 119–122 (2001)
[2] Breebaart, J., McKinney, M.: Features for audio classification. In: Proc. Philips Symposium on Intelligent Algorithms, Eindhoven (2002)
[3] Bugatti, A., Flammini, A., Migliorati, P.: Audio classification in speech and music: a comparison between a statistical and a neural approach. Applied Signal Processing (1), 372–378 (2002)
[4] Cavaco, S.: Statistical modeling and synthesis of intrinsic structures in impact sounds. PhD thesis, Carnegie Mellon University (2007)
[5] Cavaco, S., Lewicki, M.S.: Statistical modeling of intrinsic structures in impact sounds. Journal of the Acoustical Society of America 121(6), 3558–3568 (2007)
[6] Chou, W., Gi, L.: Robust singing detection in speech/music discriminator design. In: Proceedings of the Acoustics, Speech, and Signal Processing on IEEE International Conference, pp. 865–868 (2001)
[7] Chu, S., Narayanan, S., J Kuo, C.-C.: Environmental sound recognition using MP-based features. In: Proc. IEEE ICASSP, pp. 1–4 (2008)
[8] Eronen, A.: Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. Signal Processing and Its Applications 2, 133–136 (2003)
[9] Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based context recognition. Audio, Speech, and Language Processing 14(1), 321–329 (2006)
[10] Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley & sons, Ltd., Chichester (2001)
[11] Kraft, F., Schaaf, T., Waibel, A., Malkin, R.: Temporal ICA for classification of acoustic events in a kitchen environment. In: Proc. International Conference on Speech and Language Processing - Interspeech, pp. 2689–2692 (2005)
[12] Liu, Z., Wang, Y., Chen, T.: Audio feature extraction and analysis for scene segmentation and classification. The Journal of VLSI Signal Procesing 20(1-2), 61–79 (1998)
[13] Ma, L., Smith, D.J., Milner, B.P.: Context awareness using environmental noise classification. In: Proc. of Eurospeech, vol. 3, pp. 2237–2240 (2003)
[14] Nóbrega, R., Cavaco, S.: Detecting key features in popular music: case study - singing voice detection. In: Ramirez, R., Conklin, D., Anagnostopoulou, C. (eds.) Proc. of the Workshop on Machine Learning and Music of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2009)
[15] Ntalampiras, S., Potamitis, I., Fakotakis, N.: Automatic recognition of urban environmental sounds events. New Directions in Intelligent Interactive Multimedia, 147–153 (2008)
[16] Pfeiffer, S., Fischer, S., Effelsberg, W.: Automatic audio content analysis. In: Proc. of ACM International Conference on Multimedia, pp. 21–30 (1997)
[17] Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Proc. IEEE ICASSP, vol. 2, pp. 1331–1334 (1997)
[18] Tzanetakis, G., Essl, G., Cook, P.: Audio analysis using the discrete wavelet transform. In: Proc. Conference in Acoustics and Music Theory Applications (2001)