Fall 10-20-2014

Self-organizing the space of vocal imitations

Davide Rocchesso

Davide Andrea Mauro PhD
Marshall University, maurod@marshall.edu

Recommended Citation

Rocchesso, D. and Mauro, D.A. "Self-organizing the space of vocal imitations." XX CIM Colloquio di Informatica Musicale, Rome (Italy), 20-22 October, 2014.

SELF-ORGANIZING THE SPACE OF VOCAL IMITATIONS

Davide Rocchesso, and Davide Andrea Mauro
Department of Architecture and Arts, Iuav University of Venice
`roc, dmauro @iuav.it`

ABSTRACT

The human voice is a powerful instrument for producing sound sketches. The sonic space that can be spanned with the voice is vast and complex and, therefore, it is difficult to organize and explore. In this contribution, we report on our attempts at extracting the principal components from a database of 152 short excerpts of vocal imitations. We describe each excerpt by a set of statistical audio features and by a measure of similarity of the envelope to a small number of prototype envelopes. We apply k-means clustering on a space whose dimensionality has been reduced by singular value decomposition, and discuss how meaningful the resulting clusters are. Eventually, a representative of each cluster, chosen to be close to its centroid, may serve as a landmark for exploring the sound space.

1. INTRODUCTION

The EU project SkAT-VG [1] (Sketching Audio Technologies using Vocalizations and Gestures, 2014-2016) is aimed at finding ways to exploit voice and gestures in sonic interaction design. Research in SkAT-VG proceeds along three directions: (i) improving our understanding on how sounds are communicated through vocalizations and gestures; (ii) looking for relations between vocal/gestural primitives and the physical characteristics of sound-producing phenomena; (iii) designing tools for converting vocal and gestural actions into parametrized sound models.

In this paper we describe a research exercise that may be ascribed to the first direction of the SkAT-VG project. We try to see how the space of vocal imitations could be rearranged and simplified to highlight clusters of sounds that are acoustically similar. We also check if the clusters, produced by algebraic and algorithmic manipulations, make sense to humans as well.

The paper is organized as follows: we first introduce the sonic space we want to be able to explore and understand. Then in sections 3 and 4 we tackle the problem of how to encode and efficiently manipulate the data in order to organize them. We present then perspectives on how to relate the automatic analysis with the subjective classifications obtained from human subject.

2. THE VOCAL SONIC SPACE

In a sense, the human voice has for acoustic communication a role similar to what the hand and pencil have for visual communication. Humans use their voice for verbal communication as well as for non-verbal acoustic expression, similarly to the hand which is used both for writing and for drawing. Just as the hand and pencil are extensively used for visual sketching, the voice has potential to be exploited for sketching sounds. Indeed, sketching comes before verbal – oral or written – expression in development of both the human species and the human individuals [1].

In order to devise tools that facilitate sound design by vocal sketching we must gain a better understanding of what the voice can do and how vocalizations are interpreted by listeners. The space of voice-produced sounds need to be described both in acoustic and in articulatory terms. We need to know the characteristics of a comprehensive repertory of vocal sounds and how these can be achieved by our voice organ. From a sound design perspective, it is particularly useful to organize the vocal sound space on a low-dimensional layout whose navigation can be facilitated by landmarks, or sounds that represent distinct neighborhoods. The purpose of this study is to construct such a layout automatically from a database that significantly spans the possible non-verbal uses of the human voice.

A database of 152 audio segments were manually extracted from the Fred Newman's repertory of vocal imitations described in his book [2] and included in the companion CD. The segments were all $500\ \mathrm{ms}$ long (22050 samples at 44100 samples/s) and were taken to represent a single sound event or process. There is a degree of arbitrariness in this operation, as some events may be the result of a concatenation of articulatory actions of a shorter length, but for the scope of this study each audio segment may be considered to include a single vocal utterance.

The idea of using landmarks to facilitate navigation in the sound design space was previously explored in the context of parametric sound synthesis [3, 4], and auditory representations were used both to give a visual snapshot to each sound and to compute distances that would allow locating new sounds in the map. In this work we aim at extracting landmarks as representatives of clusters.

3. REDUCING DIMENSIONALITY: A COMPACT DESCRIPTION OF SOUNDS

Reducing the dimensionality seems to be a reasonable approach to organize a sonic space. A classic way to do that is

[1] `http://www.skatvg.eu/`

by means of Principal Component Analysis (PCA), which is based on Singular Value Decomposition (SVD).

Taken our reference database of 152 audio segments, a PCA on the raw audio files would reveal that as many as 110 singular values need to be retained to account for 90% of the energy. The basis vectors, while being recognizable as vocal sounds, are not easily associated to different articulatory or acoustic characteristics. The situation does not change much if we apply PCA on the (Fourier, or Wavelet) transformed version of the audio segments, in the sense that the number of dimensions that would retain most of the signal energy is almost as large as the number of exemplars in the database. Things get better if we giveup invertibility and, for example, apply PCA on the magnitude Fourier spectrum. In this case 95 singular values are sufficient, yet still too many, to account for 90% of the energy. Giving up invertibility means that it would not be possible to select a point in the reduced sonic space and have the sound produced by inverse-transforming. Still, it would be possible to localize a given sound in the space and, for example, to interpolate between landmarks to synthesize a similar sample.

In the area of music information retrieval a lot of research has been devoted to extract audio descriptors (or features) that could concisely represent sound and music [5]. Several software libraries are available to easily extract brightness, spectral flux, and other descriptors from a given soundfile, and to collect statistical information from them. For this study, we have been using the MIR toolbox [6], and we applied some of its feature extractors to summarize each of our audio segments with statistical information. In particular, we used the median and interquartile range values (as recommended in [5]) of spectral flux, centroid, roughness, flatness, entropy, skewness, and RMS energy computed over 18 windows spanning the $500\mathrm{ms}$ of each audio segment.

In addition to the statistical audio features, we added some features that would account for the temporal morphology of each audio segment. The idea is that, for example, such features would mark a clear difference between a sustained noise and an impulsive click. However, there is the problem of where short temporal events actually occur in time, as it should be irrelevant if an impulsive click occurs at time $100\mathrm{ms}$ or $300\mathrm{ms}$ in the considered time span. In order to account for possible elastic deformations of time, Dynamic Time Warping (DTW) is used to compare distances between the extracted RMS profile and each of a set of prototype temporal envelopes, namely upward slope, downward slope, up-down profile, and impulses.

All collected features are non-negative real numbers, but their range and units are quite different from each other. For the subsequent step of PCA, we perform a normalization to the maximum value of each feature in our population of samples. Still, most of the distributions are heavily skewed toward zero. In order to obtain feature distributions that more closely resemble a gaussian we distort the distribution of values of each feature by its cumulative histogram, a cheap trick that is called histogram equalization in image processing. As an example, Figure 1 shows the distribution of median and interquartile range values of spectral flux after histogram equalization.

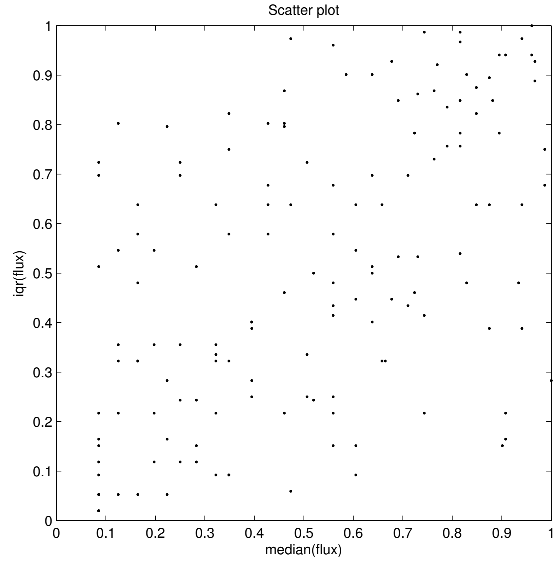

Figure 1. Distribution of median and interquartile range values of spectral flux after histogram equalization

Before the extraction of principal components, the mean is subtracted from the distribution of each feature, and the distribution is further normalized to range between -1 and 1. Then, the thin SVD is computed on the matrix $B \in \mathbb{R}^{m \times f}$, where $m = 152$ is the number of audio segments and f is the number of features:

$$B = USV'. \tag{1}$$

$S \in \mathbb{R}^{f \times f}$ is the diagonal matrix of singular values in descending order, $U \in \mathbb{R}^{m \times f}$ is the matrix of orthonormal basis vectors (principal components) that best represents the set of audio segments (described as features) in a L^2 sense. The $i-$th row of U expresses the $i-$th audio segment as a set of coefficients of a combination of principal directions, or “feature modes”. These modes are expressed as columns of $SV' \in \mathbb{R}^{f \times f}$.

To reduce dimensionality, we retain only columns 1 to l of matrix U, corresponding the l largest singular values, or to the most prominent feature modes.

4. CLUSTERING

In general, clustering in the PCA-reduced subspace is more effective than doing it in the original space, because the subspace of $l + 1$ cluster centroids is spanned by the first l principal directions of data [7].

For our database of audio segments, each summarized by the 18 features described in Section 3, the singular values are plotted in Figure 2.

With $l = 2$ (two principal components) a run of k-means clustering with a variable number of clusters returns a squared sum of errors as represented in Figure 3. With such low value of l, the extraction of three clusters is particularly effective, and such clusters can be displayed in the 2-D space of principal components, as in Figure 4. In

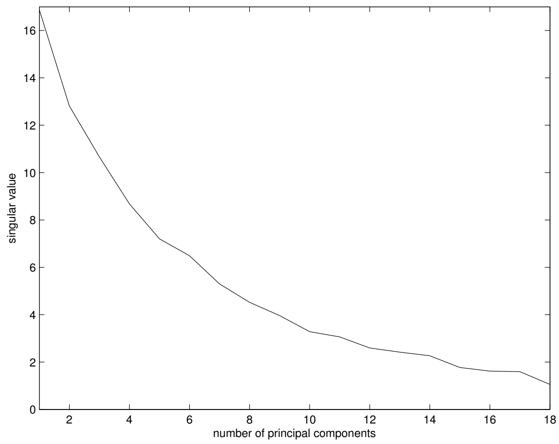

Figure 2. Singular values of matrix B of descriptions of audio segments

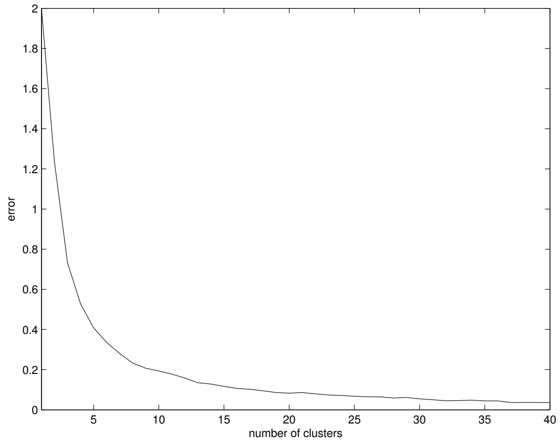

Figure 3. Squared sum of errors as a function of number of clusters ($l = 2$)

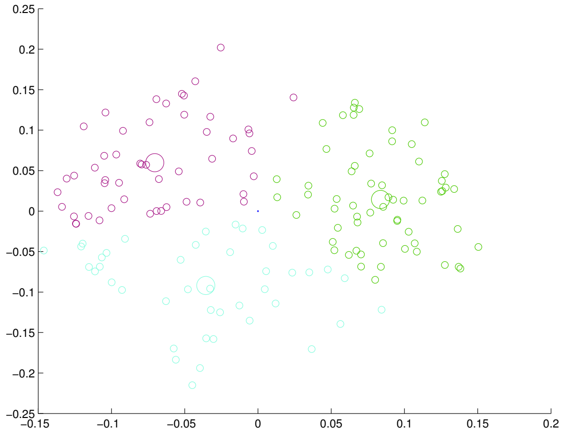

Figure 4. Three clusters in the space of the two principal components

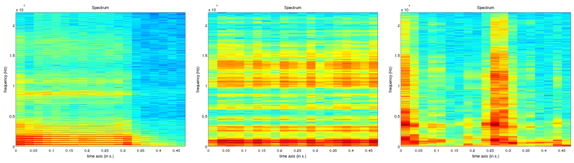

Figure 5. Spectrograms of representatives of clusters 1 (60 elements), 2 (42 elements) and 3 (50 elements)

that figure, the larger circles correspond to the cluster centroids, which ideally should be selected as representatives of each cluster. In practice, since resynthesizing a vocalization that corresponds to such centroids is not possible, we can choose the closest member as a cluster representative. The spectrograms of such representatives are depicted in Figure 5. The first is described as imitation of a trumpet, the second is a prototype of “glottal fry”, and the third is a “tongue flop” that could be used to imitate horse steps. Given such a relatively small number of clusters compared to the number of elements and the vague nature of the terms and categories that can be used to describe sounds it is not easy to interpret them. In the first cluster we have sounds that are (mostly) pitched or exhibit an intonation. The second cluster contains sounds that are continuous. Finally the third cluster encompasses sounds that are characterized by an impulsive behavior or a temporal evolution.

By visualizing the cells of the equalized, centered, and normalized matrix of features we can make sense of how clustering operates on audio descriptors projected along principal components. Such visualization is displayed on Figure 6, where a row-wise sorting has been performed to group all audio segments belonging to the same cluster. In the cells, deep blue corresponds to -1 and deep red corresponds to +1 as normalized feature values.

Even more meaningful is connectivity analysis [7]. Figure 7 shows, discretized to a binary color, the matrix $U_2 {U_2}'$, where U_2 are the first two columns of U, with their rows sorted according to the extracted clusters. Apart from the three clusters, which are clearly visible, elements of “contamination” between and within clusters are also visible. This indicates the opportunity of refined clustering, either increasing l, or looking for more than $l + 1$ clusters, or running PCA and clustering hierarchically on each cluster. In figure 7, the degree of connectivity is $c = 65\%$, i.e.,

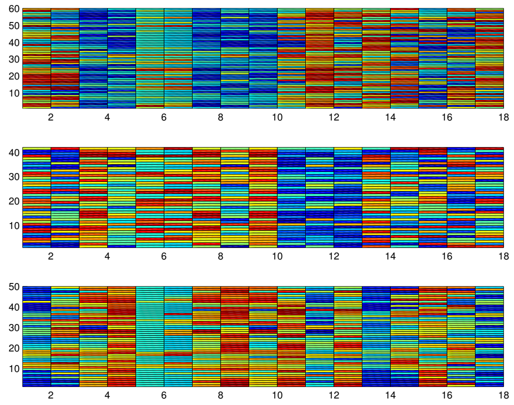

Figure 6. How clusters are derived from features (one per column)

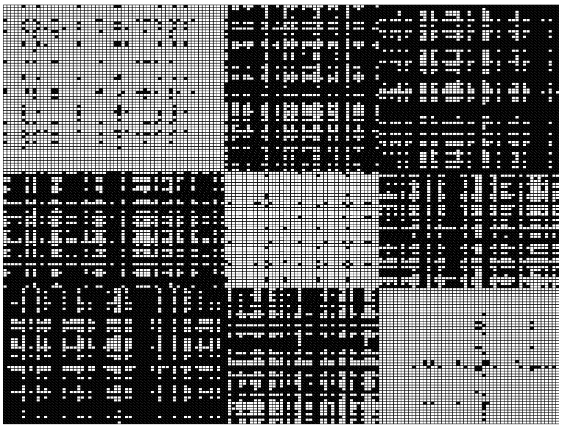

Figure 7. Connectivity of 152 audio segments in two principal components

65% of the white cells belong to the three squares on the diagonal, showing strong connection within clusters.

5. HOW WOULD A HUMAN DO?

Considering a small (i.e., 3) number of clusters, we asked three colleagues, not involved in this research exercise, to memorize the three cluster representatives and then to assign each of the remaining 149 sounds to one of the representatives. From these associations, we computed the confusion matrix and the clustering accuracy for each subject. The three subjects showed the values of accuracy: 0.48, 0.53, 0.65, where a random assignment would return a value 0.33 of accuracy. For example, for the subject that is the closest to automatic clustering, the confusion matrix is $C = \begin{bmatrix} 46 & 13 & 1 \\ 6 & 24 & 12 \\ 10 & 11 & 29 \end{bmatrix}$, where element $c_{i,j}$ represents the number of audio segments that have been assigned to cluster i by the machine and to cluster j by the human. It is an interesting coincidence that the accuracy expressed by this subject (65%) is the same as the degree of connectivity expressed in figure 7.

6. POSSIBLE EXTENSIONS

The degree of connectivity c depends on the number l of principal components and on the number k of clusters. The maximum value $c = 100\%$ is obtained for $l = 1$ and $k = l + 1 = 2$, where the clear separation into two clusters seems to be largely determined by spectral centroid and flatness. Increasing the number of clusters by just one ($l = 1, k = 3$) gives a much more confused picture ($c = 0.62$).

In general, subdivision into k clusters is best done on $l = k - 1$ principal components. With this constraint, and for $l = 1, 2, 3, 4, 5$ we get degrees of connectivity $c = 1, 0.65, 0.49, 0.39, 0.33$, respectively. In all cases, the prototype sounds (cluster representatives) found are perceptually distinct from each other, and they may well serve the purpose of automatically finding landmarks in the space of vocal imitations.

Further explorations of the space are currently being pursued, especially experimenting with hierarchical clustering, to see if more meaningful subdivisions will emerge. So far, relatively little attention has been payed to the features, which were chosen from a set of standard audio features used for musical signals extended with signatures of temporal envelope. The fact that the sounds are all of vocal origin should be exploited to include specific features that come from the literature of speech and voice analysis.

7. ACKNOWLEDGMENTS

The authors are pursuing this research as part of the project SkAT-VG and acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 618067.

8. REFERENCES

[1] M. Changizi, *Harnessed: How Language and Music Mimicked Nature and Transformed Ape to Man*. Perseus Books Group, 2013.

[2] F. Newman, *MouthSounds: How to Whistle, Pop, Boing, and Honk... for All Occasions and Then Some*. Workman Publishing, 2004.

[3] C. Drioli, P. Polotti, D. Rocchesso, S. Delle Monache, K. Adiloglu, R. Annies, and K. Obermayer, "Auditory representations as landmarks in the sound design space," in *Proceedings of Sound and Music Computing Conference*, 2009.

[4] K. Adiloglu, C. Drioli, P. Polotti, D. Rocchesso, S. Delle Monache, *et al.*, "Physics-based spike-guided tools for sound design," in *Conference on Digital Audio Effects*, 2010.

[5] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[6] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects*, pp. 237–244, 2007.

[7] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, (New York, NY, USA), pp. 29–, ACM, 2004.