Fall 10-7-2015

Analyzing and organizing the sonic space of vocal imitation

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

D. Rocchesso

Recommended Citation
Mauro D.A., Rocchesso D. "Analyzing and organizing the sonic space of vocal imitation." Audio Mostly 2015, October 7-9, 2015, Thessaloniki, Greece.

Analyzing and organizing the sonic space of vocal imitations

D.A. Mauro, and D. Rocchesso
Iuav University of Venice
Department of Architecture and Arts
Venice, Italy
{dmauro, roc}@iuav.it

ABSTRACT

The sonic space that can be spanned with the voice is vast and complex and, therefore, it is difficult to organize and explore. In order to devise tools that facilitate sound design by vocal sketching we attempt at organizing a database of short excerpts of vocal imitations. By clustering the sound samples on a space whose dimensionality has been reduced to the two principal components, it is experimentally checked how meaningful the resulting clusters are for humans. Eventually, a representative of each cluster, chosen to be close to its centroid, may serve as a landmark in the exploration of the sound space, and vocal imitations may serve as proxies for synthetic sounds.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing

General Terms

Experimentation

Keywords

Vocal imitations, Clustering, Landmarks, PCA

1. INTRODUCTION

In this contribution, we investigate how the space of vocal imitations could be arranged and simplified to highlight clusters of sounds that are acoustically similar. We also assess if the clusters, produced by mere algebraic and algorithmic manipulations, make sense to humans as well. Prototype sounds are automatically selected to represent clusters, and human participants are requested to label each of the remaining sounds as being perceptually closer to one of the prototypes. Between subject consistency is measured and the low-dimensional space is partitioned according to the preferences of participants.

AM '15, October 7-9, 2015, Thessaloniki, Greece.

DOI: http://dx.doi.org/10.1145/2814895.2814921.

Section 2 introduces the sonic space of vocal imitations and its possible uses. Sections 3 and 4 respectively show how to encode and group samples of vocal imitations in order to organize them. In section 5 we check how automatic clustering compares with human performance for the task of classifying a whole set of vocal imitations based on similarity to the extracted cluster prototypes. Finally, section 6 indicates how this work will be extended and used for sound design.

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 or imitating sounds. Indeed, sketching comes before verbal – oral or written – expression in development of both the human species and the human individuals [3]. Recent research has shown that vocal imitations can be more effective than verbalizations at representing and communicating sounds [13]. Such natural capabilities are being exploited for sound retrieval and synthesis [2]. The European project SkAT-VG is investigating the use of voice and gesture as intuitive means for the selection and control of sound models in sonic interaction design [19].

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 needs 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 explore the construction of 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 [17] and included in the companion CD. The segments were all 500 ms long (22050 samples at 44100 samples/s) and were taken to represent a single sound event or process. The length has been chosen in order to try to

accomodate for different phenomena. There is still a degree of arbitrariness in this operation, as some events may be the result of a concatenation of articulatory actions of a shorter time span, but for the scope of this study each audio segment may be considered to include a single utterance.

Since the audio segments were extracted from a comprehensive set of examples of a renown professional vocal artist, they are likely to represent well the possibilities of human voice. In general, we would like to be able to browse collections of vocal samples, and to do that it is desirable to organize them on a two-dimensional surface, where new vocal imitations could be added and possibly used as proxies for non-vocal or synthetic reference sounds. In a practical application for sound designers, we may want to navigate the sonic space of a given sound synthesis model, and vocal imitations may be used to refer to underlying synthetic sounds.

Tools for sonic browsing on two dimensions were proposed in the past [7]. The idea of using landmarks to facilitate navigation in the sound design space was explored in the context of parametric sound synthesis [5, 1], 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 the present study, we show how a low-dimensional space of vocal imitations, each possibly corresponding to an underlying synthetic sound, can be automatically arranged and partitioned, with landmarks automatically extracted as representatives of clusters.

Relevant related work is [20], where a free-sorting task on 150 non-vocal sound effects, assigned to several subjects, produced dissimilarity matrices to train an automated classifier via multidimensional scaling. Categorization via manual grouping was done for everyday sounds in selected contexts, such as cars [15]. For kitchen sounds [9], the four main categories of solids, electricals, gases, and liquids were found, and they were largerly confirmed when subjects were requested to sort imitations of such sounds [12]. The organization of sound material into spatial layouts for performance control was investigated in [16], where mixtures of Gaussians were used to achieve continuous interpolation in the sonic space. The mapping between vocal postures/gestures and sound-synthesis parameters is an active research topic in sound and music computing, where several machine-learning techniques can be exploited [6].

3. REDUCING DIMENSIONALITY: A COMPACT DESCRIPTION OF SOUNDS

Digital signals are described by sequences of many values, and reducing the dimensionality is a necessary step in order to organize a sonic space. A classic way to do that is by means of Principal Component Analysis (PCA), which is based on Singular Value Decomposition (SVD) [10].

Attempting a reduction of dimensionality on the raw audio files or on their invertible transformations (Fourier, or Wavelet) is not successful. That is why more compact descriptions of sounds are conveniently adopted, even if they do not allow to reconstruct the original signals. However, in a sonic space where landmarks are associated with instances of sound models, it would be possible to localize a given sound in the space and to interpolate between neighboring landmarks to synthesize a new sample, even without direct reconstruction from descriptors.

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 [18]. Several software libraries are available to easily extract brightness, spectral flux, and other descriptors from a given soundfile, and to collect statistical descriptors from them. For this study, we have been using the popular MIR toolbox v.1.5 [11] under Matlab R2010b, and we applied a number of its feature extractors to summarize each of our audio segments with statistical information. In particular, we used the median and interquartile (IQR) range values (as recommended in [18]) of spectral flux, centroid, roughness, flatness, entropy, skewness, and RMS energy computed over 18 windows spanning the 500ms-duration 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 100ms or 300ms 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 a number of templates. The set of prototypical temporal envelopes is constituted by: upward slope, downward slope, up-down profile, and impulses. As compared to the study on morphological profiles conducted on 55 environmental sounds by Minard et al. [14], we used four of the six dynamic profiles that resulted from manual clustering by their pool of experts. Among the many other possible descriptors that could be used, those exploiting the nature of vocal sounds are particularly interesting, and will be briefly considered in section 6. However, in this study the organization of the sonic space, the extraction of prototype sounds, and the subjective tests are voice agnostic. Table 1 lists the features used in this study.

1 2	Flux	median IQR	Distance between consecutive spectral frames
3 4	Centroid	median IQR	The first moment of a spectral frame
5 6	Roughness	median IQR	Estimation of sensory dissonance
7 8	Flatness	median IQR	Indicates whether the spectrum is smooth or spiky
9 10	Entropy	median IQR	The relative entropy of a spectral frame
11 12	Skewness	median IQR	A measure of symmetry of a spectral frame
13 14	RMS	median IQR	The global energy of a spectral frame
15	Upward		Upward slope
16	Downward		Downward slope
17	Up-down		Up-down profile
18	Pulses		Train of pulses

Table 1: The eighteen features considered in the study.

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 evenly span the unit interval we distort the distribution of values of each feature by its cumulative histogram (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 $f = 18$ is the number of features:

$$B = USV'. \quad (1)$$

$S \in \mathbb{R}^{f \times f}$ 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 to the l largest singular values, or to the most prominent feature modes. For our database of audio segments, each summarized by the 18 features of Table 1, the decay of singular values is relatively slow, thus not giving an obvious cutoff for l. Still, a meaningful and practical navigation of the sonic space can only be afforded by a low-dimensional space. In particular, the first two principal components are the ones that would afford effective browsing [7], even though they explain less than one third of the variance for this set of sounds.

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 [4]. Particularly interesting is the case of two principal components ($l = 2$), because that gives a bi-dimensional space that is easy to navigate, as if it was a map displaying a set of landmarks. 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. Figure 1 displays the clusters of 60 (red), 42 (green), and 50 (blue) elements, as well as the six largest principal-component loading vectors (two-component reduction of the columns of SV'). The anti-diagonal of this space is roughly aligned with the median centroid, or brightness of sound. Although the m audio segments do not tend to cluster in three distinct groups, the clustering procedure provides a three-fold subdivision of the sonic space. In Figure 1, 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 2. The first can be 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 (red) cluster we have sounds that are (mostly) pitched. The second (green) cluster contains sounds that are continuous and noisy. Finally the third (blue) cluster encompasses sounds that are characterized by an impulsive behavior or a temporal evolution. The three classes of vocal imitations roughly correspond to the categories of instrument-like, motor, and impact sounds as they emerged from the analysis of a free-sorting task on 83 sounds of car interiors, air-conditioning units, car horns, and car doors [15]. The classes could also be put in correspondence with the categories of electricals, gases/liquids, and solids, as they emerged from categorization of kitchen sounds and of their imitations [9, 12].

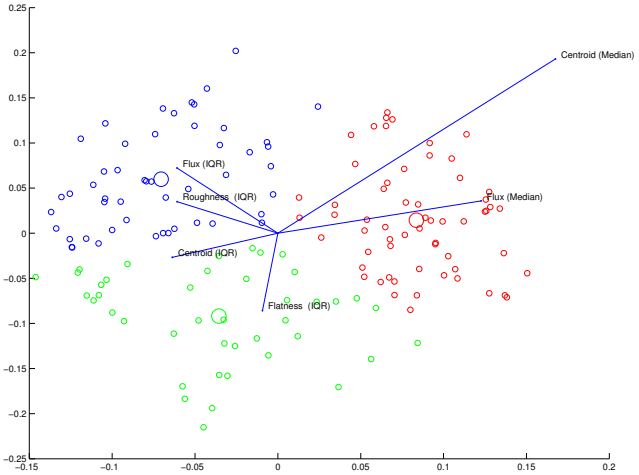

Figure 1: Three clusters in the space of the two principal components

Even more meaningful is connectivity analysis [4]. Figure 3 shows, discretized to a binary color, the matrix U_2U_2', where U_2 are the first two columns of U, with their rows sorted according to the extracted clusters. By discretizing such matrix to binary values, the three clusters as well as the elements of "contamination" between and within clusters can be made visible. In this clustering the degree of connectivity is $c = 0.65$, i.e., 65% of the active cells belong to the three squares on the diagonal of U_2U_2', thus showing strong connection within clusters.

4.1 Different clustering techniques

It is possible to replace the k-means clustering with other different techniques that enable us to highlight different perspectives on data. We report results for hierarchical, Fuzzy C-means and GMM clustering:

- Connectivity for Hierarchical= 0.58
  - Cophenetic correlation coefficient: 0.69 [1]
- Connectivity for Fuzzy C-Means= 0.64
- Connectivity for GMM= 0.57

With hierarchical clustering we can plot the dendrogram for the linkage. Looking at this graph depicted in Figure 4, it is possible to gain some insight on the nature of grouping.

[1] As a comparison it is 0.95 in [12]

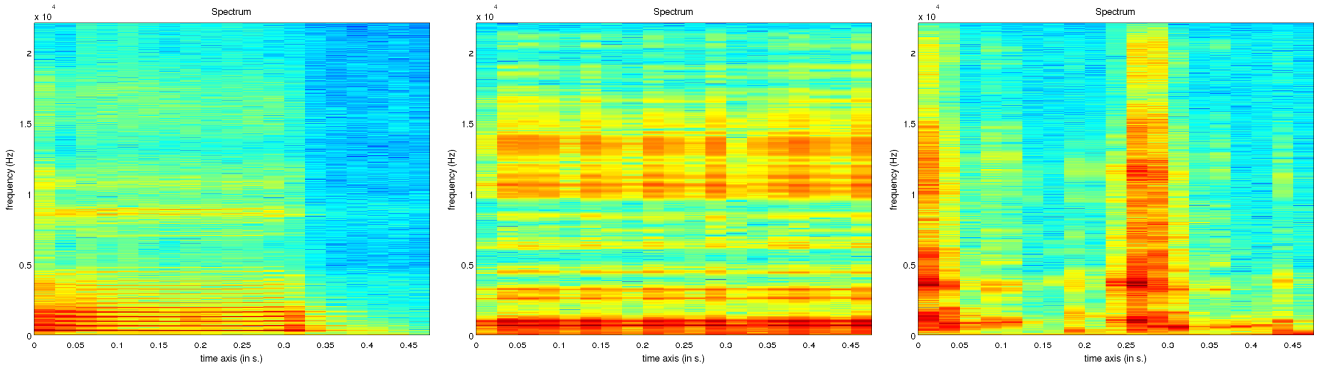

Figure 2: Spectrograms of representatives of clusters 1 (60 elements, red cluster), 2 (42 elements, green cluster) and 3 (50 elements, blue cluster)

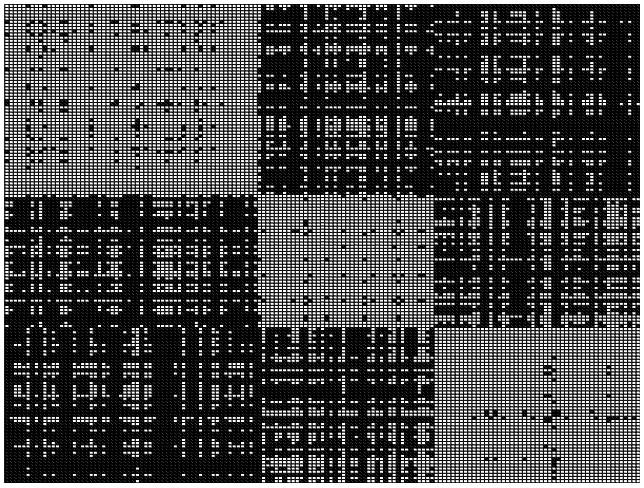

Figure 3: Connectivity of 152 audio segments in the two principal components with k-means clustering

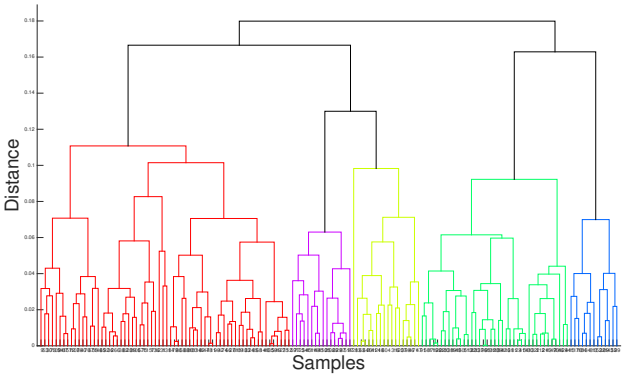

Figure 4: Dendrogram for hierarchical clustering.

If the distances are very small (i.e. all the grouping occurs on the lower part of the dendrogram) it means that we have a good fit of the data. This can be summarized by the Cophenetic correlation coefficient. One of the drawbacks of this approach is that it does not provide "prototypes" for each cluster, thus requiring to separately calculate the barycenter. By cutting the dendrogram at different levels we get different numbers of clusters. In this case, it seems that the most natural cardinalities of clusters are 3, 4, or 7. As a comparison we present in Figure 5 the results of hierarchical clustering with 4 clusters.

Fuzzy C-Means and GMM (Gaussian Mixture Model) can conversely provide a degree of membership for each sound to each of the clusters thus allowing to handle situations where a sound can not be clearly positioned in one of the classes.

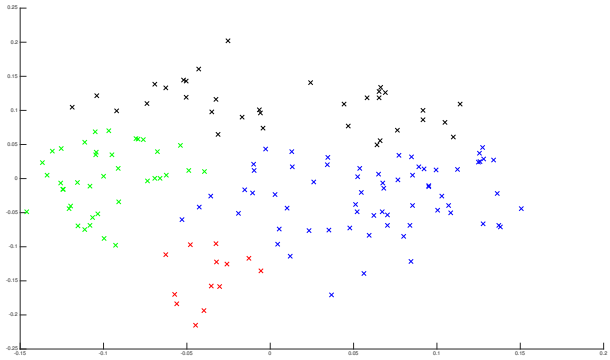

Figure 5: Hierarchical clustering.

5. HOW WOULD A HUMAN DO?

Considering a small (i.e., 3) number of clusters, we asked 26 listeners (15 experts in sound and music computing and 11 naive, 21 male and 5 female, age ranging between 18 and 54 years), not involved in this research, to use a web application to perform the following task: Listen to the three cluster representatives and then 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, as compared to the machine-provided clusters. Subjects showed values of accuracy ranging from 0.40 to 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 (accuracy is 0.65), 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. The mean accuracy for the 26 subjects is 0.50, which is significantly larger than 0.33 (one-tailed t-test, t(25)=13.88, $p < 0.01$). The mean accuracy for the 15 expert subjects is 0.54, while that of the non-experts is 0.47. The difference between the mean accuracies of the two subgroups is small yet significant (one-tailed t-test, t(24)=2.67, $p < 0.01$), thus showing that expert subjects are slightly closer to the machine in labeling sounds according to three prototypes.

5.1 Agreement between subjects

In order to see how human experts agree with each other in the proposed classification task we considered the array of labels (cluster numbers) that each participant assigned to the audio segment. For each of the 325 pairs that could be formed out of the 26 participants, we computed the agreement using the inter-rater agreement statistic (Cohen's Kappa) between the two arrays of assignments. The measured mean agreement is 0.43, which could be labeled as fair-to-moderate. This value is significantly larger than 0.26 (labeled as fair), i.e. the mean agreement between each subject and the machine-provided labeling (one-tailed t-test on two unpaired samples, $t(349) = 4.29$, $p < 0.01$). This gives a measure of how far machine clustering is from the grouping consensus achieved between people.

5.2 Partitioning the sonic space

Having asked 26 participants to label the 152 audio segments by similarity to the 3 prototypes extracted by the automatic clustering procedure, we could collect empirical probabilities for each of the three classes. For each class, we counted the percent number of times that class was chosen for a given audio segment. A probability surface was obtained for each class by K-nearest neighbor regression (with a smoothing of $K = 20$), so that a Bayesian decision could be taken for each point of the plane, simply by choosing the largest of the three probabilities at that point. The resulting regions are portrayed in Figure 6. This partition of the sonic space, as derived from the labeling exercise, can be compared to the distribution of clusters of Figure 1. Some overlap between the green and red regions is apparent. Since these regions are respectively associated with the "glottal fry" and with the "trumpet" vocal prototypes, such region of confusion may be due to sounds with both a rough and a tonal structure.

It is also possible to compare the clustering responses for each single subject and rate them according to internal and external validity indices [8]. External validity indices assume that the true labeling is known (in our case we use the automatic clustering as the baseline) while internal indices, normally used to evaluate different clustering algorithms or different values of parameters (e.g. k in k-means clustering), only exploit the data. In general, we found that the participants who label the audio segments more similarly to automatic clustering (by comparing external indices) are also the ones that score higher in internal indices, thus suggesting that their responses might rely on the features that are exploited in the automatic clustering. As an example the "best" subject scored 2.13 in the Davies-Bouldin (DB) index, while the "worst" scored 10.77[2].

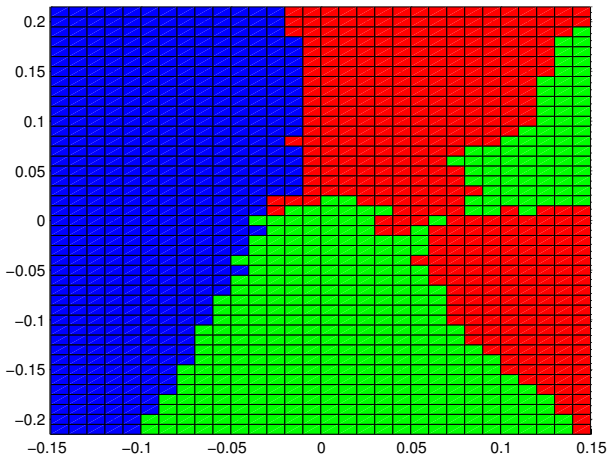

Figure 6: Bayesian subdivision of the sonic space by similarity to three given sound prototypes. Decision boundaries drawn after a labeling exercise with 26 subjects.

5.3 Consensus clustering

Closely related to Bayesian probability is the concept of consensus clustering [8]. It refers to the situation in which a number of different (input) clusterings have been obtained for a particular dataset and it is desired to find a single (consensus) clustering which is a better fit in some sense than the existing clusterings. Consensus clustering for unsupervised learning is analogous to ensemble learning in supervised learning and, interpreted as an optimization problem known as median partition, has been shown to be NP-complete. We based our implementation on the KCC algorithm presented in [21]. Given the formulation of the problem it is interesting to let the algorithm grow a different number of clusters with respect to the original set[3]. The reason behind this approach is that a number of subjects might systematically express a different subdivision of the original data thus highlighting the need for more categories. We present in Figure 7 the same Internal Indices proposed in Section 5.2, for a varying number of clusters, showing that deriving more clusters does not necessarily leads to better results.

[2]DB index is defined as a function of the ratio of the within cluster scatter, to the between cluster separation, a lower value will mean that the clustering is better.

[3]In our case this means to have more than 3 clusters as output.

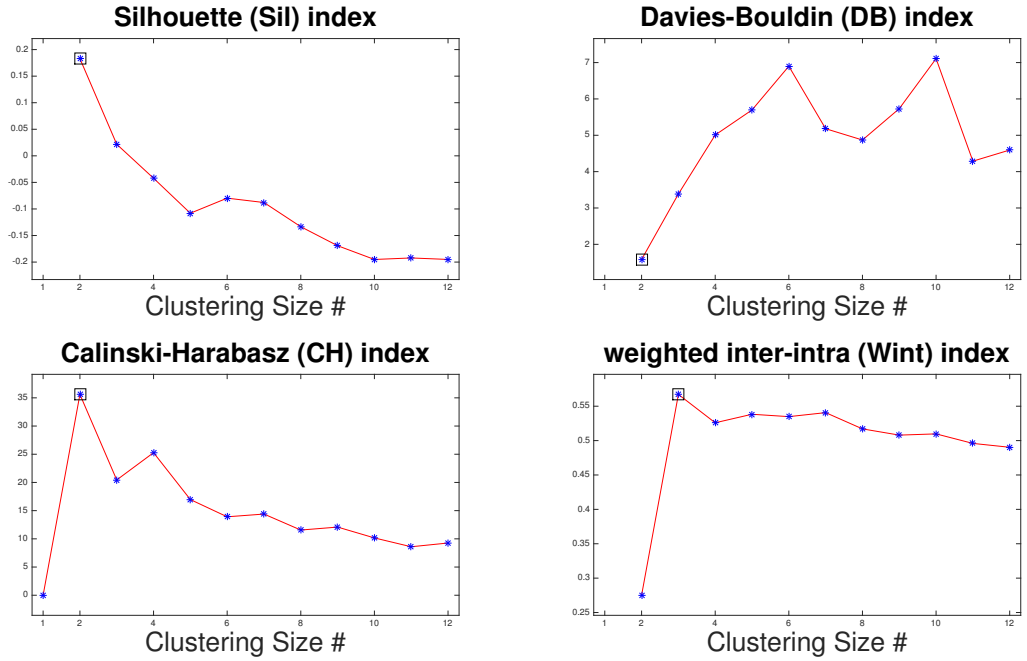

Figure 7: Internal indices for various KCC sizes.

6. CONCLUSIONS AND EXTENSIONS

In this work we have shown how dimensionality reduction can be applied to a set of vocal imitations, each represented by a feature vector, so that the sounds can be automatically arranged on a low-dimensional space and a few representative prototypes can be extracted. The process is based on fairly standard techniques of singular vector decomposition and clustering. 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 (or 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. Prior literature on perceptual grouping have shown that three or four categories can be used to partition the space of everyday sounds, or their imitations, in given contexts. The two-dimensional space is particularly attractive for sound design, because it can be used as a sonic map where a few landmarks are highlighted. We have shown how human subjects tend to partition the two-dimensional space of vocal sounds when they are asked to refer to three automatically extracted prototypes. In future work, we are going to use vocal imitations to access the sonic space of a given sound synthesis model, where landmarks will be associated with both a synthetic sound and its vocal imitation, and new synthetic exemplars could be located on the plane, either by spatially placing them [5] or by new vocalizations.

In this work, relatively little attention has been payed to the quality of descriptors, 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. It is possible that pitch (melodic) profiles, which turned out to be not important for the categorization of environmental sounds [14], may be relevant for a more robust construction of a sonic space of vocal imitations. In any case, the results of section 5 give a bound to the improvements that could possibly be achieved, as they are limited by the agreement that human experts show in assigning labels to sounds.

7. ACKNOWLEDGMENT

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. The authors wish to thank Fabio Pastori for the web interface for the experiment.

8. REFERENCES

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

[2] M. Cartwright and B. Pardo. Vocalsketch: Vocally imitating audio concepts. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, april 2015.

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

[4] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-first*

International Conference on Machine Learning, New York, NY, USA, 2004.

[5] 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.

[6] S. Fasciani and L. Wyse. Mapping the voice for musical control. Technical report, Arts and Creativity Lab, National University of Singapore, 2013.

[7] M. Fernström and E. Brazil. Sonic browsing: An auditory tool for multimedia asset management. In *Proceedings of the International Conference on Auditory Display*, Espoo, Finland, july 2001.

[8] A. Guénoche. Consensus of partitions: a constructive approach. *Advances in data analysis and classification*, 5(3):215–229, 2011.

[9] O. Houix, G. Lemaitre, N. Misdariis, P. Susini, and I. Urdapilleta. A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52–80, 2012.

[10] J. N. Kutz. *Data-driven modeling & scientific computation: methods for complex systems & big data*. Oxford University Press, 2013.

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

[12] G. Lemaitre, A. Dessein, P. Susini, and K. Aura. Vocal imitations and the identification of sound events. *Ecological psychology*, 23(4):267–307, 2011.

[13] G. Lemaitre and D. Rocchesso. On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, 135(2):862–873, 2014.

[14] A. Minard, N. Misdariis, O. Houix, and P. Susini. Catégorisation de sons environnementaux sur la base de profils morphologiques. In *10ème Congrès Français d'Acoustique*, Apr. 2010.

[15] N. Misdariis, A. Minard, P. Susini, G. Lemaitre, S. McAdams, and E. Parizet. Environmental Sound Perception: Metadescription and Modeling Based on Independent Primary Studies. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[16] A. Momeni and D. Wessel. Characterizing and controlling musical material intuitively with geometric models. In *Proceedings of the 2003 Conference on New Interfaces for Musical Expression*, NIME '03, pages 54–62, Singapore, Singapore, 2003. National University of Singapore.

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

[18] 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*, 130(5):2902–2916, 2011.

[19] D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, and P. Boussard. Sketching sound with voice and gesture. *interactions*, 22(1):38–41, Jan. 2015.

[20] G. P. Scavone, S. Lakatos, P. R. Cook, and C. Harbke. Perceptual spaces for sound effects obtained with an interactive similarity rating program. In *Proceedings of International Symposium on Musical Acoustics*, 2001.

[21] J. Wu. K-means based consensus clustering. In *Advances in K-means Clustering*, pages 155–175. Springer, 2012.