OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 64, No. 7/8, July/August 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0026



# Organizing a Sonic Space through Vocal Imitations

DAVIDE ROCCHESSO[1], DAVIDE ANDREA MAURO[1], AND CARLO DRIOLI[2]
(roc@iuav.it) (dmauro@iuav.it) (carlo.drioli@uniud.it)

[1] *Iuav University of Venice, Italy*
[2] *University of Udine, Italy*

A two-dimensional space is proposed for exploration and interactive design in the sonic space of a sound model. A number of reference items, positioned as landmarks in the space, contain both a synthetic sound and its vocal imitation, and the space is geometrically arranged based on the acoustic features of these imitations. The designer may specify new points in the space either by geometric interpolation or by direct vocalization. In order to understand how the vast and complex space of the human voice could be organized in two dimensions, we collected 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 has been experimentally checked how meaningful the resulting clusters are for humans. The procedure of dimensionality reduction and clustering is demonstrated in the case of imitations of engine sounds, giving access to the sonic space of a motor sound model.

## 0 INTRODUCTION

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, similar 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 [4]. Recent research has shown that vocal imitations can be more effective than verbalizations at representing and communicating sounds [16]. Such natural capabilities are being exploited for sound retrieval and synthesis [3]. 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 [25].

In this article we investigate how the vocal mimicking capabilities of humans may be exploited to access and explore a given sonic space. In this context, a sonic space is the space where the sounds produced by a given sound model can be distributed. Essentially, the multidimensional range of possibilities of a sound generator defines its sonic space.

The driving idea of this work is that a collection of exemplar sounds produced by the sound model could be coupled to corresponding vocal imitations, thus providing landmarks to navigate the sonic space. The spatial organization of the exemplars should result in a low-dimensional layout where such landmarks can prompt exploration by vocal imitation or by geometric interpolation.

The construction and use of a two-dimensional space is explained in Sec. 1, where the synthesis-analysis-resynthesis loop involving human imitations is introduced. Since vocal imitations would be used to orient the user in a low-dimensional space of sounds, it is interesting to understand what is the span of the space of vocal imitations at large and to check if humans can make sense of operations of dimensionality reduction and clustering performed in that space. This is the scope of Sec. 2, which investigates how the space of vocal imitations could be arranged and simplified to highlight clusters of sounds that are acoustically similar. 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-subjects consistency is measured and the low-dimensional space is partitioned according to the preferences of participants. Finally, Sec. 3 provides a practical example of two-dimensional spatial organization for a model of engine sounds. A demonstrative interface is illustrated, which starts from a set of synthetic/imitative exemplar couples, and it gives the possibility to access unexplored points of the sound model space by vocal imitation.

Tools for sonic browsing on two dimensions were proposed in the past [9]. The idea of using landmarks to facilitate navigation in the sound design space was explored in the context of parametric sound synthesis [7, 1], and auditory representations were used both to give a visual snapshot to each sound and to compute distances that would

Table 1. The relevant sets and their meanings.

| Set | Meaning |
|---|---|
| $S$ | prototype sounds of the synthesizer |
| $I$ | vocal imitations of sounds in $S$ |
| $B$ | trajectories of features of imitations in $I$ |
| $P$ | trajectories of parameters of synthetic sounds |

allow locating new sounds in the map. As compared to those studies, here 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.

In relevant related work [26], a free-sorting task on 150 non-vocal sound effects produced dissimilarity matrices to train an automatic classifier via multidimensional scaling. Categorization via manual grouping was done for everyday sounds in selected contexts, such as cars [20]. For recorded kitchen sounds [12], the four main categories of solids, electricals, gases, and liquids were found, and they were largely confirmed when subjects were requested to sort imitations of such sounds [15]. The organization of sound material into spatial layouts for performance control was also investigated [21], and mixtures of Gaussians were proposed 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 [8]. The Spiral Discovery Method [5] is an algebraic approach that gives a practical tradeoff between complexity and interpretability that is inevitable when defining a relationship between the parameter space and some perceptual space of reduced dimensionality.

## 1 A PLANE FOR MODEL SOUNDS AND THEIR IMITATIONS

The general problem that we are addressing is that of organizing and accessing the sonic space of a given sound model. This is a situation sound designers are often confronted with when they have to restrict their attention to a certain class of sounds or when they want to exploit a certain sound synthesizer. The entities that we are referring to are found in the sets described in Table 1. Given a sound synthesis model (e.g., a model for motor sounds), this gets represented through a set of samples $S$, or prototype sounds of the synthesizer. Since vocal imitations are better than words at describing sounds [16], we propose to "label" each sample by a vocal imitation. An imitation $i$ is produced for each $s \in S$, thus producing a set $I$ of imitations, having the same cardinality as $S$.

Vocal imitations are complex sound objects that can be described by a set of features, which can be quite numerous. In a screen-based interface, the set $I$ should be projected on the plane by a transformation $\mathcal{T}_I$, that includes a dimensionality reduction on the set $I$. Each $s \in S$ will follow its corresponding $i \in I$ in the projection on the plane.

The sound projection plane will be obviously displayed on the screen where each couple $(i_n, s_n)$ will be highlighted as a graphical element (a dot or a more informative glyph). Simple interactions through conventional input devices can be programmed with these items on the plane. For example, with a two-button mouse the following actions can be programmed:

- For any item on the plane: Left click plays back the imitation $i$; Right click plays back the synthetic sound $s$;
- A click on an empty area of the plane produces a new couple $(i_n, s_n)$, where $s_n$ is obtained by parametric interpolation and immediately played back, and the user is prompted for a new (optional) imitation $i_n$. As soon as such imitation is provided, the couple $(i_n, s_n)$ is moved to the point of the plane that corresponds to $\mathcal{T}_I(i_n)$;
- A new imitation $i_n$ can be produced without looking at the plane. The transformation $\mathcal{T}_I$ is automatically applied so that $i_n$ gets projected on the plane, and $s_n$ is produced by parametric interpolation based on neighbouring samples.

### 1.1 Controlling Sound Models with the Voice

Given a sound synthesis model represented through a set of samples $S$, assume we know the temporal trajectory of model parameters $p$ (a multivariate time series) that produce each sample $s \in S$. Let $P$ be the set of parameter trajectories. We can ask the sound designer to produce a vocal imitation $i$ for each $s \in S$, so that a set $I$ of imitations gets formed.

Assume we have a set of feature extractors that operate in real time on short-term slices of the audio input. Each $i \in I$ produces the temporal trajectory $b$ of a feature vector. The features extracted from imitations can be used to control the parameters of sound synthesis. Each model parameter is a hypersurface on the space of imitation features, and each sample is associated to a trajectory on that surface. If, for a given parameter $p_j$, we interpolate (or regress) a surface from a set of trajectories we get a map $p_j = m(b)$.

In this way, it would be possible to recreate the synthetic sounds of $S$ by using the imitations $I$ as control signals. The whole chain resembles a classic synthesis-analysis-resynthesis loop, with the human imitator as a within-loop active agent:

$$S \overset{imitation}{\longrightarrow} I \overset{feature-extr.}{\longrightarrow} B \overset{mapping}{\longrightarrow} P \overset{resynthesis}{\longrightarrow} S \qquad (1)$$

In the chain 1, the set of imitations $I$ is derived from the corresponding model sounds $S$ (for whom we know the parameter trajectories $P$), feature trajectories $B$ (the set of all elements $b$) are extracted from each imitation, and a mapping from vocal audio features to model parameters is derived by interpolation or regression given the couples $(b, p)$.

At the center of chain 1 there is the human individual, with her preferences, limitations, and idiosyncrasies. This

makes the couples ($b$, $p$) highly subjective but ensures the highest level of embodyment of the sonic space, as its structure is directly organized as an egocentric frame of reference [22].

It should be noted that the vocal features are "instantaneous" and so are the parameters. This does not mean that there is no memory in the sound synthesis process. In fact, the sound model may be time- or state-dependent. For example, friction may sound different depending on how we reach a given configuration of parameters.

Sec. 3 presents an example realization of the proposed interactive sonic space for a motor sound model. Beforehand, we need to show how the projection may actually work and how the space distribution of vocal samples is representative of human thinking in sound.

## 2 THE VOCAL SONIC SPACE

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. 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 section 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.

When imitating everyday sounds, humans often use vocal mechanisms that are never found in spoken languages [11], and the construction of a vocal sonic space can only start from the collection of a set of significant examples. A database of 152 audio segments were manually extracted from the Fred Newman's repertory of vocal imitations described in his book [23] 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 was chosen in order to try to accommodate for different vocal 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, for the purposes outlined in Sec. 1, we would like to be able to browse collections of vocal samples organized on a two-dimensional surface.

### 2.1 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) [13].

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 [7]. 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 [24]. 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 [14] 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 [24]) of spectral flux, centroid, roughness, flatness, entropy, skewness, and RMS energy computed over 18 windows spanning the 500 ms-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 100 ms or 300 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 a number of templates. The prototypical temporal envelopes are: 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. [19], 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 Sec. 4. However, in this study the organization of the sonic space, the extraction of prototype sounds, and the subjective tests are voice agnostic. Table 2 lists the features used in this 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).

Table 2. The eighteen features considered in the study.

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

Before the extraction of principal components, the mean is subtracted from the distribution of each feature, and the distribution is further normalized to range betweenn $-1$ 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:

$$B = USV'. \tag{2}$$

$S \in \mathbb{R}^{f\times f}$ 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 2, 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 [9], even though they explain less than one third of the variance for this set of sounds. In fact, the slow decay of the singular values (diag$(S) = [16.9, 12.8, 10.7, 8.7, 7.2, 6.5, 5.3, 4.5, 4.0, 3.3, 3.1\ldots]$) shows that there is no obvious cutoff point for dimensionality reduction.

## 2.2 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 [6]. 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. Fig. 1 displays the clusters of 60 (red), 42 (green), and 50 (blue) elements extracted with k-means clustering, 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 of spectral 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 Fig. 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 item as a cluster representative. The spectrograms of such representatives are depicted in Fig. 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. The three prototype vocal imitations are obtained by distinct source types, such as vocal-fold phonation, glottalic myoelastic oscillation, and tongue percussion [11]. Given

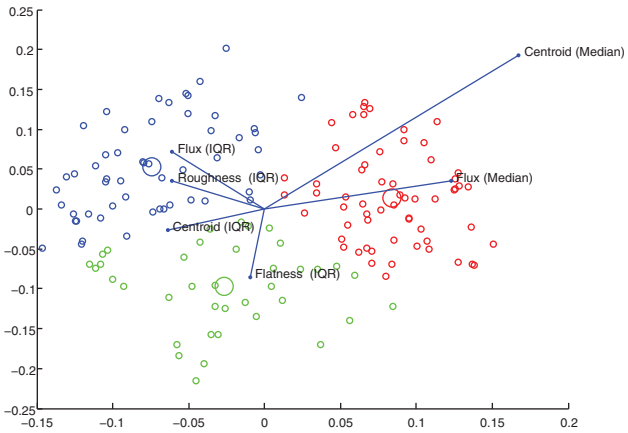



Fig. 1. Three clusters in the space of the two principal components.

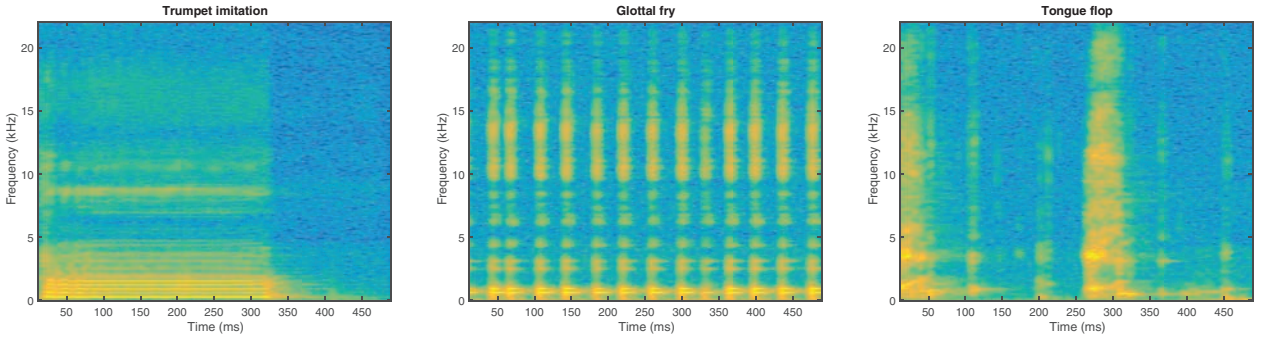



Fig. 2. Spectrograms of representatives of clusters 1 (60 elements, red cluster), 2 (42 elements, green cluster) and 3 (50 elements, blue cluster).

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 mostly contains sounds that are continuous and noisy. Finally the third (blue) cluster encompasses sounds that are mostly 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 [20]. 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 [12, 15].

Even more meaningful is connectivity analysis [6], which looks for diagonal blocks in the matrix $U_2U_2'$, where $U_2$ are the first two columns of $U$, with their rows sorted according to the extracted clusters. In this clustering the degree of connectivity is $c = 0.65$, i.e., 65% of the active cells belong to the three blocks on the diagonal of $U_2U_2'$, thus showing strong connection within clusters.

#### 2.2.1 Different Clustering Techniques

It is possible to replace the k-means clustering with other different techniques (see [18] for details) that enable us to highlight different perspectives on data. We report connectivity $c$ for hierarchical, Fuzzy C-means, and GMM clustering:

- Hierarchical: $c = 0.58$;
- Fuzzy C-Means: $c = 0.64$;
- GMM: $c = 0.57$.

With hierarchical clustering we can plot the dendrogram for linkage, which can give some insight on the nature of grouping, but cluster "prototypes" can be obtained only by separate computation of the barycenter. 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 cannot be clearly positioned in one of the classes.

### 2.3 How Would a Human Do?

Having shown how a machine can distribute and cluster voice samples on a plane, it is interesting to see if and how humans agree on grouping the samples around prototypes. 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 3 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 (subject 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.

#### 2.3.1 Agreement between Subjects

In order to see how humans 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

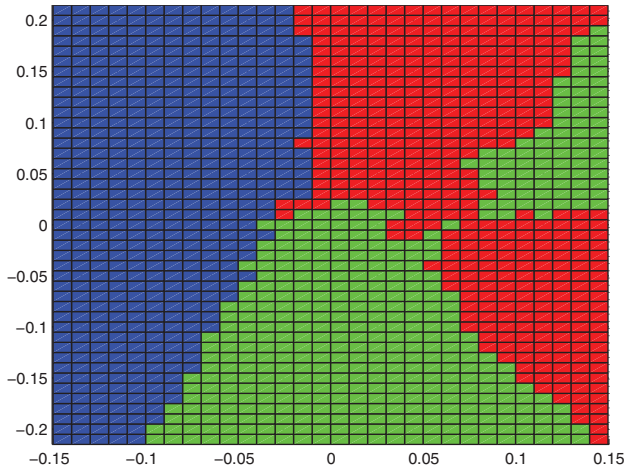



Fig. 3. Bayesian subdivision of the sonic space by similarity to three given sound prototypes. Decision boundaries drawn after a labeling exercise with 26 subjects.

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.

### 2.4 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 3 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 Fig. 3. This partition of the sonic space, as derived from the labeling exercise, can be compared to the distribution of clusters of Fig. 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 [10]. 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[1].

Closely related to Bayesian probability is the concept of consensus clustering [10]. 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 that 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. Based on an implementation of the KCC algorithm [27], we let the algorithm grow the number of clusters to see if a number of subjects systematically expressed a different subdivision of the original data, thus highlighting the need for more categories. Indeed, the analysis of four different internal indices showed that deriving more clusters does not necessarily lead to better results (see [18] for details).

## 3 EXAMPLE: MOTOR SOUNDS

In this section we illustrate a practical realization of the plane described in Sec. 1, where the samples $S$ come from a physics-based model of motor sounds [2]. This is an engine sound simulator able to effectively reproduce a wide range of four-stroke engine sounds. It can be configured and controlled in terms of physically informed high level parameters, including the number of cylinders, the size of various components (muffler, main pipes, inlet, outlet), and the revolutions per minute of the crankshaft (RPM). The implementation of the synthesis model is available as a Cycling'74 Max external, allowing for real-time control through a set of parameters. In the following, we use a subset of seven representative parameters.

To access and interact with the sonic space we collected the set $I$ of imitations corresponding to the synthetic samples, and a graphical user front-end (GUI) was designed using the Matlab framework, making it able to communicate with Max externals through the OSC protocol. Its main frame represents the 2D sonic space where a collection of sounds is organized by dimensionality reduction and 2D projection from the corresponding vocal imitations, as described in Sec. 2.

Moreover, for each reference sound the corresponding trajectory of model parameters $p$ is known and stored, and new synthetic sounds can be created in the 2D space by feeding the synthesizer with a mixture of the parameters of the reference sounds. This is achieved in our implementation by an interpolation scheme based on the Delaunay triangulation on the reference sound positions [1]. The interpolation is first performed in the Matlab client, the new parameters are then sent to the Max synthesis server, and finally the new sound is loaded and represented in the Matlab client GUI.

[1] 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.

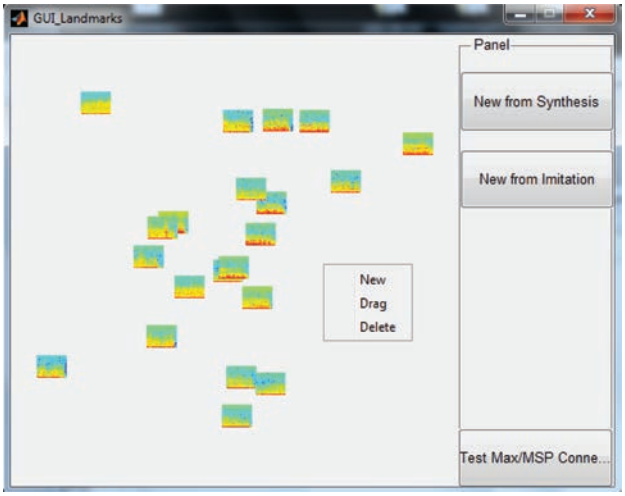



Fig. 4. The Matlab GUI and the organization of the set of engine sound imitations on the 2D projection space.

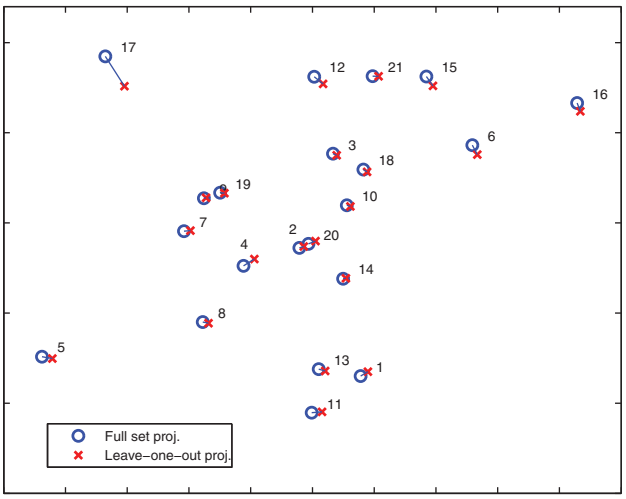



Fig. 5. The 2D projection performed on the whole imitation data set (circle markers), and the position of each imitation sample $i_n$, $n = 1...21$, when mapped through the projection matrix computed on the set $I - \{i_n\}$ (cross markers).

Given these tools at hand, the construction and population of the sonic space can be described as follows:

1. A set $S$ of 21 reference sounds is created using the synthesis model;
2. A set $I$ of vocal imitations is collected by imitating the 21 reference synthesis sounds;
3. The dimensionality reduction is performed on the set of imitation sounds, and their projection in the 2D space is rendered in the GUI;
4. New synthesis sounds are created by selecting new points in the sonic space (thus triggering the generation of new parameters by interpolation), or by providing new imitations, which are projected in the sonic space and are used to generate the new synthesis parameters by interpolation.

The dimensionality reduction is based on the preliminary feature extraction discussed in Sec. 2. As a result of steps 1–3, the 21 imitation sounds are projected in the 2D sonic space, forming a set of reference imitation sounds, each of which is linked to a synthesis sound and its parameter set. Fig. 4 shows the Matlab GUI and the projection of the 21 engine sound imitations on the 2D sonic space. Our choice here was to graphically represent each sound in $S$ by its spectrogram.

To test the 2D projection operator based on the acoustic cues of choice, a leave-one-out cross-validation test was performed. This is a procedure which, for every test element (i.e., one of the $N$ samples in the set) computes its position using the projection model established from the $N - 1$ other elements. The comparison with the positions given by the reference projection model, computed by using the whole $N$-samples data set, provides a clue of the consistency of the projection model with respect to a new imitation. Fig. 5 illustrates how the majority of the test samples gets projected consistently with the full-set case.This test tells us that if we compute a projection model on a given imitation set and then acquire a new imitation, its position will be consistent with the projections provided by the model built using this sample as well.

It would be interesting to perform a similar consistency check on the model sounds, to see if a new sound, synthesized by interpolation on three neighboring landmarks, turns out to be closer to those synthetic neighbors than to any other landmark. Such test turns out to be deceptive, for the following reasons: (1) there are no guarantees and no criterion ensuring that each imitation is closer to its reference model sound than to any other; (2) if the same dimensionality reduction process would have been applied to reference model sounds instead of imitations, their distribution on the plane would have been quite different; (3) the construction of the nonlinear multivariate $B \rightarrow P$ mapping, as part of the proposed loop (Eq. 1), is a challenging task which is here shortcut by local linear interpolation. This third point, in particular, requires further exploration in future work. In any case, if the new synthetic sound is found to follow its imitated counterpart as soon as it becomes available, the problem of positioning it on the plane is overcome, as its location will be computed using the imitation, and such new position has been shown to be consistent in relation to the other imitations.

## 4 CONCLUSIONS

We have proposed to represent the sonic space of a sound model as a plane where a number of prototype synthetic sounds are positioned based on vocal imitations. Each sound landmark is a couple synthesis/imitation, and its spatial organization is based on dimensionality reduction on the set of available imitations, each represented by a high-dimensional feature vector. The process is based on fairly standard techniques of singular vector decomposition and clustering.

To check the consistency and feasibility of such imitation-based space organization, we performed dimensionality reduction and clustering on a large range of vocal productions. We found that the prototype sounds (or cluster representatives) are perceptually distinct from each other, and they may well serve the purpose of landmarks in the space of vocal imitations. 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. This experiment gives us a measure of how meaningful the machine distribution and grouping of vocal sounds are to humans, and it confirms two facts: (i) that humans are able to effectively use the acoustic and articulatory cues at their disposal to associate sounds to given prototypes; and (ii) that the acoustic/articulatory cues used in the automatic clustering process are sufficiently consistent with the cues used by humans to categorize acoustic phenomena.

A plane is preferred to higher-dimensional spaces because it is more easily accessed, navigated, and organized. Even if the percentage of variance explained by the first two components is rather low, the automatic clustering is consistent with human clustering behaviors found in the literature and with results of the labeling experiment. Nonetheless, the limited descriptive capacity of the two dimensions does limit the consistency of the interpolation/resynthesis step of the Synthesis-Imitation-Resynthesis model.

As an example application, we used vocal imitations to access the sonic space of a motor sound synthesis model. Here, landmarks are associated with both a synthetic sound and its vocal imitation, and new synthetic exemplars can be positioned on the plane, either by spatially placing them 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 [19], may be relevant for a more robust construction of a sonic space of vocal imitations. Marchetto and Peeters [17] developed some audio descriptors that capture the morphological aspects of sounds and that proved to be effective to recover categories of vocal imitations. In any case, the results of Sec. 2.3 give a boundary to the improvements that could possibly be achieved, as they are limited by the agreement that human experts show in assigning labels to sounds.

## 5 ACKNOWLEDGMENT

The authors DR and DAM have been pursuing this research as part of the project SkAT-VG and acknowledge the financial support of the Future and Emerging Technologies (FET) program 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 described in Sec. 2.3.

## 6 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*, Graz, Austria (2010).

[2] S. Baldan, H. Lachambre, S. D. Monache, and P. Boussard "Physically Informed Car Engine Sound Synthesis for Virtual and Augmented Environments," in *Proc. 2nd Workshop on Sonic Interactions in Virtual Environments (within IEEE VR 2015)*, Arles, France (2015).

[3] M. Cartwright and B. Pardo "Vocalsketch: Vocally Imitating Audio Concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 43–46, New York, NY, USA (2015).

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

[5] Á. B. Csapó and P. Z. Barony "The Spiral Discovery Method: An Interpretable Tuning Model for Coginfocom Channels," *J. Advan. Computational Intel. & Intelligent Informatics*, vol. 16, no. 2, pp. 358–367 (2012).

[6] 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).

[7] 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*, Porto, Portugal (2009).

[8] S. Fasciani and L. Wyse "Mapping the Voice for Musical Control," Technical Report, Arts and Creativity Lab, National University of Singapore (2013).

[9] 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 (2001 July).

[10] A. Guénoche "Consensus of Partitions: A Constructive Approach," *Advances in Data Analysis and Classification*, vol. 5, no. 3, pp. 215–229 (2011).

[11] P. Helgason "Sound Initiation and Source Types in Human Imitations of Sounds," in *Proceedings of FONETIK*, pp. 83–88, Stockholm, Sweden (2014).

[12] O. Houix, G. Lemaitre, N. Misdariis, P. Susini, and I. Urdapilleta "A Lexical Analysis of Environmental Sound Categories," *J. Experimental Psychology: Applied*, vol. 18, no. 1, pp. 52–80 (2012).

[13] J. N. Kutz *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems and Big Data* (Oxford University Press, 2013).

[14] O. Lartillot and P. Toiviainen "A Matlab Toolbox for Musical Feature Extraction from Audio," in

*Proceedings of the International Conference on Digital Audio Effects*, pp. 237–244, Bordeaux, France (2007).

[15] G. Lemaitre, A. Dessein, P. Susini, and K. Aura "Vocal Imitations and the Identification of Sound Events," *Ecological Psychology*, vol. 23, no. 4, pp. 267–307 (2011).

[16] G. Lemaitre and D. Rocchesso "On the Effectiveness of Vocal Imitations and Verbal Descriptions of Sounds," *J. Acous. Soc. Amer.*, vol. 135, no. 2, pp. 862–873 (2014).

[17] E. Marchetto and G. Peeters "A Set of Audio Features for the Morphological Description of Vocal Imitations," in *Proceedings of the International Conference on Digital Audio Effects*, Trondheim, Norway (2015).

[18] D. A. Mauro and D. Rocchesso "Analyzing and Organizing the Sonic Space of Vocal Imitations," in *Proceedings of the Audio Mostly Conference on Interaction With Sound*, AM '15, pp. 23:1–23:7, New York, NY, USA (2015).

[19] 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* (2010 Apr.).

[20] 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 J. Audio, Speech, and Music Processing* (2010).

[21] 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, pp. 54–62, Singapore, Singapore (2003).

[22] N. Navolio, G. J. Lemaitre, A. Forget, and L. M. Heller "The Egocentric Nature of Action-Sound Associations," *Frontiers in Psychology*, vol. 7, no. 231 (2016).

[23] F. Newman *MouthSounds: How to Whistle, Pop, Boing, and Honk... For All Occasions and Then Some* (Workman Publishing, 2004).

[24] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals," *J. Acous. Soc. Amer.*, vol. 130, no. 5, pp. 2902–2916 (2011).

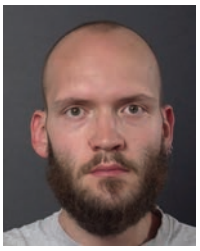
[25] D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, and P. Boussard "Sketching Sound with Voice and Gesture," *Interactions*, vol. 22, no. 1, pp. 38–41 (2015 Jan.).

[26] 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).

[27] J. Wu *Advances in K-means Clustering: A Data Mining Thinking* (Springer, 2012).

## THE AUTHORS


Davide Rocchesso


Davide Andrea Mauro


Carlo Drioli

Davide Rocchesso received the Laurea in Ingegneria Elettronica degree from the University of Padova, Italy in 1992, and the Ph.D. degree from the same university in 1996. In 1994 and 1995 he was visiting scholar at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University. Between 1998 and 2006 he was with the Computer Science Department at the University of Verona, Italy, as an Assistant and Associate Professor. Since 2006 he has been with the Iuav University of Venice, as Associate Professor. He is member of the Faculty Committee of the Ph.D. program in computer science of the University Ca' Foscari in Venice. Rocchesso has been the coordinator of project IST 2000-25287 (SOb - the Sounding Object), and local coordinator of the project NEST 29085 (CLOSED - Closing the Loop Of Sound Evaluation and Design), and of the Coordination Action IST-FET 2004-03773 ($S2S^2$ - Sound-to-Sense; Sense-to-Sound). He has been the Chair of the COST Action IC-0601 (SID - Sonic Interaction Design) and he is currently coordinating the project FP7-ICT-2013-C FET 618067 (SkAT-VG - Sketching Audio Technologies using Vocalizations and Gestures). He is Associate Editor for the *International Journal of Human-Computer Studies*.

•

Davide Andrea Mauro is a researcher in music informatics, 3D audio, recording and reproduction techniques. His Ph.D. thesis involved the implementation of real-time spatialization systems combined with head-tracking techniques. He also carried out research for hearing aids and cooperated in a study for assessing tinnitus in normal-hearing population. Mauro developed IEEE-1599 tools for transcription and encoding of live performances in sub-symbolic formats. In 2014–2105 Mauro worked as a research assistant at the Iuav University of Venice, as part of the SkAT-VG project.

•

Carlo Drioli is an assistant professor at the Department of Mathematics and Computer Science of the University of Udine. He received the Laurea degree in electronic engineering and the Ph.D. degree in electronic and telecommunications engineering from the University of Padova, in 1996 and 2003, respectively. From 1996 to 2001, he has been a researcher with the Centro di Sonologia Computazionale (CSC) of the University of Padova in the field of sound and voice analysis and processing. From 2001 to 2002, he was a visiting researcher at the Royal Institute of Technology (KTH), Stockholm, Sweden, with the support of the European Community through a Marie Curie Fellowship. In 2003–2004 he was with the Department of Phonetics and Dialectology of the Institute of Cognitive Sciences and Technology of the Italian National Research Council (ISTC-CNR), where he pursued research on voice processing and emotional speech synthesis. From 2005 to 2011 he joined the University of Verona, Department of Computer Science, as a research assistant and adjunct professor. His current research interests are in the fields of multimedia signal processing, sound and voice coding by means of physical modeling, speech analysis and synthesis. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), of the Acoustical Society of America (ASA), and of the International Speech Communication Association (ISCA).