SOUND RETRIEVAL FROM VOICE IMITATION QUERIES IN COLLABORATIVE DATABASES

DAVID S. BLANCAS[1], JORDI JANER[2]

[1] *Audiovisual Systems Engineering Degree, Universitat Pompeu Fabra, Barcelona, Spain*
davidsanchezblancas@gmail.com
[2] *Music Technology Group (MTG), Universitat Pompeu Fabra, Barcelona, Spain*
jordi.janer@upf.edu

The retrieval of sounds in large databases can benefit from novel paradigms that exploit human-computer interaction. This work introduces the use of non-speech voice imitations as input queries in a large user-contributed sound repository. We address first the analysis of the human voice properties when imitating sounds. Second, we study the automatic classification of voice imitations in clusters by means of user experiments. Finally, we present an evaluation that demonstrates the potential of a content-based classification using voice imitations as input queries. Future perspectives for using voice interfaces in sound assets retrieval are exposed.

INTRODUCTION

Collaborative databases are online libraries whose content is uploaded by their registered users. Once these repositories get popular, they become a huge file structure that must be organized to have fast access to desired resources. For instance, manual tagging, or folksonomy-based tagging [1] has been the most adopted technique to organize data, and some tools have been developed to minimize the error that allowing users to choose tags produces. Some studies such as [2] are based on automatic classification of sounds in a user-contributed library using the signal content to overcome these drawbacks. The collaborative repository of *freesound.org* [3] has tools that guide users to find or discover new sounds based on a similarity measure obtained by using multidimensional distances of acoustic feature vectors. This paper is motivated by the quest of novel user interaction paradigms for sound content retrieval in large collaborative databases. In particular, voice imitations are here used to enhance text queries providing a more direct control to the user.

The first goal of this study is to analyse the signal description of human voice imitating a set of sound examples. This description is obtained from low-level signal features, and they are supposed to classify new input sounds within a fixed taxonomy. Besides the information that could be obtained from signals, this work analyses the human behaviour when imitating different types of sounds. Imitating sounds has always been a way to describe sounds in a way that listeners or readers could understand an acoustic event or its source. Humans are not used to produce pure imitations, and commonly employ onomatopoeias instead.

Onomatopoeias are defined as words that try to imitate or represent sounds in a specific language, so they are culture-related and not necessarily linked to the actual acoustic content of sounds. Furthermore, the variety of sounds of a given category (e.g. cat) cannot be distinguished by a single onomatopoeia (e.g. "meow"), while at the same time users shall not be restricted to only search sounds that have a clear and known language representation. Our goal is based on building imitations models, whose robustness will be evaluated. A prototype that uses the *freesound.org* API [4] will be presented to show the effectiveness of this method.

To our knowledge, there are not many references in the literature about the description of this type of vocalizations, as imitations have been treated as onomatopoeias in most of the cases. In fact, sound signals generated by humans have been mostly studied to build voice models for speech-to-text or text-to-speech applications. These studies are based on understanding signals that represent phonetics, as they are the roots of communication and language [5][6]. However the use of voice can go beyond this idea. In fact, the aim of this work is to give another perspective to the capabilities of the human voice. In this direction, the ability of humans to discriminate and classify sounds gathered from previous recorded non-speech sound imitations has already been studied [7], but in this report this classification will be made automatically.

On the other hand, onomatopoeias have been also used for audio retrieval purposes applied to music [8], although the use of spoken queries has a critical culture drawback. Applications where non-speech voice is the interaction controller, also for musical purposes, have been developed to demonstrate the voice potential in instruments imitation [9].

Finally, in [10] a study about non-speech voice imitations introduces the idea of observing the relation between human discrimination of imitations and machine learning algorithms classification. As there is a close relation between imitations and original sounds, and they used a general taxonomy, - solid, liquid, gas, electric-, applicable results to develop further development or applications where imitations are one of the interaction tools were not obtained. In our work, we aim to fill the gap between voice interaction and sound retrieval. Fig. 1 shows all steps involved in our system, and details are explained throughout the following sections.

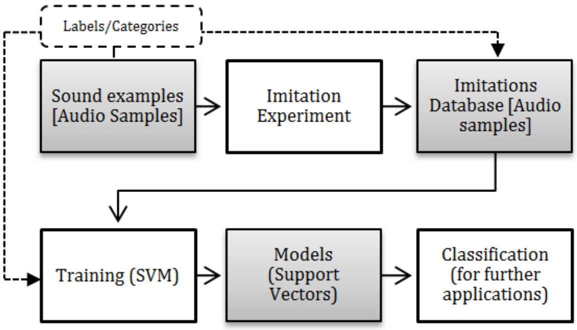
Labels/Categories
Sound examples [Audio Samples]
Imitation Experiment
Imitations Database [Audio samples]
Training (SVM)
Models (Support Vectors)
Classification (for further applications)

Figure 1: Workflow of the content-based classification of voice imitations.

1 DESCRIPTION OF VOICE IMITATIONS

The purpose of this section is to understand the semantics that can be found on voice imitations using low-level signal descriptors and their relation with the original imitated sounds.

1.1 Analysing voice imitations

First of all, we reviewed the classification experiment described in [10]. They provide 72 recorded sounds from humans imitating short acoustic references (not more than five seconds of duration) from kitchen gadgets, within a classification of solid-liquid-gas-electric. In a first approach the process of extracting descriptors was replicated, choosing an algorithm to build a machine learning structure and observing the classification results in a test environment.

1.1.1 Feature Extraction

The Timbre Toolbox [11] was used to extract 472 signal features, including global descriptors and statistics of time varying descriptors from temporal and spectral domains. There are two main assumptions before using features from imitations signals generated by users, which might have direct impact on this study:

- Some descriptors can be related to the phonetics in human speech, as different phonemes are likely to be used in the imitation process. The use of phonemes only allows imitating a limited range of sounds, and it depends on the language knowledge of the user. For that reason, users are expected to use any vocal sound that they are able to produce.
- There is a relation between the content of original and imitation sounds, especially in temporal features, as in most of the cases users try to imitate the exact temporal structure of the sound. This can be taken into account when selecting the sounds to imitate, ensuring other differentiations between sounds.

1.1.2 Automatic Classification of Voice Imitations

We use the Weka toolbox [12] to study the performance of the descriptors on various machine learning algorithms. For this purpose, sounds from [10] were manually annotated using the taxonomy solid – liquid – gas – electric, and two classification algorithms were tested to observe the accuracy.

Weka also offers the possibility to perform feature selection. The results improve dramatically when using the features chosen by the Correlation-based Feature Selection algorithm [13] (see Table 1). In addition, a 10-fold Cross Validation method was used to have several test and training datasets from a very rich amount of signals, both for the feature selection and the classification process. Support Vector Machine (SVM) was the chosen classification algorithm because of its higher accuracy.

Algorithm	*Accuracy 472 features*	*Accuracy 31 features*
SVM	77.78%	88.89%
Naïve Bayes	65.28%	80.56%

Table 1: Classification accuracy of sounds in [10] using Support Vector Machine (SVM) and Naïve Bayes algorithms.

1.2 Relation between original and imitated sounds

We observed that the classification results are quite similar to the ones obtained in [10], which seems sufficient for such a classification problem of kitchen sounds. Nevertheless our approach targets the retrieval of a wider scope of sounds in a user-contributed database as *freesound.org*. In our approach, a text query determine the sound category, and provides an initial filtering of retrieved sound by tags. The term **category** in the context of voice imitations is introduced to be used in the whole study as the conceptual or semantic connector for a group of sounds, which will be our proposed filter.

1.2.1 Categories and Sound Sources

The timbre characteristics of a given sound might be hard to imitate with our voice. This is visible when two imitations are similar but the original sounds are acoustically different (e.g. impact sounds). In this study, the semantic context is narrowed to a predefined category in order to reduce these types of errors. A category can be related to an entity or source that produces sounds. Some variables are defined before acquiring the corresponding imitations (see Fig. 2 for an interpretation):

- Users do not have the same imitation skills. The experience includes knowing several languages with their associated phonemes, but also the ability to perform non-phonetic sounds.
- The importance of knowing the source is addressed to relate the mechanism that produces specific acoustic events. Also the experience of listening to this source can be important to perform a closer imitation to the original sounds.
- Imitations must not be directly similar to the original sound but to the concept of the sound production. The use of the acoustic resources that can be produced with the voice can be enough to understand a selection of sounds inside a category and this is something that users are expected to use unconsciously.

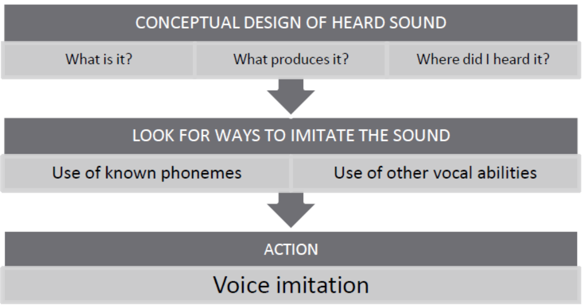
CONCEPTUAL DESIGN OF HEARD SOUND
What is it?
What produces it?
Where did I heard it?
LOOK FOR WAYS TO IMITATE THE SOUND
Use of known phonemes
Use of other vocal abilities
ACTION
Voice imitation

Figure 2: Voice Imitation Process.

2 SOUND IMITATION USER EXPERIMENT

One of the main objectives of this experiment is to study the user behavior when producing imitations of several levels of difficulty. It shall confirm the human capabilities to produce several sounds and the possibility of developing a real application that benefits from voice imitation queries. As this study was a proof of concept, reference sounds were given to the experiment subjects, as at first we wanted to know their response and behavior when facing different types of sounds to imitate.

2.1 Selection of Categories and Sounds

We chose a reduced set of categories from the collaborative database of *freesound.org*. The selection of categories (See Table 2) was made following this criterion:

- They represent a variety of acoustic categories, including tonal and noisy sounds: animal voice, mechanics, electronics and impacts.
- Availability of multiple examples of sounds for each category in *freesound.org*.
- Sounds within a category represent several concepts. In some cases the conceptual difference is more important than the acoustic difference.

Extending the number of classes would require recording a larger number of vocal queries, with longer user sessions or a larger number of participants, which was beyond the scope of this work.

Category	*Sounds / Subcategory*
Cat	Meowing, purring, yelling
Dog	Barking, whining, breathing
Car	Hand brake, door closing, horn, ignition
Drums	Bass drum, crash cymbal, hi-hat, snare

Table 2: Categories and sounds to be imitated

2.2 Experiment Protocol

17 subjects of different sex and age participated in the experiment sessions. They were given indications to know the steps they had to follow. An assistant supervised the experiment to observe and analyze the interaction and behavior. Users had to follow a number of steps in front of a Graphical User Interface containing buttons to play the sounds (see Table 2), and to record each sound with a limited duration of 5 seconds. There were three main tasks:

- Read the instructions and understand the use of imitations instead of onomatopoeias.
- Listen to each sound to understand the concept and the source. A trial imitation was not required but recommended.
- Record the imitation for each single sound. The recordings could be repeated.

It is important to have into account that imitating given sounds would not really produce robust models in a real scenario for all categories, as new input queries would pretend to be similar to one of these models. As the approach of this study is to understand the interaction using voice imitations, we propose for further research the generation of models for a concept (e.g. dog barking, not the specific sound of dog barking taken from *freesound.org*), involving a more complex and detailed work on the machine learning part.

2.3 Discussion

Users enjoyed the experience of imitating sounds as they were involved in a novel experiment, combining sound perception and production skills. For this reason, there was an extra motivation to improve the imitations the better they could; as a consequence, a powerful database of sound imitations was created (sounds are available online at [14]). There are some remarks to make about the sounds and user behavior:

- Cats and dogs sounds were easy to imitate. It is obvious if a relation between animals sound systems is made, with a quite similar sound production by using lungs air to make the vocal folds vibrate. In addition, users were asked if they had pets in their homes, as listening to the animals constantly can make the imitation process easily; this affirmation was true for the users who had a dog.
- It is interesting to remark that users tended to use onomatopoeias at the beginning of the imitation, but they realized the duty and importance of imitating the sound as it was heard. In most of the cases users concluded that imitations can be extremely more useful than onomatopoeias to represent sounds.
- In general, pitched sounds were easier to imitate. In fact, humans use tonal variations most of the time when talking or producing any sound. That is why car and drums classes were much more difficult to imitate, as the noisy constitution of the signal complicated the relation to the sound source.
- The age difference among users was also significant. For instance, younger (less than 20 years old) and older (more than 50 years old) users made the best imitations so far. In the first case, the easy capacity to learn makes the ability of imitating a potential tool for experiment; in the second case, there was less concern to produce perfect sounds, so the first idea of representing a specific sound that came to their minds was acted out. In fact, the experiment has demonstrated that imitating sounds without spending too much time in thinking how to do it can be useful because the imitation is more natural.

3 A PROTOTYPE FOR VOICE IMITATION QUERIES

3.1 Model Training

We obtained a dataset of tagged imitations from the experiment described in the previous section. Next we carried out feature selection for all categories using the Correlation-based Feature Selection algorithm. SVM was used as the classification algorithm. We evaluated the performance of the automatic classification considering 10-fold cross validation (see Table 3 for the results).

Category	*Accuracy*
Cat	90.20%
Dog	100%
Car	89.71%
Drums	77.49%

Table 3: Classification Results for Recorded Imitations

As it was expected, descriptors that explain the spectral content are the ones that have more importance in all categories. This fact explains that users have the ability to produce several types of sounds with their voice, producing complex spectral shapes, even when using non-phonetic sounds; thus, features like the spectral crest or spectral variation appear in every output. Next we discuss the results of the evaluation for the four categories in our study.

Cat: As the third sound of the cat was the most difficult to imitate (yelling), it is the one that has more classification errors, - more types of imitations were performed -, and also there are false positives with the first sound, as they are quite similar in tonal terms. The second sound is correctly classified for each trial, as the purring sound is far from the other two sounds.

Dog: The three chosen sounds of this category were acoustically and conceptually very different, so the imitations are also well classified, obtaining the best percentage of accuracy, i.e. users understood the semantic, source and action perfectly.

Car: Despite of being a category with difficult sounds to imitate, the results are quite good. There are only wrong classifications between sounds that had similar acoustic features (first two sounds correspond to impact sounds, the other two sounds to a more tonal structure).

Drums: This category obtained the lowest accuracy. It is mainly caused by the similarity between the imitations, as they form a huge variety of impacts. The variety of frequencies used to imitate the sounds but also the duration of the impacts have determined these errors.

The classification results shown in Table 3 are supposedly enough to classify correctly most of the new queries in our prototype. As we mentioned before, the selection of the categories is critical when designing such a search paradigm from vocal imitations queries.

3.2 Integration in an Online Database

At this point, we computed models for each of the selected categories to classify new imitation queries. We used Matlab to implement a multiclass SVM classifier. Additionally, any metadata related to the sound (e.g. tags, description or duration) was obtained from the *freesound.org* API. Our prototype allows searching sounds from a voice imitation query. The workflow of our system can be broken down in two main steps:

- The classification algorithm evaluates the vector of attributes of the sound imitation input in the N-dimensional space of attributes for the corresponding category - which is related to an initial text tag introduced manually by the user (the category). The cluster corresponding to a given sound category (e.g. "meow") is supposed to be obtained after the classification.
- The imitation cluster has the information of the original sound, and it can be used to make requests to the *freesound.org* API. For instance, the information/ID of the similar sounds to the obtained cluster is desired. These similar sounds are computed at the time the sounds are uploaded to *freesound.org*, where low-level descriptors are used. The problem here is that semantics are not taken into account in this search. As categories are one of the main concepts of this study, words related to the field, (e.g. "cat" and "cats" for *cat* category) are used to filter the list of sounds obtained from the API. As a result, the system returns a much narrowed list of sounds.

3.3 Evaluation

The generated models refer to specific sounds that were imitated in our experiment, so new input queries should be related to some of the imitated sounds to obtain the best results. By the time imitations were used to support text queries, they were also made after listening to the reference sound. This certainly stands far from the reality, where the imitation query made by the user will be always unknown. This prototype has been made to demonstrate that additional solutions can help to find sounds in a collaborative database, so we propose to use the corresponding API in a different way. In further development, some chosen sounds would define the concepts within a category; the model would be generated from humans imitating the source having into account this concept.

To evaluate our prototype we observed the number of sounds retrieved by *freesound.org* given three different user inputs: "Category" (text field with the category), "Category + Specific Tag" (text field with the category plus text field with the action or entity inside the category) and "Category + Imitation" (text field with the category plus a voice imitation to explain the action or concept). We assume that the system gives the correct classification in all cases (see Table 4 for results).

One of the goals of our system is to filter in a meaningful way the number of retrieved results obtained from a text-only query in freesound.org. As we observe in Table 4, we reduce the list of results to 5 and 20 in each category respectively.

Sound	Category	Category + *Specific Tag*	Category + *Imitation*
Cat meowing	667	208	5
Drums hihat	16001	599	20

Table 4: Number of sounds in *freesound.org* output list

A listening test to the retrieved sounds demonstrated the total consistency of the results. Although it seems that the small number of results is produced by a too specific search, it must be said that *freesound.org* itself returns a bunch of false-positives sounds by default, mainly caused by the user-made tagging system and the pre-computed signal similarity. For instance, 20 sounds are given by the *freesound.org* API when asked for similar sounds to the *cat meowing* sound used, while only 5 of them corresponded to the category *cat,* which are the ones obtained with our prototype. Even though the desired output is returned, there is a clear reliance on the tagging accuracy, which is assumed to be correct when users choose the category of the sound, but it may add false negatives to the output.

Further work shall address a more extended user evaluation of the search prototype. The user evaluation shall involve an online survey to gather a sufficient number of participants.

4 CONCLUSIONS

Although prior works in the use of imitations for audio retrieval were not especially encouraging, we have demonstrated that giving a twist to their use can produce satisfactory results. The concept of category applied to a set of sounds with a semantic relation can be formulated as a pattern recognition problem. The variety of sound imitations is constrained by the limitations of the human voice production system. Therefore the chosen strategy in this work is to select a reduced and coherent set of sounds (or sound concepts) to be imitated, as a way to address the problem. The original audio is only used in the automatic classification step. Later we focus on the user experiments, studying the human abilities when interacting with voice imitations. The gathered results from the imitation experiment explain that humans have quite suitable abilities to imitate sounds, at least an acoustic connection with the source. In that sense, it has been tested that knowing the source helps to generalize the model, as users tried to imitate the mechanics that produced the sounds. In addition, the fact of doing something different with the human voice is quite interesting; therefore, there is a possibility of developing applications using this type of interaction.

Regarding audio retrieval, we showed that combining the current tools of *freesound.org* with the imitations as input queries the search improves substantially, reducing the number of sounds in the output list.

As further research in this direction, we propose adapting classification algorithms to these types of sounds, and finding high-level descriptors that better explain the imitations. An extensive evaluation focused on the user interaction is foreseen, including the analysis of the imitation of more sound categories, experimenting with more users and produce sounds from a given concept. Moreover, audio descriptors of the imitation can be further used in this same context to improve the retrieval results. For instance, sounds could be sorted according to certain acoustic characteristics of the audio query such as energy temporal evolution, duration, etc.

REFERENCES

[1] Font F, Serrà J, Serra X; *Folksonomy-Based Tag Recommendation for Online Audio Clip Sharing*. 13th International Society for Music Information Retrieval Conference (ISMIR 2012).

[2] Roma G, Janer J, Kersten S, Schirosa M, Herrera P. *Content-Based Retrieval From Unstructured Audio Databases Using An Ecological Acoustics Taxonomy*. The 16th International Conference on Auditory Display (ICAD-2010).

[3] Universitat Pompeu Fabra, 2013. *Freesound.org. Repository of sounds under the Creative Commons license.* [Online]. Available: http://www.freesound.org.

[4] Universitat Pompeu Fabra,2013. *Freesound API documentation.* Available at: http://www.freesound.org/docs/api/.

[5] Won-Ho S, Byoung-Soo L, Yun-Keun L, Jong-Seok *L. Speech/Non-Speech Classification Using Multiple Features for Robust Endpoint Detection*. ICASSP'00, Istambul; 2000.

[6] Zetterholm E. *The significance of phonetics in voice imitation*. Proceedings of the Eight Australian International Conference on Speech Science and Technology 2000:342-347.

[7] Lass N, Hinzman A, Easthman S, Wright T, Karen M, Bartlett B. et al. *Listeners' discrimitaiton of real and human-imitated animal sounds*. Perceptual and Motor Skills 1984; 58:453-454.

[8] Gillet O, Richar G. *Drum Loops Retrieval from Spoken Queries.* Journal of Interlligent Information System 159-177; 2005.W

[9] Janer, J. *Singing-Driven Interfaces for Sound Synthesizers*. PhD, Universitat Pompeu Fabra; 2008.

[10] Lemaitre G, Dessein A, Susini P, Aura K. *Vocal imitations and the identification of sound events.* Ecological Psychology 2011 Nov;23 (4):267-307.

[11] Peeters G, Giordano BL, Susini P, MisdariisN, McAdams S. *The timbre toolbox: extracting audio descriptors from musical signals*. J AcoustSoc Am 2011; 130(5):2902-2916.

[12] Machine Learning Group, University of Waikato. *WEKA, University of Waikato*. 2013; Available at: http://www.cs.waikato.ac.nz/ml/weka/index.html

[13] Hall M. *Correlation-based Feature Selection for Machine Learning*. PhD, University of Waikato; 1998.

[14] S. Blancas D, Project WebSite, Available at: http://davidsz.weebly.com/sound-search-in-user-contributed-databases-using-voice-imitations.html