Non-speech voice for sonic interaction: a catalogue

Alan Del Piccolo & Davide Rocchesso

Journal on Multimodal User Interfaces

ISSN 1783-7677

J Multimodal User Interfaces
DOI 10.1007/s12193-016-0227-6

ONLINE FIRST
ISSN 1783-7677
Journal on Multimodal User Interfaces
Special Issue: Emotion Recognition in the Wild
Guest Editors: Abhinav Dhall · Roland Goecke · Tom Gedeon · Nicu Sebe
Springer

Springer

J Multimodal User Interfaces
DOI 10.1007/s12193-016-0227-6

SURVEY

Non-speech voice for sonic interaction: a catalogue

Alan Del Piccolo[1] · Davide Rocchesso[2]

Received: 17 December 2015 / Accepted: 15 July 2016

Abstract This paper surveys the uses of non-speech voice as an interaction modality within sonic applications. Three main contexts of use have been identified: sound retrieval, sound synthesis and control, and sound design. An overview of different choices and techniques regarding the style of interaction, the selection of vocal features and their mapping to sound features or controls is here displayed. A comprehensive collection of examples instantiates the use of non-speech voice in actual tools for sonic interaction. It is pointed out that while voice-based techniques are already being used proficiently in sound retrieval and sound synthesis, their use in sound design is still at an exploratory phase. An example of creation of a voice-driven sound design tool is here illustrated.

Keywords Voice · Sonic interaction · Information retrieval · Sound synthesis · Sound design

1 Introduction

The voice is the primary communication channel among humans. While speech is considered to be the most important form of voice communication, non-speech voice as well is a means to convey a wide array of information. We use it when we are lacking for words to describe something, and it also represents a pre-speech, natural and immediate form of expression. In everyday life we often describe an object by mimicking its sonic behaviour by means of non-speech voice.

Interacting with sound objects by means of the voice can be more natural and effective in several contexts. For instance, the retrieval of a sound from a collection, whether of musical nature or not, usually requires the production of some verbal description that can be matched with textual data that has possibly been attached to each sound. Conversely, mimicking and imitating sounds are typical actions that are intuitively performed by means of non-speech voice. They require no production or recollection of verbal information and thus, provided that adequate techniques to match the voice to the sounds are made available, vocal imitation is a potentially effective and immediate retrieval strategy.

Modifying a pre-existent sound is a task that is usually burdened with the learning process of a sound editing tool. The direct use of the voice to control the features of a sound, on the other hand, may relieve the user from such effort. Moreover, the nuances of a sonic idea are likey to be best captured by not having to translate them into words or manual gestures. The spontaneity of vocal exploration is valuable when devising a new sound from scratch, and it may therefore be exploited in sound design activities.

This article offers an overview of current developments in the use of non-speech voice in three main sound-related contexts:

- Sound retrieval,
- Sound synthesis and control,
- Sound design.

Sound retrieval consists in searching a sound, or more generally an audio track, amongst those contained in a collection. Sound synthesis refers to the procedure of generating sounds by using different techniques (e.g. additive synthesis,

✉ Alan Del Piccolo
alan.delpiccolo@gmail.com

Davide Rocchesso
roc@iuav.it

[1] University "Ca' Foscari" of Venice, Venice, Italy

[2] Iuav University of Venice, Venice, Italy

Published online: 22 July 2016

granular synthesis etc.), all of which are performed using electronic hardware and/or software. In this context, control refers to the ability of manipulating the parameters of the synthesis tools during the creation and/or the reproduction of the sounds. Sound design is a wide discipline involving heterogeneous activities (e.g. sound acquisition, sound manipulation, sound generation) whose generic outcome is an audio product with specific aesthetic and functional qualities. Sound design usually refers to the whole process, starting from the early sketching of sonic ideas to the final refinements of the product; as such, sound retrieval and sound synthesis may well take part in a sound design process.

A homogeneous comparison of current developments in the use of non-speech voice across the three contexts is only partly possible. While some topics are common (e.g. vocal signal feature selection), each context presents different degrees of development and peculiar issues.

In sound retrieval several strategies of use of the voice have been developed, as well as methods for extracting significant information from both the vocal signal and the collected sounds. Humming, singing, whistling, and more articulated expressions such as onomatopoeias can all be used to mimic the features of a sound. Low-level to high-level strategies can then be adopted to infer the relevant information, discarding the inaccuracies caused by the limits of human recollection and phonatory apparatus. Diverse approaches and algorithms to match the information extracted from the voice and the sound collection have been devised, some involving machine learning techniques.

In sound synthesis the evolution in the manipulation of the singing voice by means of effects has driven the creation of voice-based interfaces for digital music instruments or, more generally, sound production tools. While the pitch and loudness of a vocal sound are most easily mappable to analogous features of a synthesized sound, more advanced solutions have been developed to allow for more effective control. The voice can either retain the role of a sound source to be manipulated, or become a control tool which can be used alongside traditional manual interaction, or as an alternative to it.

The discipline of sound design, regardless the heterogeneity of approaches and techniques that are usually exploited by professionals, is likely to take advantage of the use of non-speech voice. Thanks to the immediacy of vocal sketching, sonic ideas can be readily made concrete. However, the process of sketching a sound prototype currently suffers from the lack of a tool as quick and versatile as the visual tools that are commonly available for graphic and product design. While a structured use of non-speech voice in such context is still missing, partly due to the lack of an engineered approach to the discipline, past and present research focus on exploiting non-speech voice to perform fast prototyping in sonic interaction and to facilitate the communication of audio concepts. Prior to that, understanding which sonic ideas can be adequately conveyed by means of vocal imitation helps to define the most effective fields of application.

It must be pointed out that, albeit the three abovementioned contexts include most of the current research over the use of non-speech voice for sonic applications, several other contexts are being studied. For instance, onomatopoeia words have been experimentally used for clustering audio tracks according to their similarity to such words [65,66]. Moreover, non-speech vocalizations have been also used for identifying audio tracks within a mix by means of imitation [61].

In this survey, non-speech voice will often be referred to as "vocalization" for the sake of simplicity. Several different definitions can be found for this term, e.g. "the sound made by the vibration of vocal folds modified by the resonance of the vocal tract",[1] or the act of changing a consonant sound to a semivowel or vowel.[2] Conversely, in this article we indicate with "vocalization" non-speech voice at large, encompassing sounds produced and modified by means of different phonatory mechanisms, provided that such sounds do not organize in words of a spoken language.

The catalogue is structured in three main sections, namely Sects. 3–5, one per each context of use of vocalization for sonic interaction. Section 2 briefly elaborates on the motivations of the present article. Section 3 deals with sound retrieval: Sects. 3.1–3.3 present a variety of strategies that have been explored for addressing the different stages of a vocalization-based sound retrieval system: the selection of the type of vocal input, the selection and extraction of relevant features, and the strategies for matching vocal query to auditory tracks within a collection. Section 3.4 presents the peculiar features of several retrieval systems or related researches, which are also summarized in Table 1. Section 4 deals with sound synthesis and control: Sect. 4.1 overviews the features that are typically extracted from the voice for such purposes; Sect. 4.2 describes the different conceptual and practical roles of the voice; Sect. 4.3 introduces the problem of mapping, while in Sect. 4.4 a large set of examples is displayed. Table 2 summarizes such examples in the light of the aforementioned characteristics. Section 5 deals with sound design by summarizing several past and current studies that have been considering the use of non-speech voice for such task. An example of creation of a voice-driven sound design tool is presented, which is part of the EU SkAT-VG project.[3] Conclusions and final remarks are collected in the final section of this paper.

[1] http://www.audioenglish.org/dictionary/vocalization.html.

[2] http://www.oxforddictionaries.com/definition/english/vocalize.

[3] http://www.skatvg.eu.

2 Motivations and related work

Beside clinical purposes, the human voice is usually studied in the contexts of music (singing) or language (speech). In singing, the main mechanism for generating sounds is to produce an airstream from the lungs outwards, which passes through the glottis (the opening between the vocal folds) and the vocal tract (larynx, pharynx and mouth). The vibration of the vocal folds shapes the sound of the airstream into a fundamental frequency and a number of higher harmonic partials. The vocal tract acts as a resonator: It increases or decreases the amplitude of each partial depending on its shape [67].

Such mechanism of sound production is enriched or varied in many ways in spoken language. For instance, some consonants require the speaker to stop the airstream at some point of the vocal tract, instead of modulating it. Nasal consonants, on the other hand, require some air to be expelled through the nose during sound radiation [35].

Yet, the human voice possesses much wider capabilities than what can be usually heard in speech or singing. Not only the mechanisms of producing a sound by means of mouth and vocal tract exceed in number and variety those involved in such contexts (e.g. percussive sounds, like by clashing teeth, are rarely found in speech [26]); such mechanisms can also be controlled with a degree of precision that is comparable, if not superior, to that of the hands when manipulating a concrete object. Such power and versatility, together with the aforementioned immediacy and naturalness, spur the use of non-speech voice as an alternative to traditional hand-based interaction. Sonic applications, in particular, prove to benefit from the use of the voice, in that a sound-to-sound mapping prevents the conversion of sonic ideas to intermediate representations, whether textual or visual. Thus, the immediacy of sonic ideas is preserved.

To our knowledge, no attempt has yet been made to assemble an overview encompassing all of the potential uses of non-speech voice in a sonic interaction scenario. Conversely, several studies tackled subsets of this subject, and as such they partly inspired the present catalogue.

Several works partially overviewed the use of vocalization in the sound retrieval context. Starting from two decades in the past, specific techniques such as Query by Humming have been periodically summarized [20,44,73], thus testifying an evolution in the discipline.

Experiments in practical use of vocalization for driving sound synthesis benefit from the improvements in the analysis of the vocal signal. Most commercial applications, however, still nowadays use only the pitch and loudness information of the vocal signal. Conversely, current research involves the extraction of many features from the voice, and consider machine learning techniques for tackling the increased complexity of mappings between vocal features and synthesis parameters. An overview of projects regarding such topic has been written by Fasciani [14,15].

Until recently, the discipline of sound design has not yet considered the use of non-speech voice, apart from a few exploratory studies and projects [11,71]. Actually, design culture is still overly visual, and only in recent years the practices of basic sound design and sound sketching have been introduced in design literature [9]. A more effective use of the voice in research-through-design practices would help raising the status of sound design within design disciplines.

3 Sound retrieval

Libraries of sounds, both proprietary (e.g. sound packs for music composition applications) and open-access (collaborative databases, such as Freesound [18]), usually contain thousands of samples. Text-based queries rely on the tagging of all samples in a collection with adequate metadata. This is not always the case, since tagging is often performed manually and is therefore costly. Besides, such approach requires the searcher to be able to express verbally the auditory properties of the desired sample. More natural and efficient query techniques are consequently advocated.

The use of the voice represents a promising way to facilitate the sound retrieval task. Humans commonly rely on vocal imitations when lacking for words to express sounds [38]. Such natural interaction modality prompts a retrieval mechanism that is based on perceptual features rather than linguistic representations, and as such must depend on audio content, not on descriptive metadata [3].

Most of the studies and implementations regarding sonic retrieval deal with music, a more common task than generic audio retrieval, resulting in the commercial success of applications such as Shazam [59] and SoundHound [62]. Music Information Retrieval (MIR) usually relies on the tracking of high-level features such as melodies and rhythms, which provide a common starting point for the analysis and the matching of queries and samples, and allow for several assumptions that ease such tasks. Conversely, different features may be relevant for different generic semantic categories of audio signals (e.g. animal sounds, mechanical sounds etc.). A search engine for arbitrary sounds is therefore difficult to realize. Environmental Sound Recognition (ESR) is a recent field of research that aims at the assignment of everyday sounds, other than speech of music, to predefined categories. Such definition represents an issue in the first place: A semi-automatic classification, performed for instance by machine learning algorithms relying on generic audio features, is likely to produce different categories than a classification built on perceptual basis [4]. An accurate selection of the audio features to be extracted from the sound dataset is therefore crucial.

Vocalization-based queries are usually an application of Query by Example (QbE): A searcher attempts to mimic the desired sound with her voice. The system then extracts a set of features from the vocal imitation and performs a search by similarity by comparing the values of such features with those of the items in the collection.

It is important to notice the human tendency to imitate the sound source, that is the mechanism generating the sound, rather than a precise sound [3]. This attitude relies on an intuitive categorization of sounds [36], and benefits from different kinds of contextual information.

3.1 Category-dependent feature selection

Most of the music-related uses of vocalization rely on melody extraction, and consequently on the ability of the searcher to recall and reproduce melodies. Singing (often by replacing the lyrics of a tune with onomatopoeias), humming and whistling are all natural tools which can be exploited for MIR tasks. Common issues in tracking the melody from a pitched voice concern the limits of a person in both recalling and reproducing a melody correctly. Most of the consequent inaccuracies deal with the pitch of notes (transposition) and the tempo. Conversely, pitch contours and rhythmic patterns are generally reproduced with sufficient accuracy [49].

Vocalization-based MIR often employs also non-pitched components of the vocal emission, such as rhythmic patterns. Such features can be used as a support in identifying melodies [8], or independently when querying rhythm-based databases such as drum loop collections [34].

When querying generic audio collections, such as databases of environmental sounds, no simplifying assumptions (e.g., the intention to imitate a melody) can be made. The voice signal is therefore analysed as a generic input audio stream, by extracting low-level features, such as Root Mean Square (RMS) level, spectral centroid, Mel Frequency Cepstral Coefficients (MFCCs), etc.. A selection of relevant features is then performed heuristically, and is typically category-dependent [74]. In presence of multiple sound categories within the source collection, the user is sometimes asked to explicitly select the category before performing the query, to increase the accuracy of results.

3.2 Vocal query strategies

Humming and singing are the types of vocalization that are most frequently used to reproduce melodies for MIR. Since the same pitch tracking techniques are usually adopted for both modes, they are described as Query by Singing/Humming (QbSH), or simply Query by Humming (QbH). Whistling is used as well, and presents the advantages of being gender-independent and to present only dominant frequency and overtones [50], while humming and singing produce more complex signals.

Rhythmic sounds and patterns can be queried by imitating the drums with percussive vocal sounds. Early experiments focused on imitating a single drum sound, namely an indian tabla [21], but the possibility of matching the imitation of different drums has been investigated as well [34]. Such research has been inspired by the technique of imitating drums with the voice named “beatboxing”. Its use in MIR has been consequently named Query by Beatboxing by some researchers [34].

Non-music sound retrieval generally makes no assumptions on the sound objects that are going to be imitated (e.g., a melody, or a rhythm), and consequently on the vocal strategy that is going to be used. The effectiveness of the query then depends on either the definition of a specific sound category that has to be queried [3] or on performing an efficient clustering of the sound collection, e.g., based on the perceptual similarity of sounds [74].

The use of onomatopoeias is generally stigmatised, since they are language- and culture-dependent [3], and therefore the matching patterns that may be identified are not generalizable. This accounts especially for the unconstrained use of onomatopoeias, i.e. when the searcher is allowed to emit arbitrary utterances. Nonetheless, experiments have been performed within the limited scope of a single language, namely Japanese, in conjunction with verbal queries [29,70]. Conversely, the restricted use of onomatopoeias can be convenient in that it facilitates the recognition of atomic segments of a vocalization. In melodic queries, the use of a single onomatopoeia (e.g. “la”) for singing a tune facilitates the detection of note onsets [49]. In rhythmic queries, matching schemes between a set of onomatopoeias and a set of different drums [22], or for different stroke techniques over a single drum [21], have been used for similar tasks. At any rate, onomatopoeias are iconic representations of sounds rather than their imitation, and as such their use in sound retrieval differs from the QbE paradigm.

3.3 Matching strategies

Due to the nature and limitations of the phonatory apparatus, vocal sounds are intrinsically different from the sounds that the users are trying to imitate. Sound databases are therefore pre-processed in order to extract features that enable the comparison between the vocal query and the sound collection. A representation of each audio track is created from such features, and the matching takes place between pairs of representations instead of actual sounds.

The procedure of creating concise representations of sounds is often called “audio fingerprinting” in that it returns a unique description of the salient features of each audio track. The requirements of such description, named fingerprint, include “discrimination power over huge numbers of other fingerprints, invariance to distortions, compactness, computational simplicity” [5]. Audio fingerprinting was developed in the context of exact matching, e.g. to match two copies of the same sound, one of which has been degraded by noise or filtering. This context is exemplified by the commercial application Shazam, whose purpose is to identify a song by matching a low-quality excerpt of it (for instance played on a radio) with a database containing an uncorrupted version of the song. Nonetheless, this strategy may be applied to other sound retrieval techniques, as long as both the query and the sounds from the collection share the same type of audio fingerprint.

One strategy for creating a fingerprint is to extract many different descriptors, both low- and high-level. This approach may be computationally expensive, therefore a selection of the most relevant features is necessary. Such a selection represents the main issue of this strategy, since it may not result equally effective across sound categories. Once the features have been extracted, their trajectories can be used for defining general trends within single sounds (constant, up, down, fluctuating). Hidden Markov Models (HMMs) can be used for indexing sounds according to such trajectories. This solution presents the advantage of being robust to time-warping distortion [74].

Another strategy is to detect the spectral peaks of the sounds, thus obtaining a fingerprint based on the energy distribution in the domains of frequency and time [72]. Such method has been developed for the aforementioned Shazam, and it may be merged with a vocal input such as humming [72]. Anyhow, spectral peak extraction proved to be particularly effective for sounds that show some regularity, such as tunes or rhythms, while it is unreliable for detecting more complex events [46]. More techniques for audio fingerprinting involve the use of low-level features such as spectral flatness [27] or differential energy flux [24].

When tracking melodies, several implementations take advantage of the human capability of recalling the pitch contour of a tune, namely the relative pitch changes between successive notes (“up”, “down”, “stable”). The resulting description of a tune is named Parsons code and enables the comparison between a vocal query and a collection in the form of string matching [20]. Absolute pitch values are commonly computed via auto-correlation [69]. Other strategies involve heuristics and statistical methods based on HMMs. An additional assumption that can be used is that the searcher is likely to imitate the vocal track of a song, rather than other melodic lines. The collection is then pre-processed in order to identify the vocal tracks, whose Parsons codes are then extracted. One strategy to extract the vocal track from a multi-track recording is to use information about the placement of each instrument in the stereo mix [19].

A more advanced approach to the use of time-related information includes the use of relative temporal changes of timbre. It requires the computation of many low-level features accounting for the “spectral shapes” of the searched sounds, and the computation of the Pareto optimality with respect to the distances from the values of the query sound’s features [12]. Several optimizations, such as the estimation of approximate distances beforehand, are implemented to reduce the computational cost.

As aforementioned, pitch and tempo information of the vocal query are generally unreliable, and they commonly require transposition and time warping in order to be used for exact matching. The monophony of the vocal query is another limitation for the effectiveness of retrieval, since most music presents different, overlapping instrument tracks. This problem has been addressed in the relatively simplified case of drum loop collections by using a Discrete Wavelet Transform (DWT) for discerning the different drum sounds by their frequency bands [34].

Concerning similarity measurements, several distance metrics have been evaluated [44] in QbH scenario: Euclidean Distance (ED), Dynamic Time Warping (DTW), and K-Nearest Neighbour (k-NN). Each technique has been applied in turn to MFCCs, Linear Predictive Coefficients (LPCs) and Linear Predictive Cepstral Coefficients (LPCCs) to assess which combination would produce the best retrieval accuracy. DTW was reported to perform slightly better than ED and k-NN, and it is used for direct matching of waveforms also by others [76]. As a drawback, DTW is basically quadratic in time, although linear approximations of the algorithm exist [58]. Conversely, ED and k-NN are linear. Another strategy employs HMM-forward algorithms in order to match note transitions and recognize a melody [60].

A novel approach to feature extraction and selection consists in using a neural network, in this case a Stacked Auto-Encoder (SAE), to deduce features that are more apt to represent vocal imitations than pre-chosen features such as MFCC’s [75].

3.4 Examples

3.4.1 Query by singing/humming

Musipedia [43] is an online search engine for MIR. It is based on pitch contour search and uses Parsons code to encode the music pieces, therefore it relies only on the identification of melodies. A recommended alternative to humming is singing a tune by using a single onomatopoeia. Rhythm-based search is also provided, but not by means of vocalization.

Tunebot [28], another QbH-based retrieval system, matches sung queries to musical themes. It uses the distance between pairs of contiguous notes. Thus, singing in a different key than the recording may still generate a match. The

Table 1 Non-speech voice for sound retrieval

	Vocal features	Vocal strategy	Matching strategy
Musipedia	Pitch	Humming, whistling	Parsons code
Tunebot	Pitch, rhythm	Singing, humming	Note intervals + rhythmic ratios
SoundHound/Midomi	Pitch, rhythm, pause locations, speech and phonetic content	Singing, humming	Match a compact representation of features
Tuneserver	Pitch	Whistling	Parsons code
Query by beatboxing	Low-level features (temporal, spectral and cepstral)	Vocal imitation of drum sounds	Match features
Generic sound retrieval from voice imitation queries	Low-level features (temporal, spectral and cepstral)	Generic non-speech voice	Match features

intervals are also unquantized to allow for non-standard tunings, given the rareness of perfect pitch in users. In addition to note intervals, Tunebot uses the rhythmic ratios between notes as well. By using the ratios instead of the actual length of the notes, the similarity measure is not affected by the tempo of the performance. A weighted string alignment algorithm matches queries and targets by using the information about note intervals and rhythmic ratios. Additionally, a database of user-contributed examples enables a progressive refinement of the results. The database consists of unaccompanied ("a cappella") melodies, each of which has been manually associated with a song. The comparison between two unaccompanied vocal tracks is then easier to perform than between a vocal track and a full band recording.

SoundHound/Midomi use MARS (Multimodal Adaptive Recognition System) technology that allows to match human voice to other human voices. This implies a pre-processing of all songs in the database for extracting the vocal signal; additionally, a user-contributed database of sung melodies is also present [2]. MARS extracts various features from the human voice including pitch variation, rhythm, pause locations, speech and phonetic content, and matches existing voices to find the information. It then adapts to the query by estimating which features are more important than others, e.g. if the query is in the form of humming, speech content is ignored. Conversely, if the user sings the lyrics as well, the search takes into account speech and phonetic content as well [42].

3.4.2 Query by whistling

The Tuneserver system [50] enables the users to whistle a melody to retrieve a song. The analysis of whistling is simplified compared to other vocal emissions: The produced sounds are less user-dependent than those of singing, and their frequency spectrum is simpler than that of other types of vocalization. The pitch is identified by detecting maximum energy frequency, then pitch transitions are converted to a Parsons code. The Parsons code is matched with those of a database of about 10,000 themes of classical music.

3.4.3 Query by beatboxing

Kapur et al. [34] developed a classification system that automatically identifies the individual vocal imitations of drum sounds to match them to actual drum sounds. The system focuses on three of the most common drum pieces, namely bass drum, snare drum and high-hat. In addition, the automatic detection of the tempo of beatboxing enables the dynamic browsing of a music database, in a more general "query by rhythm" approach.

The time and frequency characteristics of the vocal imitations cannot be directly compared to those of the corresponding real drum sounds. Consequently, the imitated sounds must be identified by means of a preliminary audio feature extraction and classification phase. Several features have been considered, both single- and multi-dimensional:

- *Time domain features* Zero crossings, RMS energy and ramp time;
- *Spectral domain features* Spectral flux, centroid and rolloff;
- MFCCs;
- LPCs;
- *Wavelet-based features* Means and standard deviations of wavelet coefficients for each sub-band of the wavelet decomposition of the signal.

Zero-crossing, centroid and rolloff performed better than RMS and ramp time among the single-dimensional features, while LPCs and MFCCs performed better than wavelet-based features. The main issue when querying drum loops is the monophony of the vocal input, since a drum loop typically presents several overlapping sounds. As already mentioned, this problem is addressed by separating the single drum sounds that compose a loop by means of a DWT.

3.4.4 Generic sound retrieval from voice imitation queries

Blancas and Janer [3] explored the issues and potentialities of vocal queries for the retrieval of generic sounds. In particular, the capabilities to produce several sounds and the tendency to reproduce sound production mechanisms rather than the original sounds were addressed. A subset of the Freesound collaborative sound database was used. Such collection is provided with tools supporting similarity measurements based on multidimensional distances of acoustic feature vectors. For this purpose, each sound that is uploaded to the database is automatically analyzed by means of Essentia [13], a library for audio analysis, in order to extract a large number of descriptors [57]. The different types of descriptors are grouped as follows:

- Low-level descriptors, which are computed from the magnitude spectrum. Early experiments involved 13 MFCCs as descriptors;
- Tonal descriptors, which are extracted from chroma features;
- Rhythm features, which are extracted from onset detection;
- sfx features, which are aimed at sound effects and describe some properties of the spectrum and the temporal evolution of pitch and amplitude.

The analysis files are then indexed by a Gaia instance, an in-memory database that contains the vector space defined by statistics of all the considered descriptors.

Test users were asked to imitate reference sounds belonging to four different categories (cat, dog, car, drums). Among the sizeable number of features that were extracted, a subset was chosen to drive a support vector machine (SVM) classification algorithm.

Early evaluations stressed the difficulty by users to imitate some timbral characteristics, and the consequent necessity of limiting the semantic context in order to reduce errors. A preliminary filtering of the sound category to be searched (e.g., sounds related to a single source or entity) was therefore performed via text query.

The behaviour of users during the imitations was observed, resulting in a series of remarks:

- Pitched sounds were easier to imitate than noisier ones, since the use of tonal variations is most common in humans when talking or producing any sound;
- The familiarity with a sound source might play a role when imitating a sound related to it;
- Instinctive imitations generally yielded better results than ponderated ones;
- The initial tendency to use onomatopoeias was quickly replaced by the use of imitations, thus confirming the greater expressive power of the latter.

A prototype for voice imitation queries based on SVM classification was devised. The voice imitation was meant to be an addition to a text query to explain an action or concept. The prototype worked in two steps:

1. The classification algorithm evaluated the vector of features of the imitation and returned a cluster corresponding to a sound category (e.g., “meow”);
2. The cluster was then used to query the Freesound database via the API it provides, which enable a similarity search based on low-level descriptors (see above).

The descriptors of the vocal imitations that proved to be more important were those related to spectral content, e.g., spectral crest and spectral variation. The overall performance of a query formed by a text query plus a vocal imitation was relevantly better than the text query alone, thus proving the effectiveness of overcoming the limitations of human phonatory system by selecting a reduced and consistent set of sounds.

4 Sound synthesis and control

The immediacy of vocalization has perhaps been the main reason for the exploration of its use for controlling sound synthesis devices or applications. While a conscious and effective use of the voice clearly benefits from study and practice, instinctive vocalization requires no training, unlike the use of most interfaces for traditional, hand-based interaction.

The use of vocalization for sound synthesis is mostly concerned with music applications. An intuitive, primary goal is the extension or augmentation of the singing vocal signal. The need for achieving a wider palette of timbres motivates the switch from singing voice modification by means of audio effects to the control of synthesis modules that replace the original sound source, while retaining some similarity with it. This enables to cross the borders of the music application context, which is related to the singing voice, and prompts both the use of various types of vocalization and the application to different sound synthesis scenarios.

These two approaches—voice as an input signal and voice as a control tool—often intermingle, for example during live music performances: In such circumstances sound generation and sound manipulation processes may be continuously alternated.

It may be argued that, as a control tool, the voice is essentially a replacement for manual control. As such, it may bring no added value to the interaction other than extending the number of variables that can be managed simultaneously. For example, while the user's hands may be still in charge of the main controls, e.g. by interacting with an instrument interface, the voice can represent "spare bandwidth" to be used for handling additional parameters while the hands are busy [15]; this use is important for real-time control situations such as live performances. Conversely, when retaining the role of a sound producing instrument, the richness of a vocalization is hardly replicable by means of other expressive tools, and its possibilities have not been fully explored yet.

A clear advantage of the use of vocal over manual control is the intrinsic multidimensionality of the former. It has been shown that controlling a two-dimensional parameter space by means of vocalization requires little effort, and even managing a three-dimensional space is possible, albeit harder [48].

Several issues arise in the use of vocalization in sound synthesis and control. The first one, which is common to all uses of vocalizations, concerns the limitations of the human voice, such as monophony or the boundaries of the comfortable pitch range [63]. On the other hand, the number of variables of the vocalization that one is able to modulate simultaneously and independently is also a matter of research [37]. The most investigated issue, however, is the mapping problem, namely what features of the voice shall be used to control which parameters of the sound synthesis, how to match different timbral spaces, etc.. A sound classification of the currently explored solutions has been provided [15], which will be summarized later in this section.

4.1 Vocal features

Pitch and loudness (or energy) are the voice features that are most commonly used for sound synthesis and control, since their modulation is common in everyday use of the voice. These two features are directly mappable onto the same parameters within the sound synthesizer, and as such they have been used for decades even in commercial products [56].

Derivative features that can be extracted from pitch and loudness information are the pitch contour, the pitch bending and the attack detection. They are usually exploited to convert the vocal signal into MIDI controls for a synthesizer, such as note values and velocity [10]. Yet, MIDI is often deemed to be insufficient for capturing the richness of the vocal signal [32].

Other high-level features deal with the identification of vocal nuances such as changes in the vowel formants, thus incorporating aspects of speech analysis [47]. Along the same line, the vocal signal can be segmented into syllables, from which timbre features are extracted [33].

The use of low-level features extracted by means of frequency analysis is also contemplated [31,51]. However, as the number of considered features and their abstraction increases, it is a common approach to resort to probabilistic descriptions [47] and to machine learning [15] in order to relieve the users from the burden of managing excessively complex mappings.

4.2 Roles of the voice

Commercial applications tend to preserve the role of the voice as an input stream [23], because they are often oriented to singers/performers. This is the case of a "voice transformation" scenario, in which the voice can become an augmented instrument, or *hyperinstrument* [33], similarly to adding sound effects or sequencing possibilities to an acoustic instrument.

The voice can be also used for the initial selection, via vocal imitation, of which synth or sound model to choose [6]. Such use overcomes the number and complexity of modern synthesizer controls, which are usually an obstacle to an artist's creative flow, in that it focuses on the perceptual aspects of sound rather than on the controls of the sound model.

All of the aforementioned solutions require the control of synths and the manipulation of the original signal to be done manually in a more traditional way. Conversely, vocal control is advocated by other researchers in order to infuse expressivity and dynamics to an electronic music performance [14,30] thanks to the human ability in controlling vocal timbres. In such contexts, the timbre may depend entirely on the sound generator, yet the extraction and use of an adequate set of vocal features and their use for controlling the model parameters enables to shape the sonic result accordingly to the original intentions behind the vocalization.

4.3 Mapping strategies

When using the voice to control the synthetic model of a music instrument, the two sonic spaces (the one of the voice and the one of the imitated instrument) have to be matched [33].

The mechanisms of vocalization (see Sect. 2) and the human phisiology that determines them define the limitations of the sonic space of the voice. The timbral possibilities, as well as the capabilities of modulation, causes the sonic space of the human voice to be much smaller than that of most generic synthesizers. Other limitations include:

- *Comfortable pitch range* The human voice can be used for prolonged periods only within a subrange of one's vocal frequency range [63];
- *Monophony* The human voice is basically monophonic, consequently it usually cannot originate different pitches simultaneously. However, different mechanisms (e.g. myoelastic and turbulent) can operate concurrently.

Matching techniques are therefore necessary, which often involve a dimensionality reduction of the synthesizer parameter space.

A second issue is the multiplicity of parameters at both ends. The number of features that can be extracted from a vocal signal can be significant, especially when dealing with low-level features. On the other hand, the parameters of a sound synthesis engine can be equally numerous. As a consequence, the variety of many-to-many mappings is likely to become unmanageable by humans.

Two main strategies for coping with mapping issues have been summarized by Fasciani and Wyse [15]:

- "Explicit mapping" Synthesis is based on the vocal imitation of the acoustic characteristics of the desired output sound;
- "Generative mapping" Machine learning techniques are adopted to establish the relationships between performer actions and instrument parameters.

The imitation-based approach consists of vocally imitating the acoustic characteristics of the desired output sound, and as such it is easy to interpret. Conversely, only the synthesis parameters that are related to such characteristics are available to the user. Moreover, mappings are mostly inflexible and little adaptable to the specific vocal characteristics of different users (it is only possible to set tuning thresholds and scaling values).

On the other hand, the adoption of supervised or unsupervised machine learning to establish the relationship between performer actions and instrument parameters may lead to a facilitation of mapping definition, and to a user adaptive instrument interface.

4.4 Examples

4.4.1 Extending voice-driven synthesis to audio mosaicing

Janer and de Boer [33] devised a system that uses vocal signals as a tool for audio mosaicing. Audio mosaicing consists in concatenating micro-segments of songs, or other audio, to match the original signal, here the voice. Rhythm, tone

Table 2 Non-speech voice for sound synthesis

	Vocal features	Mapping strategy
Extending voice-driven synthesis to audio mosaicing	RMS energy, rhythm, spectral centroid, spectral flatness, MFCC's, harmonic pitch class profiles (HPCP's)	Generative
Auracle	(Linear prediction of) zero-crossing rate, F0, frequency and bandwidth of F1 and F2	Explicit
Voice-controlled plucked bass guitar	Pitch, RMS energy, spectrum-based timbral descriptors	Explicit
Singing-driven interfaces for sound synthesizers	Pitch, amplitude, F1, F2	Explicit
A voice interface for sound generators	50 features including pitch, RMS energy, LPC's, MFC's	Generative
Billaboop	Zero-crossing rate, high-frequency content, spectral centroid, spectral roll-off, high frequency content, overall band energy variation	Generative
Native Instruments' The Mouth	Pitch, note onsets	Explicit
The gesture follower	– (Features can be selected arbitrarily)	Generative
The Wekinator	– (Features can be selected arbitrarily)	Generative
The singing tree	Pitch, amplitude, noisiness, brightness, formant structure etc.	Explicit
Wahwactor	Low frequency spectral energy	Explicit
Discreet	Pitch, amplitude, note onsets	Explicit
Synthassist	Pitch, amplitude, spectral centroid, spectral spread, spectral kurtosis etc.	Explicit
Intuitive sound design using vocal mimicking	Pitch, amplitude	Explicit

and timbre of the output sound are controlled by the performer's voice. The voice signal is segmented according to phonetic variations based on musical functions (attack, sustain, release, etc.). The user selects a small corpus of audio sources. For each target segment a list of similar units in the audio sources are selected using a distance measure. The similar units are then randomly chosen and concatenated in a sound loop, possibly consisting of several layers to produce a richer sound. To achieve a higher level of control, mapping functions between voice and audio sources are learnt by means of supervised training methods, such as Gaussian Mixture Models (GMMs). Given the differences among users in imitating sounds, it is preferred to let the system learn the mapping functions for every user.

4.4.2 Auracle

Auracle [51], a voice-controlled online collaborative instrument, features a linear prediction of low-level features such as fundamental frequency (F0), RMS, zero-crossing rate, and the frequencies and bandwidths of the first two formants (F1, F2). It segments analysis frames into gestures, computes features of those gestures, and classifies them. Data from each musician are not directly mapped to the parameters of different instruments, but merged to control a single sound synthesis system. The feature envelopes are classified using Principal Component Analysis (PCA) for dimensionality reduction, followed by a neural network. An example of use considers amplitude and the envelopes of fundamental and formants to control the analogous parameters in a physical model synthesizer.

4.4.3 Voice-controlled plucked bass guitar

In a sound synthesizer controlled via singing voice [31], descriptors are extracted from the vocal signal via Short-time Fourier Transform (STFT) and classified in four groups related to their use for control: Excitation (F0 and energy), Vocal Tract (vowel formants), Voice Quality and Context. The mapping is structured in two layers, which adapt the input parameters to different instruments and different bass guitar synthesis techniques:

- The first layer matches the energy derivative to the note onsets and the pitch of the voice to the instrument pitch;
- The second layer depends on the synthesis technique and matches different parameters accordingly. Two models have been developed:
    - For a physical model, the voice energy envelope onset detection triggers the string excitation, and the pitch defines the length of the string delay line;
    - For a spectral morphing synthesis, the sound is generated by concatenation of spectral frames from a database, storing information about spectrum, harmonic peaks, pitch, dynamic and attack type. The pitch and harmonic peaks from the voice are used to query the database to retrieve the nearest frames, while a feature named "attack unvoiceness" operates a selection between fingered and sharp slap attack.

4.4.4 Singing-driven interfaces for sound synthesizers

In his Ph.D. thesis, Janer [32] proposed the imitation of the sound of a musical instrument by means of the user's voice. The main control strategy that has been investigated is the temporal segmentation of the vocal signals based on syllables. The consequent mapping is perceptually clear:

- Voice pitch and loudness control the analogous parameters of the instrument;
- A single continuous value derived from F1 and F2 controls the timbre modulation.

4.4.5 Making music through real-time voice timbre analysis

In his Ph.D. thesis, Stowell [64] explored the possibility of controlling a musical instrument by means of the user's voice. The focus of such research was on timbral control, and how the timbral qualities of the vocal input could be extracted and used to control the synthesis. The case study referred to beatboxing, as an example of the timbral diversity of the human voice. Many spectral features were analysed to infer the most suitable ones with respect to perceptual relevance, robustness and independence. Although spectral centroid and spectral 95-percentile resulted the most relevant and robust, a wider set of features was adopted and then reduced via PCA to enable the mapping between the voice's and the instrument's timbre spaces. Unsupervised and supervised machine learning was employed to deduce mappings between vocal features and synthesis parameters automatically.

4.4.6 A voice interface for sound generators

Fasciani and Wyse [14] investigated the use of vocalization to control sound synthesis parameters for a DMI in real time, to generate a many-to-many mapping automatically, and to adapt synthesis parameters to perceptual sound features. The prototype allows users to dynamically modify the timbre of the synthetic sound by using the dynamics in their voice. A voice analyzer module computes vectors of 50 features, including features both in time and spectrum domain such as: RMS, pitch, zero-crossing rate, spectral centroid, spectral flux, spectral deviation, Mel spectrum deviation and centroid, spectral flatness coefficients, LPC and MFC coefficients, formant frequencies and magnitude. The system automatically analyzes the voice to detect stable and dynamic features, links the DMI's control parameters to the perceptual features of the synthetic sounds, and maps the vocal features to the controls. A dimensionality reduction via PCA is applied to the set of perceptual features.

4.4.7 Billaboop

Billaboop [25] is a Virtual Studio Technology (VST) plugin which matches onomatopoeic vocal beatboxing into synthetic or sampled drums. Sounds are triggered depending on the onsets that are detected in the vocal signal. The onset detection algorithm is based on the variations of High Frequency Content and overall Band Energy. Three target sound classes have been selected: bass drum, snare drum and cymbal. The choice of the sound to be played is operated through a decision tree classifier. The vocalizations for the training set were provided by a single performer. Since the sounds used in beatboxing may vary between different performers, a supervised classifier ensures adaptivity.

4.4.8 Pitch-based commercial applications

A plethora of DMIs and applications that rely almost exclusively on pitch detection are present on the market.

Native Instruments' The Mouth [23] detects the pitch of an incoming audio signal. The signal is auto-tuned to a musical scale of choice, or to notes coming from a MIDI device. The auto-tuned signal triggers a synthesizer which, together with the gate parameters, allows the production of additional melody and harmony. Besides the synthesizer sound, it is possible to mix in unprocessed audio, a vocoder and effects. It can also take the articulation of the voice and use it to modulate the sound. It presents two operating modes:

- In Pitch mode, The Mouth analyzes pitch artifacts to auto-tune the incoming signal relative to the selected musical scale or MIDI input. The snapped pitch of the input is fed into the synthesizer and vocoder which adds additional melody and harmony;
- In Beats Mode, The Mouth processes drum patterns acting upon the transients and frequencies of incoming audio.

Roland SPV-355 [56], Bitspeek [45], and DigitalEar [10] capture the voice' loudness and pitch, and map them onto the loudness and pitch of a digital instrument. Additional features are polyphonic pitch tracking, pitch bending, and attack detection.

4.4.9 The singing tree

In this installation [47] ten dynamic parameters are extracted from a singing voice and used to control a set of MIDI instruments. Amplitude, pitch contour, brightness, noisiness and formants of the voice signal are determined by means of time-, frequency- and cepstrum-domain analysis. Such features are converted into meaningful inputs for the music generation engine, named Sharle, allowing for the control of scale, key, tempo and more parameters. The pitch values control the progress of a musical sequence, while loudness, formants, cepstra and their deviations are mapped to multiple parameters using probabilistic algorithms. Fuzziness is introduced to reduce the determinism of the sonic result and foster a creative use of the interface.

4.4.10 Wahwactor

In the Wahwactor [40] the central frequency of the resonant filter of a guitar wah-wah effect is controlled by the musician's voice transition between the phonemes */u/* and */a/*. Preliminary study findings suggest that the spectral energy in the [500, 1500] Hz band yields the smoothest and most stable response in comparison to spectral centroid, cepstra and others.

4.4.11 Discreet (voice control project)

In this project [30] samples are manipulated in response to vocal sounds. Vocal features are extracted by means of FFT spectral analysis to enable pitch and amplitude tracking, onset detection for both pitched and unpitched sounds, division between pitched and unpitched sounds, calculation of the length of a single sound and of a phrase and more. The vocal signal is consequently broken down in four discrete timbre types, namely "syllables" (short tone sounds), "vowels" (long tone sounds), "consonants" (short noise sounds) and "breaths" (long noise sounds). The sound generation is realised with granular synthesis using recorded cello samples. Four grain types were developed, which selectively trigger based on the content of the input, and each of these respond to input data in a different way. The voice timbre types are recorded into a loop. A random selection of samples is read from it, each of which triggers a sound event, namely a grain. By adjusting the length of the recording loop, the performer can adjust the control between immediate responsiveness and randomness.

4.4.12 Synthassist

Vocal imitation is used in this project [6] for querying a database of settings for audio synthesizer modules. The goal is to let users focus on high level features of the desired sound rather than on low-level controls. A first set of queries, which have some of the characteristics of the target sound, is provided by the user. Given these examples, an interactive refinement process is started, where the system presents sounds for the user to rate. The system progressively refines its estimate of the desired concept and learns which audio features are important to the user. Finally, Synthassist returns the selected synthesizer parameters.

A small number of high-level features are extracted from the vocal signal (which can also be a small sound excerpt other than the voice): pitch, loudness, inharmonicity, clarity, spectral centroid, spectral spread, and spectral kurtosis. Each query and search key is then represented as fourteen time series: Seven of them are the actual values of each feature at a given time ("absolute features"), while seven of them capture the relative changes in time of each feature ("relative features"). The database is searched by calculating the distance from the query to each search key (DTW is used to cope with different sound lengths).

4.4.13 Intuitive sound design using vocal mimicking

The goal of the project by Wake et al. [71] is to enable users to edit the characteristics of a sampled sound by using vocalizations as a tool to transform its pitch and amplitude envelope. The changes in vocal pitch and amplitude are in fact used to process the base sound. The purpose is to enable sound editing without dealing with visual waveform representation. It is meant to be a tool for sound designers, who often produce various sounds with different timbre but the same outward form (length, rhythm, pitch etc.) for comparison. The time required for producing such test sounds is sensibly reduced in comparison to traditional wave editors.

5 Vocalization for sound design

Regardless of the different contexts of use, which make it difficult to achieve an unambiguous definition of the discipline itself, sound design stands apart from the design of other design disciplines. Whether for artistic or industrial purposes, the goal of realizing an invisible and intangible product such as a sound makes it difficult for designers to use typical practices such as prototyping and collaborative design.

Sonic iteraction dsign (SID) [17,52,55] is a field of research that aims at conveying information, meaning and aesthetic qualities mainly through sound within interactive applications or products. One of the motivations behind such research is the lack of readily available tools to produce early sketches of sounds and sonic behaviors quickly and effectively, just as a visual designer would do with paper and pencil. This lack of tools reflects not only in the productivity of the designers, who are not able to validate a draft until late in the design process, but especially in the loss of rapidity in materializing a creative impulse. Therefore many subtleties of the original idea might get lost along the way.

The voice represents a suitable tool for sketching sounds [53]. Its use is natural and immediate in humans, and as such it overcomes the possible limitations in technical knowledge of design tools. Besides, the voice can be exploited simultaneously with manual interaction. In particular, non-verbal vocalizations are often employed by humans to express concepts or sonic ideas that are hard or impossible to describe by the means of words [38], and they are largely free from linguistic and cultural dependencies.

The degree of control of the voice apparatus is by no means inferior to the degree of control of the hands. Yet, the lack of suitable interfaces puts the voice in disadvantage when compared to hands in tasks that involve the fine control of many parameters. That is why non-speech voice appears to be more suitable for the early stages of the sound design process (sketching) rather than for sound refinement and prototyping.

This section summarizes several past and current studies that have been considering the use of non-verbal vocalization for sound design. Among such studies, the current SkAT-VG project (2014–2016) aims at the creation of sound sketching tools based on voice and gesture. A description of the research involved in the creation process is presented in the form of an example of development of such a tool.

5.1 Vocal sketching: a prototype tool for designing multimodal interaction

Tahiroğlu and Ahmaniemi [68] investigated the use of vocal sketching in the context of multimodal interaction. A series of experiments was conducted, in which users were asked to vocally imitate the expected auditory behaviour of an interactive object while manipulating its tangible interface. The imitations were meant to be produced according to the shape, affordance and functionality of the object, and to the specific gesture that was performed on it (moving, squeezing, stroking). In this way, the sonic characteristics of the device, as well as the information that would have been conveyed by the sounds, were meant to be outlined before the realization of actual audio synthesis modules. At the same time, the experiment intended to highlight the expectations regarding the auditory behaviour of the device, derived from its shape and affordance. In particular, the coupling between manual gestures and vocal sounds was analysed.

The sounds sketched by participants mostly fell within three wide categories: real world (e.g. elevator sound), synthetic (e.g. sound effects), and abstract sounds. The movements triggered changes in the character of the sounds, e.g. the pitch was moved from low to high in response to a vertical movement of the device, while circular movements and horizontal rolling gestures were coupled with continuous pitched sounds.

The strong interconnection between sounds and interaction modalities is one of the major findings of this study. For instance, the duration of the sound was always the same as that of the gesture. Moreover, specific gestures prompted specific sounds, e.g. a shaking gesture was mostly accompanied by percussive vocal sounds.

5.2 Using vocal sketching for designing sonic interactions

Ekman and Rinott [11] investigated vocal sketching as a methodology to approach sound design. In particular, the difficulties encountered by non-experts during the early stages of a sound design process were addressed. Vocal sketching was meant to help users when thinking and communicating about sonic ideas. A workshop was held in which participants were asked to sketch the sonic behaviour of objects by using only their voice.

The sounds that were produced were mostly complex, often with an organic origin. Such sounds would have required high technical expertise to be produced by synthesis and to be used interactively in a design.

Several limitations of vocalization were reported to have an impact on the design process:

- The monophony of the voice, which makes harmonies or polyphonies possible to perform only in groups;
- The difficulty in articulating specific complex sounds;
- The difficulty in controlling single acoustic features;
- The limits of the breath cycle, which makes long continuous sounds impossible to produce.

As a general remark, vocal sketching was attested to drive design particularly towards sounds that are hard to produce by means of current tools.

5.3 VOGST project

Franinović et al. [16] built on the findings by Ekman and Rinott to develop a tool enabling the use of voice and gesture to sketch and improvise sonic interaction. The goal of such research was to overcome the limitations of the technical knowledge of designers and artists while sketching interactive sound concepts. The main focus was the process of designing the interactions between gesture and sound, and how to facilitate such task.

Several problems were addressed in the process:

- *Ergo-audition* The human voice is heard differently by the person producing the sound and the one hearing it. This may represent an obstacle in the communication of sonic ideas;
- *Further refinement of vocal sketches* Visual sketches, such as pencil lines over a paper sheet, can be redrawn, corrected and changed at will. Conversely, the ways of achieving such elasticity with vocal sounds and gestures are to be investigated.

The resulting tool, namely a simple abstract object capable of capturing both voice and gesture, was tested by interaction designers in a workshop, to elicit possible design problems and to specify the iterative process of prototyping.

5.4 VocalSketch: vocally imitating audio concepts

Cartwright and Pardo [7] collected thousands of crowd-sourced vocal imitations of a large set of heterogeneous sounds, together with data representing the participants' ability to correctly identify such imitations. The goal was to build a data set to "help the research community understand which audio concepts can be effectively communicated with this approach" [7].

Four categories of sounds were devised: "everyday", "acoustic instruments", "commercial synthesizers" and "single synthesizer". Users were asked to produce a vocal imitation starting from either a sound label or a reference sound. Users were discouraged to use onomatopoeias.

The "everyday sounds" category was the one which was most effectively communicated with vocal imitations. Authors argued that this may have been due to the familiarity of such sounds, but also to the ease in reproducing them. Indeed, the vocal imitations of sounds that are easily producible by the voice (e.g., yawning), or that present peculiar time-varied characteristics (e.g., police siren) were those which were recognized with the highest accuracy. Conversely, sounds consisting of many overlapping sonic events (e.g., glass breaking) led to the least accurate recognition.

As a general remark, inaccuracies mostly led to a description of similar or possibly more general concepts than the one that had been imitated. Authors argue that, as a consequence, more information might be needed for disambiguation. Such information may be provided verbally by users.

5.5 SkAT-VG project

The SkAT-VG project [53] aims at enabling designers to use their voice and gestures to produce sketches of the auditory aspects of an object, whether an industrial product or an artistic effort. The final goal of the project is to devise a system that interprets the users' intentions through their vocal sketches, and consequently selects appropriate sound synthesis modules which enable iterative refinement and col-

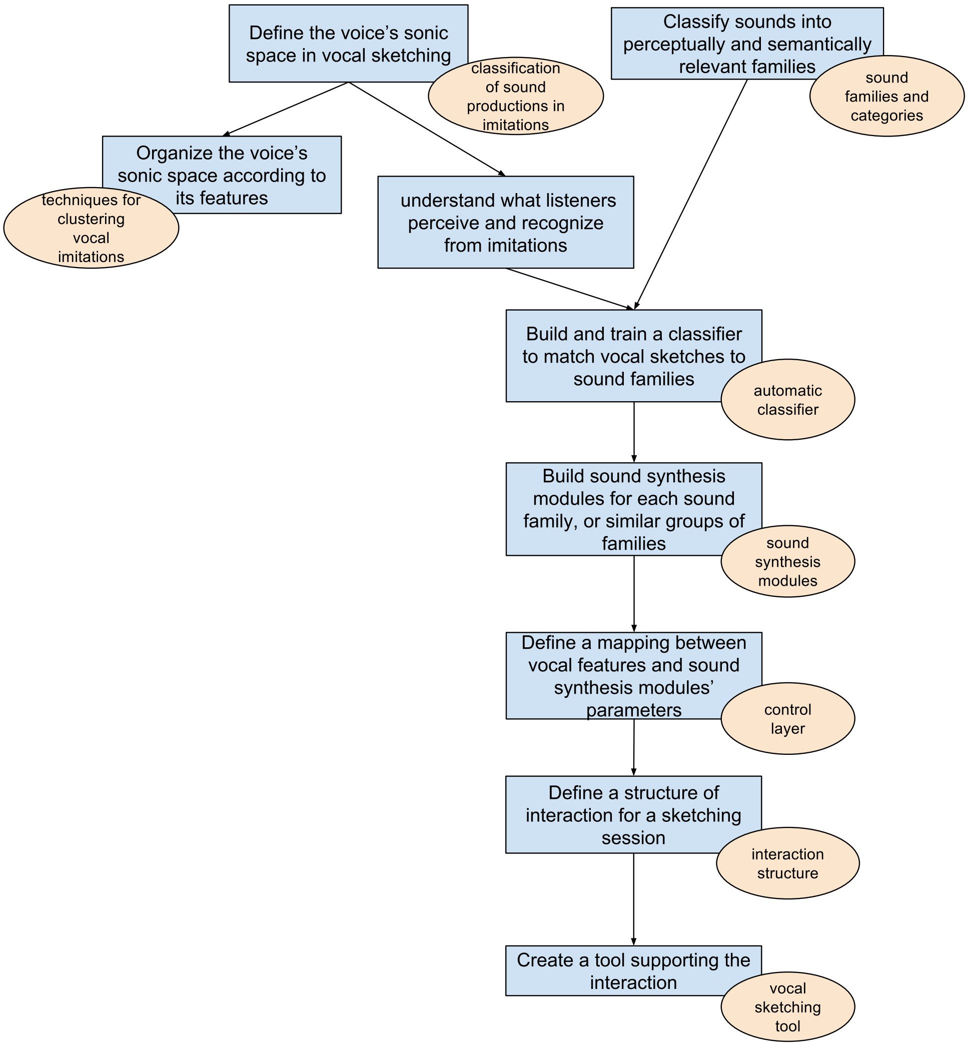
Define the voice's sonic space in vocal sketching
classification of sound productions in imitations
Classify sounds into perceptually and semantically relevant families
sound families and categories
Organize the voice's sonic space according to its features
techniques for clustering vocal imitations
understand what listeners perceive and recognize from imitations
Build and train a classifier to match vocal sketches to sound families
automatic classifier
Build sound synthesis modules for each sound family, or similar groups of families
sound synthesis modules
Define a mapping between vocal features and sound synthesis modules' parameters
control layer
Define a structure of interaction for a sketching session
interaction structure
Create a tool supporting the interaction
vocal sketching tool

Fig. 1 Phases of the creation of a vocal sketching tool

laborative design. A block diagram depicting the phases of the creation of the vocal sketching tool and the involved research is shown in Fig. 1.

The research behind the SkAT-VG project involves different disciplines such as phonetics, machine learning and interaction design. The preliminary steps include:

- The identification of the vocal sonic space in this context, which has been found to exceed that of spoken language, in that phonatory mechanisms that are rare or unused in language find use in vocal imitation [26];
- The classification of target sounds on both perceptual and semantic basis. A set of 26 sound categories, organized in three main families (Abstract, Interaction, Machine) and limited to the context of product sound design, has been experimentally defined [39];
- The implementation of an automatic classifier to couple the user's vocal sketch to a sound category: Specific audio descriptors that directly highlight the morphological aspects of sound have been shown to be acceptably accurate in the classification of vocal imitations [41].

A prototypical tool named miMic has been devised for enabling such activities [54]. It consists of a microphone which has been augmented with two latching buttons and an inertial measurement unit. The designer is meant to use miMic both as a tool for recording sketches ("select mode", activated by the first button) and as a controller for interacting with the synthesis model ("play mode", activated by the second button). In the play mode, the user uses both voice and gestures to affect sound synthesis: The envelopes of the voice and movement features are mapped to synthesis parameters through a control layer. Although the microphone is connected to a computer for computation and visual display of information, keyboard-based interaction is thus almost totally avoided in favour of a more natural, spontaneous interaction.

To clarify the structure of a sketching session, an example is provided here below. Since one of the application fields that have been investigated is the production of combustion motor sounds by driving sound synthesis modules [1], the context might be the creation of sounds for a not-yet-existent wheeled motor vehicle:

0. The designer **selects** a sound model or a mixture of sound models. The selection is operated as follows:
   (a) The user presses "select" and performs a vocal sketch into the microphone;
   (b) The system analyzes the sketch and classifies it into a sound category, or a mixture of weighted categories (e.g., combustion engine and wind);
   (c) The system returns the sound synthesis modules that are relevant for the chosen sound categories.
1. The user presses "play" and begins to vocally **mimic** the desired sound; thus, the designer directly drives the sound synthesis, validates the model selection and familiarizes with the synthetic sonic space;
2. Further on, by creatively using the voice beyond mere imitation, the user extensively **explores** the possibilities of the synthesizer(s). By moving the microphone in various ways (rotating, tilting, swinging) the user is able to manipulate additional features of the sound;
3. The designer manipulates on the computer the individual sound synthesis parameters, to gradually and iteratively **refine** a sketch until obtaining a sound prototype.

6 Conclusions

In this paper we presented a survey of the current state of the art in the use of non-speech voice for sonic interaction. Whenever possible, general strategies and techniques have been summarized. The three contexts of use that have been considered here (information retrieval, sound synthesis and control, and sound design) present several common issues in the use of non-speech voice. The first issue concerns the extraction and the selection of the audio features from a vocal signal. The purpose is to convey the most information to interpret the user's intentions. Controllability of such features is important as well. The second issue concerns the control of sound synthesis modules, and consists of the mapping of the audio features to the controls of each module.

Non-verbal vocalization has been already extensively investigated in the context of sound retrieval, especially in MIR, in which the narrowed field of research allows for assumptions that ease the task. Conversely, research in generic sound retrieval has to deal with classification issues originated by the difficulty of accurately reproducing a sound with one's voice. Consequently, matching the vocal imitation to a cluster of sounds within a generic sound collection is still unviable. Work-around solutions include the manual selection of a sound category to be queried. On the other hand, a promising way to address classification is to extend the notion of semantic categorization, which has proven to be effective in humans, to both the machine classification of vocal queries and to the clustering of the target sounds.

In the context of use of the voice as an input and a control tool for sound synthesis, almost all of the examples apply to music, which is due to the central role of the voice in such context. Several different approaches can be identified from the shown examples, from a more analytical, "holistic" [30] use of low-level features of the voice, which is pursued to retain as much of the original characters of the voice as possible, to simpler pitch-based applications, which are common in commercial music production tools. The mapping between vocal features and synthesis controls depend on the multiplicity of both, and while the challenge is to retain perceptual relevance in the mapping, machine learning solutions can be adopted to achieve otherwise too complex control patterns.

Unlike sound retrieval and synthesis, sound design is a discipline in which the use of vocalization has not been much considered. Proof is the total lack of fully developed, professional tools for vocal-based sound design. One of the main reasons may reside in the ambiguous definition of the discipline itself, which leads to a plethora of different techniques and tools, thus showing a general lack of an engineered approach to sound design. As such, the introduction of a new possibility such as the use of vocalization inevitably involves an attempt to find a common ground between the different approaches and techniques, upon which to create a tool that is effective in most situations. The few studies described in this article outline the potentialities of vocalization for sound design, especially in the sketching phase. Among those studies, the current SkAT-VG project takes a multi-disciplinary approach to devise effective tools for supporting the sound design activity.

Acknowledgments The Project SkAT-VG received 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.

References

1. Baldan S, Delle Monache S (2014) His engine's voice: towards a vocal sketching tool for synthetic engine sounds. In: Proceedings of the XX Colloquium on Musical Informatics (CIM). Università IUAV di Venezia, Rome

2. Ballinger RA (2015) Midomi.com: A unique way to search music. Tech. rep., Tangient LLC. http://hum-socsci-tech.wikispaces.com/Midomi.com+-+A+Unique+Way+to+Search+Music. Accessed 10 December 2015
3. Blancas DS, Janer J (2014) Sound retrieval from voice imitation queries in collaborative databases. In: Proceedings of AES 53rd Conference on Semantic Audio. Audio Engineering Society, London
4. Bountourakis V, Vrysis L, Papanikolau G (2015) Machine learning algorithms for environmental sound recognition: Towards soundscape semantics. In: Proceedings of the Audio Mostly 2015 on Interaction With Sound (AM 15). ACM, Thessaloniki
5. Cano P, Batlle E, Kalker T, Haitsma J (2002) A review of algorithms for audio fingerprinting. In: Proceedings of the 2002 IEEE workshop on Multimedia Signal Processing. IEEE, St. Thomas, Virgin Islands
6. Cartwright M, Pardo B (2014) Synthassist: Querying an audio synthesizer by vocal imitation. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Goldsmiths, University of London, London, pp 363–366
7. Cartwright M, Pardo B (2015) Vocalsketch: Vocally imitating audio concepts. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI'15. ACM, New York, pp 43–46
8. Chai W, Paradiso JA, Schmandt C (2000) Melody retrieval on the web. In: Proceedings of ACM/SPIE Conference on Multimedia Computing and Networking. San Jose, California, pp 226–241
9. Delle Monache S, Rocchesso D (2014) Bauhaus legacy in research through design: the case of basic sonic interaction design. Int J Design 8(3):139–154
10. Digital Ear convert audio (.wav) files to midi. (2015). http://www.digital-ear.com. Accessed 10 Dec 2015
11. Ekman I, Rinott M (2010) Using vocal sketching for designing sonic interactions. In: Proceedings of the 8th ACM Conference on Designing Interactive Systems, DIS '10. ACM, New York, pp 123–131
12. Esling P, Agon C (2013) Multiobjective time series matching for audio classification and retrieval. Audio, speech, and language processing. IEEE Trans 21(10):2057–2072. doi:10.1109/tasl.2013.2265086
13. Essentia (2014) Open source C++ library for audio analysis and audio-based music information retrieval. http://essentia.upf.edu. Accessed 10 Dec 2015
14. Fasciani S, Wyse L (2012) A voice interface for sound generators: adaptive and automatic mapping of gestures to sound. In: Proceedings of the International Conference on New Interfaces for Musical Expression. University of Michigan, Ann Arbor, Michigan
15. Fasciani S, Wyse L (2013) Mapping the voice for musical control. Tech. rep., National University of Singapore, Singapore
16. Franinović K (2014) Vocal Sketching Workshops—VOGST. IAD Interaction Design, Zürcher Hochschule der Künste. http://blogs.iad.zhdk.ch/vogst/. Accessed 10 Dec 2015
17. Franinović K, Serafin S (eds) (2013) Sonic interaction design. MIT Press, Cambridge
18. Freesound.org (2015). http://freesound.org. Accessed 10 Dec 2015
19. Fu L, Xues X (2005) A new spectral-based approach to query-by-humming for MP3 songs database. In: Proceedings of The Second World Enformatika Conference (WEC'05). Istanbul, Turkey, pp 117–120
20. Ghias A, Logan J, Chamberlin D, Smith BC (1995) Query by humming: Musical information retrieval in an audio database. In: Proceedings of the Third ACM International Conference on Multimedia, MULTIMEDIA '95. ACM, New York, pp 231–236
21. Gillet OK, Richard G (2003) Automatic labelling of tabla signals. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003). ISMIR, Baltimore, Maryland
22. Gillet OK, Richard G (2005) Drum loops retrieval from spoken queries. J Intell Inf Syst 24:159–177
23. Gover D (2010) User's manual for Native Instruments The Mouth, Product Version: 1.0. Native Instruments GmbH, Berlin, Germany
24. Haitsma J, Kalker T, Oostveen J (2001) Robust audio hashing for content identification. Int Workshop Content Based Multimed Index 4:117–124
25. Hazan A (2005) Performing expressive rhythms with Billaboop voice-driven drum generator. In: Proceedings of the 8th Int. Conference on Digital Audio Effects. Madrid, Spain, pp 20–23
26. Helgason P (2014) Sound initiation and source types in human imitations of sounds. In: FONETIK. Stockholms Universitet., Stockholm
27. Herre J, Allamanche E, Hellmuth O (2001) Robust matching of audio signals using spectral flatness features. In: IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA 2001). IEEE, Mohonk, pp 127–130
28. Huq A, Cartwright M, Pardo B (2010) Crowdsourcing a real-world on-line query by humming system. In: Proceedings of the 7th Sound and Music Computing Conference (SMC10). SMC, Barcelona
29. Ishihara K, Nakatani T, Ogata T, Okuno HG (2004) Automatic sound-imitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes. In: PRICAI 2004: Trends in Artificial Intelligence. Springer, Auckland, pp 909–918
30. Janaway N (2007) Voice control: how feasible is the use of the voice as a controller of electronic sound environments? Msc thesis, University of Edinburgh
31. Janer J (2005) Voice-controlled plucked bass guitar through two synthesis techniques. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Vancouver, pp 132–135
32. Janer J (2008) Singing-driven interfaces for sound synthesizers. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona
33. Janer J, de Boer M (2008) Extending voice-driven synthesis to audio mosaicing. In: Proceedings of the 5th Sound and Music Computing Conference (SMC05). SMC, Berlin
34. Kapur A, Tzanetakis G, Benning M (2004) Query-by-beat-boxing: Music retrieval for the DJ. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004). Barcelona
35. Ladefoged P (2005) Vowels and consonants, 2nd edn. Blackwell Publishing, USA
36. Lemaitre G, Dessein A, Susini P, Aura K (2011) Vocal imitations and the identification of sound events. Ecol Psychol 23(4):267–307
37. Lemaitre G, Jabbari A, Houix O, Misdariis N, Susini P (2015) Vocal imitations of basic auditory features. In: 169th Meeting of the Acoustical Society of America. Journal of the Acoustical Society of America, Pittsburgh
38. Lemaitre G, Rocchesso D (2014) On the effectiveness of vocal imitations and verbal descriptions of sounds. J Acoust Soc Am 135(2):862–873
39. Lemaitre G, Voisin F, Scurto H, Houix O, Susini P, Misdariis N, Bevilacqua F (2015) A large set of vocal and gestural imitations. Tech. rep., Deliverable 4.4.1 of the Project SkAT-VG, FP7-ICT-FET-618067. http://www.skatvg.eu
40. Loscos A, Aussenac T (2005) The wahwactor: a voice controlled wah-wah pedal. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Vancouver, pp 172–175
41. Marchetto E, Peeters G (2015) A set of audio features for the morphological description of vocal imitations. In: Proceedings of the 18th Int. Conference on Digital Audio Effects (DAFX'15). Trondheim, Norway
42. Minsook M (2015) Multimodal adaptive recognition system and sound2sound: a music information retrieval application. Tech. rep., Tangient LLC. http://hum-socsci-tech.

wikispaces.com/Multimodal+Adaptive+Recognition+System+%26+Sound2Sound+-+A+Music+Information+Retrieval+Application. Accessed 10 Dec 2015

43. Musipedia (2015). http://www.musipedia.org. Accessed 10 Dec 2015
44. Nagavi TC, Bhajantri NU (2012) An extensive analysis of query by singing/humming system through query proportion. Int J Multimed Appl 4(6). doi:10.5121/ijma.2012.4606
45. Sonic Charge—Bitspeek (2015). http://soniccharge.com/bitspeek. Accessed 10 Dec 2015
46. Ogle JP, Ellis DPW (2007) Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2007. Honolulu
47. Oliver W, Yu J, Metois E (1997) The singing tree: design of an interactive musical interface. In: Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques. The Netherlands, Amsterdam, pp 261–264
48. Orio N (1997) A gesture interface controlled by the oral cavity. In: Proceedings of the 1997 International Computer Music Conference. International Computer Music Association, San Francisco, pp 141–144
49. Pardo B, Shifrin J, Birmingham W (2004) Name that tune: a pilot study in finding a melody from a sung query. J Am Soc Inform Sci Technol 55(4):283–300
50. Prechelt L, Typke R (2001) An interface for melody input. ACM Trans Comput Hum Interact 8(2):133–149
51. Ramakrishnan C, Freeman J, Varnik K (2004) The architecture of Auracle: a real-time, distributed, collaborative instrument. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Hamamatsu, Japan, pp 100–103
52. Rocchesso D (2011) Explorations in sonic interaction design. Logos, Berlin
53. Rocchesso D, Lemaitre G, Susini P, Ternström S, Boussard P (2015) Sketching sound with voice and gesture. ACM Interact 22(1):38–41
54. Rocchesso D, Mauro DA, Delle Monache S (2016) miMic: The microphone as a pencil. In: Proceedings of the 10th International Conference on Tangible, Embedded and Embodied Interaction, TEI 2016. ACM, Eindhoven, The Netherlands
55. Rocchesso D, Serafin S (2009) Sonic interaction design. Int J Hum Comput Stud 67(11):905–906
56. Roland (1983) Owner's manual for Roland P/V synthesizer SPV-335. Roland Corporation, Osaka, Japan
57. Roma G, Serra X (2015) Querying freesound with a microphone. In: WAC—1st Web Audio Conference. IRCAM/Mozilla, Paris
58. Salvador S, Chan P (2004) Fastdtw: toward accurate dynamic time warping in linear time and space. In: KDD workshop on mining temporal and sequential data (TDM-04), pp 64–74
59. Shazam (2015). http://www.shazam.com. Accessed 10 Dec 2015
60. Shifrin J, Pardo B, Meek C, Birmingham W (2002) Hmm-based musical query retrieval. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '02. ACM, New York, pp 295–300
61. Smaragdis P, Mysore GJ (2009) Separation by "humming": User-guided sound extraction from monophonic mixtures. In: Applications of Signal Processing to Audio and Acoustics. WASPAA'09. IEEE Workshop on. IEEE, pp 69–72
62. SoundHound (2015). http://www.soundhound.com. Accessed 10 Dec 2015
63. Sporka AJ (2008) Non-speech sounds for user interface control. Ph.D. thesis, Czech Technical University in Prague
64. Stowell D (2010) Making music through real-time voice timbre analysis: machine learning and timbral control. Ph.D thesis, Queen Mary University of London, London
65. Sundaram S, Narayanan S (2006) Vector-based representation and clustering of audio using onomatopoeia words. In: Proceedings of Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium. Arlington
66. Sundaram S, Narayanan S (2007) Analysis of audio clustering using word descriptions. In: Acoustics, Speech and Signal Processing. ICASSP 2007, vol 2. IEEE International Conference on, pp II–769–II–772
67. Sundberg J (1977) The Acoustics of the Singing Voice, Scientific American offprints, vol 356. W.H Freeman, New York
68. Tahiroğlu K, Ahmaniemi T (2010) Vocal sketching: a prototype tool for designing multimodal interaction. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, no. 42 in ICMI-MLMI '10. ACM, New York, pp 42:1–42:4
69. Unal E, Narayanan SS, Shih HH, Chew E, Kuo, CCJ (2003) Creating data resources for designing user-centric front-ends for query by humming systems. In: Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR). Berkeley, pp 475–483
70. Wake SH, Asahi T (2001) Sound retrieval with intuitive verbal descriptions. IEICE Trans Inf Syst 84(11):1568–1576
71. Wake SH, Fukuzumi S, Asahi T (2001) Intuitive sound design using vocal mimicking. IEICE Trans Inf Syst E84–D(6):749–750
72. Wang A (2006) The shazam music recognition service. Commun ACM 49(8):44–48
73. Weinstein E (2005) Query by humming: a survey. Tech. rep., NYU and Google
74. Xue J, Wichern G, Thornburg H, Spanias A (2008) Fast query by example of environmental sounds via robust and efficient cluster-based indexing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2008. Las Vegas, pp 5–8
75. Zhang Y, Duan Z (2016) Imisound: An unsupervised system for sound query by vocal imitation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Shanghai
76. Zhu Y, Shasha D (2003) Warping indexes with envelope transforms for query by humming. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, SIGMOD '03. ACM, San Diego, pp 181–192