A set of audio features for the morphological description of vocal imitations

Enrico Marchetto, Geoffroy Peeters

A SET OF AUDIO FEATURES FOR THE MORPHOLOGICAL DESCRIPTION OF VOCAL IMITATIONS

Enrico Marchetto
UMR STMS IRCAM-CNRS-UPMC
Paris, France
enrico.marchetto@ircam.fr

Geoffroy Peeters
UMR STMS IRCAM-CNRS-UPMC
Paris, France
geoffroy.peeters@ircam.fr

ABSTRACT

In our current project, vocal signal has to be used to drive sound synthesis. In order to study the mapping between voice and synthesis parameters, the inverse problem is first studied. A set of reference synthesizer sounds have been created and each sound has been imitated by a large number of people. Each reference synthesizer sound belongs to one of the six following morphological categories: "up", "down", "up/down", "impulse", "repetition", "stable". The goal of this paper is to study the automatic estimation of these morphological categories from the vocal imitations. We propose three approaches for this. A base-line system is first introduced. It uses standard audio descriptors as inputs for a continuous Hidden Markov Model (HMM) and provides an accuracy of 55.1%. To improve this, we propose a set of slope descriptors which, converted into symbols, are used as input for a discrete HMM. This system reaches 70.8% accuracy. The recognition performance has been further increased by developing specific compact audio descriptors that directly highlight the morphological aspects of sounds instead of relying on HMM. This system allows reaching the highest accuracy: 83.6%.

1. INTRODUCTION

1.1. Using vocal imitations as sketches

In typical design approaches (whether in architecture, products, etc.), the very first step is often a "sketch", that is a simple graphical representation of the target. This initial sketch is a useful tool to enhance communications between designers and stakeholders. In the case of sound design, professionals often use vocal imitations to add more detail to the sound description [1], trying to transmit to their interlocutor the main cues of their sound idea [2]. "Vocal imitations" can therefore be considered as the sound design "sketches".

The goal of our current project, the SkAT-VG project, is to expand this vocal imitation idea toward a better sound design tool [3, 4]. The resulting device should be able to translate a vocal (and gestural) cue into a novel and pertinent synthetic sound. The interactive sound design begins with a phase in which the user produces an imitation and the system provides some draft sounds; in a second phase the latter are then interactively refined, again using voice and gestures. This paper addresses the task of automatic recognition of imitation, presenting different strategies to do that.

1.2. The recognition task and imitation dataset

In order to study the mapping between voice and synthesis parameters, the inverse problem is first studied. A set of reference synthesizer sounds (*stimuli*) have been created and each sound has been imitated by a large number of people. Each reference synthesizer sound belongs to one of the six following morphological categories: `up`, `down`, `up/down`, `stable`, `impulse`, `repetition`. Each of the categories is represented by two reference sounds, and each reference sound is imitated by 50 subjects.

The six categories are abstract and are defined as:

Up: Sounds which have an increasing profile in terms of spectral content and/or loudness, thus expressing a kind of rising;

Down: Opposite of the previous one, these stimuli present a downward profile;

Up/down: Sounds with non-monotonic profiles: can be described as combination of the previous two, the stimuli profile moves upward and then downward;

Impulse: This class contains sounds with very short duration and sharp attach and decay;

Repetition: Sounds which are composed by the repetition, with varied rhythmic patterns, of short and almost impulsive ones;

Stable: Longer sounds, with almost flat pitch and loudness profiles.

The goal of this paper is to study the automatic estimation of these six morphological categories from the vocal imitations of their reference sounds.

The study of the perception of vocal imitations (how do people choose their strategy to imitate a sound, how consistent are the imitations and how these imitations are recognized) have already been the subject of the paper [5] and will be the subject of further papers in the framework of the SkAT-VG project.

1.3. Related works

A well-known approach to allow time series recognition is the extraction of low level signal descriptors, which are then modeled using Hidden Markov Models (HMM) [6, 7]. This approach will be used to create our base-line system. In speech recognition [8, 9, 10], the best results are obtained combing language models (based on grammars) and acoustical models [11]. Unfortunately, abstract sounds are not bound to any grammar and a language model cannot be used in our case.

Another closely related topic is the recognition of "words for sounds", such as onomatopoeias. Proposed approaches to this problem, linked to speech recognition, still rely on phonemes [12] or lexical cues [13]. There are also examples of features clustering and modeling [14], which are related to our base-line system and to the first methodology that we propose. Description of sounds in

terms of morphological profiles has been initially proposed by P. Schaeffer works [15]. The automatic estimation of these profiles for abstract sounds has been previously studied by [16] and [17] which also propose dedicated descriptors.

2. AUTOMATIC RECOGNITION OF VOCAL IMITATIONS

In this section, we propose three methods to automatically recognize the six morphological categories indicated in sec. 1.2.

- The first method relies on the extraction of a set of instantaneous audio features $d_i(k)$, $i \in [1, ..., I]$ over time k. Each category is modeled by its own hidden Markov model.
- The second method uses the same instantaneous audio features $d_i(k)$, which are quantified into symbols to be used as input for discrete hidden Markov models.
- The third method does not rely at all on hidden Markov models, but models the time evolution directly in the audio features. We therefore denote them by "morphological audio descriptors" as in Peeters et al., 2010 [17].

Before applying one of these methods, we first detect the non-silent regions (named "active regions" in the following) using standard methods such as [18, 19, 20]. This leads to a set of N' Active Region(s) $A = \{[b_r, f_r] : r \in [1, ..., N']\}$ where b_r and f_r are their starting and ending time.

2.1. Base-line system using Local Trend descriptors

We extract 6 instantaneous audio features $d_i(k)$, $i \in [1, ..., 6]$ where k denotes the time frame number. In order to smooth the variation of $d_i(k)$ over time, a low-pass filter is applied (zero-phase filter). The first 4 are standard audio features: the spectral centroid, spectral spread, sprectral rolloff and the pitch. They are computed using standard techniques [21] and using the Swipep algorithm [22] for the pitch[1].

Given that one of the important specificities of the morphological categories relates to the temporal evolution of the spectrum content, we also propose two new audio features: "LPC-min" and "Spectral-peak-min". The novel features are defined as follows:

LPC-min: A one-pole preemphasis filter is applied. Low-order LPC is used to estimate the position of the single most prominent formant[2]. The prediction coefficients are converted into formant frequencies F_ρ, where ρ is the formant index [23, 24]. Only the frequencies $F_\rho > 20$Hz are kept. The LPC-min value is measured in Hz, and is defined as the minimum F_ρ.

Spectral-peak-min: From the energy spectrum (computed as the square DFT) we select the 5 most important frequency bins. The Spectral-peak-min is defined as the lowest frequency among these 5 frequencies; it is thus measured in Hz.

[1]We used the Swipep algorithm since it has state-of-the-art performances and is readily available on-line. Spectral centroid, spread and rolloff are computed using well-known formulas.

[2]Here the objective is not a full-fledged formants tracking, but a robust analysis of the energy location among frequencies, complementary to spectral centroid.

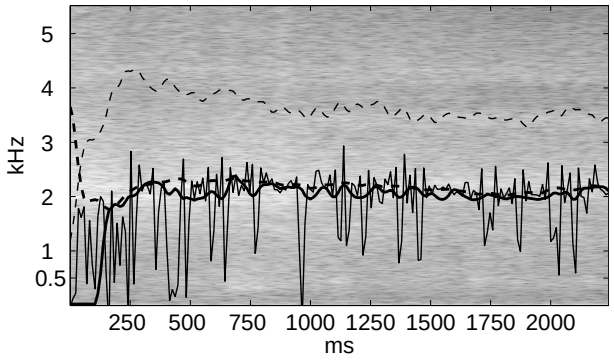

Figure 1: *Comparison between LPC-min (dashed bold line), Spectral-peak-min (bold line), spectral centroid (dashed line) and pitch by Swipep (thin line). This vocal imitation is noisy with a stable formant around 2kHz. Swipep does not detect any pitch, giving unreliable information, and the spectral centroid is moved toward higher frequencies. Both LPC-min and Spectral-peak-min are instead detecting and following the formant.*

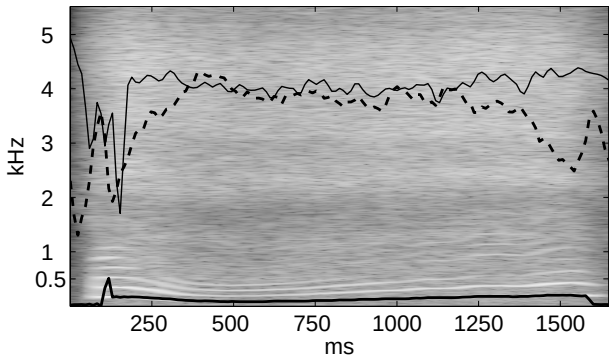

Figure 2: *Comparison between LPC-min (bold line), spectral centroid (dashed bold line) and Spectral-peak-min (thin line); pitch is not reported because perfectly matches LPC-min. This vocal imitation is harmonic but presents also noise in higher frequencies. LPC-min is clearly detecting the pitch; Spectral-peak-min is following the energy of noise, with better accuracy than centroid.*

It should be pointed out that these descriptors could have overlapping meanings, and are used together to reinforce the information. Spectral centroid and LPC-min could be similar on noisy sounds, but when a strong formant is present the centroid may loose meaning compared to LPC-min (Fig. 1). Spectral centroid and Spectral-peak-min could also be similar on noisy signal, but when a strong partial exists at the pitch the Spectral-peak-min is better at measuring it (Fig. 2).

The six categories to be recognized relate to evolution of values over time, hence we compute the derivative of each $d_i(k)$. The derivative $d'_i(k)$ is found by linear regression on the local values (5 points on the left and 5 points on the right of k). We finally normalize their range to $[-1, 1]$ using arctangent mapping: $d''_i(k) = 2/\pi \arctan(d'_i(k))$. This completes the computation of the Local Trend descriptors, exemplified in Fig. 3.

For each of the six categories c, we define a continuous hidden Markov model $\mathcal{M}_c$. Each $\mathcal{M}_c$ represents the transition between

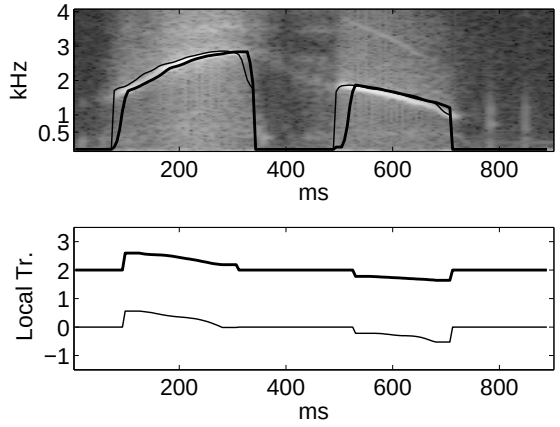

Figure 3: *Example of Local Trend descriptors on* `up/down` *imitation. Topmost panel is the signal spectrogram with spectral centroid (thin line) and Spectral-peak-min (bold line). The bottom panel shows the same two descriptors after Local Trend computation (scale is shifted for Spectral-peak-min for clarity).*

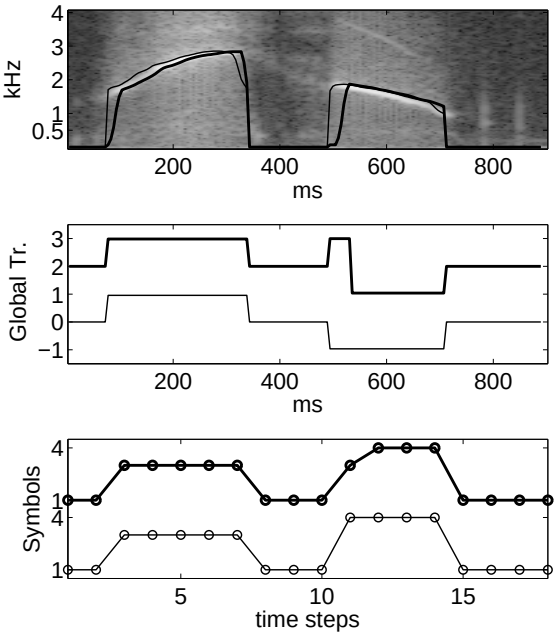

Figure 4: *Example of Global Trend descriptors on* `up/down` *imitation. Topmost panel is the signal spectrogram with spectral centroid (thin line) and Spectral-peak-min (bold line). Middle panel shows the same two descriptors after Global Trend computation (scale is shifted for Spectral-peak-min for clarity). In the bottom panel the two descriptors are quantized into symbols and downsampled (once more Spectral-peak-min has been shifted).*

a set of $S = 4$ states. The emission probability (probability of emitting $d_i''(k)$ given state s) is modeled as a mixture of $M = 8$ Gaussians and diagonal covariance matrix. The HMMs are created in a supervised way. To understand this, let's consider the case of the `up/down` category. This one can be represented as the transition from a state "silent" to state "up" to state "down" and back to state "silence". In the same spirit the `up` category can be represented as the succession of states "silence", "up", "silence". We therefore define 4 states: "silence", "up", "down", "stable". The training of the six HMMs is done in two stages:

- We first train the observation probabilities. This is done independently of $\mathcal{M}_c$. Indeed, given that a state (such as "up" in the above example) can be shared by different $\mathcal{M}_c$, we train the emission probabilities using descriptors from the `up`, `down` and `stable` classes, plus an added silent class.
- The transition probabilities are trained for each category.

Considering that the number of self-transition $s(t+1) = s(t)$ is much larger than the non-self ones $s(t+1) \neq s(t)$, we found the training of the last difficult. In order to circumvent this, we decided to decimate over time the descriptors time series by a factor of 3. This allowed to increase the performances. Also, rather counterintuitively, better results were obtained by forcing the HMM training to *not* update the emission probabilities mixtures (thus only updating the transition matrix).

2.2. Global Trend descriptors

The same 6 audio features of sec. 2.1 are used as underlying time series for the Global Trend descriptors: spectral centroid, spectral spread, spectral rolloff, pitch, LPC-min and Spectral-peak-min. Instead of using them directly as input to a continuous HMM, we convert them to symbols.

For each descriptor $d_i(k)$ and each active region r we apply the following procedure:

- We compute the linear regression over the region r. We denote by α_i^r its angular coefficient and by ϵ_i^r its prediction error.
- If ϵ_i^r is larger than a specific threshold K_1[3], r is split into two regions at the position of the maximum value of $d_i(k)$ and a two-piece minimum least-square linear regression is computed. We denote by α_i^{r1} and α_i^{r2} the corresponding angular coefficients. This process will allow us to discriminate between monotonic classes (such as `up` or `down`) and `up/down`.
- The quantized time serie $e_i(k)$ corresponding to $d_i(k)$ is then built. It has the value $e_i(k) = 0$ for k corresponding to silent part, $e_i(k) = \alpha_i^r$ for k corresponding to region r (or α_i^{r1} for region $r1$ and α_i^{r2} for region $r2$).

We then convert the values of $e_i(k)$ to symbols using the following rules:

$$e_i'(k) = \begin{cases} 1 & \text{if } |e_i(k)| \leq 10^{-7}; \\ 2 & \text{if } 10^{-7} < |e_i(k)| \leq 0.1; \\ 3 & \text{if } e_i(k) > 0.1; \\ 4 & \text{if } e_i(k) < -0.1. \end{cases} \quad (1)$$

The four symbols {1,2,3,4} express the overall condition of the time serie, respectively: silence, stable (small angular coefficient),

[3] K_1 has been optimized by grid-search. We use a value of $3 \cdot 10^3$.

upward (large positive angular coefficient), downward (large negative angular coefficient).

As a final step, the descriptors series $e'_i(k)$ are decimated over time by taking one value every 10 frames. It should be noted that we have chosen to not apply any antialiasing process, since we have found by experiment that active regions shorter than 10 samples usually correspond to errors. In Fig. 4 we illustrate the Global Trend descriptors.

Training. Since the time series $e'_i(k)$ are symbols (unordered values) we model them using discrete hidden Markov models. Each descriptor i is modeled by its own HMM $\mathcal{M}_{c,i}$.

For a given class c and a given descriptor i we denote by $E_{c,i}$ its emission matrix (with size $S \times O$, where S is the number of hidden states and O the number of symbols), by $T_{c,i}$ the transition matrix (with sizes $S \times S$) and by $S_{c,i}(k)$ the decoded states at frame k.

In order to train the emission matrix for class c and descriptor i, we concatenate all descriptors of the sounds belonging to c into a matrix D_c. Each row of D_c corresponds to a given descriptor i.

We define a function $\mathcal{F}$ that normalizes an input matrix such that its rows elements sum up to 1.

The training algorithm is the following:

1. Set the initial value for the emission matrix to $E_{c,i} = \mathcal{F}(I + 0.05)$, where I is the diagonal identity matrix. This almost associates each state to a single symbol, but does not exclude the possibility for each state to emit each possible symbol.
2. Exploiting the previous association, estimate $T_{c,i}$ by accumulating all the emissions transitions found in row i of D_c into the matrix T. Obtain the transition matrix for i as $T_{c,i} = \mathcal{F}(T)$.
3. Use $T_{c,i}$ and $E_{c,i}$ to estimate the hidden states $S_{c,i}(k)$ by Viterbi decoding.
4. Re-estimate $E_{c,i}$ by counting the number of times each state generates each emission, and again normalizing by $\mathcal{F}$.

In principle, the re-estimation procedure (points 2.-4.) can be iterated, but early experiments showed little performance improvements.

Classification. In order to classify an unknown sound represented by its time series matrix D_*, we decode each row i of D_* using a specific class model $\mathcal{M}_{c,i}$ and the Viterbi decoding algorithm. Each model $\mathcal{M}_{c,i}$ provides a likelihood $l_{c,i}$. The final class label is found as $x = \text{argmax}_{c\in[1,6]} \sum_i l_{c,i}$.

2.3. Morphological descriptors

We introduce here a new set of *morphological* descriptors. These are crafted to compactly describe the structure of the signals present in the dataset.

Each audio file is represented by these descriptors using a vector with 8 components:

- Ψ_1, Ψ_2 and Ψ_3 measure repetitions or patterns in the serie;
- Ψ_4, Ψ_5 and Ψ_6 describe the active region(s);
- Ψ_7 and Ψ_8 are related to the global signal trend.

The descriptors embed directly the time evolution of the signal. Because of this they can be used for classification without requiring the use of HMM for temporal modeling. Moreover the Ψ_i values lay in homogeneous ranges. Because of this we will use for classification a simple k-Nearest Neighbor algorithm with Euclidean distance in the following.

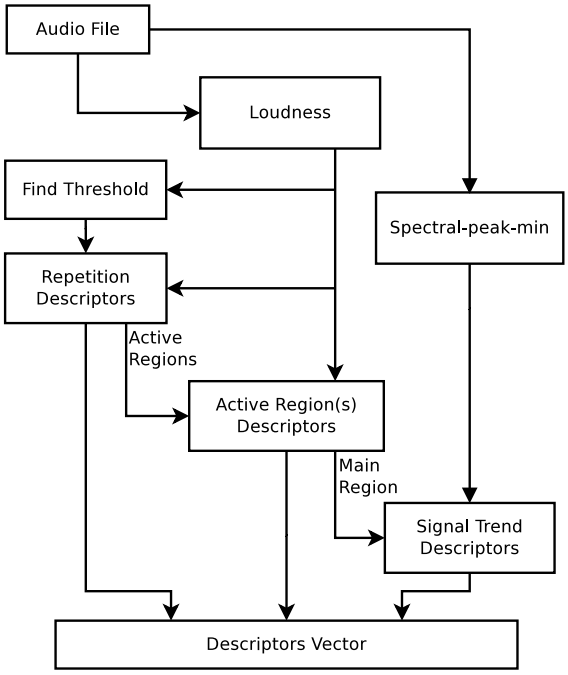

Figure 5: *Computation of morphological descriptors.*

Fig. 5 points out the main steps of the computation.

First the non-silent regions $A = \{[b_r, f_r] : r \in [1, ..., N']\}$ are detected.

For two consecutive active regions, r and $r+1$, we define the duty-cycle of r as $u_r = (f_r - b_r)/(b_{(r+1)} - b_r)$ (see Fig. 6). In this, we assume that every active region is followed by a silent one. When adjacent regions are in the same state (silent or non-silent) we simply merge them, thus enforcing the active/non-active alternance. The first two pattern descriptors Ψ_1 and Ψ_2 are defined as the mean and the standard deviation of non-silent regions duty-cycles u_r, $r \in [1, N']$. `impulse` and `repetition` classes are expected to have small values of Ψ_1. In the opposite, the other classes (stable, up, down, up/down) will have a single long active region with a large value of Ψ_1. Ψ_2 measures the regularity of the repeated patterns. In the case of a single active region, $\Psi_2 = 0$.

We define the "importance" i_r of an active region as the product between its length l_r and its mean loudness m_r (over the active region duration). Both l_r and m_r are normalized in the range $[0, 1]$ for each signal (the value of 1 is assigned to the longest and the loudest regions respectively).

The third descriptor Ψ_3 is the number of active regions which have *Importance* i_r above a threshold K_2. A value of $K_2 = 0.25$ is chosen in order to correspond to the product of half-range normalized values of l_r and m_r. Ψ_3 is computed as:

$$\Psi_3 = \frac{\arctan\left(\text{card}\left(\{i_r \text{ such that } i_r > K_2\}\right) - 1\right)}{(\pi/2)} \qquad (2)$$

The threshold on i_r allows the rejection of very short active

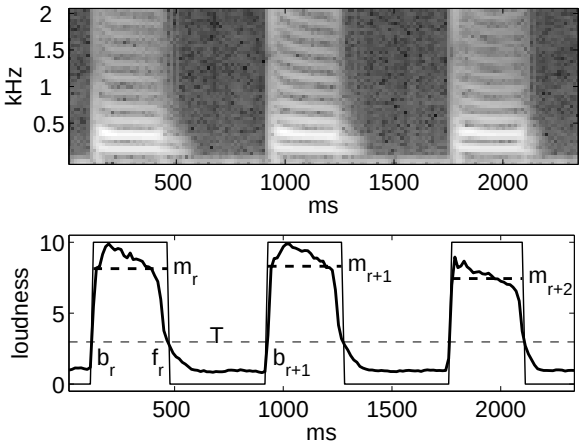

Figure 6: *Computation of descriptors Ψ_1, Ψ_2 and Ψ_3 on a* `repetition` *imitation. The loudness profile is showed (bold line), along with threshold T (dashed line) and active region detection (thin line); mean loudness value m_r (dashed bold line) is used to compute importance i_r.*

regions, or with very low loudness level. For single-region signals Φ_3 is equal to 0, while for signals with three or more regions (typical of `repetition` category) Ψ_3 is above 0.7.

Descriptors Ψ_4 and Ψ_5 focus on the main region R only, that is the non-silent region with highest importance i_R.

The Ψ_4 descriptor is the duty cycle of the main region *defined on the whole signal*: $\Psi_4 = (f_R - b_R)/N$, where b_R and f_R are the start and the end of R and N is the total signal length. Ψ_4 is expected to discriminate `impulse` and `stable` classes.

The descriptor Ψ_5 is computed on the loudness time serie in the main region $d_L(k)$, $k \in [b_R, f_R]$, which has length l_R. The original serie and its half-length circularly-shifted copy are used:

$$\Psi_5 = \sum_{k=1}^{l_R} \left[d_L(k) - d_L\left((k + \frac{l_R}{2}) \bmod l_R \right) \right]^2 \qquad (3)$$

Ψ_5 is thus the energy of the difference between $d_L(k)$ and its shifted copy. The descriptor is then normalized in the $[0, 1]$ range using $\arctan$, as for Ψ_3. Ψ_5 improves the discrimination between classes which have flat or non-flat evolution, such as `up/down` *vs* `stable`. The descriptor Ψ_6 is defined as the sum of the active regions lengths l_r, minus a constant γ. Ψ_6 is then normalized by arctangent, such as in (2), and γ is optimized to have the "shift" of the arctangent function improving the discrimination between `impulse` and `stable` (or `repetition`) classes.

The Ψ_7 and Ψ_8 morphological descriptors have been developed to measure the slope of the signal. The aim is to evaluate the slope between the beginning and the middle, and between the middle and the end, of a given time serie.

Spectral-peak-min is taken as an underlying descriptor: only its values in the main region R are considered, deleting those below 40Hz as they are unreliable. To overcome boundary effects, Spectral-peak-min is observed within three windows of 11 samples taken at the beginning, the middle and the end of the region

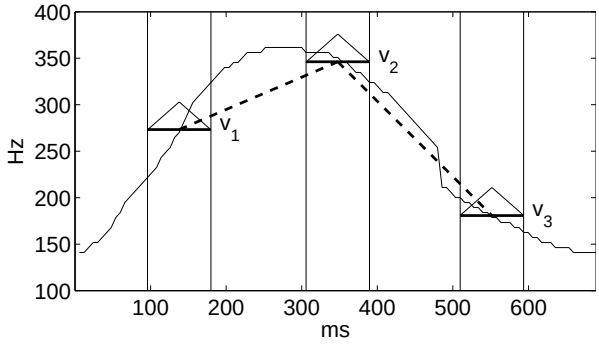

Figure 7: *Computation of descriptors Ψ_7 and Ψ_8 on an* `up/down` *imitation. Spectral-peak-min is showed (thin line), with 3 windows centered at $1/5$, $1/2$ and $4/5$ of its total length. Mean values v_j (bold line) are used to find Ψ_7 and Ψ_8, which are proportional to the slopes (dashed bold line).*

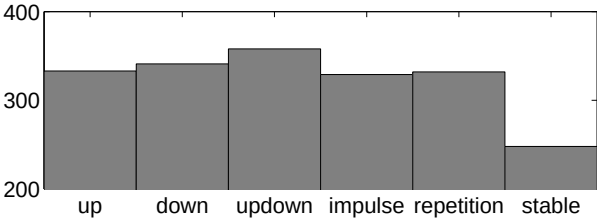

Figure 8: *Distribution into the six categories of the dataset.*

(see Fig. 7). In each window, the Spectral-peak-min is weighted by a triangular window function and then the average is computed. This leads to three mean values: $V = [v_1, v_2, v_3]$. The trend descriptors Ψ_7 and Ψ_8 are found as:

$$\Psi_7 = \frac{v_2 - v_1}{v_1} \qquad \Psi_8 = \frac{v_3 - v_2}{v_2} \qquad (4)$$

and normalized using again the $\arctan$ function. Local windows are used in order to smooth the signal, possibly generated by noisy time series, and the triangular functions give more importance to the central part of the window. Ψ_7 and Ψ_8 measure the evolution in time of the signal: they discriminate between `up`, `down` and `up/down`.

3. EVALUATION

3.1. Description of the train and test sets

The dataset used in this paper comes from a perceptual experiment. In this experiment, 50 French subjects were asked to imitate sounds. For each of the 6 classes described in Sec. 1.2, two reference sounds have been selected. After listening to one of the reference sounds (without knowing its class), subjects were asked to imitate it using only voice (VO) or using voice and gesture (VG). In each case, the subject could do several trials. In theory, each class is therefore represented by 1000 audio examples: 2 (reference sounds) times 50 (number of subjects) times 2 (VO and VG) times the number of trials of each subject (ranging from 1 to 5). In practice, considering the variability of the number of trials of each

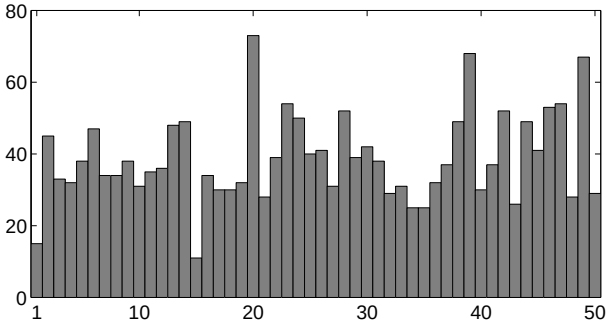

Figure 9: *Number of recordings per subject.*

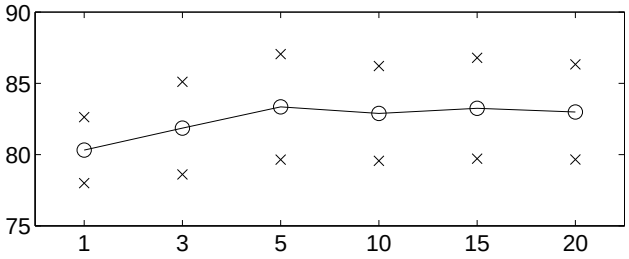

Figure 10: *Recognition accuracy in % (line and circles) as function of k value in k-NN. Results are averaged by 5 folds, and the ± 1 standard deviation intervals are marked by $\times$.*

subject, the number of audio examples is lower. We represent the number of audio examples per class in Fig. 8. The average number of imitations provided by each subject is about 40, with large variations as showed in Fig. 9. On average, every subject provided 1.6 trials per stimulus (instead of 5).

The total size of our test-set is 1941 audio files, rather equally distributed among the six categories (except `stable`, see Fig. 8). In the following experiments all the available data is exploited.

3.2. Comparison of the recognition methods

We have described three recognition methods in this paper:

Local Trend: the descriptors rely on local variations of the signal, and are modeled using continuous HMMs;

Global Trend: the main evolutions of the signal are measured by discretized descriptors, modeled using discrete HMMs;

Morphological: the descriptors directly represent both the signal shape and its time evolution; the recognition of classes by the Morphological descriptors relies on the k-Nearest Neighbor algorithm (see sec. 2.3).

The results are presented in Table 1. All the figures have been obtained using 5-folds crossvalidation, selecting train and test set in order to not have the same subject in both. Recall and precision values are given for each of the three methods and for each class. The mean recall and precision, and the overall accuracy, are given at the bottom of Table 1.

The Global Trend descriptors obtain an accuracy of 70.8%, a 28.5% improvement compared to the base-line system (Local Trend, 55.1%). The Morphological descriptors give the best performances, with 83.6% accuracy (51.7% improvement over base-line). Fig. 10 shows the values of accuracy obtained by the system when using different values for k. The best value of $k = 5$ has been selected for the presented results.

Table 1: *Recognition results by different methods, averaged over the 5 crossvalidation folds.*

Methods	Local Trend		Global Trend		Morphol.	
Measures	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
`up`	80.8	83.9	83.2	81.8	**87.7**	79.6
`down`	**88.5**	43.9	76.2	76.3	71.5	73.7
`up/down`	38.2	53.0	39.1	57.2	**76.3**	76.2
`impulse`	25.3	54.1	60.3	79.1	**91.5**	91.9
`repetition`	29.5	35.5	78.0	62.4	**90.3**	93.2
`stable`	72.9	99.2	**97.2**	70.0	85.8	92.4
Average	55.9	61.6	72.3	71.1	**83.9**	84.5
Avg. Accuracy	55.1		70.8		**83.6**	

It is interesting to look at the results class-by-class. `up` and `down` classes are well recognized by all methods. There are instead classes for which Morphological descriptors are better, such as `repetition`. The opposite case also happens, as the recall for `stable` is better using Global Trend; for `down` the best recall is given by Local Trend, but with poor precision.

The overall conclusion which can be drawn is that Morphological descriptors have a better performance because they work rather well on all classes, giving comparable recall/precision. This is not the case for Local and Global Trend, which have instead one or more classes with particularly bad results.

3.3. Discussion of the results

The Local Trend method has good performances on the `up`, `down` and `stable` classes, but not on the remaining ones. It has been verified that `up/down` recordings are often recognized as `down` (which has in fact a low precision). A possible explanation for the confusion arises by observing the spectrograms of the recordings: it has been found that many subjects prepare the downward slope in `down` by first producing a rising profile (see Fig. 11). Moreover, while `up/down` is recognized as `down`, the confusion with `up` is less frequent: in `up/down` the downward phase is usually stronger, thus justifying the observations.

The global shapes of `up`, `down` and `stable` are well modeled. Classes `impulse` and `repetition` are instead problematic: the transition matrices of the HMMs do not succeed to capture the temporal cues of the signals, and this has a bad influence on classes which are defined relying on that.

The purpose of Global Trend methodology is to provide better modeling of the overall temporal evolution of the signal, hence avoiding these shortcomings. The quantization of the descriptors in four symbols has the effect of keeping only the most relevant information, and the decimation enforces the transition matrices to describe the temporal cues of the signals. Despite having still poor results on `up/down`, Global Trend outperforms Local exactly because it improves the recognition of `impulse` and `repetition`.

The Morphological descriptors have been designed to embed the main characteristics of the classes into a compact and effec-

Table 2: *Confusion matrix for the morphological descriptors.*

Classes	1	2	3	4	5	6
up - 1	**292**	20	13	2	2	4
down - 2	32	**244**	49	9	6	1
up/down - 3	15	47	**273**	11	6	6
impulse - 4	4	11	8	**301**	3	2
repetition - 5	9	7	6	5	**300**	5
stable - 6	16	3	11	0	5	**213**

tive representation. Among their good performances, it should be pointed out the improved discrimination between `down` and `up/down` compared to the previous methods. However, the confusion matrix in Table 2 shows that the issue is still present. Similarly there is confusion between `up` and `down`. `repetition` is well recognized, thanks to the presence of the specific Ψ_3 descriptor. Both `impulse` and `stable` are confused, even if not to a large extent, with `up`, `down` and `up/down`. This could be explained by the fact that the system identifies rising or falling cues even in the `impulse` and `stable` imitations, and provides a classification according to this.

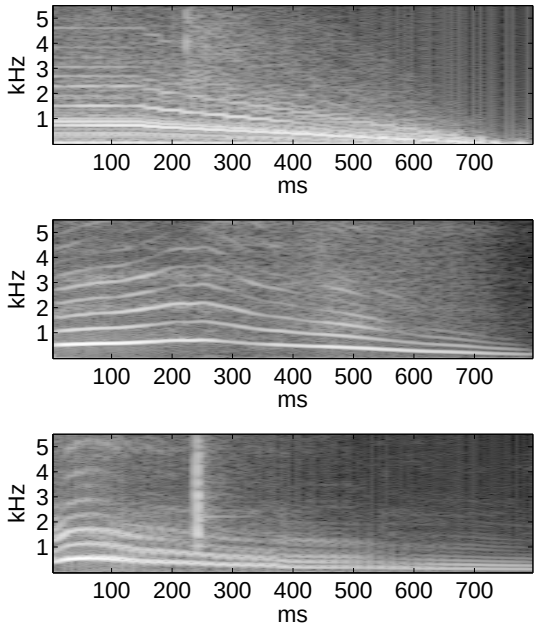

Figure 11: *Example of imitations labeled as class `down`. Topmost panel is one of the two reference stimuli for `down`. The following two panels are imitations of the stimulus: both begin with a rising pitch, followed by the expected lowering; these imitations are likely to be recognized as class `up/down`.*

3.4. Discussion on the dataset

The dataset classes have already been introduced in Sec. 1.2. In the following we give details about the dataset definition which may have an impact on the recognition. The classes `up`, `down`, `up/down` and `stable` are all defined by their evolution in time; this is not the case for `impulse` and `repetition`, which are instead defined by their duration and rhythm, respectively. In other words, the stimuli are labeled according to different domains. Moreover `up` and `down` are semantically portions of `up/down`: it is thus likely to have confusion between these three classes.

There are several factors which lead to a strong intra-class variability. Each class is defined by 2 different reference sounds, and each of the 50 subjects provides many recordings. The imitations are done by non-experts: the same stimuli is thus imitated using different strategies, and sometimes the imitations contents can be conflicting.

There are, finally, some issues related to the content of the dataset, that is to the vocal imitations. The definition of the classes, although very clear, can not be translated straightforwardly into acoustic cues: there are examples in which the subject imitates a complex stimulus by rising the pitch and lowering the first formant (thus decreasing the spectral centroid). The imitation has thus ambiguous meaning because its descriptors will have opposite evolutions. Many recordings suffer from boundary effects: the imitator could begin to produce sounds during the preparation phase of the articulation, introducing spurious effects at the begin of the recording. Similar artifacts can also be spotted at the end of some sounds.

The recognition system therefore has to cope with all the exposed issues of the dataset. The descriptors have to provide the main cues of the signal even in case of noisy or unreliable underlying time series (loudness, spectral centroid, etc.). This situation motivates the use of different descriptors with similar meanings, in order to reinforce the extracted information.

4. CONCLUSION

A dataset has been compiled, in the context of the SkAT-VG project, in which a large number of subjects have imitated sounds which belong to six morphological categories.

We have presented three methodologies to automatically recognize these morphological categories.

Our base-line system uses Local Trend descriptors, which are designed to measure the local behavior of the time serie. The descriptors are then modeled by continuous hidden Markov models with 4 states. The system reached an accuracy of 55.1%. An improved approach is based on Global Trend descriptors, which express the evolution of the signal along its whole duration. This second set of descriptors is modeled by discrete hidden Markov models, using 4 states and 4 emitted symbols. The obtained accuracy is 70.8%. Our third approach for the automatic recognition is based on the Morphological descriptors, which are designed to give a compact representation of the time series evolution. These descriptors do not need temporal modeling, such as HMM, and have low dimensionality. The classification is thus done using the k-Nearest Neighbor algorithm. This system has provided the best recognition accuracy of 83.6%.

The automatic classification on the given dataset proved to be a non-trivial task, despite the apparently clear definition of the classes. The manual check of many recordings within the dataset has confirmed a degree of confusion between certain classes, due to objective spectrograms similarities. A notable example is the pair `down` and `up/down`. Moreover the dataset is made by imitations of reference sounds, and imitators use different strategies to render the same stimulus: this rises the intra-class variability.

The non-optimal recognitions results are in part justified by these findings.

The presented set of Morphological descriptors goes toward the characterization of sounds by their acoustic *shape*. Future work will involve the organization of the descriptors in a systematic topology, similarly to [15]; this may point out shortcomings of the proposed set. Adding new descriptors could hence encompass a broader set of signal categories.

The proposed Morphological descriptors could be applied to other datasets, with classes defined in different ways. The integration of the Morphological descriptors with other, more traditional, ones is another topic of future studies, fostering recognition accuracy in many contexts.

5. ACKNOWLEDGMENTS

This work was supported by the 7th Framework Programme of the European Union (FP7-ICT-2013-C FET-Future Emerging Technologies) under grant agreement 618067. The audio dataset as well as the choice of the categories and the annotations used in this paper have been created by the Perception and Sound Design team of IRCAM. We are grateful to the reviewers and to S. Perry for their valuable contributions and corrections.

6. REFERENCES

[1] G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.

[2] G. Lemaitre, A. Dessein, K. Aura, and P. Susini, "Do vocal imitations enable the identification of the imitated sounds?," in *Proceedings of the 8^{th} annual Auditory Perception, Cognition and Action Meeting (APCAM 2009)*, Boston, MA.

[3] G. Lemaitre, A. Dessein, P. Susini, and K. Aura, "Vocal imitations and the identification of sound events," *Ecological Psychology*, vol. 23, no. 4, pp. 267–307, 2011.

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

[5] Guillaume Lemaitre and Davide Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.

[6] Lawrence Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, New York: Marcel Dekker, 2000.

[8] Lawrence R. Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, vol. 14, PTR Prentice Hall Englewood Cliffs, 1993.

[9] Biing-Hwang Juang and Lawrence R Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, vol. 1, 2005.

[10] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[11] George Saon and Jen-Tzung Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 18–33, 2012.

[12] Kazushi Ishihara, Tomohiro Nakatani, Tetsuya Ogata, and Hiroshi G Okuno, "Automatic sound-imitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes," in *PRICAI 2004: Trends in Artificial Intelligence*, pp. 909–918. Springer, 2004.

[13] Shiva Sundaram and Shrikanth Narayanan, "Vector-based representation and clustering of audio using onomatopoeia words," in *Proceedings of the American Association for Artificial Intelligence (AAAI) symposium series. Arlington, VA*, 2006.

[14] S. Sundaram and S. Narayanan, "Classification of sound clips by two schemes: Using onomatopoeia and semantic labels," in *Multimedia and Expo, 2008 IEEE International Conference on*, June 2008, pp. 1341–1344.

[15] P. Schaeffer, *Trait des objets musicaux*, Paris: Seuil, 1966.

[16] Julien Ricard and Perfecto Herrera, "Morphological sound description: Computational model and usability evaluation," in *Audio Engineering Society Convention 116*. Audio Engineering Society, 2004.

[17] Geoffroy Peeters and Emmanuel Deruty, "Sound indexing using morphological description," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 675–687, March 2010.

[18] Marijn Huijbregts and Franciska de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[19] M. Ito and Robert W. Donaldson, "Zero-crossing measurements for analysis and recognition of speech sounds," *Audio and Electroacoustics, IEEE Transactions on*, vol. 19, no. 3, pp. 235–242, 1971.

[20] Javier Ramırez, José C. Segura, Carmen Benıtez, Angel De La Torre, and Antonio Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[21] Geoffroy Peeters, "A large set of audio features for sound description (similarity and classification) in the Cuidado project," Cuidado project report, IRCAM, 2004.

[22] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 124, pp. 1638–1652, 2008.

[23] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 135–141, Apr 1974.

[24] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag, 1976.