Syllabling on instrument imitation: case study and computational segmentation method

Jordi Janer
Music Technology Group, Pompeu Fabra University, Barcelona
jjaner at iua.upf.edu - http://www.mtg.upf.edu

Alicia Peñalba
Department of Musicology, University of Valladolid
apenalba at mpc.uva.es - http://www.uva.es

In: K. Maimets-Volk, R. Parncutt, M. Marin & J. Ross (Eds.)
Proceedings of the third Conference on Interdisciplinary Musicology (CIM07)
Tallinn, Estonia, 15-19 August 2007, http://www-gewi.uni-graz.at/cim07/

Background in Audio Processing. Voice has been widely studied in the Audio Processing field, where researchers have principally addressed jointly analysis/synthesis with the aim of creating human-like singing synthesis. However, an appropriate analysis of the voice signal, let us also explore other musical uses. One example are voice-controlled digital synthesizers that use nonsense syllables. Looking at the literature, Sundberg [1] has studied the use of syllables in spontaneous nonsense text singing, which revealed some points about the choice of syllables in syllabling. In the context of Music Information Retrieval, studies such as [2] addressed the use of syllables in Query by Humming systems. Concerning automatic syllabling analysis, we should refer to research in Automatic Speech Recognition (ASR), where reliable systems use Machine Learning, combining acoustic models with language models.

Background in Music history, ethnomusicology and education. Nonsense text singing is often referred as voice instrumental. The voice can represent the aesthetic goal itself where all the possibilities of the voice are used with expressive intentions. Some manifestations are found in contemporary classical music, such as Carl Orff´s use of the voice and Arnold Schoenberg´s ”Sprechstimme”. Luciano Berio and Steve Reich used the voice in aleatory music. In traditional cultures, nonsense voice is used in Carnatic music of South India, Tuvinian throat singing and Hasidic Jews in ”nigunim”. Popular music, mainly jazz, uses the voice as an instrument, such as famous Louis Armstrongs and Ella Fitzgerald´s ”scat singing”[3] and hip-hop uses ”beatboxing” which involves creating beats, rhythms, vocal scratching and melodies using the human voice. The voice can also be employed to imitate an instrument in pedagogy such as in teaching percussion in Cuban music [4] and ”Vayttari” Indian music[5], [8] where a set of syllable commands are used in the pedagogy of percussion.

Aims. Our aim is to extend the research on syllabling toward a phonetical code for music articulation. This can be applied later, as suggested, to voice-driven synthesis. We analyze the syllabling produced in an educational context. More precisely, when the teacher gives indications to students by singing, mimicking instrument articulations with nonsense syllables. The experiment data are manually annotated recordings of several master classes, covering various instruments. An additional goal is to develop a computational method for an automatic segmentation of syllabling.

Main contribution A first part studies the relationship of syllable choice and the type of rhythmic articulation it tries to imitate in clarinet classes of different teachers. Despite cultural differences, we found some constants. Phonemes are found to have different functions depending on the position and the type. Most syllables start with a consonant to determine the articulation: [ta] for normal attack, [da] for a softer attack and [pa] for staccato. The most common vowel is [a]; [i] is used in high pitch sound and [o] in dark timbers. At the end of the syllable, [m] is used in long sounds. In the second part, we analyzed syllabling for other instruments. Our data set consisted of 604 syllables from 82 recordings. Results indicate that regardless of instrument and subject, a small number of syllables is used. The most uttered syllable is [ta](20%), used at the beginning of note group or in staccato articulations, followed by the syllable [ra](16%). The second contribution is the implementation of an automatic syllable segmentation system based on acoustic signal analysis. An acoustic front-end extracts low-level descriptors, from which a segmentation module uses heuristic rules to perform syllable segmentation and phonetic-based classification.

Implications. Digital instruments can be controlled by a wide variety of interfaces. By studying the syllabling in instrument imitation, we intend to build accurate tools to control DMI’s by exploiting human voice nuances. In the process of designing voice-driven synthesizers, musicology is needed in signal processing methods to define appropriate mappings from the voice descriptors to the synthesized instrument sound. There are constants in culture ways of imitating instruments and the examples presented in this paper so we can approach to more intuitive ways of mapping.

Syllabling is nonsense text singing that is widespread over cultures. This paper describes a case study of syllabling on instrument imitation. An additional interest of this study is to apply the results to improve voice-driven synthesizers. The presented case study addresses syllabling on instrument imitation in an educational context.

Introduction

Syllabling and instrument imitation

From an audio processing perspective, research on the singing voice has been principally directed toward the generation of a human-like artificial singing. In this paper, we combine audio processing techniques with fundamentals of musicology to study a very particular case of singing: syllabling in instrument imitation.

Our aim is to extend the research on syllabling toward a phonetical code for music articulation in instrument imitation. This can be applied later, as suggested, to voice-driven synthesis, as in [12].

To our knowledge, first scientific publications on syllabling refer to the work by [1]. This preliminary study looked at the choice of syllables in informal nonsense text singing for six short melodic excerpts.

As they report, most subjects used the same small set of syllables, being the syllable [da][1] employed in the 34% of the cases and the syllable [di] in the 13%. Another result of this study is that the voiced consonant [m] was used to infer micropauses, and was often followed by the voiceless stop consonant [p]. The reason of this later case was note grouping. In the action of syllable choicing, in addition to articulatory convenience, the study revealed that it also has a musical meaning.

Also in a context of Query-by-Humming, we find work related to syllabling, here referred as *Query-by-Voice*, e.g. [2]. In order to transcribe the voice query, they consider a syllable with the following structure: the onset (consonant), the nucleus (vowel) and the code (final consonant).

From another perspective, Patel and Iversen [8] identify coincident acoustic features in the sounds of Tabla drums form North India and the syllables used to imitate these sounds. They found that there are similarities in spectral centroid, rate of amplitude envelope decay, duration between the releases of consonants in a cluster, fundamental frequency and the influence of aspiration on the balance of low vs. high frequency energy in a vowel in eight vocables and their corresponding drum sounds. They also demonstrate that naïve listeners could match onomatopoeia and their corresponding drum sounds quite easily.

In contrast to Sundberg's study, where subjects where requested to sing a musical score with nonsense syllable, we analyze the syllabling produced in an educational context. More precisely, when the teacher gives indications to students by singing, mimicking instrument performances with nonsense syllables. The experiment data are manually annotated recordings of several master classes, covering various instruments. An additional aim is to develop a computational method for automatic syllabling analysis.

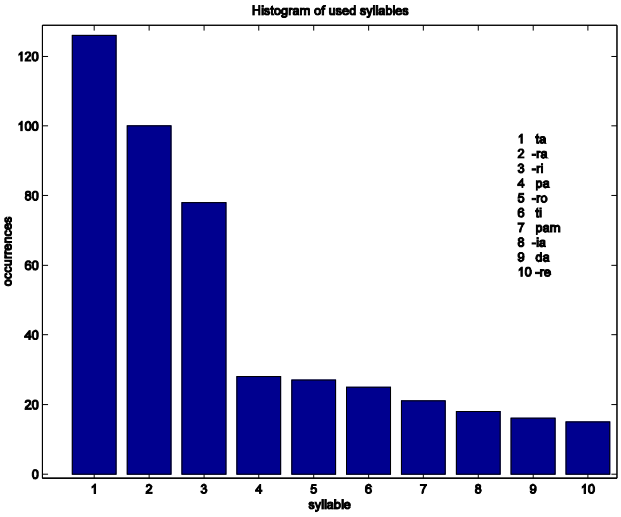

Figure 1. . Histogram of the choice of syllable, showing only the ten most used syllables in our data set with the corresponding transcription. Syllables:*1- ta, 2-ra, 3-ri, 4 pa, 5-ro, 6-ti, 7-pam, 8-ia, 9-da, 10-re*.

[1] Syllables are transcribed used the SAMPA that is a machine readable variant of the International Phonetic Alphabet. http://www.phon.ucl.ac.k/home/sampa/

Voice instrumental

Nonsense text singing is often referred as voice instrumental or onomatopeia. The voice can represent the aesthetic goal itself where all its possibilities are used with expressive intentions. Some manifestations are found in contemporary classical music, such as Carl Orff´s use of the voice and Arnold Schoenberg´s "Sprechstimme". Luciano Berio and Steve Reich used the voice in aleatory music. In traditional cultures, nonsense voice is used in Carnatic music of South India, Tuvinian throat singing and Hasidic Jews in "nigunim". Popular music, mainly jazz, uses the voice as an instrument, such as famous Louis Armstrongs and Ella Fitzgerald´s "scat singing"[3] and hip-hop uses "beatboxing" which involves creating beats, rhythms, vocal scratching and melodies using the human voice.

The voice can also be employed to imitate an instrument in pedagogy such as in teaching percussion in Cuban music [4] and "Vayttari" Indian music [5] [8] where a set of syllable commands are used in the pedagogy of percussion. Pekin opera [11] percussion sounds and Japanese Noh flute[9] are also characterized with some kind of speech sound symbolism.

Case study

Description of the experiment

Instrument imitation with voice is as a multifaceted topic, which might encompass areas such as musical acoustics, musicology or phonetics. Also from a social and cultural point of view, it has its significance since most people have in some occasion imitated a musical instrument by singing.

Being aware of this, we have to stress that this is a preliminar and rather constrained study. Many aspects remained unaddressed here, for instance, cultural differences in the choice of syllables, both within western traditions and compared to non-western traditions.

This section presents two complementary case studies that refer to syllabling on instrument imitation. The first looks at the phonetics employed by performance teachers in an educational context.

Experiment data

The experiment data consist of manually annotated recordings of several master classes, covering various instruments (see Table 1). Our data set consisted of 82 recordings with a total number of 604 syllables. The annotation process consisted in transcribing the sung syllables.

For a first part of the study only a subset of clarinet imitations is used. For the second part, we made an analysis of the complete data set.

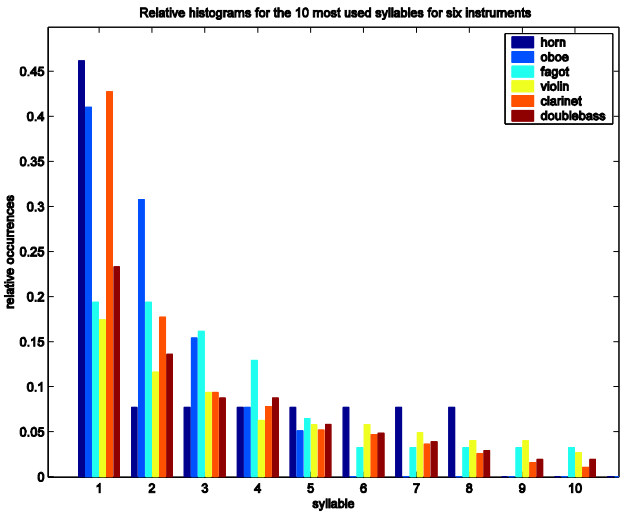

Figure 2. Relative histogram of syllable choice of six different instruments.

Instrument	num. instances	%
violin	224	37
clarinet	192	22.7
double bass	103	17.1
oboe	39	6.5
fagot	31	5.1
horn	13	2.2

Table 1. Strucutre of the dataset of our study, with the percentage of the master class recordings for different instruments.

Results

Results derive from two distinct analysis. A first part studies the relationship of syllable choice and the type of rhythmic articulation it tries to imitate in clarinet classes of different teachers. In this case, only a subset of recordings are considered. Despite cultural differences, we found some constants. Phonemes are found to have different functions depending on the position and the type.

Imitation tries to reproduce grouping of musical sounds. In each group we found that phomens were set in different positions. Most syllables start with a consonant. [ta], [da] and less current [pa]. They define articulation; [ta] for normal attack, [da] for softer attack and [pa] for staccato. Vowels constitute the center of the syllable. [a] is the most common choice, however [i] and [o] can also be found; the former in high pitches and the latter in dark timbers. At the end of the syllables [m] is used for long sounds resonance. When two or more syllables are linked in groups, [ra] is often used in quick linking ("tara tara"). The [r] is a linguapalatal liquid. In the one hand, the explanation of its use is that it does not stop the air and on the other hand, because of the point of articulation, it allows velocity for the tongue, as it happens in double and triple tounguing with consonants [k] or [g] in wind instruments such as the flute.

In the second part of the study, we analyzed recordings of syllabling in several instruments master classes. Results indicate that regardless of instrument and subject, a small number of syllables is used. The most uttered syllable is [ta](20%), used at the beginning of note group or in staccato articulations. It is followed by the syllable [ra](16%).

From a qualitative analysis of these results, we extract two conclusions. First, that a reduced set of syllables are used, as depicted in Figure 1. Second, by looking at Figure 2, we can conclude that there is a constant in the number of syllables used for different instruments of our data set.

From this analysis, we find interesting commonalities with phonetics. It occurs, for instance, in the choice of the vowel [i] for high pitched notes. In fact, this relates to the concept of *intrinsic pitch*, widely known in phonetics [13]. Intrinsic pitch is the average fundamental frequency for a given vowel. Studies demonstrate that, it has a frequency of 186 Hz for the vowel [i], which is around two semitones higher than the intrinsic pitch for the vowel [a]. This would explain that the [i] vowel is chosen unconsciously for reaching high pitches while saving effort.

Discussion

Results indicate that regardless of instrument and subject, a small number of syllables is used. The most uttered syllable is [ta](20%), used at the beginning of note group or in staccato articulations, followed by the syllable [ra](16%) used as shown before in groupings. Nevertheless, attending to the instruments, there are some exceptions. For example, we observed that horn does not use [ra] [ri] and [ti] as much as the others. However, horn uses [ia] and [pam] much more that usual. Despite these results we cannot extract any conclusion yet as they might depend on the type of the pieces analysed.

Survey on instrument imitation

The second, and still ongoing, experiment consists of a web survey [2], in which subjects are requested to transcribe the phonetics used to imitate a set of musical phrases of three instruments: bass guitar, saxophone and violin. Participants input the transcription of their imitation, typing the answer on a web questionnaire. Results are analyzed by means of text processing techniques.

Our goal is to identify quantitatively, to only which syllables are most used in instrument imitation, but also whether the results are subject-dependent or instrument-dependent.

Automatic Segmentation

[2] This web survey will be opened until the end of the CIM'2007 conference (August 2007). The results will be publicly available online after the conference on http://www.mtg.upf.edu/~jjaner/scatit

The second contribution is the implementation of an automatic syllable segmentation method based on acoustic signal analysis. An acoustic front-end extracts low-level descriptors such as loudness, and mel-frequency cepstrum coefficients. Next, a segmentation module uses heuristic rules to perform syllable segmentation before phonetic classification.

In the context of instrument imitation, singing voice audio signal has a distinct characteristic in relation to traditional singing. Principal musical information involves pitch, dynamics and timing; and those are independent of the phonetics. The role of phonetics is reserved for determining articulation and timbre aspects. For the former, we will use phonetics changes to determine the boundaries of musical articulations. For the latter, phonetic aspects such as formant frequencies in vowel can be used to alter timbre in the synthesis (e.g. brightness). Unlike in speech recognition, a phoneme recognizer is not required and a more simple classification will fulfill our needs.

In Phonetics, phonems are classified attending to various aspects, e.g. from the acoustic properties of the articulatory gestures. A commonly accepted classification based on the acoustic characteristics consists of six broad phonetic classes [7]: vowels, semi-vowels, liquids and glides, nasals, plosive, and fricatives.

Nevertheless, we might consider a new phonetic classification that is better suited to the acoustic characteristics of voice signal in our particular context. As we have previously introduced, a reduced set of phonemes is mostly employed in syllabling. Furthermore, this set of phonemes tends to convey musical information. Vowels constitute the nucleus of a syllable, while some consonants are used in note onsets (i.e. note attacks) and nasals are mostly employed as codas. Our proposal envisages different phonetic categories resulting from a classification based on its musical function: attack, sustain, release, articulation (articuatory, ligature), other (additional).

imitation. This table comprises a reduced set of phonemes that are common in various languages.

Method description

Our method is based on heuristic rules. In a first stage, it looks at the timbre changes in the voice signal, segmenting it according to the phonetic classification mentioned before. In a second stage, it uses a state transition model that takes into account the behavior in instrument imitation. This process aims at locating phonetic boundaries on the syllabling signal. Each boundary will determine the transition to one of the categories showed in Table 2. This is a three steps process:

- Extraction of acoustic features.
- Computation of a probability for each phonetic class based on heuristic rules.
- Generation of a sequence of segments based an a transition model (see Fig. 4).

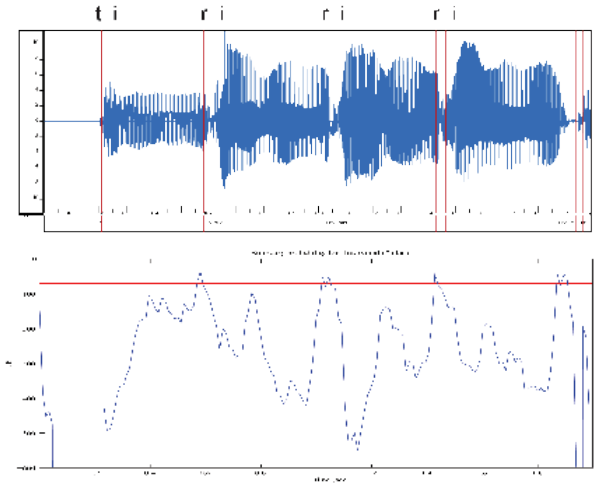

Figure 3. Syllabling Segmentation (from top to bottom): phonemes, waveform, labels and boundary probability for *articulation* class (horizontal line representing the threshold b_{thres}).

Acosutic features are computed frame by frame, with a window size of *1024* and a hop size of *512* samples at *44100 Hz*. This segmentation algorithm is designed for a real-time operation in low-latency conditions. The list of low-level features includes: *Energy*, *DeltaEnergy*, *Mel-Frequency Cepstral Coefficients* (MFCC), *DeltaMFCC*, *Pitch* and *Zero-crossing rate*. *DeltaMFCC* is computed as the sum of the absolute values of the MFCC coefficients derivative (13 coeffs.) with one frame delay.

From the acoustic features, we use a set of heuristic rules to calculate boundary probabilities for each phonetic class. Unlike for an offline processing, in a real-time situation, this algorithm is currently not able to distinguish between *Articulation* and *Release* phonetic classes.

We compute at each frame k a boundary probability for each phonetic class j, $p_j(x[k]) = p(B_j|x[k])$. At each frame, to decide if a boundary occurs, we take the maximum of all four probabilities ($p(B \mid x[k])$) and compare it to a empirically determined threshold b_{thres}.

Finally, in order to increase robustness when determining the phonetic class of each segment in a sequence of segments, we use a state transition model.

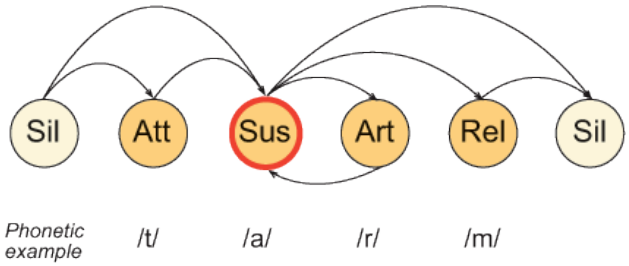

Figure 4. Model for the segment to segment transition for the different phonetic classes.

The underlying idea is that a note consists of an onset, a nucleus (vowel) and a coda. In addition, a group of notes can be articulated together, resembling *legato* articulations on musical instruments. Thus, we need to identify these grouped notes, often tied with liquids or glides. Figure 4 describes the model for the allowed boundary transitions.

Evaluation

With the proposed method, we are able to segment effectively phonetic changes and to describe a voice signal in the context of instrument imitation as a sequence of segments.

An evaluation of the algorithm was carried out by comparing automatic results with a manual annotated ground truth. The ground truth set consists of 94 syllabling recordings. Syllabling examples were voice imitations by four subjects of sax recordings with an average duration of *4.3 sec*. For the evaluation, we consider *vowel onsets*, since it corresponds to the beginning of a musical note. The averaged results for the complete collection is shown in Table 3.

	Mean	**Stdev**
Correct detections (%)	90.68	15.38
False positives (%)	13.99	52.86

Table 3. Averaged results of the onset detection compared to a ground-truth collection of 94 files. The average time deviations was *-4.88 ms*.

Conclusions

Musicians make use of syllabling in various situations. This paper has enlightned some properties of syllabling in an educational context. There is a clear link between phonemes, their acoustic properties and their musial function in syllabling. Based on this results, we have proposed an automatic segmentation algorithm. Due to its focused goal, it attains better results than general-purpose onset detection algorithms and query-by-humming front-ends.

Finally, we would like to stress the interest of this work in further research on voice-driven musical synthesis. Musicological aspects help signal processing methods in defining appropriate mappings from the voice features to the final synthesized instrument sound.

Acknowledgments. This research has been partially supported by the ***e-Content plus project VARIAZIONI***[3]. Authors would also like to thank all participants in the syllabling recordings and Uli Reich at the Universität zu Köln for his comments.

References

[1] Sundberg, J. (1994), *Musical Significance of Musicians' Syllable Choice in Improvised Nonsense Text Singing: A Preliminary Study,* Phonetica, vol. 54,.

[2] Lesaffre, M. et al. (2003), *The MAMI query-by-voice experiment: collecting and annotating vocal queries for music information retrieval*. Proceedings of ISMIR 2003, London.

[3] Kernfield, B. (1988*). The New Grove Dictionary of Jazz. 2 vols.*, New York: Grove Dictionaries of Music, Inc.

[4] Gómez, Z. and Eli, V. (1995). *Música latinoamericana y caribeña*. La Habana: Pueblo y Educación.

[5] Hitchcock, H. (1986). Wiley and Stanley Sadie, eds. *The New Grove Dictionary of American Music*. 4 vols. New York: Grove's Dictionary of Music, 1986.

[6] Touliatos, D. (1989), *Nonsense Syllables in the Music of the Ancient Greek and Byzantine Traditions*", Journal of Musicology, Vol. 7, No. 2 (Spring, 1989)

[7] Lieberman, P. and Blumstein S.E. (1986). *Speech physiology, speech perception, and acoustic phonetics*, Cambridge University Press.

[8] Patel, A.D. & Iversen, J.R. (2003). *Acoustic and perceptual comparison of speech and drum sounds in the North Indian tabla tradition: an empirical study of sound symbolism.* Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 2003, pp.925-928.

[9] Hughes, D. (2000). *No nonsense: the logic and power of acoustic-iconic mnemonic systems*, British Journal of Ethnomusicology, vol. 9, no. 2, pp. 95 122, 2000.

[10] Kippen, J. (1988). *The Tabla of Lucknow: A cultural analysis of a musical tradition*, Cambridge: Cambridge University Press, 1988.

[11] Li, D. (2001). *Onomatopoeia and Beyond: a Study of the Luogu Jing of the Beijing Opera*, Ph.D. dissertation,UCLA, 2001.

[12] Janer, J. (2005). *Voice-controlled plucked bass guitar through two synthesis techniques*, Proceedings of 2005 International Conference on New Interfaces for Musical Expression; Vancouver, Canada.

[13] Laver, J. (1994). *Principles of Phonetics*, Cambridge University Press, 1994.

3 http://www.variazioniproject.org