INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Expressive control of singing voice synthesis using musical contexts and a parametric *F0* model

Luc Ardaillon, Celine Chabot-Canet, Axel Roebel

IRCAM - UMR STMS (IRCAM - CNRS - Sorbonne Universités)
Paris, France

luc.ardaillon@ircam.fr, celine.chabot-canet@univ-lyon2.fr, axel.roebel@ircam.fr

Abstract

Expressive singing voice synthesis requires an appropriate control of both prosodic and timbral aspects. While it is desirable to have an intuitive control over the expressive parameters, synthesis systems should be able to produce convincing results directly from a score. As countless interpretations of a same score are possible, the system should also target a particular singing style, which implies to mimic the various strategies used by different singers. Among the control parameters involved, the pitch (*F0*) should be modeled in priority. In previous work, a parametric *F0* model with intuitive controls has been proposed, but no automatic way to choose the model parameters was given. In the present work, we propose a new approach for modeling singing style, based on parametric templates selection. In this approach, the *F0* parameters and phonemes durations are extracted from annotated recordings, along with a rich description of contextual informations, and stored to form a database of parametric templates. This database is then used to build a model of the singing style using decision-trees. At the synthesis stage, appropriate parameters are then selected according to the target contexts. The results produced by this approach have been evaluated by means of a listening test.

Index Terms: singing voice synthesis, singing style, *F0* model

1. Introduction

While state-of-the-art singing voice synthesis systems (SVS) already provide a sufficient quality for a use in creative and commercial applications [1], synthesized voices lack of expressivity compared to real voices. SVS, beyond generating a natural-sounding voice timbre, must also reproduce the various expressive intentions of real singers. For this purpose, an appropriate control model is required, which, from an input score and text, generates all the synthesis control parameters. The main parameters to be considered are : fundamental frequency (*F0*), intensity, phonemes durations, and voice quality. Other style-specific characteristics such as rhythmical variations or ornemental notes are also important features, but can be described in the symbolic domain, and should thus better be handled at the score level.

Although each of those parameters should be considered, the *F0*, which conveys the melody and many stylistic characteristics [2, 3, 4, 5], should be modeled first. The main interest in using SVS is to give the user a complete control over the synthesis. For this reason, it is advantageous to use a parametric model allowing to characterize expressive fluctuations of the *F0* in an intuitive way. In [6], such a model has been proposed. A first study showed that this model can produce synthetic *F0* curves comparable to real ones, for various singing styles. However, no automatic way for choosing the parameters was given.

Various approaches for the control of the *F0* for singing voice synthesis exist. Rule systems such as [7] benefit from musical knowledge, and can be progressively improved in an analysis-by-synthesis procedure, but are not flexible, as thorough studies are necessary to define rules for each singing style. Conversely, HMM-based approaches [8] can model new singing styles by choosing an appropriate learning database, without requiring specific knowledge. Such systems also allow a high-level control of the synthesis (e.g. interpolating between singing styles), but don't provide any local control of the expressivity. This approach also requires an important quantity of learning data. The unit-selection approach in [9] allows, by using real contours extracted from recordings, to model a singing style with few data, while avoiding oversmoothing problems related to the statistical modelisation of HMMs. But this method also doesn't provide any control for modifying the result. In [10], the authors propose to model expressive *F0* variations by selecting parametrized templates (called "vocal expressions") in a library of examples extracted from commercial recordings. In [9] and [10], units and templates are chosen according to the similarity between their original contexts and the target contexts of the synthesis score, using cost functions or pre-defined rules. Thus, only a restricted and fixed set of contextual informations is used. But their relative importance may vary from one singer to another, which can't be considered with those approaches.

We propose here a new approach based on the selection of parametrized *F0* templates for generating expressive interpretations and model singing styles. In order to include in the modelization a rich contextual description, and adapt to the variable importance of those contexts in the interpretative choices of singers, we propose to use a context-clustering technique such as used in HMM-based methods [11, 12], using decision trees. In a first learning stage, the model parameters are extracted from recordings for a given singing style, along with their associated contexts, to form a database of parametrized *F0* templates. Decision trees are then built in order to automatically organize those templates according to their original contexts. For the synthesis, those trees are then used to choose in the database the more appropriate parameters, according to the target contexts. This approach is also used to predict the phonemes durations for the synthesis. A listening test has been conducted in order to assess the first results obtained by the proposed method.

This article is organized as follows : In section 2, the used *F0* model is presented, and the procedure for extracting the parameters from recordings is explained in section 3. Section 4 and 5 present the methods used to build the style models and

Research supported by the ANR project "ChaNTeR" (ANR-13-CORD-011).

http://dx.doi.org/10.21437/Interspeech.2016-1317

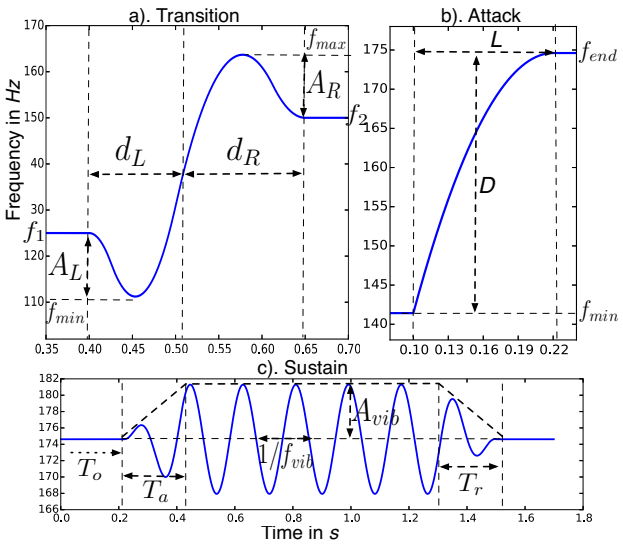

Figure 1: *Parametrization of the F0 model segments*

choose the parameters for the synthesis. The results of the evaluation are then discussed in section 6, and our conclusions are presented in section 7.

2. F0 Model

In [6], a parametric *F0* model for singing voice has been described, which decomposes the *F0* curve into several additive layers : a melodic and vibrato layer (modelling the expressive componants of the *F0*), as well as a jitter and a micro-prosodic layer (modelling the uncontrolled variations induced by the voice mechanism). The melodic layer is temporally segmented into 5 elementary segments : silences, attacks, sustains, transitions, and releases. Figure 1 summerizes the parameterization of those segments. Attacks and releases are parametrized by their length L (s) and depth D (cents). The transitions are controlled by 4 parameters : d_L and d_R (*s*) which set the durations of the left and right parts of the transition, and A_L and A_R (*cents*) which set the amplitudes of possible inflexions on the left ("preparation") and on the right ("overshoot"). For sustain segments, the vibrato is characterized by its frequency f_{vib} (*Hz*) and an Attack-Sustain-Release (ASR) amplitude curve with a global amplitude A_{vib} (*cents*), an attack time T_a (*s*), a release time T_r, and an offset time T_o to allow a delay between the start of a note and the start of the vibrato.

In order to generate the F0 curve, the sequence of those segments is first determined from the notes of the score, and their parameters are fixed. The curve is then generated using B-splines whose knots positions and weights are deduced from those parameters, as explained in [6].

3. Model parameters extraction

The first step in modelling a singing style is to extract the model parameters from recordings. For this purpose, the *F0* curve of a song is first segmented, in a semi-automatic procedure, into the elementary units of the model (attacks, sustains, transitions, releases, and silences). Conversely to [9], we want to model the singing styles from commercial recordings, with original lyrics. Thus, in order to reduce the impact of the micro-prosody on the parameters estimations, a linear interpolation of the *F0* is done on each voiced fricative and plosive, using the phonetic segmentation. The curve is then smoothed by low-pass filtering before extracting the parameters of each *F0* segment.

For attacks and releases, D is computed as the distance, in *cents*, between the minimum value f_{min}, and the final (respectively initial) value f_{end} (resp. f_{start}): $D = 1200 \cdot log_2(\frac{f_{min}}{f_{end}})$. The length L is given by the *F0* segmentation.

For sustains segments, a low-pass filtering is first applied, to isolate the vibrato and center it around 0 by subtracting the filtered curve to the original. For this purpose, an FIR filter is used, whose coefficients are made of a hanning window of size $2 \cdot sr \cdot T_{max}$ as in [13], where sr is the curve's sampling frequency, and T_{max} is the maximal presumed period of the vibrato. The extremums of the vibrato curve are then extracted, and an ASR amplitude profile is fitted to those extremums, using a grid-search procedure. The vibrato periods are computed from the distance between the extremums, and the vibrato frequency f_{vib} is set as the inverse of the mean period. For long notes, only the central third of the vibrato cycles are used.

For transition segments, the middle of each transition is determined by the extremum of the derivative, in the segment, with a sign matching the transition's direction. From the *F0* segmentation, this central point allows to determine the length of the left part d_L and right part d_R of the transition. The amplitudes of the preparation and overshoot A_L and A_R are computed as the distance, in *cents*, between the extremum frequency f_{min} (respectively f_{max}), on each side of the transition, and the frequency f_1 (respectively f_2) at the segment's boundary.

4. Modeling singing styles

In [14], the authors suggest that "personal style is conveyed by repeated patterns of [some] features occurring at characteristic locations". Modeling a singing style would thus imply being able to capture these features along with the characteristic contexts where they occur. Figure 2 gives an overview of the proposed approach. In the following section, we first describe the corpus we used in this study. Then, the contextual informations used in our modelization are presented, and the construction of the style models is explained.

4.1. Corpus and annotations

In order to study and model different singing styles, a corpus has been created with the help of a musicologist. In order to benefit from well-know cultural references and previous musicological studies [3], we chose to base our work on commercial recordings of singers representative of different styles, rather than making new dedicated recordings from unknown singers. Our system primarily targeting the French language, we selected 4 famous French singers from the 20th century: Edith Piaf, Sacha Distel, Juliette Greco, and François Leroux. The choice of those 4 singers has been encouraged by the fact that they all recorded an interpretation of the same song *"Les feuilles mortes"* ("*Autumn leaves*" in English) in their own singing style, offering us a common reference for comparison.

For each singer, a single song (of 2 to 3.5 minutes), different from "*Les feuilles mortes*", has been chosen and annotated. The annotations contain : a phonemes segmentation, the *F0* curve of the voice, and a midi annotation aligned on the recording. Similarly to [10], the recordings used include instrumental accompaniement, and the annotations have thus been manually corrected, using the audiosculpt software [15], in order to obtain reliable data. The *F0* correction can be done by drawing over the most visible voice's harmonics on the spectrogram.

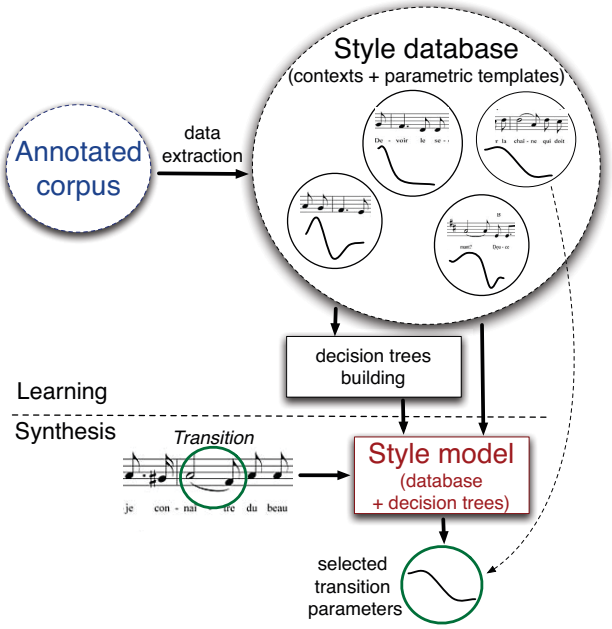

Figure 2: *Overview of the proposed method*

4.2. Contexts used

In current approaches, the contexts used are often limited to the current and neighbouring notes. It seemed important to us, though, to take into account also wider contextual informations, at the phrase level, which may influence on the interpretative choices of the singers. The contexts considered in this work for the modelization of the *F0* are listed below :

- Current note: pitch; duration; consonants' length (for transitions); vowel's length (for sustains); caducity (see below);
- Previous and next notes: pitch; duration; interval and duration's difference with current note; caducity;
- Phrase: temporal (first, last or penultimate note) and melodic (highest or lowest note, peak or valley) positions of current and neighbouring notes in the musical phrase.

This list has been established with the help of a musicologist, from empirical observations. The caducity of a note is a French concept, defined by a syllable usually not prononced in speech, sung at the end of a word. We define a musical phrase as the set of notes comprised between 2 rests. Other contexts, such as the position of a note inside a bar, are currently not considered as they cannot be deduced from the available annotations.

For modelling the phonemes' durations, some additionnal phonetic-related contexts are also considered:

- The number of successive consonants in the note
- The position of the considered consonant in a consonants' cluster (first, last, or central)
- The identity of the previous, current, and next phonemes

4.3. Decision-tree based learning

As we try to model a singing style from a single song, HMM-based approaches are not suitable. The authors in [8] use a database of 25 songs, and 60 in [12]. The unit or parametric template selection-based methods used in [9] and [10], requiring fewer data, seem thus more appropriate in our case. With such approaches, the units are chosen according to their original contexts, which should match at best the target contexts. But the notion of "closest context" implies a distance relationship, which is difficult to define, because of the variable dimensions and relative importances of the contexts. [9] and [10] base their selection on cost functions and empirically-defined rules, which are not very flexible for including new contexts or taking into account the variable importance of some contexts from one style to another. Another solution, in the case of a parametric modelization, is to use the "best" contexts based on their influence on the choice of the model parameters. For this purpose, we use here a context-clustering method based on decision trees, such as those used in HMM-based approaches for speaking [16] or singing [8] voice.

After extracting the parameters and phonemes durations from the corpus, we store them in a database along with their original contexts, as illustrated in figure 2 for transitions. From this database, some decision-trees are built using the CART algorithm [17]. The criterion used for choosing the question at each node of the decision trees is the minimization of the mean squared error [18]. Thus, the databases constituted for each singing style, along with the trees built from those databases, form our models of the singing styles.

4.3.1. F0 model

For a better coherence between the *F0* model parameters of each segment, they are not considered independentely, but are tied together as parametrized templates. Thus, a multi-output decision tree [19] is built for each segment's type. All contexts listed above for the *F0* are used for building the trees, except for "attacks" and "releases" segments for which only the pitch and length of the current note (respectively first and last note of a musical phrase) are considered. For transition segments, the note considered as current is the one at the right of the transition. Also, unvoiced transitions, as they don't have a continuous *F0* curve to fit the model parameters, are left apart. Finally, model parameters having different dimensions (lengths in *s*, amplitudes in *cents*, and frequency in *Hz*), they are first normalized by their maximum value so that all values lie in the same 0-1 range, to avoid favoring one parameter over another. In order to avoid possible overfitting when constructing the trees, a stopping criteria has to be defined. After some tests, we chose to keep at least 5% of the total number of the learned segments, with a minimum of 2, in each tree leaf.

4.3.2. Phonemes durations model

Besides the F0, consonants durations vary in a non-linear way with the tempo and are also often used by singers as an expressive mean to accentuate some notes. In order to model those effects, the same method is applied. For this purpose, the durations are directly extracted from the phonetic segmentations of the corpus. As we can't expect to meet each phoneme in an important variety of contexts in a single song, phonemes are grouped by phonetic classes with other phonemes having similar articulatory characteristics (and thus durations), for a better contexts coverage. During the learning stage, a tree is built for each phonetic class. The phonemes classes we used are (with the corresponding phonemes in SAMPA notation): voiced fricatives (v,z,Z); unvoiced fricatives (f,s,S); voiced plosives (b,d,g); unvoiced plosives (p,t,k); nasals (m,n); semi-vowels (w,j,H); R; and l. The phoneme's identity then becomes itself a context used in the tree building for each class.

5. Synthesis

5.1. Choice of parameters

Consonants durations being part of the contexts used for the *F0* modelling, they need to be fixed first. For each consonant in the input phonetized text, the tree of the corresponding class is

read from its root to a leaf, following the path matching the target contexts. A value is then randomly selected in the database among those associated to the same leaf.

For the *F0*, the sequence of segments of the model is first determined according to the notes in the score, as explained in [6]. For each segment, the same procedure is then used to select a set of parameters. The random selection process allows to keep a part of variability in the generated interpretations.

5.2. Additional rules and corrections

The overshoot and preparation in transitions are usually merged with the micro-prosodic inflexions of consonants. To avoid unnatural placement of the transitions regarding the text pronounced, the following rules are applied :

- Upward transitions start at the time of the first consonant
- Downward transitions end on the vowel's onset

Once the parameters have been chosen and those rules applied, it is possible that some segments of the *F0* model overlap. In such case, the duration parameters of those segments are reduced so that the end of the left segment matches the start of the right segment. Also, the total consonants' durations for a given note is limited to 85% of the note's length.

6. Evaluation

The proposed approach has been implemented in the synthesis system described in [6]. A singing style model has been built for each of the four singers of our corpus, and a subjective evaluation has been conducted by means of a listening test.

6.1. Test design

The main goal of our work being the modelization of singing styles, a first test aimed at measuring the recognition rate of the style of synthesized singing. For this purpose, the chorus of *"Les feuilles mortes"* has been split into 4 parts. For each part, the original interpretations in 2 styles were first presented. Then, a synthesis produced using one of the 2 corresponding style models was presented, and the user had to guess which style was used for the synthesis, in an ABX testing procedure.

A second test was designed to assess the gain in expressivity when using the proposed approach to predict parameter values from contexts, compared to a default configuration using only the mean values of a singing style. The listeners were asked to compare those 2 configurations, presented in random order, and rate their preferred interpretation on a 0-3 scale, based on the *perceived expressivity* (also defined as *liveliness*, or *musicality*) of the synthesis, in a standard CMOS procedure [20].

The 2 tests have been conducted on 22 participants listening with headphones or earphones through a web interface. The 2 female and the 2 male styles of our corpus were used for both tests. As the random selection process may lead to different interpretations for the style models, each synthesis was run 2 times and the one judged as best was kept. The sounds used in this evaluation can be found on the web page at the url [21].

6.2. Results and discussion

The results of the first test are shown in table 1. The overall mean recognition rate is only 58.9% but gets up to 76.3% for the *Piaf* style, which suggests that the *F0* variations and/or phonemes durations are characteristic features for this singing style, while those features are not sufficient to recognize well other styles. The percentages of good answers for the presented A-B pairs of styles are also given in the second row, and the p-value obtained for the *Piaf-Greco* pair indicates a significant result in regard to the random hypothesis. The different original interpretations of the song presented for this test had important rhythm and pitch differences, which may also help to explain the low rating obtained for the *Leroux-Distel* pair, as it was difficult for the listeners to focus only on the *F0* and phonemes durations, although we used an average score for the synthesis to avoid favoring one style due to similarities in the score itself.

style	all	Leroux	Distel	Piaf	Greco
recognition rate	58.9%	52.3%	55.9%	76.3%	54%
		54.2%		63.6%*	
p-value		0.363		0.024*	

Table 1: *Singing style recognition rates (*significant results)*

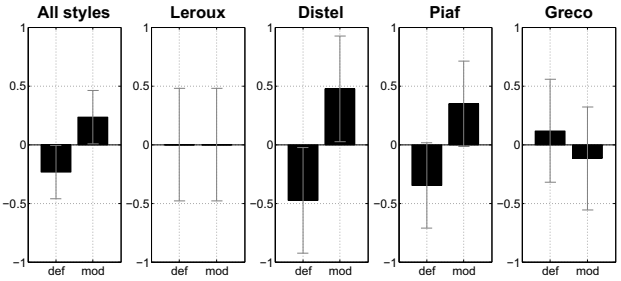

Figure 3: *CMOS for default settings (def) vs style models (mod)*

Figure 3 shows the result of the second CMOS test, with confidence intervals of 95%. Although not very strong, a positive tendency in favor of the proposed model is observed for the global result. However, the results for each style differ a lot. Especially, the preference is quite clear for the *Distel* and *Piaf* styles, while no significant difference is observed for the *Leroux* or *Greco* styles. One reason may be that some singers, like *Leroux*, who is a lyrical singer, use relatively constant parameters (steady vibrato, ...) which are thus easier to model using default values. We may also expect the results to get better using longer test sounds, as the repetition of the same parameters may become more obvious for the default configuration. But longer sounds are harder to memorize to assess the differences, and we chose to keep them relatively short (around 15*s*). The contexts of the learned song for each style may also not cover well all target contexts for the synthesis, and we can expect that the results improve by using more data.

However, considering that each singing style is built from a single song, and that we only model the *F0* and phonemes durations, many other stylistic aspects being left apart in this study, the obtained results are encouraging.

7. Conclusion

In this paper, we presented a new approach to singing-style modelization for expressive singing voice synthesis. The evaluation showed that this approach can, in some cases, improve the perceived expressivity of the synthesis by using contextual informations, compared to a default configuration, without additional input from the user. Although the overall recognition rate of the singing styles is relatively low, a significant rate was achieved for the *Piaf-Greco* pair, which shows that we managed to capture some of the stylistic characteristics allowing to discriminate those 2 styles. But the results also indicate that *F0* and phonemes durations are not sufficient to recognize a style very well, and encourages us to pursue our research by including more features in our modelization in future work. The methods presented have been used to participate to the Interspeech 2016 *"Fill-in-the-gap"* singing synthesis challenge.

8. References

[1] H. Kenmochi *et al.*, "Singing synthesis as a new musical instrument." in *ICASSP*, vol. 2012, 2012, pp. 5385–5388.

[2] T. Saitou and M. Goto, "Acoustic and perceptual effects of vocal training in amateur male singing." in *INTERSPEECH*. Citeseer, 2009, pp. 832–835.

[3] C. Chabot-Canet, "Interprétation, phrasé et rhétorique vocale dans la chanson française depuis 1950: expliciter l?indicible de la voix," Ph.D. dissertation, Université Louis Lumière-Lyon II, 2013.

[4] T. Kako, Y. Ohishi, H. Kameoka, K. Kashino, and K. Takeda, "Automatic identification for singing style based on sung melodic contour characterized in phase plane." in *ISMIR*. Citeseer, 2009, pp. 393–398.

[5] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 519–530, 2007.

[6] L. Ardaillon, G. Degottex, and A. Roebel, "A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls," in *Interspeech 2015*, Dresden, Germany, Sep. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01251898

[7] G. Berndtsson, "The kth rule system for singing synthesis," *Computer Music Journal*, vol. 20, no. 1, pp. 76–91, 1996.

[8] T. Nose, M. Kanemoto, T. Koriyama, and T. Kobayashi, "Hmm-based expressive singing voice synthesis with singing style control and robust pitch modeling," *Computer Speech & Language*, vol. 34, no. 1, pp. 308–322, 2015.

[9] M. Umbert, J. Bonada, and M. Blaauw, "Generating singing voice expression contours based on unit selection," in *Proc. SMAC*, 2013.

[10] Y. Ikemiya, K. Itoyama, and H. G. Okuno, "Transferring vocal expression of f0 contour using singing voice synthesizer," in *Modern Advances in Applied Intelligence*. Springer, 2014, pp. 250–259.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[12] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An hmm-based singing voice synthesis system." in *INTERSPEECH*, 2006.

[13] A. Roebel, S. Maller, and J. Contreras, "Transforming vibrato extent in monophonic sounds," in *International Conf on Digital Audio Effects*, 2011, pp. 1–1.

[14] C. Shih and G. Kochanski, "Prosody control for speaking and singing styles." in *INTERSPEECH*, 2001, pp. 669–672.

[15] C. Picasso, "Audiosculpt software," http://forumnet.ircam.fr/product/audiosculpt-en/, 2016.

[16] N. Obin, "Melos: Analysis and modelling of speech prosody and speaking style," Ph.D. dissertation, Université Pierre et Marie Curie-Paris VI, 2011.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees belmont," *CA: Wadsworth International Group*, 1984.

[18] "sklearn decision trees documentation," http://scikit-learn.org/stable/modules/tree.html, 2016.

[19] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[20] I. Recommendation, "1284-1: General methods for the subjective assessment of sound quality," *International Telecommunications Union, Geneva*, 2003.

[21] "demo page." [Online]. Available: http://recherche.ircam.fr/anasyn/ardaillon/IS2016/listTest/demo.php