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

Bertsokantari: a TTS based singing synthesis system

Eder del Blanco[1], *Inma Hernaez*[1], *Eva Navas*[1], *Xabier Sarasola*[1], *Daniel Erro*[1,2]

[1]AHOLAB Signal Processing Laboratory, UPV/EHU, Bilbao, Spain
[2]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
ederdelblanco@gmail.com, {inma.hernaez,eva.navas}@ehu.eus

Abstract

This paper describes the implementation of the Aholab entry for the Singing Synthesis Challenge: Fill-in the Gap. Our approach in this work makes use of an HTS based Text-to-Speech (TTS) synthesizer for Basque to generate the singing voice. The prosody related parameters provided by the TTS system for a spoken version of the score are modified to adapt them to the requirements of the music score concerning syllables duration and tone, while the spectral parameters are basically maintained. The paper describes the processing details developed to improve the quality of the output signal: the syllable timing, the generation of the intonation with vibrato and the manipulation of the model states. In this entry, the lyrics have been freely translated into Basque and the rhythm has been adapted to a Basque traditional rhythm.

Index Terms: speech synthesis, singing synthesis, human-computer interaction

1. Introduction

In the last years synthetic singing voice generation has raised a lot of research and commercial interest. Nowadays, as happens with speech [1, 2], mainly two main techniques are applied to generate the singing voice: unit selection synthesis [3] and statistical parametric synthesis [4]. Both techniques rely on a corpus, and the quality and variety of the recordings used to build the system have a critical influence on the final result. A good natural singing database which covers the whole spectrum of musical expression is thus needed to produce a pleasant synthetic singing voice [5]. To this day, such a database does not exist for Basque language. For this reason, in this work we show how a spoken database has been used to synthesize singing voice in Basque.

Transforming speech into singing is not trivial as sung and spoken voices exhibit important differences [6]. From the prosodic point of view, in singing voice the intonation is determined by the melody and rhythm specifications and not by the text structure or the characteristics of the language. Moreover, rhythm is synchronized with respect to vowel onsets [7] instead of the beginning/ending of the syllables. Regarding the phonetic content, vowels represent a high percentage of the acoustic content of the sung signal, and the presence of long sustained vocalic segments is frequent. As for the acoustic properties of the signal, the sung voice usually exhibits higher intensity with a suitable laryngeal phonation mode [8], and specific phenomena like vibrato [9] or the so-called singer's formant [10].

The system described in this paper is the initial outcome of an investigation carried out to characterize and be able to synthesize a traditional Basque singing style: *Bertsolaritza* [11]. Using our Basque TTS system as starting point, we have implemented several new functionalities to read music files, impose a specified rhythm and melody to the generated speech, and mimic some of the acoustic characteristics of singing voices. The resulting system, Bertsokantari, has been applied to synthesize a customized version of the song "Autumn Leaves", which has been submitted to the Singing Synthesis Challenge. The remainder of this paper is structured as follows. Section 2 introduces our Basque version of the score provided by the challenge organizers. Section 3 describes the general structure of the system and the modifications that have been made to account for the different characteristics of the sung voice. Finally, section 4 discusses the advantages and limitations of the current approach along with the future lines of work.

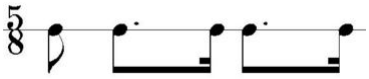

Figure 1: *Zortziko rhythm*

2. Basque Version of Autumn Leaves

The song "Autumn Leaves" has been adapted in order to fit it into a *zortziko* rhythm. This Basque word *zortziko*, which can be translated as *of eight*, refers nowadays mainly to a song written in a irregular 5/8 measure (see figure 1). A *zortziko* also describes a melodic unit composed by eight measures. Finally, the same word also refers to a stanza of eight verses very much used in *Bertsolaritza*, a popular improvised singing style with old tradition in the Basque Country [12]. Given that the authors are presently working on the development of a *Berstolaritza* database [11] and that the proposed score presents a regular eight measures structure, we performed the adaptation of the score to the zortziko rhythm. The Basque lyrics also corresponds to the particular rhythm of the *zortziko major* with 10 syllables at even lines and 8 syllables at odd lines (which is also the distribution of syllables in the English version).

3. Description of the synthesis system

3.1. System overview

'Bertsokantari' is a singing voice synthesis system based on the TTS system for Basque AhoTTS [13, 14]. It uses the song information contained in an XML music score to produce the synthetic singing signal. The general architecture of the system is shown in figure 2.

The main synthesis process is performed sentence by sentence, where a sentence is delimited either by an orthographic period found in the score, or by the musical rests. The text obtained from the score lyrics is sent to the linguistic processor

http://dx.doi.org/10.21437/Interspeech.2016-1123

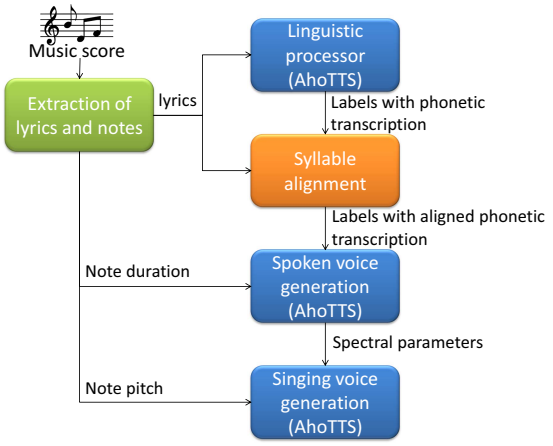

Figure 2: *Structure of Bertsokantari system.*

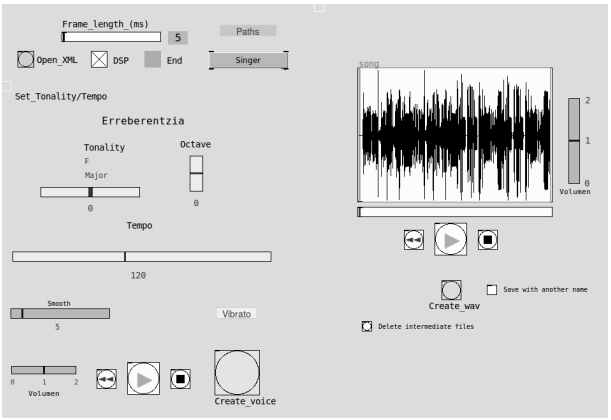

Figure 3: *User interface.*

of AhoTTS, where labels containing syllables and their corresponding pronunciations are produced. The syllable stream so produced must then be aligned with the score. In this way, three parallel streams are generated, containing pitch, duration and lyrics of every note. These streams are obtained for the whole song before proceeding with the synthesis process.

The waveform generation module of the system is based on the hidden semi Markov model (HSMM) based approach [15]: during training, the correspondence between text labels and acoustic features is modeled through HSMMs; during synthesis, a parameter generation engine [16] calculates the most likely acoustic feature trajectories given the input text labels. The specific acoustic features used by AhoTTS are those provided by the vocoder presented in [17], namely the logarithm of the fundamental frequency ($\log f_0$), a Mel-cepstral representation of the spectral envelope, and the so-called maximum voiced frequency (see [17] for details). Although Bertsokantari has been designed to use any AhoTTS-compatible voice as input, in this particular work HSMMs were trained from a speech database composed by 2000 short phonetically-balanced utterances spoken by a professional Basque male speaker. The sampling frequency of the recordings was 16 kHz, so Bertsokantari sings at 16 kHz sampling frequency too.

The synthetic singing voice is obtained in two steps: first, a spoken version of the score is obtained, with the correct rhythm but incorrect pitch (phone durations are imposed according to the input scores but there is no way to impose a pitch contour under the conventional parameter generation framework). Then the pitch stream obtained from the melody overwrites the spoken version stream, and after a post-processing of the Mel cepstral coefficients (details are given in section 3.5), all the parameters are sent to the vocoder to produce the final singing voice.

In the following sections, details about the most important components, procedures and settings are provided.

3.2. User interface

The user interface has been built using Pure Data [18] which has already been successfully used to control singing voice synthesis systems [19].

Through the user interface the selected XML score is opened and loaded. The original tonality and tempo of the score will be shown, and using sliders we will be able to:

- Select the singer voice.
- Fit the song octave and tonality in semitones.
- Change the tempo.
- Set the pitch smoothing level.

Moreover, the attributes of the vibrato (described in section 3.4) can be adjusted in detail:

- The maximum amplitude.
- The no-vibrato initial interval duration.
- The fade-in time.
- The fade-out time.
- The period of the vibrato.
- The minimum duration of a note to apply vibrato.

By using the whole song information, and with the help of the graphical interface, the user can listen to a midi preview of the melody and can adjust the tempo and octave before starting the voice synthesis process. If a modification of those parameters takes place once the synthesis operation has begun, the modification of the parameters will take effect on the next sentence to synthesize. This way a pseudo real-time modification of the main parameters is possible.

The singing result will start playing as soon as the first sentence is ready, i.e. there is no need to wait until the whole song has been synthesized to start listening.

3.3. Syllables and timing

It has been reported that, in singing, note onsets are located at vowel onsets rather than at consonant onsets [7, 20, 21]. The phonemes of the lyrics must be distributed between the notes such that the transitions between notes coincide with the onset of the vowel or set of vowels. In this way, considering one syllable, the consonantal phonemes located before the first vowel will be pronounced within the previous note interval. The result is a redefinition of the borders of the syllables.

After this redistribution of the phonemes in the new syllables, the duration of each note must be distributed among the phones therein. This is done using the generic durations predicted by the linguistic module as a starting point. If the lengthening needed imposed by the score is smaller than 30%, the lengthening will be equally distributed inside the syllable. If it is higher than 30%, then the vowel will account for 90% of

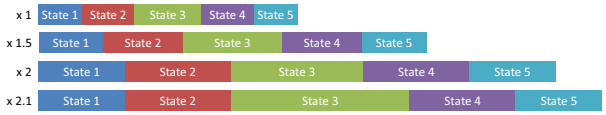

Figure 4: *Distribution of phone duration among states for several lengthening factors.*

the enlargement. In any case, the unvoiced sound durations are never modified.

In standard HSMM-based speech synthesis systems [22], 5 states per phone are considered, and when phone durations are specified as input, the corresponding state durations are calculated statistically. In Bertsokantari, the statistical parameter generation has been modified in such manner that the duration of a phone, usually much longer in singing than in speech, is concentrated mostly on the central state which can be considered to be the most stable and the best articulated one. More specifically, phone durations are distributed among states proportionally to their expected mean duration, except when the lengthening factor is greater than 2; beyond that limit, only the central state is lengthened. This is illustrated by figure 4. Note that this strategy is easy to apply in an HSMM framework while it would be much harder to apply under modern deep learning based generation paradigms [23].

3.4. Generation of the intonation

The intonation curve is obtained directly from the musical note. The jumps from one note to the next are smoothed through a cosine function [7]. The $\log f_0$ information contained in the model is only used to take the voiced/unvoiced decision at every frame in accordance with the new durations.

Vibrato is a musical feature, not present in spoken speech that adds expressiveness to the singing voice and is usually modeled as an amplitude and frequency modulated signal [24]. Vibrato is not among the distinctive characteristics of the *bertsolaritza* style; when present, it is considerably weaker than in other more classical styles. Nevertheless, in order to avoid the beats and buzzing produced by a sustained synthetic vowel, we have implemented a simple vibrato model according to the following expression (see figure 5):

$$\log f_0(t) = \log F_0 + a\frac{\log 2}{12}A(t)\sin(2\pi f_v t) \qquad (1)$$

where F_0 is the pitch value derived from the score. Parameters a and f_v control the modulation depth and frequency, respectively. For instance, a modulation depth $a = 1$ would introduce a variation of ± 1 semitone. In this version of the system, a has been empirically adjusted to 0.44. The modulation frequency has been set to $f_v = 5$ Hz. Function $A(t)$ is a 4th degree parabolic function implementing a smooth transition of the envelope. The default values for the remaining vibrato parameters have been set to *fade-in* = 150 ms, *fade-out* = 75 ms and *no-vibrato* = 75 ms.

3.5. Spectral transformations

It is commonly known that one of the most prominent spectral differences between singing voice and spoken voice is *spectral tilt*. In other words, singing voice is usually produced in a more pressed phonation, which results in a relatively higher amount of energy at mid-high frequencies. This enhancement of mid-high frequencies can be implemented as a filter, i.e. an additive term in cepstral domain that can be summed either to the HSMM states or to the sequence of parameter vectors generated by the statistical engine. In this work we choose the second strategy because it preserves the original model, thus allowing the generation of both speech and singing from the same model. The response of the filter used in this work, a deterministic one inspired by [25, 26], is depicted in figure 6. Note that this filter can be applied regardless of the input voice.

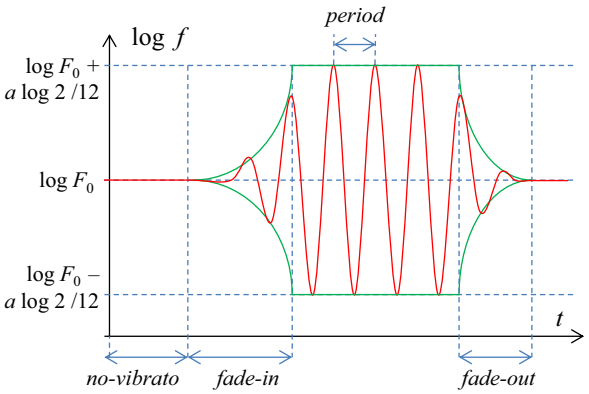

Figure 5: *Schematic description of the implemented vibrato.*

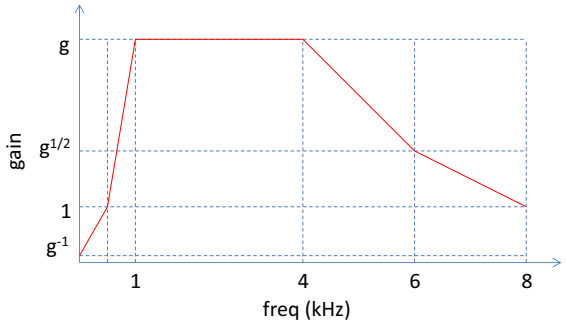

Figure 6: *Mid-high frequency enhancement filter. For the particular voice used in this challenge, parameter g is set to* 2.

4. Discussion and future works

This version of our singing synthesis system Bertsokantari is a preliminary work that allows us to obtain singing voice without the availability of a specific singing voice database during training. The system has not been formally evaluated yet. Its performance is illustrated by the multimedia material submitted along with this paper.

Considering the simple processing applied to the spoken voice we are not dissatisfied with the results. However, most of the parameters (vibrato and spectral filtering) were adjusted just by carefully listening to the output, and they may not produce similar results when used with another spoken voice model. In fact, the selection of the spoken voice was one of the most critical decisions: several candidates were considered, and an artificial spectral manipulation was applied to the selected one to improve its pleasantness. Such a manipulation was applied directly to HSMMs (more details on how to manipulate HSMMs can be found in [27]) and is not part of the singing synthesis process itself.

One of the disadvantages of using read (spoken) utterances to train a singing system is that continuous speech is, in general, less articulated than singing. As a result, some of the sustained vowels sung by the system are not natural enough. This phenomenon is particularly audible near the sentence endings, which reveals another possible source of articulation inaccuracies: in read speech, sentence endings may contain some degree of vocal fry, which often misleads the acoustic analysis made by the vocoder. Unfortunately, sentence endings are especially prominent in singing because the corresponding note is usually long. Also, as we are imposing an external melody to speech parameters generated in an almost-standard way, one more reason for misarticulation is the pitch contrast between the musical scores and the parameters generated from the HSMMs. Indeed, in some parts of the song we are altering the "spoken" pitch by a very large factor. For a more natural output, pitch alteration by a large factor should be accompanied by a proper spectral (Mel-cesptral) manipulation. Alternatively, the system could be instructed to take pitch contrast into account in some manner when selecting the sequence of Mel-cepstral HSMM states for generation.

Another problem found during the development of Bertsokantari is that the rhythm specified by the music score breaks the consistency between the trajectory of the Mel-cepstral parameters and their global variance, which is one of the most relevant aspects considered during parameter generation [16]. Although the magnitude of this problem vanishes when synthesizing utterances that are long enough, a robust solution would imply the use of post-filtering techniques instead of global variance enhancement.

Despite these issues, which will be addressed in the near future, we would like to remark that all the modifications proposed in this work, including those related to spectral tilt correction or state durations, have been implemented so as to make the resulting system compatible with both speech and singing. For example, tilt correction is performed just by enabling a specific flag of the vocoder. As for state durations, while it is true that we have modified the standard parameter generation engine to concentrate elongations in the central state of the phones, this is done beyond a certain elongation factor that is never reached in normal speech. Thus, the basic components of the TTS system and the requirements of the input voices have not been altered; on the contrary, we have enriched the TTS with new singing functionalities that could notably increase its expressiveness in applications that combine speech and singing, like storytelling.

As mentioned at the beginning, the final goal of our work is the synthesis of the *Bertsolaritza* style of singing. This traditional Basque style differs notably from both classical and modern singing styles. We are currently preparing a suitable dedicated database to improve the performance of Bertsokantari [11]. Furthermore, the final system is to be integrated with an artificial verse improvising module to build *Bertsobot* [28].

5. Conclusions

This paper has presented Bertsokantari, a singing synthesizer built from a classical HSMM-based TTS, AhoTTS. Taking an XML music score and any AhoTTS-compatible voice as input, the system imposes the specified rhythm to the synthetic speech, concentrating the possible elongations in the most stable part of the phones, and overwrites the generated pitch contour by the specified one. It also applies a manually-tuned vibrato and a spectral transformation that compensates the spectral tilt differences between speech and singing. The naturalness of the resulting singing voice can be judged as intermediate, the main observable artifacts being related to the imperfect articulation of sustained vowels.

6. Acknowledgments

The authors want to thank the adaptation of the song to the rhythm of the zortziko measure and the accompaniment to Andoni Elías and to Arantza Hernáez. This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), the Basque Government (ElkarOla project, KK-2015/00098) and the Spanish Ministry of Economy and Competitiveness (RESTORE project, TEC2015-67163-C2-1-R).

7. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[2] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] H. Kenmochi and H. Ohshita, "VOCALOID-commercial singing synthesizer based on sample concatenation," in *Interspeech*, 2007, pp. 3–4.

[4] K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-Based singing voice synthesis and its application to Japanese and English," in *ICASSP*, 2014, pp. 265–269.

[5] M. Umbert, J. Bonada, and M. Blaauw, "Systematic database creation for expressive singing voice synthesis control," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013, pp. 213–216.

[6] M. Garnier, N. Henrich, M. Castellengo, D. Sotiropoulos, and D. Dubois, "Characterisation of voice quality in Western lyrical singing: From teachers' judgements to acoustic descriptions," *Journal of interdisciplinary music studies*, vol. 1, no. 2, pp. 62–91, 2007.

[7] J. Sundberg, "The KTH synthesis of singing," *Advances in Cognitive Psychology*, vol. 2, no. 2, pp. 131–143, 2009.

[8] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "Glottal open quotient in singing: measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *J. Acoust. Soc. America*, vol. 117, no. 3, pp. 1417–1430, 2005.

[9] I. Arroabarren, "Signal Processing Techniques for Singing and Vibrato Modeling," Ph.D. dissertation, Universidad Publica de Navarra, 2004.

[10] J. Sundberg, "Level and center frequency of the singer's formant," *Journal of Voice*, vol. 15, no. 2, pp. 176–186, 2001.

[11] X. Sarasola, E. Navas, D. Tavárez, D. Erro, I. Saratxaga, and I. Hernáez, "A singing voice database in Basque for statistical singing synthesis of bertsolaritza," in *LREC 2016*, Portoroz, Slovenia, 2016, pp. 67–70.

[12] J. Garzia, "History of improvised bertsolaritza: A proposal," *Oral Tradition*, vol. 22, no. 2, pp. 77–115, 2007.

[13] Aholab, "AhoTTS – TTS for Basque and Spanish." [Online]. Available: https://sourceforge.net/projects/ahotts

[14] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez, "HMM-based Speech Synthesis in Basque Language using HTS," in *FALA 2010*, Vigo, Spain, 2010, pp. 67–70.

[15] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[16] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[17] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal Sel. Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[18] M. Puckette, "Pure Data : another integrated computer music environment," in *International Computer Music Conference*, 1996, pp. 37–41.

[19] M. Astrinaki, A. Moinet, N. D'Alessandro, and T. Dutoit, "Pure Data External for Reactive HMM-based Speech and Singing Synthesis," in *16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth (Ireland), 2013, pp. 78–81.

[20] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[21] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system." in *Interspeech 2006*, Pittsburgh, PA, USA, 2006, pp. 2274–2277.

[22] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Speech Synthesis Workshop*, 2007, pp. 294–299.

[23] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, 2015.

[24] I. Arroabarren, X. Rodet, and A. Carlosena, "On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1413–1421, 2006.

[25] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012, pp. 635–638.

[26] D. Erro, T.-C. Zorila, and Y. Stylianou, "Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications," *IEEE/ACM Trans. Audio, Speech, & Lang. Process.*, vol. 22, no. 12, pp. 2101–2111, 2014.

[27] D. Erro, I. Hernaez, E. Navas, A. Alonso, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños, R. Perez-Ramon, M. Sulir, X. Tian, X. Wang, and J. Ye, "ZureTTS: Online platform for obtaining personalized synthetic voices," in *Proc. eNTERFACE'14*, 2014.

[28] A. Astigarraga, M. Agirrezabal, E. Lazkano, E. Jauregi, and B. Sierra, "Bertsobot: the first minstrel robot," in *6th International Conference on Human System Interactions (HSI-2013)*, Sopot (Poland), 2013, pp. 129–136.