A singing voice database in Basque for statistical singing synthesis of *bertsolaritza*

Xabier Sarasola[1], Eva Navas[1], David Tavarez[1], Daniel Erro[1,2], Ibon Saratxaga[1], Inma Hernaez[1]

[1] AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2]Basque Foundation for Science (IKERBASQUE), Bilbao, Spain
{xabier.sarasola, eva.navas, david.tavarez, daniel.erro, ibon.saratxaga, inma.hernaez}@ehu.eus

Abstract

This paper describes the characteristics and structure of a Basque singing voice database of *bertsolaritza*. *Bertsolaritza* is a popular singing style from Basque Country sung exclusively in Basque that is improvised and a capella. The database is designed to be used in statistical singing voice synthesis for *bertsolaritza* style. Starting from the recordings and transcriptions of numerous singers, diarization and phoneme alignment experiments have been made to extract the singing voice from the recordings and create phoneme alignments. This labelling processes have been performed applying standard speech processing techniques and the results prove that these techniques can be used in this specific singing style.

Keywords: Singing synthesis database, Basque oral resource

1. Introduction

Singing voice synthesis has attracted a lot of research and commercial attention in the last years. Two main techniques have been applied to the generation of singing voices: unit selection synthesis (Kenmochi and Ohshita, 2007) and hidden Markov model (HMM) based synthesis (Nakamura et al., 2014). Both of them are based on corpus and the quality of the produced voice depends heavily on the variety and quality of the corpus. Therefore it is essential to have a suitable database of singing voice that is sufficiently representative to obtain good synthesis results. However, there are no such appropriate databases publicly available for any language. Though a systematic procedure to create a database for expressive singing voice synthesis has been proposed (Umbert et al., 2013), in general, each group performs its own recordings, such as those used in the works for Japanese (Saino et al., 2006) (Oura et al., 2010), for Spanish (Janer et al., 2006) and for French (d'Alessandro et al., 2014).

The final aim of our work is to develop an HMM based singing synthesis system suitable to sing *bertsos*, a very popular singing style in the Basque Country. There has been a previous first attempt at creating a *bertso* singing synthesis system in the past (Astigarraga et al., 2013), using a read speech database. Since the quality of the synthesized voices strongly depends on the training data, the results achieved by this system are far from satisfactory in terms of naturalness and voice quality. Therefore a new resource has to be developed in order to achieve suitable results. In this paper we present this new database and the results of the first segmentation experiments.

The paper is structured as follows: section 2 briefly introduces *bertsolaritza* and the particularities of this style. Section 3 describes the new singing database. In section 4 the segmentation experiments made on this new database are detailed and, finally, section 5 presents the conclusions and the pending works for the future.

2. *Bertsolaritza*

Bertsolaritza is the art of improvised sung poetry, very popular in the Basque Country. It is generally performed a cappella and on-stage but it is also performed sitting at the table at dinners or lunches. There is an official national competition with several qualification rounds that end in a final competition every four years. This event is followed by thousands of people, both live and by television. The verses are called *bertso* and the singers that improvise them are the *bertsolaris*. Traditionally *bertsos* were sung by men but there is an increasing number of young female *bertsolaris* today.

In the shows, the host calls to one or more *bertsolaris* and gives them a theme for the *bertsos*. Singers have a limited time to create the *bertsos* and to start singing that varies from 5 seconds to 2 minutes depending on the exercise type. There are three types of exercises:

- Exercise in pairs: The first *bertsolari* has 30 seconds to start with the *bertso* and the answer by the second *bertsolari* must start in less than 10 seconds from the end of the the first *bertso*.
- Exercise with point: The host sings a single point and the singer has to answer with the rest of the *bertso*. As the answer is shorter it has to be very fast, they only have 5 seconds to answer the point.
- Alone exercise: The *bertsolari* has 2 minutes to start singing the first *bertso* and may use a 30 second interval between *bertsos*.

The structure of the *bertso* in terms of rhyme and number of syllables obeys rules that singers must follow when they improvise. *Bertsos* are composed by points as shown in Figure 1. A point is the phrase that goes from a rhyme-word to the next rhyme-word. Normally rhyme-words are the last word of the even lines. Different structures of *bertso* are defined changing the length of the points and the total number of points. The most common metres in improvised *bertsolaritza* are the next ones:

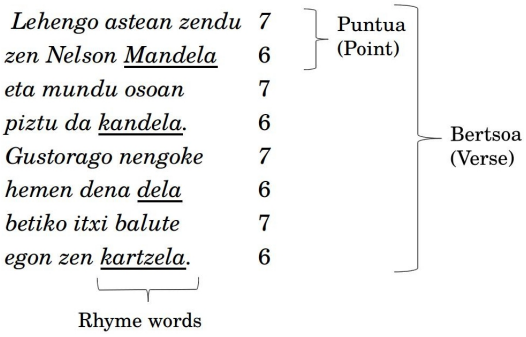

Figure 1: Structure of a *Zortziko* minor

- *Zortziko* major: Odd lines have 10 syllables and even lines have 8 syllables. The *bertso* has 4 points.
- *Hamarreko* major: Odd lines have 10 syllables and even lines have 8 syllables. The *bertso* has 5 points.
- *Zortziko* minor: Odd lines have 7 syllables and even lines have 6 syllables. The *bertso* has 4 points.
- *Hamarreko* minor: Odd lines have 7 syllables and even lines have 6 syllables. The *bertso* has 5 points.

Each of the metres has a set of defined melodies to choose, but it is not prohibited to create new melodies for the competitions. In fact, new good melodies can improve the punctuation the jury gives to a *bertso*.

3. *Bertsolaritza* singing database

The new *bertsolaritza* singing database for synthesis is composed of two parts, as shown in Figure 2: on the one hand the audio recordings of the *bertsos* together with phonetic and orthographic transcriptions of these recordings and additional information on each recording; on the other hand the music scores with the common melodies used in *bertsos*.

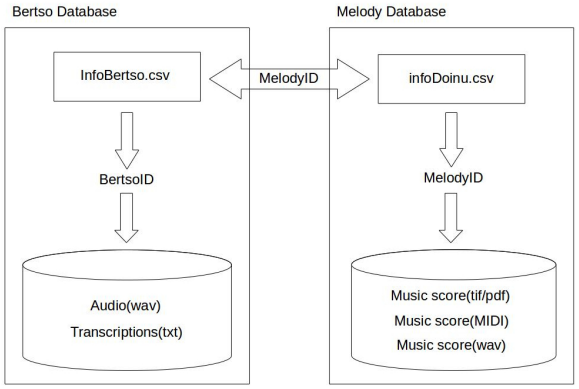

Figure 2: Structure of the *bertsolaritza* singing database

3.1. Bertso recordings

The recordings have been downloaded in mp3 format from the online site devoted to *bertsolaritza*, Bertsozale Elkartea (www.bertsozale.eus). They correspond to different competitions and shows held from 1979 to 2014. The retrieved material is presented in Table 1. Only alone exercises and exercises with point have been considered.

Total duration	6105 min
Total number of recordings	2094
Recordings from male singers	1860
Recordings from female singers	234
Total number of singers	189
Number of male singers	158
Number of female singers	34

Table 1: Downloaded resources

The mp3 files have been converted to Windows PCM format, with 44100 Hz sample rate and 16 bits/sample. The gender of the singers has been manually labelled in each recording. The number of recordings per singer is variable ranging from a maximum of 93 to a minimum of one. The total length of the recordings of each singer varies between 1 and 314 minutes, as shown in Figure 3. The recordings include pauses, speech by the host and applauses, and approximately only half of each recording corresponds to actual singing. So, most of the singers in the database have less than 70 minutes of singing, which is the amount commonly used to to build a singing voice from scratch for HMM based synthesis (Oura et al., 2010). In fact, there are only 10 possible candidates for the creation of the singing voice, nine male and one female, although singers with less material available may be used to adapt existing singing speech models and create new voices.

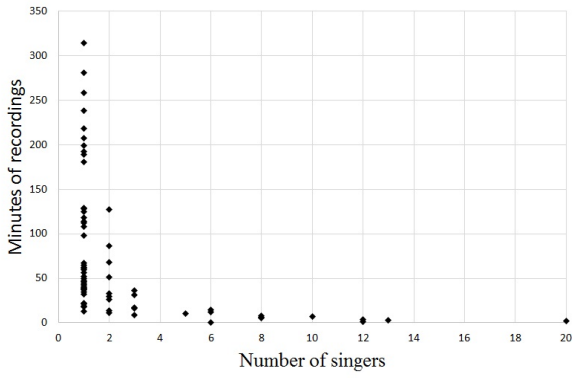

Figure 3: Distribution of recording length by singer

Together with the recordings, the following additional information about them has been downloaded to a csv file:

- BertsoID: name of the audio file without the .wav extension.
- Place and date: place and date where the recording took place.
- Singer: name of the singer.
- Melody name: actual name of the melody.

- MelodyID: identifier of the melody that allows loading information about the melody from the Melody part of the database.
- Melody metre name: name of the metre, e.g. *zortziko* major.
- Number of *bertsos*: number of *bertsos* contained in the recording.
- Singer gender: manually added.

Not every recording has all these additional data available. 1385 recordings have information about the melody and 1898 have information about the metres. In those cases where no melody information was available but the melody was known, melody identifiers had been manually assigned.
Orthographic transcriptions of the *bertsos* have also been downloaded and saved in text files with the same name of their respective audio. These transcriptions present a problem whenever there is an encore. The encores are indicated with the word *bis* at the end of the *bertso*, but the repetition in the singing can affect either only the last line of the *bertso* or the last two lines. Unfortunately there is no indication about it in the downloaded transcriptions. To be able to create the correct transcriptions an encore label that identifies the type of encore has been manually added for each melody in the Melodies part of the database.

3.2. Melodies

There are 3110 different melodies in the database. For each melody downloaded there is a MIDI, a wav and a tiff or pdf format music score file that represent the melody.
Wav Files contain piano recordings of the music score with a 44100 Hz sample frequency and 16 bit per sample. Together with the recordings, the following additional information about them has been downloaded to a csv file:

- MelodyID: identifier of the melody shared by the *Bertso* part of the database
- Melody name: actual name of the melody
- Theme: type of theme the melody is commonly used for, e.g. tragic, comic
- Melody metre name: name of the metre, e.g. *zortziko* major
- Melody metre in syllables: melody metre giving the number of syllables per line and point, not counting the encores
- Musical form: musical form of the melody shown with the standard notation
- Encore (bis): type of repetition used in the melody, added manually. It has 5 possible values
  - Repetition of a single line
  - Repetition of two consecutive lines (full point)
  - Repetition of single line with humming in between
  - Two different repetitions of a single line in the same *bertso* of two non consecutive lines
  - Two different repetitions in the same *bertso*, a single line and a point

As shown in Figure 2, using the MelodyID contained in the information about the *bertso* (infoBertso.csv), music score and recording of each melody can be retrieved from the Melody part of the database.

4. Segmentation of the recordings

To be able to use the recordings to build an HMM based singing voice, the voice of the singer must be extracted from them and then the extracted segments must be segmented at phoneme level.

4.1. Speaker segmentation

At the beginning of each recording the host introduces the *bertsolari* and sometimes he also explains the theme on which the *bertso* lyrics have to be based on. Besides, as the audio is always a recording of a live show, at the end of every *bertso* the audience claps creating noise in the recording. Both of them (speech introduced by the host and the applauses) must be removed before using the recordings to build the singing voice models.
To segment the recordings and label the singing voice of the *bertsolari* we used a classic diarization system (Luengo et al., 2010), with a post-processing of the results (Tavarez et al., 2012). The system considers Gaussian mixture models (GMMs) for the voice, silence and applauses built using 9 recordings of one singer with a total length of 33 minutes.
Standard diarization techniques applied to singing are not completely suitable, as the system tends to identify different speakers when the singer changes note. Therefore, the result of the classic diarization system has been post-processed to group all the different identifiers corresponding to the same singer.
In exercises with point, where the host starts singing the *bertso* and the *bertsolari* has to continue, the results of the automatic speaker segmentation are incorrect as the diarization system is not prepared to detect the same person singing and speaking. Therefore, this type of exercise would require a more elaborate diarization algorithm. At the moment we will not use this kind of exercise for the building of the HMM singing voice.
To evaluate the quality of the segmentation process, 30 recordings from the 6 singers with the largest amount of singing material have been manually labelled. This reference segmentation has been compared with the automatic speaker segmentation. The results show that the automatic system is able to locate all the reference labels, while it inserts some new boundaries not present in the reference. Thus, a 100% recall with a 89% precision is achieved, which produces a very good value of F-score of 93.78%.
70.76% of the correctly identified boundaries are closer to the reference location than 500 ms, which is a good result overall. The boundaries located farther than 500 ms usually correspond to the end of the *bertsos*, where an overlap of speech and applauses makes the process of identifying the boundary more difficult, even in the reference manual labelling.

4.2. Phoneme segmentation

Two singers, a male and a female have been selected for the first phoneme segmentation experiments. Using the transcriptions and part of the recordings, phoneme alignments have been obtained using triphone-based forced alignment in HTK. Recordings equivalent to 1000 phoneme labels have been separated from the training material and reserved for testing purposes. The reference alignments have been manually created for these 1000 reserved labels as well as for another 1000 labels from the training set in order to check whether using the recording to train the segmentation system has an impact in the results or not.The percentages of agreement with different tolerances are shown in Table 2 and Table 3 for the male and female singer respectively.

Seen during training	<5ms	<10ms	<20ms	≥20ms
Yes	23.12%	25.23%	34.37%	65.63%
No	23.43%	26.61%	38.96%	61.04%

Table 2: Percentages of agreement for male singer

Seen during training	<5ms	<10ms	<20ms	≥20ms
Yes	34.32%	36.62%	43.21%	56.79%
No	35.74%	36.72%	41.56%	58.44%

Table 3: Percentages of agreement for female singer

There are no differences in the segmentation accuracy results between the recordings used to train the triphone models and the new recordings of the same speaker. The results are better for the female singer because there was more speech material available to train the models. For read speech, the state of the art in phoneme level segmentation is around 80% within 20 ms from the reference segmentation (Goldman, 2011). The results achieved in our experiment are still far from this value, but are good enough to make the process of manual segmentation easier.

5. Conclusion and future work

A new singing voice database has been created for Basque singing style *Bertsolaritza*. It includes 2094 recordings from 189 different singers. The speech corresponding to singers has been automatically identified applying a diarization process that correctly detects the singing voice in the recordings, excluding the exercises with point. The phoneme segmentation has been obtained by forced alignment with results that allow the manual correction of the boundary positions and the subsequent building of the first speech models.

In the near future, better phoneme alignments will be obtained by building different models for short and long versions of phonemes. A melody detector will be developed to identify the melody in *bertsos* that only have metre defined. The metre information limits the possible melodies for the *bertso* to 150 different melodies in the worst case, therefore a detector based on the pitch curve extracted from the singing voice can provide good results for this task.

6. Acknowledgements

The authors would like to thank the Association of the Friends of Bertsolaritza, Bertsozale Elkartea, for their work in the promotion of *bertsolaritza* and the digitalization of the performances that has make the task of compiling this database to build HMM based singing voices easier.

This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), the Basque Government (Ber2tek project, IE12-333) and the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03).

7. References

Aitzol Astigarraga, Manex Agirrezabal, Elena Lazkano, Ekaitz Jauregi, and Basilio Sierra. 2013. Bertsobot: the first minstrel robot. In *6th International Conference on Human System Interaction (HSI 2013)*, pages 129–136.

Christophe d'Alessandro, Lionel Feugere, Sylvain Le Beux, Olivier Perrotin, and Albert Rilliard. 2014. Drawing melodies: Evaluation of chironomic singing synthesis. *Journal of the Acoustical Society of America*, 135(6):3601–3612.

Jean Philippe Goldman. 2011. EasyAlign: An automatic phonetic alignment tool under Praat. In *Interspeech*, pages 3233–3236.

Jordi Janer, Jordi Bonada, and Merlijn Blaauw. 2006. Performance-driven control for sample-based singing voice synthesis. In *9th International Conference on Digital Audio Effects, (DAFx-06)*, pages 41–44.

Hideki Kenmochi and Hayato Ohshita. 2007. VOCALOID-commercial singing synthesizer based on sample concatenation. In *Interspeech*, pages 3–4.

Iker Luengo, Eva Navas, Ibon Saratxaga, Inmaculada Hernáez, and Daniel Erro. 2010. AhoLab Speaker Diarisation System for Albayzin 2010. In *FALA 2010*, pages 393–396.

Kazuhiro Nakamura, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. 2014. HMM-Based singing voice synthesis and its application to Japanese and English. In *ICASSP*, pages 265–269.

Keiichiro Oura, Ayami Mase, Tomohiko Yamada, Satoru Muto, Yoshihiko Nankaku, and Keiichi Tokuda. 2010. Recent development of the HMM-based singing voice synthesis system-Sinsy. In *7th ISCA Workshop on Speech Synthesis (SSW7)*, pages 211–216.

Keijiro Saino, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, and Keiichi Tokuda. 2006. An HMM-based Singing Voice Synthesis System. In *Interspeech*, pages 2274–2277.

David Tavarez, Eva Navas, Daniel Erro, and Ibon Saratxaga. 2012. Strategies to Improve a Speaker Diarisation Tool. In *8th international conference on Language Resources and Evaluation (LREC)*, pages 4117–4121.

Marti Umbert, Jordi Bonada, and Merlijn Blaauw. 2013. Systematic database creation for expressive singing voice synthesis control. In *8th ISCA Workshop on Speech Synthesis (SSW8)*, pages 213–216.