An open dataset for vocal music transcription

Marcos Woitowitz[1*], **Helena de S. Nunes**[1], **Rodrigo Schramm**[1]

[1]Music Department, Universidade Federal do Rio Grande do Sul
Rua Senhor dos Passos, 248 – Porto Alegre, RS

marcoswoitowitz@gmai.com, helena.souza.nunes@ufrgs.br, rschramm@ufrgs.br

Abstract

This work presents an audio dataset which is designed to support the development of techniques for multi-pitch detection and voice assignment applied to audio recordings containing performances with multiple singers. The proposed dataset contains recordings of popular Brazilian songs, performed by non-professional vocal quartets. Besides the mix down with the complete ensemble, the dataset also contains each vocal part recorded in separated tracks, with its frame-based pitch ground truth and music score.

1. Introduction

The technology of automatic music transcription has evolved significantly; however, a process capable of converting the audio into symbolic representation of a music score is still considered a major challenge, particularly in the transcription of recordings with multiple singers [1]. A growing number of techniques for the transcription of polyphonic signals have been proposed over the last decade [2], especially some more recent, involving machine learning, such as Deep Learning [3]. One of the major challenges for the development of these techniques is the lack of databases for training, testing and algorithm evaluations. This work describes the process of creating a new database of this type, with audio recordings of SATB vocal samples.

2. Materials and Methods

The database described here is composed of 114 audio files, captured at the Center for Electronic Music (CME) of UFRGS, between March and April 2017. For the dataset recordings, we have used ten small excerpts (between four and eight measures) of Brazilian choral works with four voices. These pieces include various musical components and have different characteristics (triplets, punctuated rhythms, tempo changes, alternating metric bars, distinct melodic extension, etc.). The set of choral pieces is: Cabocla Bonita (P. A. Amorim), Canário Terra (F. Matos, 2008), Dies Sanctificatus (J. Maurício, 1793), Divertimento Coral (E. Aguiar, 1950), Final (Guerra-Peixe, 1973), Lá Vai Eu (Guerra-Peixe, 1973), Minha Namorada (C. Lyra, V. de Morais, D. Cozzela, 2009), Muiraquitã (P. Amorim, A. Diniz, T. Sias, 2015), Padre Nosso (G. Velasquez, 1908) e Suíte Pescadores (D. Caymmi, 1957).

The group of interpreters was formed by seven undergraduate students from the third semester of the UFRGS Music Course: a soprano, an alto, three tenors and two basses. All had some previous experience in singing (choir participation and / or accompanied singing groups), but only the soprano was a professional soloist. No person had previous experience in studio work.

Fig 01 – Recording Session

The audio setup used to capture the voices was: a Shure SM58Beta microphone, a Shure SM57

[*]Supported by CNPq – Proc. 145275/2016-7

microphone, an M-AUDIO FAST TRACK ULTRA preamp (interface) and the Cockos Reaper v5.27 multi track system software. All recordings were sampled at a rate of 44100Hz / 24bit, and the audio was recorded in WAV format. On each recording session, in addition to the metronome, the performers could listen to a previously recorded piano base, corresponding to the melodic line of their own vocal part. The goal was to ensure that during the recordings each voice had a reference melody to aid in the tuning.

Five recording sessions were held, each lasting five hours. In the first session, a period of fifteen to twenty minutes was provided, so that the students could know the score, in a process of sight-reading. The following sessions were held at intervals of one week so that everyone would study the scores in advance. The quality of each of these two outcomes was very different. The recording process was performed in two stages: 1) sample collection using piano accompaniment as the melodic reference; and 2) sample collection without melodic reference. Both stages have used a stereo metronome inserted into the headset. Each voice was recorded individually, using two takes in sequence: after the recording with metronome and reference melody (first take), the singer restarted singing the same passage (second take), with the metronome, but without the melodic reference.

The first results showed many mistakes and high time consumption, requiring on average three sessions (lasting around fifteen minutes) for each person. Throughout the sessions, due to the study of the scores and to the own experience acquired, a gradual reduction of this effort was observed to the finalization of the audio files, passing for an average of two takes and seven minutes per excerpt / singer. We have identified the users have more difficulties in the maintenance of time than in the precision of the tuning. The recordings have no cut or assembly of takes; On account of a more compromising error, the session was cancelled, and the singer repeated the recording from the beginning.

In addition to the original sound material, the dataset also contains the music score for each piece, and the respective pitch tracking and automatic melody transcription for each of the vocal parts. This annotation was made automatically using the pYIN [4] algorithm (using default parameters). The resulting files are organized, for each of the pieces and their respective voices, in the following structure:

- **song**
  - song_mix.wav
  - **soprano**
    - song_soprano.wav
    - song_soprano_pitch_track.txt
    - song_soprano_notes.txt
    - song_soprano_notes.mid

The final version of this dataset is available at http://inf.ufrgs.br/~rschramm/projects/msingers/.

3. Conclusion

The search for similar material to the one produced here, already ready to use, proved disappointing. Although individual recordings of choral piece voices have been found, they have the greater purpose of serving as support for the playing of choir's pieces, not serving as data for machine learning techniques. The construction of a database for supporting experiments of automatic music transcription and voice assignment, a direction in which this work will continue, is a necessary and urgent task.

4. References

[1] R. Schramm and E. Benetos. Automatic transcription of a cappella recordings from multiple singers. In AES International Conference on Semantic Audio, June 2017.

[2] E.Benetos,S.Dixon,D.Giannoulis,H.Kirchhoff,and A. Klapuri. Automatic music transcription: challenges and future directions. J. Intell. Inf. Syst., 41(3):407– 434, 2013.

[3] S. Sigtia, E. Benetos and S. Dixon. An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(5):927–939, May 2016.

[4] M. Mauch and S. Dixon, "pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2014.