Automatic Pitch Estimation in Choir Singing Recordings using Deep Learning Strategies

Extended Abstract

Helena Cuesta
Music Technology Group, Universitat Pompeu Fabra
Barcelona, Spain
helena.cuesta@upf.edu

Emilia Gómez
Music Technology Group, Universitat Pompeu Fabra
Barcelona, Spain
emilia.gomez@upf.edu

ABSTRACT

This work presents a data-driven method for the automatic pitch[1] estimation of *a cappella* choir singing performances. We focus on polyphonic music recordings, as choral music involves multiple singers typically grouped into four main voices (soprano, alto, tenor and bass). The task of pitch estimation becomes challenging in this context due to the variety of acoustic scenarios (from solo singers to big choirs) and the lack of annotated datasets for training and evaluation, especially for the polyphonic case. We built a dataset of choir singing that contains different types of performances (solo singers, unison [2] and choir). Then, we train several deep learning architectures to extract pitch information from monophonic singing voice signals, and adapt them afterwards to work with polyphonic signals by modeling the pitch distribution in unison performances. Preliminary experiments provide state-of-the-art accuracies in monophonic music signals and unison performances, while further parameter optimization needs to be done in the polyphonic case.

CCS CONCEPTS

• **Applied computing → Sound and music computing**;

KEYWORDS

ACM, womENcourage, MIR, pitch extraction, deep learning

1 INTRODUCTION

Choral music and singing have been studied from many different perspectives, e.g. physiological, musicological..., but there are very few studies that work both with choirs and technology. The CASAS project tries to combine Music Information Retrieval (MIR) tasks and Singing Voice Processing and Synthesis to develop new technologies for choir singing practice [3]. The present work is developed in the scope of the CASAS project and deals with pitch estimation of *a cappella* recordings, which have not yet been extensively explored. Pitch is the main musical descriptor because it is the basic element of melody, harmony, and tonality, and it can be used for further applications such as music transcription[4] or chord detection. Current state-of-the-art pitch estimation algorithms obtain good results with monophonic signals, or when there is a clear melody in polyphonic signals, e.g. leading vocals. However, they generally fail with orchestral and choral music because of their high acoustic and musical complexity compared to traditional scenarios such as piano music. Choir recordings are especially challenging because the voices may have similar pitch contours [5] and timbre properties. According to the review by Salamon et al. [10], we can distinguish three types of pitch estimation methods: salience-based [9][5][7][8], source separation-based [3][11], and data-driven [4][6][12]. The latter refers to machine learning and deep learning approaches, and it is the one that has been explored fewer times due to the lack of annotated data. For this type of approaches, big annotated datasets are necessary to train and build models. Given that deep learning has become very popular these past years in other computing fields, we decided to use it for pitch estimation. Although very few annotated data is available for polyphonic singing voice, our approach exploits monophonic singing voice, i.e. one single voice, where we find large annotated datasets created for other tasks such as speech recognition and then, we apply some restrictions and rules to extrapolate the models to work with polyphonic singing voice.

2 METHODOLOGY

This project is developed in three steps: (1) building the singing voice dataset: we record several singers from a choir[6] individually and together to obtain monophonic, unison and polyphonic performances; (2) pitch estimation in monophonic singing voice, for which three deep learning architectures are implemented: first, a multi-layered neural network similar to the one presented by Verma and Schafer [12], then, the multi-column deep neural network by Kum et al. [6]; and finally, a convolutional neural network; (3) pitch estimation in unison performances: by first analyzing the pitch dispersion in unison performances we can then tune the parameters of the networks accordingly for the networks to predict pitch in this kind of performances.

More details about the implementation, parameters, and results will be given in the full paper, but preliminary experiments with these networks in the iKala dataset [2] (monophonic singing) and TIMIT [13] (speech) showed results on par with the current state of the art methods - around 80% of pitch accuracy. Step (3) is currently in the process of evaluation with larger datasets, but first results are similar to the ones in step two.

[1] Pitch is the auditory attribute of sound according to which sounds can be ordered on a scale from low to high.

[2] Unison performances are those where all the members of a choir sing the same notes (or at a distance of one octave) at the same time.

[3] CASAS: Community-Assisted Analysis and Synthesis. http://mtg.upg.edu/projects/casas

[4] According to [1], music transcription is the process of converting an audio signal into some symbolic representation such as a score

[5] A pitch contour is a set of frequencies that define the pitch of a melody over time.

[6] The CASAS project works the Anton Bruckner Choir http://www.cdcantonbruckner.com

REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. 2013. Automatic music transcription: Challenges and future directions. *Journal of Intelligent Information Systems* 41, 3 (2013), 407–434. https://doi.org/10.1007/s10844-013-0258-3

[2] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang. 2015. Vocal activity informed singing voice separation with the iKala dataset. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 718–722.

[3] Jean-louis Durrieu and Bertrand David. 2010. Source / Filter Model for Unsupervised Main Melody. 18, 3 (2010), 1–12.

[4] Daniel P W Ellis and Graham E. Poliner. 2006. Classification-based melody transcription. *Machine Learning* 65, 2-3 (2006), 439–456. https://doi.org/10.1007/s10994-006-8373-9

[5] Masataka Goto. 2004. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication* 43, 4 SPEC. ISS. (2004), 311–329. https://doi.org/10.1016/j.specom.2004.07.001

[6] Sangeun Kum, Changheun Oh, and Juhan Nam. 2016. Melody Extraction on Vocal Segments Using Multi-Column Deep Neural Networks. *Proc. 17th International Society for Music Information Retrieval Conference* (2016). https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/119

[7] M Marolt. 2005. Audio melody extraction based on timbral similarity. *Proceedings of the 2nd Music Information Retrieval . . .* (2005). https://doi.org/10.1109/EURCON.2005.1630193

[8] Rui Pedro Paiva, Teresa Mendes, and Amílcar Cardoso. 2006. Melody Detection in Polyphonic Musical Signals: Exploiting Perceptual Rules, Note Salience, and Melodic Smoothness. *Computer Music Journal* 30 (2006), 80–98. https://doi.org/10.1162/comj.2006.30.4.80

[9] J Salamon and E Gómez. 2012. Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 6 (2012), 1759–1770. https://doi.org/10.1109/MSP.2013.2271648

[10] Justin Salamon, Emilia Gomez, Daniel P W Ellis, and Gael Richard. 2014. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine* 31, 2 (2014), 118–134. https://doi.org/10.1109/MSP.2013.2271648

[11] Hideyuki Tachibana, Takuma Ono, Nobutaka Ono, and Shigeki Sagayama. Melody Extraction in Music Audio Signals By Melodic Component Enhancement and Pitch Tracking. (????), 3–5.

[12] Prateek Verma and Ronald W. Schafer. 2016. Frequency estimation from waveforms using multi-layered neural networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-Sept (2016), 2165–2169. https://doi.org/10.21437/Interspeech.2016-679

[13] Victor Zue, Stephanie Seneff, and James Glass. 1990. Speech database development at MIT: Timit and beyond. *Speech Communication* 9, 4 (1990), 351–356. https://doi.org/10.1016/0167-6393(90)90010-7