SOURCE NUMBER ESTIMATION FOR MONOAURAL CHORAL RECORDING

Darius Petermann
Music Technology Group
Universitat Pompeu Fabra
dariusarthur.petermann01@estudiant.upf.edu

ABSTRACT

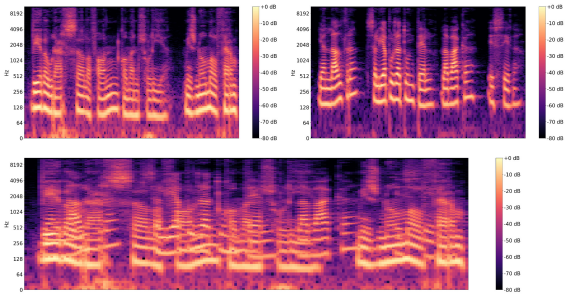
Source number estimation for single-channel recording is an important yet relatively unexplored aspect of music information retrieval (MIR). Previously, there have been several approaches proposed in order to address the issue. Some of them are directly derived from source localization methodologies and are thus relying on the spatial information the recording may convey. In the field of MIR, this task mainly relies on the fact that the spectra of common musical sources in a song are largely uncorrelated. The case of choral singing however, involves groups of people singing in harmony, which leads to higher correlation amongst the constituent signals. In this work, we propose a novel approach consisting in extracting the multiple fundamental frequency tracks from a given monoaural recording, which allows us to then deduce the underlying number of sources present in the given mixture. We demonstrate the efficacy of this approach by comparing our results with a previously-computed baseline using MFCCs. We finally conclude the paper by discussing future work and potential improvements.

Index Terms— Source number estimation, harmonic CQT, deep salience.

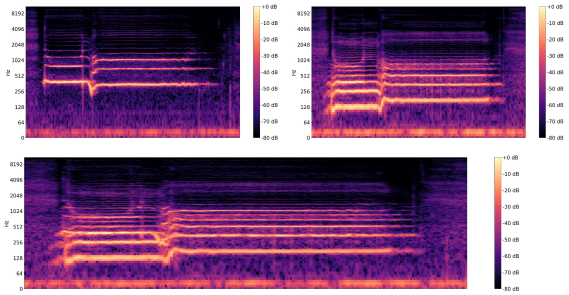
1. INTRODUCTION

Source number estimation consists in estimating the number of different sound sources that are present in an audio recording. Although this task remains fundamentally the same independently of the type of signal setting involved, the nature of the sources (e.g.: speech, music, singing voices) and their relations may entail various challenges and, consequently, may require different analysis methodologies to be used. Choir music constitutes a type of music composed for an ensemble of singers. A choir ensemble is usually structured by grouping the voices into four different sections; namely 'soprano', 'alto', 'tenor', and 'bass'. Such structural setting is usually referred to as a SATB setting. Although different variants of this structure exist, this one is arguably the most common type and the one we will be focusing on in this work. Unlike the type of signals found in speech mixtures, where the constituents' spectra happen to be largely uncorrelated and each one of them represents a distinct and unique spectral profile, the case of SATB recordings becomes much more challenging as it entails identical sources performing in harmony, leading to overlapping spectral components [5]. Beside the problem of pitch distribution across the various SATB groups arises the issue of unison within each group as well. For the scope of this work we will however restrict our research to one singer at most per group, putting aside the unison problem for future work.

The task of source number estimation has been widely investigated in various fields already, such as in source localization [4] as well as source separation [8]. Methods such as Akaike's Information Criterion (AIC) or Ris-



(a) Spectrograms of Male and Female Speakers first isolated and then mixed together.



(b) Spectrograms of Tenor and Bass parts first isolated and then mixed together.

Figure 1: Unlike Regular Speech Mixture, SATB Mixture Has its Constituent Highly Correlated

sanen's Minimum Description Length (MDL) [12] have proven to be fairly effective for the task, however, these estimation approaches, along their more recent adaptations [9], are all heavily relying on spatial diversity and are often closely coupled with the localisation problem. Since our problem specifically address the estimation task for monoaural recordings, any method relying on recordings involving a large number of microphones should be put aside. Previous work [6] proposes ways of converting single-channel recordings to multi-channel signals in order to expand the number of observable data per points, and consequently allow the use of multi-channel source separation algorithms. It is important to note that once again here, this channel expansion approach assumes the various sources to be each located at unequal distance from the microphone. Consequently this can't be feasible for our use-case as each recorded sources (i.e.: SATB groups), has been recorded separately and placed at equal distance from the microphone, thus, no spatial information are conveyed from these recordings.

To this day, very little work that has been carried in the field of MIR for choral singing specifically. In [5], the authors propose a methodology for modeling pitch contour of choral recordings, attempting in overcoming the issue introduced by the highly correlated nature of the present voices. The paper specifically tackles pitch distribution in unison, that is, within individual SATB groups. This is partly achieved by estimating the multi-F0 contour of the four choir voices (i.e.: SATB) and the one we will base our approach on.

2. DATASET

There are a very few publicly available choir music datasets, thus our choice remains limited. We opt to take advantage of the Choral Singing Dataset, described in [7], which comprises three songs and offers separate, isolated tracks for each of the individual singers for a total of 16 audio stems per song. Additionally the dataset includes F0 ground truth files. We will not be needing these files for this precise experiment, however as described later in this work, this will come handy for potential future work.

The choral singing dataset is very well suited for our experiment since the isolated track for each individual singer will allow us to proceed the same way as in [5]; by creating artificial mixes and combining various stems from different groups together (i.e.: SA, ATB, SB, etc.), and thus maximizing the amount of usable information from this limited dataset.

3. EVALUATION METHODOLOGY

To our knowledge, no previous work has investigated the task of source number estimation for choral recordings. Since this task entails additional challenges due to the nature of the sources to be analysed, we opt to build our own baseline and improve on top of it.

The dataset is split into train, validate, and test groups using a mix of section and song-based randomized split method. One full song is dedicated to the test set (including all 16 singer tracks). The two other songs are split between validation and training sets based on the set of singers; one set of singer per song is extracted and added to the validation set (i.e.: 2×4 stems), the remaining stems (i.e.: 24 stems total) are dedicated to training. Our batch generator function creates various linear combinations of stems, each 88200 samples long (i.e: 4 second long). For each mix created, the ground truth is also computed as the number of singers present in the created mixes (1 - 4).

3.1 Baseline Computation with MFCCs

Our first approach will be framed as a traditional deep learning task. We will attempt in extracting features from the previously-generated audio mixes, which could best depict the numbers of present singing voices in a given mixture. We will then train a deep neural network on those features. A popular acoustic feature amongst task involving voiced signals [13] [10] are Mel Frequency Cepstral Coefficients (MFCCs) which are coefficient that are derived directly from spectrogram Mel-Bands. MFFCs are a compact and efficient way to represent the envelope of a short time power spectrum on a logarithmic scale. Hence, MFCCs can be found to be truly effective since they approximates the human auditory system's response efficiently. In cases like ours, where the singing voice is the main and sole signal to be analysed, we should expect to find the most crucial information about the signal under 5000 Hertz.

We use the signal processing library Essentia [3] to extract the MFCCs from the mixes over all frames. We opt to set the upper bound of the analysis frequency range to 6000 Hertz. Restricting the analysis range as such will serve two purposes; first discard the non-critical information from the mixes, but most importantly put the focus on the band in which most singers' formants are concentrated. The MFCCs will then be distributed logarithmically over this range.

Our choice of coefficient number for this task is largely based on empirical evidences. We observed that restricting the number to 13, as usually seen in many speech recognition tasks, was noticeably limiting our accuracy. Increasing the number of coefficients up to 40 however happened to improve our results considerably. Figure **??** shows Mel-band spectrograms along their coefficient sets computed from our training data mixes (cherry-picked).

3.2 Deep Salience Approach

In the second stage of our experiment, we attempt to improve upon our initial baseline by framing our approach as a multi-F0 detection problem. One way to find the total number of sources present in a given recordings is to look at the recording's multi-F0 contour, which happens to be an extremely challenging task in itself. Detection of multi-F0 in polyphonic music is generally achieved by measuring the pitch salience over time of a given signal, that is the measure of the presence of a certain pitch f at time t over multiple frames. This type of representation is typically obtained by emphasizing harmonic content

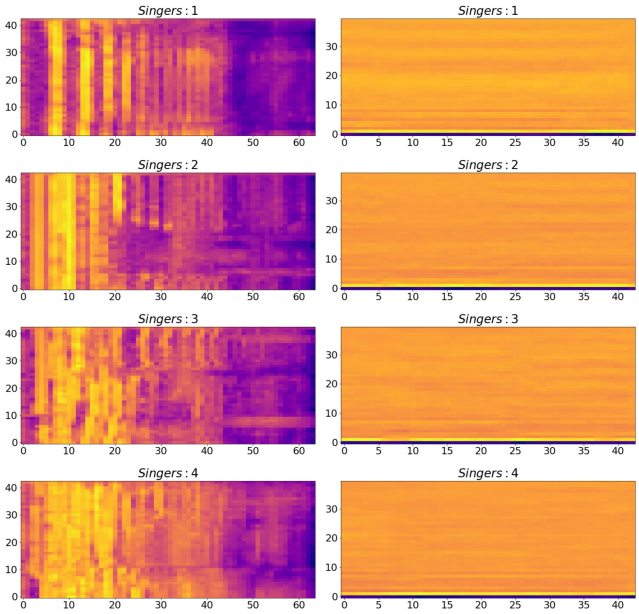


Figure 2: Mel-Spectrograms and MFCCs for Each Singer Class, Represented in the Time-Frequency Domain. The Mixes Have Been Cherry-Picked from the Training Set.

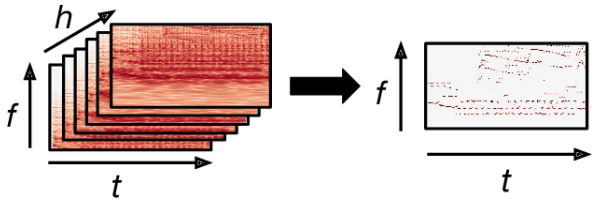


Figure 3: Data Workflow Proposed in [2] for Pitch Salience Detection.

over spectral bins and weighting the sum accordingly [11]. In [2] the authors propose to extend this idea in the context of deep learning by conditioning a network to predict the pitch salience track of a given recording using its harmonic representation. Figure 3 depicts the overall proposed workflow.

The input representation of the model is defined as the *harmonic* constant-Q transform (HCQT). CQT representations happen to be ideal in cases involving any type of audio signals as their bins are equally distributed across musical octaves. Unlike the regular CQT, which work in a two dimensional space $[t,f]$, the HCQT adds another dimension to its representation space, h, which measures the hth harmonic at frequency f and time t. The pre-trained model included as part of [2] [1] has been trained on the MedleyDB multitrack dataset [1], which targets mainly Pop/Rock genres. This isn't ideal for our task but we hope that this approach will show some preliminary improvements over our baseline, described in section (3.1), and will open doors to potential further research in this field.

[1] https://github.com/rabitt/ismir2017-deepsalience

Param.	Value
N. Octave	6
Bins/Octave	60
Harmonics	[0.5, 1, 2, 3, 4, 5]
Fmin	32.7Hz
Hop Length	256

Table 1: Configuration Used in the HCQT Computation.

4. EXPERIMENT

As a first step, we extract 40 MFCCs over 64 Mel-bands from each mixes in our train, valid, and test set. We feed the resulting vectors into a generic neural network designed using the open-source library *Keras*. The network architecture consists in three hidden layers, each of which includes dense layers composed of 20, 40, and 60 units respectively. The model is trained over 100 epochs using a batch size of size 25.

The second stage of our experiment consists in computing the pitch salience track from our mixes. We first import the pre-train model introduced in section 3.2. Since our model has already been trained, the training stage can be discarded; this allows us to jump directly onto the testing phase. We compute the HCQT representation for each mixes in our test set. Table 1 provides the parameters used in the computation of the HCQT.

There is no intuition behind the choice of this configuration beyond the fact that the same parameters have been used in [2] for both training and testing. Hence, it seems like a suitable preliminary approach to take for our task. The computation will then return the signal energy over 360 frequency bins (6 octaves at 60 bins per octave (20 cents per bin)) over N time frames. As a last step, this representation will be fed to the network which will then output the pitch salience track for the mix, that is the track for each *F0* contour detected over the analysed audio segments. In an optimal scenario, we would simply extract the number of detected *F0* in order to find the number of singers in the recording. However the task happens to be a bit more difficult that that, which is why we propose three different ways of predicting the number of sources from the predicted pitch salience tracks:

$$\text{mean}(\overline{v}) = \frac{1}{N}\sum_{i=1}^{N}\overline{v}_i \quad (1)$$

$$\max(\overline{v}) \quad (2)$$

$$\text{mostcommon}(\overline{v}) \quad (3)$$

where $\overline{v}$ is the vector carrying the various length of the multi-*F0* vectors over all frames. For example, $\overline{v} = [1, 2, 4, 1]$ means that frame n_1 has one *F0*, frame n_2 has two, and so on. In (1) we take the mean of all the numbers of simultaneous *F0's* found across all frames in the mix.

In (2) we simply take the maximum number of concurrent *F0's*.

Taking the example above, each of the methods we just described would return the following predictions: $\text{mean}(\overline{v}) = 2$, $\max(\overline{v}) = 4$, $\text{mostcommon}(\overline{v}) = 1$.

5. RESULTS AND DISCUSSION

We tested the various methodologies described in section 4 on a test set composed of 256 phrases, each 4-seconds long (88200 samples). For the MFCCs, we first trained a neural networks over 100 epochs. Figure 4 shows the evolution of the loss and accuracy function over the training process. We observe that, while the fitting on the train set shows promising loss and accuracy curves, with an accuracy reaching 85%, the valid set doesn't perform nearly as well and remains static around 60%.

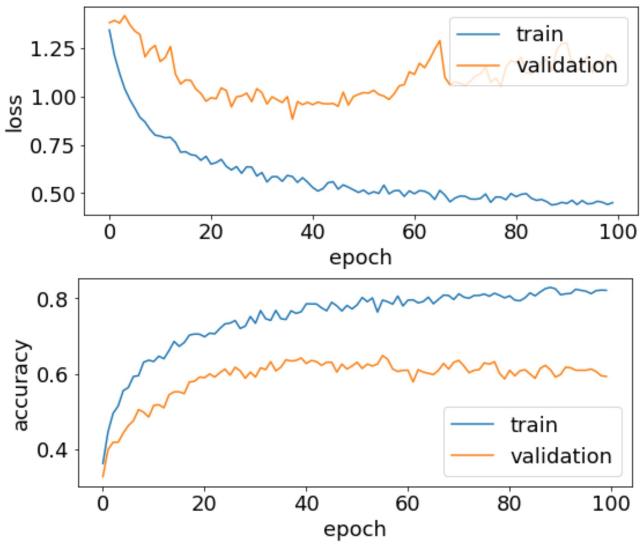


Figure 4: MFCCs: Loss and Accuracy Function on the Training and Validation Sets

Figure 5 shows the accuracy results of the four different approaches taken. We clearly see that deep salience is a more effective way of predicting the number of present source in a given choral recording, with an accuracy close to 70% against the MFCC approach which barely reaches 40%.

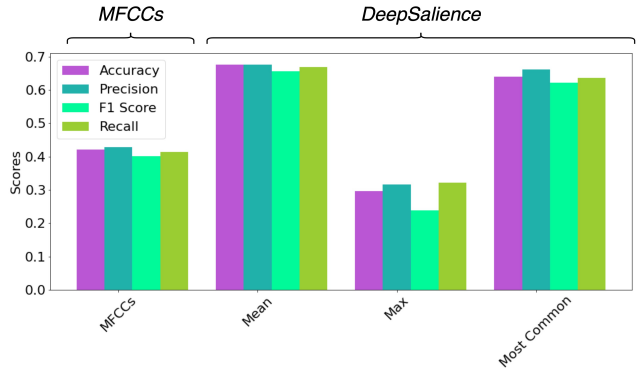


Figure 5: Classification Report for All Four Approaches.

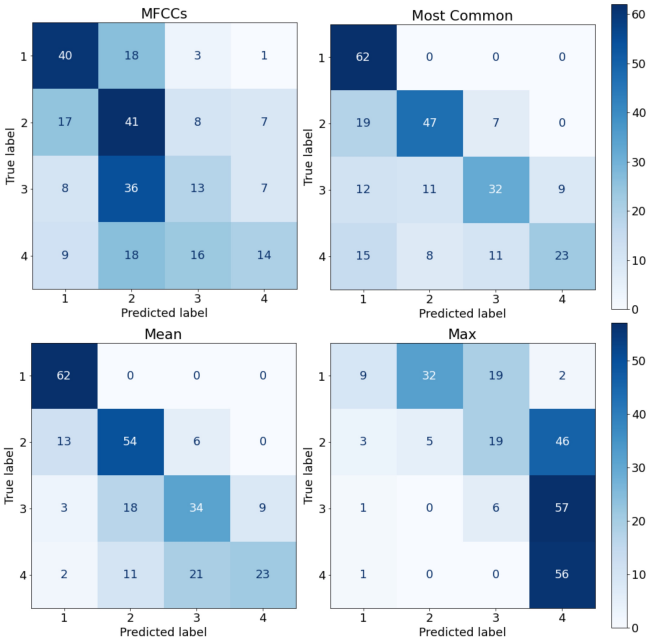


Figure 6: Resulting Confusion Matrices for MFCCs and Deep Salience (Mean, Max, Most Common) Done on 256 4-seconds Long Phrases.

Figure 6 shows the confusion matrices from the various predictions. The model trained on the MFCC features manifestly struggles to differentiate between the 2 and 3-singers recordings. Additionally, we observe that the amount of dispersion for the 4-singers recordings is substantial. For the deep salience approaches, the most successful approach (i.e.: computing the mean) shows a great amount of dispersion in the 4-singers recording, while the 1-singer ones are predicted with an accuracy of 100%. Finally, we observe that taking the maximum number of concurrent *F0's* doesn't show promising results since nearly everything is predicted as 3 or 4-singers recordings.

6. CONCLUSION AND FUTURE WORK

In this work we addressed the issue of source number estimation specifically targeted towards choral recordings. We first approached the problem by computing our initial baseline. This was done by training a neural network on 40 MFCCs extracted from various SATB mix phrases. The second stage of our experiment consisted in improving this baseline by taking a multi-*F0* detection approach: we used a deep salience model, which was pre-trained as part of [2], in order to extract the various *F0* tracks from the test phrases. The numbers of *F0* tracks detected then conveyed the number of present sources in the given mix. This latter stage of our experiment showed promising and better results over our initial baseline. The model used to predict the multi-*F0* tracks was trained on the MedleyDB dataset, which mainly targets Pop and Rock music. In order to further improve the performance of this approach, the model should be trained on choral music recordings specifically. This could help better the *F0* contour prediction and consequently improve the source number estimation process.

7. REFERENCES

[1] Rachel Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. *Proceedings - 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 10 2014.

[2] R.M. Bittner, B. McFee, J. Salamon, P. Li, and J.P. Bello. Deep salience representations for f_0 estimation in polyphonic music. Oct. 2017.

[3] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, O. Mayor, Gerard Roma, Justin Salamon, J. R. Zapata, and Xavier Serra. Essentia: an audio analysis library for music information retrieval. pages 493–498, 04/11/2013 2013.

[4] Maximo Cobos, Jose Lopez, and David Martinez. Two-microphone multi-speaker localization based on a laplacian mixture model. *Digital Signal Processing*, 21:66–76, 01 2011.

[5] Helena Cuesta, Emilia Gómez, and Pritish Chandna. A framework for multi-f0 modeling in satb choir recordings. 04 2019.

[6] Zhi Dong, Junpeng Hu, Bolun Du, and Yunze He. Improvement of source number estimation method for single channel signal. *PLOS ONE*, 11:e0164654, 10 2016.

[7] Helena Cuesta i Mussarra, Emilia Gómez Gutiérrez, Agustín Martorell Domínguez, and Felipe Loáiciga. Analysis of intonation in unison choir singing. 2018.

[8] H. Luan, H. Jiang, and X. Liu. Source number estimation in single channel blind source separation. 9:4445–4449, Oct 2010.

[9] Nilesh Madhu and R. Martin. Source number estimation for multi-speaker localisation and tracking. pages 1–5, 2018.

[10] Tushar Ratanpara and Narendra Patel. Singer identification using mfcc and lpc coefficients from indian video songs. pages 275–282, 2015.

[11] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, 20:1759–1770, 08/2012 2012.

[12] M. Wax and T. Kailath. Determining the number of signals by information theoretic criteria. 9:232–235, March 1984.

[13] Prashant P. Zirmite, Mr. Mahesh K. Patil, Mr. Santosh P. Salgar, M. Mr.Veeresh, and Metigoudar. Separating voiced segments from music file using mfcc , zcr and gmm mr . 2016.