Score-Informed Source Separation of Choral Music

Matan Gover - Music Technology - M.A. Thesis Proposal

Introduction

As a longtime choir singer and conductor, I have seen professional and amateur choirs alike spend much of their rehearsal time on part-learning. Most singers are not able to 'sight sing' perfectly from a score, and are helped by hearing a recording of their part. For this reason, conductors sometimes record *practice tracks* to help singers and save valuable rehearsal time.

In this thesis, I propose to devise a method to produce practice tracks automatically by extracting them from mixed choral music recordings. This act of 'de-mixing' a recording into tracks is known as audio source separation. Source separation enables additional applications such as fine-grained editing and analysis.

Choral music poses several challenges to source separation. The singing voice is hard to model due to its highly variable acoustic properties. Overlap between partials in the spectrum, which is ubiquitous in music, makes it difficult to differentiate between simultaneously sounding notes. Furthermore, an important goal in choral performance is achieving *blend* between singers, and separation must undo that blend. The large amount of reverberation that is often present in choir recordings acts as another obstacle. Finally, choirs are seldom recorded in a 'one voice per track' setting, and this lack of multi-track recordings makes it harder to design and validate source separation systems.

This research will attempt to overcome these challenges by taking advantage of recent research breakthroughs that have shown promising results in singing voice source separation.

Previous Work

Source separation: Numerous approaches to source separation have been proposed over the years (Vincent, Virtanen, & Gannot, 2018). One prominent technique is Nonnegative Matrix Factorization (NMF) (Lee & Seung, 1999). NMF is used to estimate audio spectrograms as a set of fixed components and their time-varying activations (Smaragdis et al., 2014). NMF in its basic form is often not sufficient to produce well-separated components; a large part of the literature has focused on formulating constraints to improve separation, such as group sparsity (Lefèvre, Bach, & Févotte, 2011), temporal continuity (Smaragdis et al., 2014), and harmonicity (Hennequin, Badeau, & David, 2010).

The musical score, where available, is an extremely valuable source of information. It can be used to set both spectral and temporal constraints using techniques such as time-frequency masking (Duan & Pardo, 2011), specially-crafted NMF initializations (Ewert et al., 2014), and instrument timbre models (Rodriguez-Serrano et al., 2015).

Recently, deep neural networks (DNNs) (LeCun, Bengio, & Hinton, 2015) have emerged as a major paradigm for source separation. They enable formulation of the source separation task holistically as minimizing the error of a mapping from mixtures to source signals (Vincent, Virtanen, & Gannot, 2018). In a recent evaluation campaign (Stöter, Liutkus, & Ito, 2018), DNN-based submissions were the most successful. Submissions included a combination of convolutional and recurrent DNNs that operate on spectrograms (Takahashi & Mitsufuji, 2017) and a convolutional neural network (CNN) that operates directly in the time domain (Stoller, Ewert, & Dixon, 2018).

One study used a CNN that operates on score-filtered spectrograms to separate orchestral music recordings (Miron, Janer, & Gómez, 2017).

Choral music: The singing voice can be described as a "voice source" generated by the vibrating vocal folds and a vocal tract "filter" that modifies it (Sundberg, 1987). Vocal and choral music possess several unique acoustic characteristics (Ternström, 2003), notably intonation, vibrato, pitch drift, and blend (Daffern, 2017; Dai & Dixon, 2017). DNNs have been used for singing voice processing for pitch estimation, source separation, and synthesis (Gómez et al., 2018).

Proposed Research

Method: In the absence of previous work on source separation of choral music, we have conducted a preliminary study to test the applicability of two state-of-the-art source separation methods to choral music using a synthesized dataset.

Using score-informed NMF (Ewert et al., 2014), we achieved separation results that were quite satisfying but contained significant vibrato-related artifacts. It is probably possible to improve results using a source-filter model, better score-based constraints, or a more powerful variant of NMF. In its current form, NMF is too limited to fully model the acoustic characteristics of the singing voice and their evolution over time.

We achieved much better results using a CNN that operates in the time-domain (Stoller, Ewert, & Dixon, 2018), matching this algorithm's success in the aforementioned evaluation campaign. With its multi-scale convolutional architecture and learned parameters, this CNN effectively models the separation task; we therefore intend to use it as a starting point for our work.

Our methodology is to progress gradually from controlled synthesized signals to real recordings, and improve the separation method as needed. Due to the complexity of choral music, in some cases even human experts cannot identify individual parts without the score. We therefore plan to integrate score information into the separation process using a dataset of pre-aligned scores. Another technique we will use to improve separation quality is adversarial training, in which a discriminator and generator are trained simultaneously (Goodfellow et al., 2014).

Datasets: The success of DNNs largely depends on the amount of training data available. For the preliminary study we synthesized a corpus of Bach chorales using a sample-based synthesizer. To achieve more realistic singing synthesis, we will examine methods such as concatenative synthesis and DNNs (Gómez et al., 2018).

Multi-track choir recordings are unfortunately very rare; we are aware of only one small dataset (Cuesta et al., 2018). We intend to gather choir practice tracks from the web[1]. In addition, we have obtained several high-quality private multi-track choir recordings with aligned scores.

Evaluation: Separation will be assessed using widely-adopted criteria: source to distortion ratio, source to interferences ratio, and source to artifacts ratio (Stöter, Liutkus, & Ito, 2018).

Summary

The expected contribution of this research is twofold. Firstly, it will enable an application to help choir musicians learn their parts. Secondly, it may lead to developments that are applicable to the source separation field as a whole.

[1] ChoralPractice (https://www.choralpractice.com/)

References

Cuesta, H., Gómez, E., Martorell, A., & Loáiciga, F. (2018). Analysis of intonation in unison choir singing. In *15th Int. Conf. Music Perception and Cognition* (pp. 125–130). Graz, Austria.

Daffern, H. (2017). Blend in singing ensemble performance: Vibrato production in a vocal quartet. *Journal of Voice*, *31*(3), 385.e23–385.e29.

Dai, J., & Dixon, S. (2017). Analysis of interactive intonation in unaccompanied SATB ensembles. In *18th Int. Society for Music Information Retrieval* (pp. 599–605). Suzhou, China.

Duan, Z., & Pardo, B. (2011). Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1205–1215.

Ewert, S., Pardo, B., Mueller, M., & Plumbley, M. D. (2014). Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine*, *31*(3), 116–124.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems 27* (pp. 2672–2680). Montreal, Canada.

Gómez, E., Blaauw, M., Bonada, J., Chandna, P., & Cuesta, H. (2018). *Deep learning for singing processing: Achievements, challenges and impact on singers and listeners*. Paper presented at the 2018 Joint Workshop on Machine Learning for Music, Stockholm, Sweden.

Hennequin, R., Badeau, R., & David, B. (2010). Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *13th Int. Conf. Digital Audio Effects* (pp. 246–253). Graz, Austria.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lefèvre, A., Bach, F., & Févotte, C. (2011). Itakura-Saito nonnegative matrix factorization with group sparsity. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing* (pp. 21–24). Prague, Czech Republic.

Miron, M., Janer, J., & Gómez, E. (2017). Monaural score-informed source separation for classical music using convolutional neural networks. In *18th Int. Society for Music Information Retrieval* (pp. 55–62). Suzhou, China.

Rodriguez-Serrano, F. J., Duan, Z., Vera-Candeas, P., Pardo, B., & Carabias-Orti, J. J. (2015). Online score-informed source separation with adaptive instrument models. *Journal of New Music Research*, *44*(2), 83–96.

Smaragdis, P., Fevotte, C., Mysore, G. J., Mohammadiha, N., & Hoffman, M. (2014). Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, *31*(3), 66–75.

Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In *IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (pp. 2391–2395). Calgary, Alberta, Canada.

Stöter, F.-R., Liutkus, A., & Ito, N. (2018). The 2018 signal separation evaluation campaign. In *Int. Conf. Latent Variable Analysis and Signal Separation* (pp. 293–305). Guildford, UK.

Sundberg, J. (1987). *Science of the singing voice*. Northern Illinois University Press.

Takahashi, N., & Mitsufuji, Y. (2017). Multi-scale multi-band DenseNets for audio source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 21–25). New Paltz, NY, USA.

Ternström, S. (2003). Choir acoustics: An overview of scientific research published to date. *International Journal of Research in Choral Singing*, *1*(1), 3–12.

Vincent, E., Virtanen, T., & Gannot, S. (2018). *Audio source separation and speech enhancement*. John Wiley & Sons.