Learning a Large-Scale Vocal Similarity Embedding for Music

Aparna Kumar [1] **Rachel M. Bittner** [1] **Nicola Montecchio** [1] **Andreas Jansson** [1] **Maria Panteli** [1]
Eric J. Humphrey [1] **Tristan Jehan** [1]

Abstract

This talk describes an approach for modeling singing voice at scale by learning low-dimensional vocal embeddings from large collections of recorded music. We derive a set of embedding spaces from different input representations and genre labels, and evaluate on both objective (ranked retrieval) and subjective (perceptual evaluation) tasks. By comparing the embeddings we show the spaces capture different relationships between genres. We find cases of tracks from different artists and genres which cluster together and through crowd sourced evaluation we confirm these groups of tracks share vocal styles. We conclude with a summary of our ongoing effort to crowd source vocal attribute labels to further refine our approach.

1. Introduction

A vast range of vocal styles and timbral qualities are used in music. A singer's voice may be raspy or clear; natural or heavily processed with audio effects; have a simple or complex, virtuosic style; they may be rapping or singing lyrically; sound old or young (Caffier et al., 2017).

Models of acoustic similarity proposed in the literature have proven to be a useful tool for music retrieval and recommendation at commercial scale; for instance, a system similar to the one described in (Van den Oord et al., 2013) is in use at Spotify in popular features that affect millions of users. Such approaches focus for the most part on modeling the overall acoustic similarity between tracks; however, the problem of measuring similarity along *specific, perceptually salient components* of music – such as the properties of a singing voice – has received considerably less attention, especially in the context of large-scale retrieval.

[1]Spotify Inc.. Correspondence to: Aparna Kumar <aparna@spotify.com>, Rachel Bittner <rachelbittner@spotify.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

2. Related Work

Early work on modeling timbre similarity between singers used various representations including source-filtering to focus the models on vocal characteristics (Kim & Whitman, 2002; Nakano et al., 2014; Fujihara et al., 2010). These works used limited data and quantified performance on an artist retrieval task. Other studies have measured overall timbral similarity between full tracks (Pachet & Aucouturier, 2004; Logan, 2005). More recently, deep neural network-based approaches (Van den Oord et al., 2013) leveraged user data, in the form of collaborative filtering vectors, as a learning objective and a proxy for similarity.

Some methods have captured pitch contour to differentiate musical genres of Flamenco singing (Salamon et al., 2012) or a large collection of global folk music (Panteli et al., 2017). To our knowledge, the latter work is the only other study to model singing voice at scale.

In speech recognition, (Li et al., 2017) trained an embedding to capture similarity between different speakers. Speaker clustering, identification and verification is related to our work with the exceptions : (1) typical speech recordings involve a single speaker with limited background noise, while in music many other sound sources (instruments, voices) occur simultaneously with the primary voice; and (2) the range of typical variations in spoken voice is much smaller than that of the singing voice.

3. Method

Our goal is to produce an acoustic model of vocal similarity that can be applied at scale, and generalized to new inputs. Similarity judgments are difficult to obtain at scale, and thus we leverage curated genre labels as a proxy. While genre classification as a task is known to be fundamentally flawed (Sturm, 2013), our goal is not to build a genre classifier, but rather exploit the correlations between genre labels and singing characteristics (Potter, 2006; Thalén & Sundberg, 2001; Pachet & Cazaly, 2000). To this end, a convolutional neural network is trained to predict genre labels; we then treat the dense representation in the penultimate layer to be our vocal style embedding.

We consider 5 input representations which include decom-

positions of the full mix : (1) mel spectrograms of full mixture, (2) mel spectrograms of source separated vocals, (3) and instrumentals (Jansson et al., 2017), (4) vocal pitch salience map with harmonics (H-VPSMs) and (5) VPSM (Bittner et al., 2017). Representations (2), (4), and (5) are specifically trained to isolate characteristics of the signal related to voice, while (4) and (5) contain non-vocal information.

We consider a dataset of 10,000 tracks from Spotify, sampled to be evenly distributed among 10 expert-labeled genres and 100 artists, resulting in a set of 20 tracks per artist and 2000 tracks per genre. Half of this dataset, partitioned by artists, is used to train the model, and the remaining half for analysis of the embeddings below. The dataset is sampled to contain a wide array of popular genres, with some genres being closely related. This is based on the hypothesis that such a dataset will help us assess the levels of granularity captured in each embedding space.

4. Results

4.1. Distances in embedding spaces

We see that the distances between genres vary across the 5 spaces, and at a high level different relationships are captured. From this, we hypothesize that each space captures different aspects of vocal similarity at different granularities, and perhaps that some aspects of vocal style can be measured in some embedding spaces, but not others. We would highlight this area as an interesting area for future investigation.

Artist retrieval is used in MIR to evaluate vocal style similarity. In a pure artist retrieval setting, each embedding space performs at least 10 times better than random.

Next, we find tracks from different artists and genres that cluster together in the embedding spaces. Through crowd sourced comparisons we confirm these tracks in these groups share distinct vocal styles with each other that are not present in the non-neighboring tracks by the same artist. We conclude artist retrieval scores can under-evaluate vocal style spaces.

4.2. Crowdsourcing to refine the embedding space

We attempt to refine our embedding space with labels that explicitly capture vocal style. As this data is not readily available, we crowd sourced for vocal attribute labels for the 10K dataset. On CrowdFlower we asked questions about the attributes to pairs of tracks. This data allowed us to construct a rank by the each of the attributes.

To begin, we measured how consistently we could recover the rank on a small set of tracks for different attributes. We selected 3 attributes with the greatest rank consistency across 3 experiments: ”rhythmic talking”, ”natural”, ”raspy”. We collected data on these attributes through selective pairwise comparisons of 10K tracks. We selected the pairs to query using an active learning framework (Chen et al., 2013) which enables us to collect a fraction of the data needed to rank at scale. The next steps for this work include training models against these labels, and collecting data on vocal attributes hypothesized to not correlate with genre.

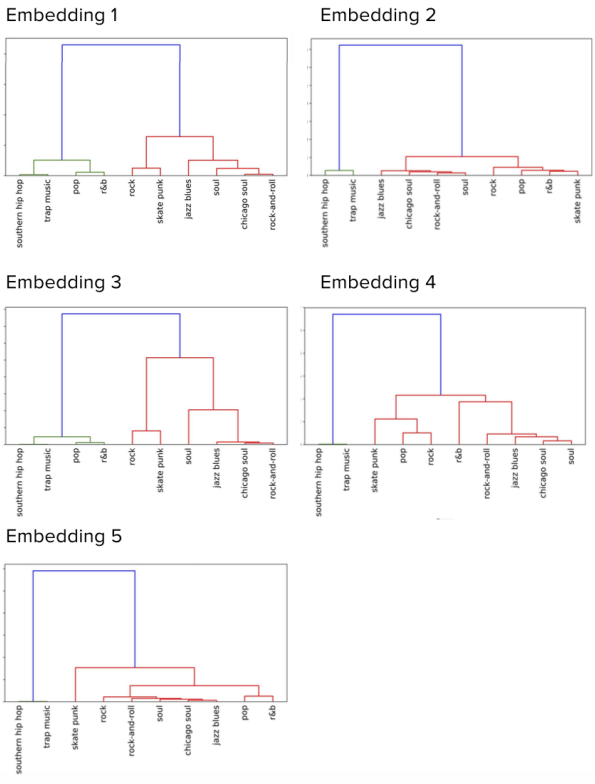

Figure 1. Genre distances in vocal style embedding spaces hierarchically clustered using cosine distance

5. Future work

In the next iteration of work we will expand the number of tracks, genres and the vocal attribute labels in our dataset to obtain more generalizable results. We will investigate the consequences of variations in vocal style embedding spaces. Finally, we leave assessing the impact vocal style similarity can have on search, recommendations and other music features to future work.

Additional details about this work are available in the slides, available at this link: `goo.gl/6LuaAt`.

References

Bittner, Rachel M., McFee, Brian, Salamon, Justin, Li, Peter, and Bello, Juan Pablo. Deep salience representations for f0 estimation in polyphonic music. In *18th International Society for Music Information Retrieval (ISMIR) conference*, 2017.

Caffier, Philipp P, Nasr, Ahmed Ibrahim, Rendon, Maria del Mar Ropero, Wienhausen, Sascha, Forbes, Eleanor, Seidner, Wolfram, and Nawka, Tadeus. Common vocal effects and partial glottal vibration in professional non-classical singers. *Journal of Voice*, 2017.

Chen, Xi, Bennett, Paul N, Collins-Thompson, Kevyn, and Horvitz, Eric. Pairwise ranking aggregation in a crowd-sourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 193–202. ACM, 2013.

Fujihara, Hiromasa, Goto, Masataka, Kitahara, Tetsuro, and Okuno, Hiroshi G. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):638–648, 2010.

Jansson, Andreas, Humphrey, Eric J., Montecchio, Nicola, Bittner, Rachel M., Kumar, Aparna, and Weyde, Tillman. Singing voice separation with deep U-Net convolutional networks. In *18th International Society for Music Information Retrieval (ISMIR) conference*, 2017.

Kim, Youngmoo E and Whitman, Brian. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, volume 13, pp. 17, 2002.

Li, Chao, Ma, Xiaokong, Jiang, Bing, Li, Xiangang, Zhang, Xuewei, Liu, Xiao, Cao, Ying, Kannan, Ajay, and Zhu, Zhenyao. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

Logan, Beth. Nearest-neighbor artist identification. *Proceeding of Music Information Retrieval Evaluation eXchange*, pp. 192–194, 2005.

Nakano, Tomoyasu, Yoshii, Kazuyoshi, and Goto, Masataka. Vocal timbre analysis using latent dirichlet allocation and cross-gender vocal timbre similarity. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 5202–5206. IEEE, 2014.

Pachet, Francois and Aucouturier, Jean-Julien. Improving timbre similarity: How high is the sky. *Journal of negative results in speech and audio sciences*, 1(1):1–13, 2004.

Pachet, François and Cazaly, Daniel. A taxonomy of musical genres. In *Content-Based Multimedia Information Access-Volume 2*, pp. 1238–1245. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

Panteli, Maria, Bittner, Rachel, Bello, Juan Pablo, and Dixon, Simon. Towards the characterization of singing styles in world music. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 636–640. IEEE, 2017.

Potter, John. *Vocal authority: singing style and ideology*. Cambridge University Press, 2006.

Salamon, Justin, Rocha, Bruno, and Gómez, Emilia. Musical genre classification using melody features extracted from polyphonic music signals. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 81–84. IEEE, 2012.

Sturm, Bob L. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, Dec 2013. ISSN 1573-7675. doi: 10.1007/s10844-013-0250-y. URL `https://doi.org/10.1007/s10844-013-0250-y`.

Thalén, Margareta and Sundberg, Johan. Describing different styles of singing: A comparison of a female singer's voice source in"classical","pop","jazz"and"blues". *Logopedics Phoniatrics Vocology*, 26(2):82–93, 2001.

Van den Oord, Aaron, Dieleman, Sander, and Schrauwen, Benjamin. Deep content-based music recommendation. In *Advances in neural information processing systems*, pp. 2643–2651, 2013.