Comparison of VGGish embeddings and perceptually-motivated features for Singing Voice Detection

Shayenne Moura
Universidade de São Paulo
shayenne.moura@usp.br

ABSTRACT

Singing Voice Detection still have place in Music Information Retrieval research, particularly with the new possibilities generated by feature learning for music content recognition. VGGish embeddings are general purpose audio features that can be used as audio descriptors for multiple tasks. We make a performance comparison of singing voice detectors using vocal VGGish embeddings and vocal perceptually-motivated features. For that end, we train Random Forest models using perceptually inspired features (MFCC, Fluctogram, Vocal Variance) and VGGish embeddings. Our results show that VGGish embeddings have classification performance metrics at least comparable to perceptually-motivated features.

1. INTRODUCTION

Singing voice detection is a task in Music Information Retrieval that aims to determine where there is vocal sources present in a sound mixture. This task is a common first step for melody extraction [1], artist recognition [2] and lyrics alignment [3] .

In this work, we are exploring the use of a learned representation for audio signals (VGGish embeddings) to generate a singing voice detector that has results comparable to the ones of a detector generated with vocal perceptually-motivated features.

2. METHOD

We compare the work of Lehner [4] that uses MFCC, Vocal Variance, Fluctogram, Spectral Flatness and Spectral Contraction (all of these audio descriptors related to voice caracteristics on audio signals) with a recent descriptor for general audio purposis VGGish embeddings [5].

We continue the experiments with VGGish embeddings described in [6]. We conduct the error analysis over the MedleyDB dataset [7]. We evaluate classification models using either MFCC or VGGish features with a random forest (RF) classifier.

We perform grid search for hyper-parameters, using 'bootstrap=[True, False]', 'max_depth=[10, 20, 30, 40]', 'max_features=['auto', 'sqrt']', 'n_estimators=[10, 35, 60, 85, 110]'.

Finally, the models are evaluated against the test sets and three distinct outputs: original output, majority vote (1 second probabilities smooth + threshold of 0.5), and optimal binarization (1 second probabilities smooth + threshold calculated from train set).

Audio features are calculated using the source code made available by the authors. All vocal features are calculated in 200 ms segments, without overlap. For VGGish embeddings, the features are calculated in 0.96 second segments, with 0.48 seconds overlap. We choose this segment size to make possible the use of VGGish default parameters.

The ground-truth was based on instrument activations, as defined in the MedleyDB dataset [7]. We consider that a 960 ms segment has singing voice if at least 300 ms of it has active voice, based on [8]. The types of singing voice included in our dataset are: female singer, male singer, vocalist and choir sources.

3. EVALUATION

3.1 Dataset

The experiments are based on the MedleyDB [7] dataset, using only tracks containing singing voice.

We selected the 61 tracks containing singing voice and split them into 10 different train and test subsets. The splits were made as follow: 70% for train subset and 30% for test subset.

To avoid the artist/album effect in our classification experiments [9], we used the medleydb API [1] to make the split with artist conditional division, i.e. the subsets do not share the same artist.

3.2 Results

We find the best hyper-parameters from the grid search phase 'bootstrap=True', 'max_depth=30', 'max_features='sqrt'', 'n_estimators=10'.

With these parameters, we train 20 models (10 for each type of descriptor) using the 10 splits from the dataset. Each singing piece is evaluated in at least one test set.

Attribution: Shayenne Moura, "Comparison of VGGish embeddings and perceptually-motivated features for Singing Voice Detection", *Extended Abstracts for the Late-Breaking Demo Session of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

[1] https://github.com/marl/medleydb

We use the accuracy, precision, recall, and F-score metric to evaluate the performance of the trained models. Table 1 presents the results resumed for the 10 test sets.

	Perceptual features models			**VGGish embeddings models**		
type	**output**	**maj vote**	**opt bin**	**output**	**maj vote**	**opt bin**
ACC	0.84	0.85	0.86	0.86	0.87	0.87
P	0.88	0.86	0.88	0.91	0.88	0.92
R	0.89	0.94	0.91	0.89	0.94	0.90
F1	0.88	0.89	0.89	0.89	0.91	0.90

Table 1. Resumed metrics for all test sets

Table 1 presents comparable metric values for the proposed models, although using VGGish embeddings models do not present any metric lower than the perceptually motivated feature models.

Table 2 shows evaluation metrics for tested pieces from different genres from the MedleyDB dataset.

genre	**Classical**	**Jazz**	**Musical Theatre**	**Pop**	**Rock**	**Singer/ Songwriter**	**World/ Folk**
	Perceptual features models						
ACC	**0.94**	0.94	**0.93**	0.83	0.83	0.80	0.97
P	0.97	0.93	0.94	0.86	0.84	0.86	0.99
R	0.94	0.98	0.99	0.93	0.88	0.87	0.97
F1	0.95	0.95	0.96	0.89	0.86	0.85	0.98
	VGGish embeddings models						
ACC	0.92	**0.97**	0.88	**0.88**	**0.86**	**0.84**	**0.99**
P	0.96	0.97	0.97	0.92	0.90	0.90	0.99
R	0.94	0.98	0.91	0.91	0.87	0.87	0.99
F1	0.95	0.98	0.93	0.92	0.88	0.88	0.99

Table 2. Resumed metrics by separated genres

From Table 2, we notice that specific genres (jazz, pop, rock, singer/songwriter, world/folk) are better classified using VGGish embeddings, while other genres (classical, musical theater) are better classified using perceptually-motivated features (see values in bold). We believe that some sound sources are frequently present when singing voice is active, and the models associate these sources with the desired objective. In the future we could train the models with specific genres and vocal targets, restricting the learning complexity.

4. CONCLUSIONS

In this abstract we describe the results of our experiments where we evaluate the use of VGGish embeddings to perform singing voice detection. We use Random Forest models to classify the singing voice segments from MedleyDB and compare the results to using a set of perceptually motivated features (MFCC, Vocal Variance, Fluctogram and confiability indicators). Our results show that VGGish embeddings have at least comparable metrics in comparison with perceptually motivated descriptors used to perform singing voice detection for Random Forest models on our test sets. For future directions, we plan to include a pitch recognition phase and use the system to perform singing voice transcription.

5. REFERENCES

[1] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[2] A. L. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *22nd Int. Conf.: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[3] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," Ohio State University Columbus United States, Tech. Rep., 2005.

[4] B. Lehner, G. Widmer, and R. Sonnleitner, "On the reduction of false positives in singing voice detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7480–7484.

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[6] S. Moura, "Singing voice detection using vggish embeddings," in *Extended abstract of 19th Int. Soc. for Music Info. Retrieval Conf.*, Paris, France, 2018.

[7] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *15th Int. Soc. for Music Info. Retrieval Conf.*, Taipei, Taiwan, Oct. 2014, pp. 155–160.

[8] P. Belin, S. Fecteau, and C. Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in cognitive sciences*, vol. 8, no. 3, pp. 129–135, 2004.

[9] B. Whitman, G. Flake, and S. Lawrence, "Artist detection in music with minnowmatch," in *Proc. of the 2001 IEEE Signal Processing Society Workshop*. IEEE, 2001, pp. 559–568.