Sensitivity to Instrumentation of a Singing Voice Detector Based on Spectral Features

Shayenne Moura[1]*, **Marcelo Queiroz**[1]

[1]Computer Music Research Group – University of São Paulo
Av. Prof. Luciano Gualberto, 1171 - Butantã, São Paulo - SP, 05508-090

shayenne@ime.usp.br, mqz@ime.usp.br

Abstract. *Detecting voice in a mixture of sound sources remains a challenging task in MIR research. The musical content can be perceived in many different ways as instrumentation varies. We evaluate how instrumentation affects singing voice detection in pieces using a standard spectral feature (MFCC). We trained Random Forest models with song remixes for specific subsets of sound sources, and compare it to models trained with the original songs. We thus present a preliminary analysis of the classification accuracy results.*

1 Introduction

Singing Voice Detection, also referred to as Vocal Detection, is the task of identifying singing voice segments in a piece of audio containing a mixture of sound sources. This task is an intermediate step in many other tasks pertaining to Music Information Retrieval, among them singing voice separation[1] and melody transcription[2].

There are several approaches in the literature in order to identify segments with singing voice [3]. In this work, we use the feature engineering approach, i.e., we use features commonly related to voice processing tasks as input for machine learning algorithms to create a model classifier. This is a preliminary experiment which uses data augmentation based on training models with different remixes of pieces, and we present an evaluation of the classification accuracy using this approach.

In order to make the experiments automatic and easy to reproduce, the scripts for the extraction of audio descriptors are available on this github link[1], and the algorithm and evaluation are available on Jupyter Notebook files.

2 Methodology

Our goal is to compare the sensitivity of classifier models to instrumental remixes, by using a standard feature (MFCC) to perform singing voice detection. We used the MedleyDB dataset [4], which contains singing voice pieces with separate tracks for each instrumental source, and created alternative remixes by combining subsets of the original instrumental tracks.

We defined four families for these remixes, with progressively fewer instruments: (1) the original mix; (2) all monophonic instruments plus drums; (3) all monophonic instruments; and (4) only instruments playing the main melody. This creates an augmented dataset in which we want to measure the performance of singing voice detectors. It should be noted that all families include all vocal tracks, and the motivation is to verify if these new remixes would make the singing voice detection easier, as well as to obtain more data for training. A 5th family, consisting of purely instrumental remixes, i.e. original mixes without vocals, was also considered to provide more training data with non-singing voice examples, in an attempt to counterbalance the 71% rate of positive examples (i.e. segments containing singing voice) in the original data.

The ground-truth was based on instrument activations, as defined in the MedleyDB dataset [4]. We consider that a 960 ms segment has singing voice if at least 50% of its length (not necessarily contiguous) has singing voice. The types of singing voice included in our dataset are: male singer, female singer, male speaker, female speaker, male rapper, female rapper, and vocalists.

Audio features are calculated using 0.96 second segments, with 0.48 seconds overlap. Specifically, we use Librosa [5] 0.6.0 to obtain MFCCs of 40 coefficients using 10 ms segments, out of which we retain the first 13 coefficients (excluding the 0th coefficient); we then summarize every 96 segments (96*10 ms) using the following summary statistics: mean, standard deviation, median, delta and double delta, in order to preserve temporal context (feature dimensionality is 13 * n_statistics).

In the experiments we used Random Forest classifiers with 100 estimators, after considering as alternatives 10, 20, 50, 100 and 150 estimators, because 100 estimators consistently produced the best results in all experimental scenarios. To evaluate detection sensitivity, we conducted a first experiment to compare the classification accuracy of models trained and evaluated within each family of remixes. In a second experiment, we wanted to verify if trained models generalized well by progressively enlarging the training data: (A) training with only the original mixes (family 1); (B) training with original mixes plus alternative remixes that include vocals (families 1+2+3+4); and (C) training with original mixes and all alternative remixes (families 1+2+3+4+5 – including purely instrumental remixes).

*Supported by CNPq.

[1]https://github.com/shayenne/VoiceDetection

3 Evaluation

3.1 Dataset

The experiments were based on the MedleyDB [4] dataset. We selected all 61 tracks containing singing voice and split them into training and test subsets. The split was defined as follows: 80% for the training subset and 20% for the test subset, leaving 48 and 13 songs for the training and test subsets, respectively. This results in 21368, and 3874 960 ms audio segments for training and test, respectively.

To avoid the artist/album effect [6] in our classification experiments, we used the medleydb API[2] to ensure that the subsets do not share the same artists, i.e. if an artist falls into the training subset, all of her songs will be in the same subset.

3.2 Results

We used accuracy to evaluate the performance of the trained models. Figure 1 presents the results within each of the four remix families, i.e. training and testing the models within a single remix family. We can verify that singing voice detection becomes more accurate when training and evaluating with a reduced subset of sound sources, as compared to using all sources in the original pieces.

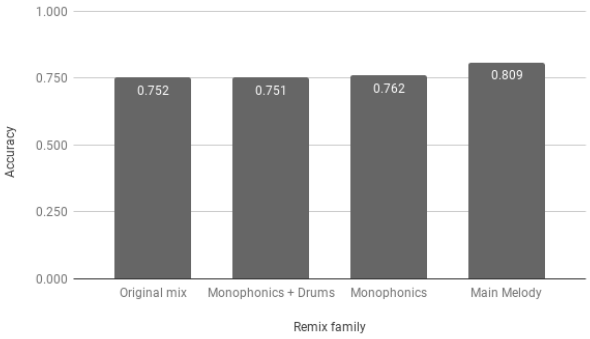

Figure 1: Accuracy using remix families

The results of generalization of our models are presented in Figure 2. In this experiment we trained the models using the three groups of training data discussed in the previous section, i.e. (A) only original mixes, (B) original + alternative remixes with vocals, and (C) original + all alternative remixes. Evaluation was made on the validation set of the original songs.

We see in figure 2 that accuracy decreased about 42% when using an augmented training set including vocal alternative remixes, in comparison to using only the original pieces in the training set. The results of classification accuracy training the model with all alternative remixes decreased yet a little bit more.

Our intuition to explain these negative results is that, even if the alternative remixes create specific contexts within which singing voice detection is slightly easier than in the original context (an interpretation endorsed by Figure 1), these contexts are possibly introducing too much

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

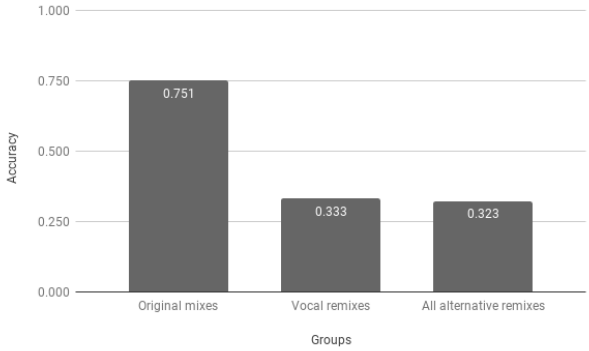

Figure 2: Accuracy in generalization of augmented training sets

dispersion in the generated MFCCs (which are well-known to reflect timbre aspects).

The confusion matrix for the last group, for instance, shows that the classifier has become substantially biased towards labelling segments as *not* containing singing voice (around 67% of all segments correspond to false negatives), suggesting that the MFCCs of the non-singing voice class overlapped most of the singing voice segments in this representation space. So, if data augmentation through diversifying instrumental variety is ever going to be useful in singing voice detection, other audio features, more directly related to the presence of voice, will necessarily have to be included.

Another observation derived from these results is the fact that adding pieces without vocals examples in the training set (in an attempt to balance the positive/negative examples) actually decreased the model ability to accurately classify singing/non-singing segments.

As future work in the direction of data augmentation techniques for singing voice detection, besides including more specific voice-related audio features in the representation space, we consider training the models with different mixes of songs, (e. g. woodwind sources, string sources), and using other classification models (as SVM, Neural Networks), within a specific set of instruments or music style/genre to be evaluated.

4 Conclusions

In this text we reported preliminary results of our experiments on evaluating the use of different instrumental remixes as a data augmentation technique for singing voice detection. We used Random Forest models to classify the singing voice segments from the MedleyDB dataset, using a standard audio feature (MFCC). Our results show that the remixes were not able to increase the classification accuracy in comparison to the use of the original pieces, but gave some insights for future improvement, such as evaluating the models trained with other groups of remixes based on instrumental families and combining MFCCs with other voice-related audio features.

References

[1] Yipeng Li and DeLiang Wang. Separation of singing voice from music accompaniment for monaural recordings. Technical report, Ohio State University Columbus United States, 2005.

[2] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, Aug. 2012.

[3] Kyungyun Lee, Keunwoo Choi, and Juhan Nam. Revisiting singing voice detection: a quantitative review and the future outlook. In *19th Int. Soc. for Music Info. Retrieval Conf.*, Paris, France, 2018.

[4] 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.*, pages 155–160, Taipei, Taiwan, Oct. 2014.

[5] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proc. 14th python in science conference*, pages 18–25, 2015.

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