AN ENSEMBLE METHOD FOR LEARNING TO EXTRACT VOCALS FROM POLYPHONIC MUSICAL AUDIO

Matt McVicar
Intelligent Systems Laboratory
University of Bristol
Bristol, United Kingdom
mattjamesmcvicar@gmail.com

Tijl De Bie
Intelligent Systems Laboratory
University of Bristol
Bristol, United Kingdom
tijl.debie@gmail.com

ABSTRACT

We present a simple method for extracting the sung voice from a polyphonic audio mixture. Our approach is to formulate singing voice separation as a classification or regression task, and make use existing systems which attempt to solve this problem. Specifically, our algorithms use a feature space where the output of existing methods is coupled with spectral features and matrix decomposition techniques. Regression and classification models are trained from this feature space to hard and soft spectral masks at each pixel in the time-frequency representation of the mixture. We hope that this will produce an ensemble method which harnesses the power of existing methods, but is also robust in cases where individual systems fail.

1. INTRODUCTION

Singing Voice Separation (SVS) is deconstruction of an audio mixture containing several sources into two elements. These two signals are the sung melody (the *vocals*) and the other contains everything else (the *background*). A popular approach to solving this problem is to first compute a short-time Fourier transform of the mixed audio. This results in a matrix $\mathbf{X} \in \mathbb{C}^{F \times T}$, where $f = 1, \ldots, F$ and $t = 1, \ldots, T$ index frequency and time respectively. A power spectrum $\mathbf{P} = |\mathbf{X}|^2$ is then computed, from which a *vocal mask* $\mathbf{M}$ of the same shape is constructed. Masks are either *hard* $\left(\mathbf{M} \in \{0,1\}^{F \times T}\right)$ or soft $\left(\mathbf{M} \in \mathbb{R}^{F \times T}\right)$ and can be inferred using expert knowledge, signal processing techniques and/or learned from data using machine learning techniques. Once a vocal mask has been computed, the Hadamark (element-wise) product of $\mathbf{M}$ and $\mathbf{X}$ is used to compute a vocal spectrum. This vocal spectrum is then inverted back to the time domain, resulting in a vocal audio track. The background audio may then be obtained by subtracting the inferred vocals from the mix.

In this submission, we take a simple ensemble classification approach to the SVS problem. For each 'pixel' $p \in \mathbf{P}$ in the power spectrum we compute a collection of features, and use these to classify how likely this pixel is to contain vocals. In the remainder of this extended abstract we outline our feature extraction methods (Section 2), classification scheme (Section 3), and give preliminary experimental results (Section 4), before finally concluding and discussing future directions of research (Section 5).

2. FEATURE EXTRACTION

Let $p \in \mathbf{P}$ be a pixel in $\mathbf{P}$. There are many possible features which may be informative of the vocal activity at p, some of which have proven efficacious in previous research. Specific examples include Harmonic Percussive Source Separation analysis [3] and Robust Principal Component Analysis [1]. However, the soft masks produced by existing systems are no doubt also extremely informative features of the vocal activity at p. These features will then be fed into classification and regression models (see Section 3), meaning that expert systems are able to 'learn' which pixels represent vocal activity. We therefore extracted the following features for each $p \in \mathbf{P}$:

- The log power at p, $\log_{10}(p)$.
- The frequency associated with p.
- 4 Gabor filters centered at p with horizontal, vertical and diagonal rotations. The filters were chosen to have spatial frequency equal to 0.8, horizontal bandwidth equal to 1 and vertical bandwidth equal to 3.
- The sparse component of a robust-PCA analysis of $\mathbf{P}$ at p.
- The harmonic component of a harmonic-percussive source separation analysis of $\mathbf{P}$ at p.
- The output of REPET-SIM [6] [1] applied to $\mathbf{P}$ at p. REPET-SIM attempts to extract the repeating background of a spectrum, prompting us to use $1-$ the value of the soft mask produced by REPET-SIM.
- The output of the deep learning system proposed by [2] [2] applied to $\mathbf{P}$ at p.

[1] http://www.zafarrafii.com/codes/repet_sim.m
[2] https://github.com/posenhuang/deeplearningsourceseparation

	SDR		SIR		SAR	
Mask type	Voice	Music	Voice	Music	Voice	Music
Hard	7.12 ± 2.47	3.80 ± 4.35	9.46 ± 8.20	19.02 ± 5.38	6.17 ± 3.47	6.20 ± 3.52
Soft	8.77 ± 4.11	5.00 ± 2.56	18.66 ± 8.73	12.18 ± 5.35	5.61 ± 3.31	5.67 ± 3.44

Table 1. Source to Distortion Ratio (SDR), Source to Interferences Ratio (SIR), Sources to Artifacts Ratio (SAR) for our experiments. All results are measured in dB relative to the true mix.

3. CLASSIFICATION MODELS

We present two models in this submission. The first learns a hard mask from the feature space to $\{0, 1\}$, for which we use logistic regression. Our second model learns a soft mask from the feature space to a ratio of the vocal energy to the total energy at each pixel in the mixture.

Our system requires ground truth masks for training, which we constructed in the following way. Given power spectra $\mathbf{P}_{\text{Mix}}, \mathbf{P}_{\text{Aca}}, \mathbf{P}_{\text{Ins}}$ representing the mixed audio, acapella and instrumental sources, hard and soft ground truth hard masks were created by element-wise comparison:

$$\mathbf{M}_{\text{hard}} = \begin{cases} 1 & \text{if } \mathbf{P}_{\text{Aca}} > \mathbf{P}_{\text{Ins}} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{M}_{\text{soft}} = \max\left(\min\left(\mathbf{P}_{\text{Aca}}/\mathbf{P}_{\text{Mix}}, 1.0\right), 0.0\right)$$

Here the soft mask is constructed from the ratio of vocal energy to total energy, normalised to be in the range $[0, 1]$. For the hard mask, a logistic regression model was then fitted from the feature space to $\mathbf{M}_{\text{hard}}$. The soft mask was mapped from $[0, 1]$ to $(-\infty, \infty)$ via the logit function. An ordinary least squares model regression model was trained from the feature space to this logit-transformed soft mask, and finally logistically transformed back to $[0, 1]$ in the testing phase after prediction.

4. EXPERIMENTS

Our model was trained using data from the publicly-available subset of the iKala dataset[3] . For scalability reasons, audio was downsampled to 16kHz, and only every 10th frame of the ground truth spectra and masks were used for training the models. Audio processing was conducted using librosa [4] and sci-kit image [7], classification was performed using scikit-learn [5], and evaluation was performed using the BSS-toolbox [8]. Evaluation was conducted using $10-$fold cross validation. 25 songs were held out for test - the remaining 227 were used for training in each fold. Mean and standard deviation of SDR (Signal to Distortion Ratio), SIR (Source to Interference Ratio), and SAR (Source to Artifacts Ratio) for both hard and soft mask versions of our method are shown in Table 1.

Comparing the results from Table 1 to the top-performing submission from MIREX 2014 (IIY2), our approach achieves superior SDR for the sung voice (8.77dB *c.f.* 4.47dB), but worse SDR for the musical background (5.00dB *c.f.* 7.87dB). In terms of SAR and SIR, our method achieves similar results. Our hard mask approach appears to perform better in terms of SAR, prompting us submit both versions for evaluation. In our final submission, we trained our model on all available data from the iKala dataset.

5. CONCLUSIONS

We presented a simple method for separating the sung voice from a polyphonic mix which harnesses the power of existing systems as input features. Internal testing revealed superior singing voice SDR compared to the cutting-edge system from MIREX 2014.

Our current system learns a map from the feature space to a mask at each pixel in the mixture spectrum independently. One may think of these as the ‘observations’ in a graphical model, with ‘hidden’ states representing the vocal activity at a pixel. We were pleasantly surprised to see that the existing system performed well, but are excited about the possibility of also training a model on the hidden chain of states, encoding the dependencies between ‘neighbouring’ pixels, where the neighbourhood may encode nearby pixels in the time, frequency, or harmonic spaces.

Acknowledgements. This work was funded by EPSRC grant EP/M000060/1. The authors would like to thank the authors of [6] and [2] for making their code available.

6. REFERENCES

[1] P. Huang, S. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 57–60. IEEE, 2012.

[2] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1562–1566. IEEE, 2014.

[3] I. Jeong and K. Lee. Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints. *Signal Processing Letters, IEEE*, 21(10):1197–1200, 2014.

[4] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and mu-

[3] `http://mac.citi.sinica.edu.tw/ikala/`

sic signal analysis in python. In *14th annual Scientific Computing with Python conference*, SciPy, July 2015.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Z. Rafii and B. Pardo. Repeating pattern extraction technique (repet): A simple method for music/voice separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):73–84, 2013.

[7] S. Van Der Walt, J. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[8] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.