SINGING VOICE SEPARATION USING DEEP NEURAL NETWORKS AND F0 ESTIMATION

Gerard Roma, Emad M. Grais, Andrew J.R. Simpson, Mark D. Plumbley
Centre for Vision, Speech and Signal Processing.
University of Surrey, UK
g.roma@surrey.ac.uk grais@surrey.ac.uk
andrew.simpson@surrey.ac.uk m.plumbley@surrey.ac.uk

ABSTRACT

Deep Neural Networks (DNN) have become a popular approach for speech enhancement, and singing voice separation. DNNs are typically trained to estimate a time-frequency mask using ground truth examples. In this submission, we combine DNN estimation as a first step with traditional refinement via F0 estimation, using the YINFFT algorithm.

1. INTRODUCTION

Single-channel source separation is generally a challenging problem, where estimates of component signals are obtained from a mixture using only one sensor. A common simplification in the case of musical audio consists on considering separation of a target signal (e.g. singing voice) against the rest (accompaniment). For pitched instruments such as the singing voice, it is common to use pitch estimation to obtain or improve separation. For example the combination of Robust PCA (RPCA) with pitch estimation via subharmonic summation (SHS) has recently achieved good results [3]. This submission explores a similar architecutre but using a supervised approach based on Deep Neural Networks (DNN) for the initial estimation. Unlike RPCA and other low-rank methods, DNN-based estimation does not rely on the specifics of pop music (i.e repetitive background), but the quality depends on the amount of available training data.

2. METHOD

2.1 Initial separation

Our algorithm works in the time-frequency domain, in this case the Short-Time Fourier Transform (STFT). We represent signals as time-frequency matrices with indices t, f. Let X be a mixture of two signals: a vocal signal V and an accompaniment signal A. An Ideal Binary Mask (IBM) can be simply defined as

$$B_{t,f} = \begin{cases} 1 & \text{if } |V_{t,f}| > |A_{t,f}| \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

These masks rely on the assumption that each time-frequency bin is perceived to belong to either V or A but not to both at the same time. The element-wise multiplication $B \odot X$ usually renders a good approximation of V. A DNN is then trained to learn the mapping from mixture spectral magnitudes to the corresponding IBMs. The result is an estimate $\hat{B}$ of the mask that allows obtaining an estimate of the target and accompaniment signals as

$$\hat{V} = \hat{B} \odot X \tag{2}$$

$$\hat{A} = |1 - \hat{B}| \odot X \tag{3}$$

The framework is similar to our previous experiments [4] but in this case for simplicity the network is trained using only 2 additional context frames around each spectral frame. Also, we use an improved version of the Stochastic Gradient Descent (SGD) including early stopping with a validation set and learning rate decay.

2.2 F0 estimation

After training the model, prediction from the DNN provides a fair approximation of the singing voice. We use the popular YINFFT algorithm [1] to obtain both a pitch contour and a pitch confidence value. The pitch confidence value is used to identify unvoiced speech regions.

2.3 Mask refinement

Using the F0 estimate from the pitch detector, a harmonic mask is constructed by computing the partials corresponding to the detected F0 in each frame. The harmonic mask is then multiplied with the initial estimate (Figure 1) in order to remove components that are too far away from the estimated harmonics (when the picth confidence is high). The complement of the harmonic mask is multiplied with the accompaniment mask estimate in order to remove the vocal harmonics.

3. EXPERIMENTAL SETUP

We used 200 clips from the iKala dataset [2] dataset for training our DNN model, and 50 more for testing. Sounds

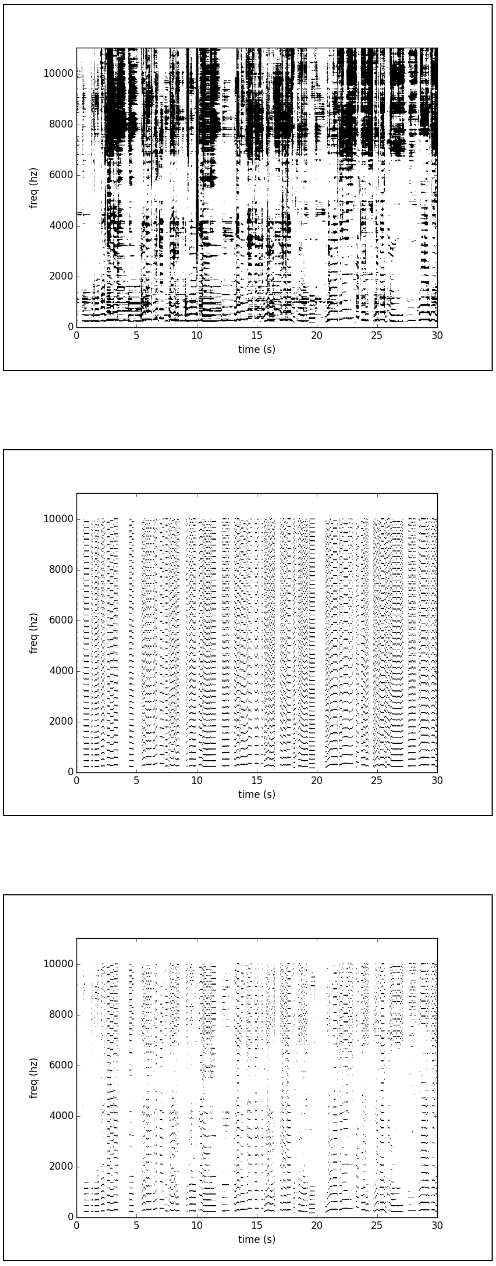

Figure 1. Estimate, harmonic and final vocal masks

were downsampled to 22050 Hz and then used to compute STFT frames using 100ms windows with 75% overlap. The resulting frames of 1025 bins were stacked using a shingling window of 3 frames. Binary masks for the output were stacked in the same way. These frames of $3\dot{1}025$ bins were used to train a DNN with 3 hidden layers. All hidden layers had the same dimension as the input and the output. The DNN was trained with stochastic gradient descent, using a momentum coefficient of 0.9. 20% of the training data was used for validation and early stopping. The learning rate was scheduled to decay after the gain in training error went below 0.0001.

For estimating the separated singing voice and accompaniment signals, the spectral magnitudes of the mixture of each unseen example was fed into the DNN to obtain an initial estimate. The estimate of the singing voice was then used to obtain an estimate of the pitch using YINFFT. An interval of 100 to 700 Hz was used to restrict the search for the pitch. From the estimate of the pitch, when the pitch confidence was above 0.5, a harmonic mask was constructed with all harmonics of the fundamental up to 10 Khz. Spectral bins closer than a fix threshold of 20Hz were added to the harmonic mask. The harmonic mask was then used to improve the separation of both singing voice and accompaniment.

In initial experiments, the harmonic mask correction step resulted in small objective improvements (around 0.5dB GNSDR). In informal subjective tests, noticeable improvements in the sound quality of estimates were perceived. Source code to reproduce our submission is available online [1] .

4. ACKNOWLEDGEMENT

This work is supported by grant EP/L027119/1 from the UK Engineering and Physical Sciences Research Council (EPSRC).

5. REFERENCES

[1] Paul M Brossier. *Automatic annotation of musical audio for interactive applications*. PhD thesis, Queen Mary, University of London, 2006.

[2] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang. Vocal activity informed singing voice separation with the ikala dataset. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 718–722. IEEE, 2015.

[3] Yukara Ikemiya, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Singing voice separation and vocal f0 estimation based on mutual combination of robust principal component analysis and subharmonic summation. *arXiv preprint arXiv:1604.00192*, 2016.

[4] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 429–436, Liberec, Czech Republic, 2015.

[1] `http://cvssp.org/projects/maruss/mirex2016/`