SOURCE SEPARATION WITH WEAKLY LABELLED DATA: AN APPROACH TO COMPUTATIONAL AUDITORY SCENE ANALYSIS

Qiuqiang Kong[1,2], Yuxuan Wang[1], Xuchen Song[1], Yin Cao[2], Wenwu Wang[2], Mark D. Plumbley[2]

[1] ByteDance AI Lab
[1]{kongqiuqiang, wangyuxuan.11, xuchen.song}@bytedance.com
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2] {yin.cao, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

Source separation is the task of separating an audio recording into individual sound sources. Source separation is fundamental for computational auditory scene analysis. Previous work on source separation has focused on separating particular sound classes such as speech and music. Much previous work requires mixtures and clean source pairs for training. In this work, we propose a source separation framework trained with weakly labelled data. Weakly labelled data only contains the tags of an audio clip, without the occurrence time of sound events. We first train a sound event detection system with AudioSet. The trained sound event detection system is used to detect segments that are most likely to contain a target sound event. Then a regression is learnt from a mixture of two randomly selected segments to a target segment conditioned on the audio tagging prediction of the target segment. Our proposed system can separate 527 kinds of sound classes from AudioSet within a single system. A U-Net is adopted for the separation system and achieves an average SDR of 5.67 dB over 527 sound classes in AudioSet.

Index Terms— Source separation, weakly labelled data, computational auditory scene analysis, AudioSet.

1. INTRODUCTION

Source separation is the task of separating sound sources in an audio recording. Source separation is fundamental for computational auditory scene analysis (CASA) [1]. In essence, CASA systems are machine listening systems that aim to separate mixtures of sound sources in the same way that human listeners do. The goal of a CASA system is to detect and separate sound sources from an audio recording. CASA is a challenging problem because there are large number of different types of sound events in the world. Sound events can occur simultaneously, leading to the well known cocktail party problem.

Source separation has been researched for several years. Early work on source separation includes non-negative matrix factorization [2, 3] by learning dictionaries for different sound sources. Sparse representations have been proposed for music separation [4]. An unsupervised method using average harmonic structure modeling was proposed for music source separation [5]. Recently neural network based methods have been used for source separation which learn a regression from the mixture of sources to an individual source. Neural network based methods include fully connected neural networks [6], convolutional neural networks (CNNs) [7, 8, 9] and recurrent neural networks (RNNs) [10, 11]. Weakly supervised source separation methods include [12, 13, 14, 15, 16, 17].

However, most source separation systems are designed to separate specific sound classes such as speech or music. In contrast, CASA is a more challenging task that requires the separation of all sound sources in the world. First, there can be hundreds of sound classes in the real world: this increases the difficulty of separating all of these sound sources. Second, previous source separation systems require mixtures and clean source pairs for training. For example, to separate vocals and accompaniment, researchers need to collect clean vocals and accompaniment sources. However, collecting these clean sources is time consuming, so the size of dataset is often limited. In addition, it is impractical to collect clean sources for such a large number of sound classes in the real world for CASA such as sounds of nature and sounds of animals. On the other hand, AudioSet is a large-scale dataset that contains 5000 hours of data with 527 sound classes. AudioSet is a weakly labelled dataset, that is, only the tags of an audio clip are provided, without the occurrence time of the sound classes, neither the clean sources.

To our knowledge, previous work has not investigated training a source separation system without clean sources. Human beings can learn to identify and separate a sound source even if they have not heard the clean source. Humans can detect sound events around them and learn what the sounds are even if the sound events are mixed with other sounds, which is the usual case in the real world.

In this work, we propose a source separation framework, to separate a large number of sound classes, that is trained with weakly labelled data only. As far as we know, this is the first attempt to solve the general source separation problem for a large number of sound classes trained with weakly labelled data. First, we train a sound event detection (SED) system on AudioSet. Then the SED system is used to detect anchor segments that is most likely contain a sound event in an audio clip. Then a regression is learned from a mixture of two randomly selected anchor segments to the target anchor segment, using the audio tagging prediction of the target anchor segment as condition.

This paper is organized as follows. Section 2 introduces the regression based source separation method. Section 3 proposes the source separation framework of a large amount of sound classes trained with weakly labelled data. Section 4 shows the experiments of source separation of AudioSet. Section 5 concludes this work.

2. REGRESSION BASED SOURCE SEPARATION

Neural network based regression methods have been used to solve music separation and speech separation [6, 7, 8, 10, 11]. Regression based source separation methods learn a mapping from a mixture of sources to a target source to be separated. We denote the individual sources as $s_1, .., s_K$, where K is the number of sound sources and each source s is a time domain signal. The mixture is denoted as x. Previous source separation systems have built regressions for each individual source s_k:

$$f(x) \mapsto s_k. \qquad (1)$$

For speech enhancement, s_k can be clean speech. For music source separation, s_k can be vocal or accompanies. In this work, we build f in the time-frequency (T-F) domain [6, 11]. The mixture x and source s_k are transformed to T-F spectra denoted as X and S_k using the short time Fourier transform (STFT). The magnitude and phase of X are denoted as $|X|$ and $e^{j\angle X}$ where $X = |X|e^{j\angle X}$. The magnitude $|X|$ is called a spectrogram. A neural network g is built to regress from a mixture spectrogram $|X|$ to an estimated source spectrogram $|\hat{S}|$. Then, the phase from the mixture is used to recover the STFT of the estimated source $\hat{S} = |\hat{S}|e^{j\angle X}$. Finally an inverse STFT is applied on $\hat{S}$ to obtain the separated source $\hat{s}$.

In this work we model the neural network g with a U-Net [9]. U-Net is a variation of CNN that consists of several encoding and decoding blocks modeled by convolutional layers. Each encoding block consists of two convolutional layers and a downsampling layer that halves the size of the feature maps. Each decoding block consists of two convolutional layers and a transposed convolutional layer. The U-Net adds additional skip connections between blocks at the same hierarchical level in the encoder and decoder. This allows low-level information to flow directly from the high resolution input to the high-resolution output.

3. SOURCE SEPARATION WITH WEAKLY LABELLED DATA

3.1. Sound event detection

Each audio clip in AudioSet is labelled with only the tags of the sound classes, without their occurrence time. To solve this problem, we train a SED system [18, 19] using weakly labelled AudioSet. The trained SED system is used to detect when a sound event happens in an audio clip. We choose the segment that most likely contains a sound event according to the SED prediction and refer to the selected segment as an *anchor segment* for a sound class in the audio clip.

To train the SED system with weakly labelled data, a log mel spectrogram is used as feature for an audio clip. Then a CNN is applied on the log mel spectrogram [18]. To predict the presence probability of sound events over time, we apply time distributed fully connected layers followed by sigmoid non-linearity to predict the presence probability of sound events over time. We denote the probability over time as $O(t) \in [0,1]^K, t = 1, ..., T$ where T is the number of time steps in the distributed fully connected layer. In training, the time distributed probability $O(t)$ is pooled to a clipwise prediction $\hat{y} \in [0,1]^K$ by a maximum aggregation function. The segment selected by the maximum aggregation has high confidence of containing the sound event to be detected [18]. Binary crossentropy $-\sum_{k=1}^{K}[y_k\ln\hat{y}_k + (1 - y_k)\ln(1 - \hat{y}_k)]$ is used for training the SED system.

In inference, an audio clip from AudioSet is used as input to the SED system. The time distributed prediction of sound events is obtained from $O(t)$. For a sound class k, we select the time step with the highest presence probability $\tau = \underset{t}{\mathrm{argmax}}\ O(t)_k$. The anchor segment is obtained by selecting a waveform segment with τ as centre. The trained SED system is from [19][1].

3.2. Source separation

One way to build the source separation system is to build one system for each individual sound event as mentioned in Section 2. However this is impractical because the number of source separation systems will increase linearly with the number of sources to be separated. In this section we propose a unified framework that can separate all sources in one system.

To begin with, we randomly select two sound classes from AudioSet. For each sound class, an audio clip containing the sound class is selected. The SED system is applied to an audio clip to select the anchor segment that contains the sound event. The selected anchor segments for the two sound classes

---

[1]https://github.com/qiuqiangkong/audioset_tagging_cnn

are denoted as s_1 and s_2, respectively. In addition to the anchor segment, a condition vector is used as an extra input to control what source to separate. For example, we denote the condition vector as c_j for source s_j. Then the proposed source separation system can be described as:

$$f(s_1 + s_2, c_j) \mapsto s_j \tag{2}$$

where $j \in \{1, 2\}$. Equation (2) shows that the separated audio depends on both the input mixture and the condition vector. The condition vector c_j controls what source to be separated.

One challenge of source separation is the aforementioned separation systems in Section 2 require both mixture and clean source for training. A clean source s_j only contains sound of one class. However, AudioSet only provides weakly labelled audio clips and there is no information when a sound event occurs. In addition, AudioSet usually contain multiple sound classes in one audio clip so is not clean. The anchor segments selected by the SED system may contain both the selected sound class and other polyphonic sound classes.

We propose that training a source separation system does not necessarily require clean sources. The solution is by setting the conditional vector c_j to reflect the target source s_j to be separated. We set $c_j \in [0, 1]^K$ that reflects the presence probability of all K sound classes in s_j. If an anchor segment s_j only contains clean source for the k-th sound class, then c_j should be one hot encoding of the k-th sound class. On the other hand, if an anchor segment s_j contains multiple sound sources then c_j should have positive values for all occurred sound sources and zero for other sound classes that do not occur in s_j. AudioSet does not provide labels that can be directly used as c_j, for the reason that the label of audio clips are in 10-second level but not in anchor segment level. Instead, we use the audio tagging prediction of s_j from the trained SED system as the condition vector c_j. This prediction can reflect the presence probability of sound events in s_j. This proposed method does not require s_j to be clean. In training, we let the system to learn the following regressions:

$$f(s_1 + s_2, c_j) \mapsto s_j \tag{3}$$

$$f(s_j, c_j) \mapsto s_j \tag{4}$$

$$f(s_j, c_{\neg j}) \mapsto \mathbf{0} \tag{5}$$

where $j \in \{1, 2\}$. The symbol $\neg j$ indicates any class index that is different from j. Equation (3) represents learning from a mixture to a separated source s_j conditioned on c_j. Equation (4) represents learning an identity mapping. That is, the system should learn to output the mixture itself if the condition describes the mixture. Equation (5) represents a zero mapping. That is, the system should output silence if it is conditioned on $c_{\neg j}$. The vector $\mathbf{0}$ is an all zero vector. Fig. 1 shows the framework of the proposed source separation with weakly labelled data method.
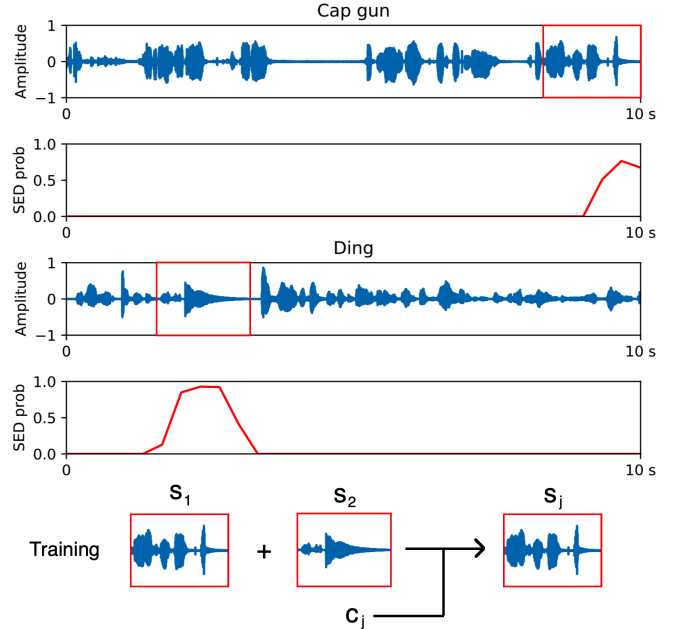
Fig. 1. The first and third rows are waveforms of two audio clips. The red rectangles are selected anchor segments described in Section 3.1. The second and fourth rows are the SED predictions of two sound classes.

3.3. Inference

In inference, we first predict the presence or absence of sound classes in an audio clip using the SED system. Only the sound classes predicted as present in the audio clip will be separated. For the k-th sound event, we use one-hot encoding as a conditional vector $c_k = \{0, .., 0, 1, 0, ..., 0\}$, where the k-th index of c_k is 1 while others 0. The design of the condition vector is to control what sound class to separate. When only one value of c_k is non-zero, the system will only output the separate source of one sound class. By choosing different condition vectors different separated sources can be obtained. Demos of the proposed system can are available[2].

4. EXPERIMENTS

We evaluate the the proposed source separation system on AudioSet. AudioSet is a large-scale audio dataset with an ontology of 527 sound classes [20] in the released version. The balanced training subset contains 22,050 audio clips. The evaluation set consists of 20,371 audio clips. Most audio clips have a duration of 10 seconds. AudioSet is a weakly labelled dataset. All downloaded audio clips are converted to monophonic with a sampling rate of 32kHz. The SED system is trained on the full training set. Log mel spectrogram of audio clips are extracted with a STFT window of 1024 samples

---

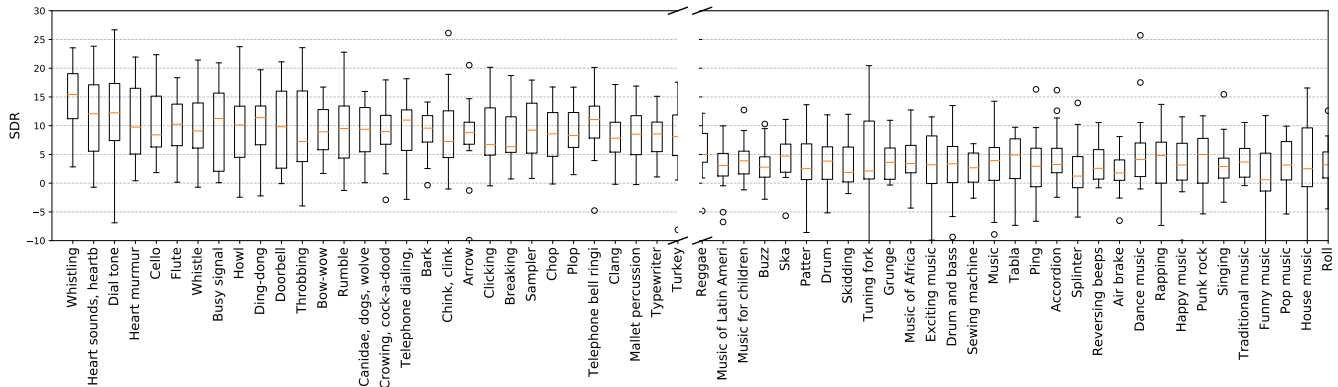[2]https://github.com/qiuqiangkong/audioset_source_separation

Fig. 2. SDRs of sound classes sorted by median value in a descending order.

and a hop size of 320 frames which leads to 100 frames in a second. The number of mel frequency bins is set to 64 following [21]. The CNN for SED consists of 13 layers. A time distributed fully connected layer is applied to the final convolutional layer to obtain the presence probability of sound events. Adam optimizer [22] with a learning rate of 0.001, β_1 of 0.9 and β_2 of 0.999 is used for training the SED system. In inference, the trained SED system is used to choose τ. The anchor segment has a duration of 1.6 seconds with τ as the centre.

For the source separation system, we apply a U-Net [9]. The input to the U-Net is the spectrogram of the mixture of two anchor segments. The spectrogram is obtained by applying a STFT on the waveform with a window size of 1024 and a hop size of 256 in an anchor segment. The input has a shape of 160×513 representing the number of frames and frequency bins. The U-Net consists of 4 encoder blocks and 4 decoder blocks. The number of feature maps in the encoder are 64, 128, 256 and 512 [23]. The number of feature maps in the decoder are 512, 256, 128 and 64. Each convolutional layer consists of a convolutional operation, a batch normalization and a ReLU non-linearity. The condition vectors is mapped to embedding vectors by a learnable matrix. The embedding vectors are added to after each ReLU operation in all layers as a bias. The condition vector controls what sources to separate. Adam optimizer with a learning rate of 0.001, β_1 of 0.9 and β_2 of 0.999 is used for training the source separation system. Mean absolute error between the estimated spectrogram $\hat{S}_j$ and the target spectrogram S_j is used as the loss function in training the network.

In inference, we use the evaluation set in AudioSet for evaluation. we randomly select and mix two anchor segments of two sound classes under SDR of 0 dB. The task is to separate the anchor segments from the mixture. We evaluate the source separation system with source to distortion ratio (SDR), source to inferences ratio (SIR) and sources to artifact ratio (SAR) [24]. The higher number indicates the better performance. Fig. 2 shows the box plot of SDRs of sound

classes sorted in a descending order. Most of sound classes achieve positive SDR indicating the effectiveness of the proposed source separation method. Some sound classes such as whistling achieves SDR of around 15 dB, dial tone of 12 dB and heart murmur of 10 dB. Fig. 2 shows that the separation difficulty is different from class to class. The proposed system modeled with U-Net achieves a SDR, SIR and SAR of 5.67 dB, 9.38 dB and 11.15 dB, respectively, averaged over all 527 sound classes in AudioSet.

5. CONCLUSION

We have presented a source separation framework trained with weakly labelled data only. The proposed source separation system is a unified system trained on AudioSet which can separate a large amount of sound classes using a unified system. The proposed system can detect and separate any sound sources in an audio recording, offering a new approach to computational auditory scene analysis. To begin with, a sound event detection system is trained to select anchor segments of sound classes in audio clips. Then, the mixture of two anchor segments and the condition vector of one anchor segment are used as input to the source separation system modeled by a U-Net. The proposed source separation system does not require clean sources for training. Overall, the system achieves an average SDR of 5.67 dB over 527 sound classes in AudioSet. In future, we will focus on improving the performance of the source separation system using weakly labelled AudioSet.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] D. Wang and G. J. Brown, "Computational Auditory Scene Analysis: Principles, Algorithms and Applications," 2006.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] J. Bruna, P. Sprechmann, and Y. LeCun, "Source separation with scattering non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1876–1880.

[4] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2009.

[5] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 766–778, 2008.

[6] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[7] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 1265–1269.

[8] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.

[9] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *International Society for Music Information Retrieval (ISMIR)*, 2017.

[10] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[11] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 261–265.

[12] N. Zhang, J. Yan, and Y. Zhou, "Weakly supervised audio source separation via spectrum energy preserved wasserstein learning," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2017, pp. 4574–4580.

[13] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2391–2395.

[14] D. Stowell and R. E. Turner, "Denoising without access to clean data using a partitioned autoencoder," *arXiv preprint arXiv:1509.05982*, 2015.

[15] E. Karamatlı, A. T. Cemgil, and S. Kırbız, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1349–1353, 2019.

[16] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 777–787, 2019.

[17] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey, "Universal sound separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 175–179.

[18] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 121–125.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv preprint arXiv:1912.10211*, 2019.

[20] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[21] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and Mark. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2014.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.