IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101260

SURVEY ON AUDIO SOURCE SEPARATION AND NOTE PROCESSING ALONG WITH EDUCATION

Minnuja Shelly[1],Safrin[2],Shreethal Janardhanan[3],Shuhaib P M[4],Gopika T G[5]

Assistant Professor, Department of Computer Science And Engineering,Universal Engineering College, Vallivattom , Thrissur, India.[1]

B.Tech Student, Department of Computer Science And Engineering,Universal Engineering College ,Vallivattom , Thrissur, India.[2,3,4,5]

Abstract: Users focus upon real-time music played to them, and at the same time, also enjoys making music. We are conducting a group survey attempting to obtain information that can help eradicate wrong assumptions in designing systems involving music-based learning systems. Our main purpose is to present an overall midi system product. In this paper, we exhibit our initial findings and analyses based on the music requests by users we have received to date. This paper also deals with the separation of music into individual instrument tracks which is known to be a challenging problem. We describe two different deep neural network architectures for this task, a feed-forward and a recurrent one, and show that each of them yields state-of-the-art results on the SISEC DSD100 dataset. The accuracy is estimated for each note played by the user.

Keywords: Video Deep Learning, RNN, LSTM, Note Music, Chord Estimation

I. INTRODUCTION

This survey is based on a music conversion and processing by source separation and providing a platform to learn piano on any song. The primary goal of the Music source separation and audio post-processing into the MIDI(MSSAPP-MIDI) project is the acquisition of real-world user data so that an empirically justifiable framework can be developed within which the sci. In the past, there had been several apps that used definite song sets which could only be used on a specific set of music. The underlying system uses infinite sets of songs and also real-time piano as input that is applied for correction details if any.

II. THEORY

A. RNN

Recurrent neural networks (RNN) is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data.It is used in the state of the art algorithm for sequential data and are used by Apple's Siri and and Google's voice search.RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory.

B. LSTM

Long short-term memory (**LSTM**) is an artificial recurrent neural network (RNN) architecture which used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).

III. RELATED WORK

Here we introduce each papers based on the technologies used in the Audio Source Separation And Note Processing Along With Education and they are arranged in technologies bases.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101260

For the problem of separating concur- rent speech through a spatial filtering stage and a subsequent time-frequency masking stage. Log-sigmoid masks are optimized to maximize the kurtosis of the separated speech. Experiments on the Pascal Speech Separation Challenge II show significant improvements compared to previous approaches. Using the kurtosis as a scale parameter for speech separation, we investigate the use of a linear constrained minimum variance (LCMV) beam former. This gave significantly better results than a super directive beam former and was only slightly inferior to the MMI beam former[1].

Taiwanese researchers have developed a sheet music generator based on deep learning techniques. The DeepSheet system is able to generate sheet music with accuracy of approximately 76%. It would be a great help if sheet music can be automatically generated for all kinds of music lovers, such as professional musicians, music enthusiasts, amateurs, and other potential users. The DeepSheet system needs to be equipped the capability to detect the pauses and chords from a given audio wave data.

The proposed system first applies a.ndouble stage HPSS as pre-processing. Features are then extracted from a filter bank on a Mel scale and supplied as input to the deep. The blocks of our system are described in more de- tails below. We are working on a deep neural network that learns how to classify binary signals. The architecture of the network is determined by the hidden layers and the number of neurons or LSTM blocks within each layer.

Chords are defined by the simultaneous sounding of two or more musi- cal notes. The interval relationships between these notes determine the type of chord. We aim to design a system capable of classifying audio frames as one of 108 chords, including 12 variations of major, minor, augmented, suspended 2nd and suspended 4th. Some systems make use of the constant-Q transform [1, 6, 8] with some of these employing a tuning algorithm to allow for differences in tuning. Other approaches calculate the chromagram directly from the dis- crete Fourier transform (DFT) of the input signal.Some existing chroma calculation techniques include all energy within a given fre- quency range in the amplitude value of a certain pitch class. Our approach identifies only the energy in the harmonics within a given range. The square root of the magni- tude spectrum is taken to reduce the amplitude difference between harmonic peaks. The chroma vector,C, is calculated by.P=12, the number of notes playing a rhythmic accompaniment, and. L(f) where L is the frame size and F is the sampling frequency.

Our approach ex- amines 2 octaves of the spectrum, between 130.81Hz and523.25Hz. We hypothesise that the majority of instruments playing a rhythmical accompaniment play the lower register of the instrument.

Each layer contains one unique neuron with a logistic activation function, whose output is an estimated probability of presence. In [20] we show that in a super-vised gradient-trained deep neural network with random weights ini- tialization, layers far from the output are poorly optimized. In our work, we extended this procedure to automati- cally learn the network architecture during the training. We used the Jamendo Corpus, a publicly avail-able dataset including singing voice activity annotations. Over-fitting is controlled by early-stopping: training starts with a step for the gradient descent$\eta = 10^{-5}$ and a momentumm= 0.9.

Sudo RM -RF: Efficent networks for universal audio sorce separation is easy to see that the proposed models can match and even out-perform the separation performance of other several state-of-the-art models using orders of magnitude less computational requirements.

On par with many state-of-the-art approaches in the literature [2, 9, 3, 6], SuDoRM-RF performs end-to-end audio source separation using a mask-based architecture with adaptive encoder and decoder basis. The input is the raw signal from a mixture $x \in R^T$ with T samples in the time-domain. First we feed the input mixture x to an encoder E in order to obtain a latent representation for the mixture

$v_x = E(x) \in R$

$C_E \times L$

Consequently the latent mixture repre- sentation is fed through a separation module S which estimates the corresponding masks $m_{b_i} \in R^{C_E \times L}$ for each one of the N sources $s_1, \cdots, s_N \in R^T$ which constitute in the mixture. The estimated la- tent representation for each source in the latent space v_{b_i} is retrieved by multiplying element-wise an estimated mask m_{b_i} with the en- coded mixture representation v_x. Finally, the reconstruction for each source s_{b_i} is obtained by using a decoder D to transform the latent- space v_{b_i} source estimates back into the time-domain $s_{b_i} = D$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101260

(vbi). An overview of the SuDoRM-RF architecture is displayed in Fig- ure 1. The encoder, separator and decoder modules are described in Sections 2.1, 2.2 and 2.3, respectively. For simplicity of our notation we will describe the whole architecture assuming that the processed batch size is one. Moreover, we are going to define some useful op- erators of the various convolutions which are used in SuDoRM-RF.

The object detector is trained for a vocabulary of C objects. In general, this detector should cover any potential sound-making object categories that may appear in training videos. The implementation uses the Faster R-CNN [36] object detector with a ResNet-101 [17] backbone trained with Open Images [26]. For each unla- beled training video, uses the pre-trained object detector to automatically2 find objects in all video frames. Then, gather all object detections across frames to obtain a video- level pool of objects. For Audio-Visual Separator the detected object re- gions to guide the source separation process. A related design for multi-modal feature fusion is also used in [13, 30, 31] for audio spatialization and separation. How- ever, unlike those models, our separator network combines the visual features of a localized object region and the audio features of the mixed audio to predict a magnitude spectro- gram mask for source separation. The network takes a detected object region and the mixed audio signal as input, and separates the portion of the sound responsible for the object. A ResNet-18 network is used to extract visual features after the 4th ResNet block with size (H/32) × (W/32) × D, where H, W, D denote the frame and channel dimensions. Then pass the visual feature.

ACKNOWLEDGMENT

We utilize this opportunity to convey our gratitude towards all those who have helped us directly or indirectly for the completion of our work. We deeply and wholeheartedly thank Dr.Sreeraj R -HOD, Computer Science and Engineering for his extreme valuable advice and encouragement. We especially thankful to our guide and supervisor Mrs Minnuja Shelly -Assistant Professor ,Computer Science and Engineering for giving us valuable suggestions and critical inputs in the preparation of this paper. We would like to extend our sincere gratitude to all faculty of Computer Science and Engineering department for the support and suggestions that helped us in the development of our work to what it is now. We thank our parents and friends for the mental support provided during the course of our work at the times when our energies were the lowest.

REFERENCES

[1] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp,Naoya Takahashi2and Yuki Mitsufuji," Improving Music Source Separation Based On Deep Neural Networksthrough Data Augmentation And Network Blending", ICASSP 2017
[2] Takeshi Hori, Kazuyuki Nakamura, Shigeki Sagayama," Music Chord Recognition From Audio Data Using Bidirectional Encoder-decoder LSTMs", Proceedings of APSIPA Annual Summit and Conference 2017
[3] J. Lee and J. Downie (2004), "Survey of Music Information Needs, Uses, And Seeking Behaviours: Preliminary Findings," ISMIR
[4] Patel, A. (2003), "Language, music, syntax and the brain," Nature Neuroscience (6): 674-681
[5] Fremerey, C., M., M., & M., C. (2009), "Towards Bridging the Gap between Sheet Music and Audio." Knowledge Representation for Intelligent Music Processing.
[6] Suzy, S., Neuman, B., & Megan, L. (2017, August 10), "The Importance of Sheet Music to Music Theory." Retrieved from TakeLessons Blog: https://takelessons.com/blog/sheet-music.
[7] Gan, T. (2005), "Musica colonial: 18th century music score meets 21st century digitalization technology," Proceedings of the 5th ACM/IEEECS Joint Conference on Digital Libraries, JCDL '05, Denver.
[8] Luth, N. (2002). "Automatic Identification of Music Notations," WEDELMUSIC.
[9] Hsu, Y. L., Lin, C. P., Lin, B. C., Kuo, H. C., Cheng, W. H., & Hu, M. C. (2017), "DeepSheet: A sheet music generator based on deep learning," Multimedia & Expo Workshops (ICMEW): 285-290, Hong Kong: IEEE.
[10] Hori, T., Nakamura, K., & Sagayama, S. (2017), "Music chord recognition from audio data using bidirectional encoder-decoder LSTMs" ,Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).
[11] Leglaive, S., Hennequin, R., & Badeau, R. (2015), "Singing voice detection with deep recurrent neural networks," Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia.
[12] Camron, G. (2016, August 12), "Recurrent Neural Networks for Beginners" – Camron Godbout – Medium. Retrieved from https://medium.com/@camrongodbout/recurrent-neural-networks-for-beginners-7aca4e933b82.
[13] Hochreiter, S., & Schmidhuber, J. (1997), "Long short-term memory," Neural Computation, 9 (8): 1735-1780.
[14] James, R., audio_segment - pydub 0.9.5 documentation. (Pydoc.io) Retrieved from https://www.pydoc.io/pypi/pydub0.9.5/autoapi/audio_segment/index.html
[15] Abien, F. A. (2018), "Deep Learning using Rectified Linear Units (ReLU)," arXiv preprint arXiv:1803.08375 .
[16] Diederik, P. K., & Ba, J. (2015), "Adam: A Method for Stochastic Optimization," 3rd International Conference for Learning Representations. San Diego.
[17] Toroghi, R. M., Faubel, F., & Klakow, D. (2012), "Multi-channel speech separation with soft time-frequency masking," SAPA-SCALE Conference.
[18] Takuya, F. (1999), "Realtime chord recognition of musical sound: A system using common lisp music," Proceedings of the 25th International Computer Music Conference: 464-467.
[19] Raphael, C. (1999), "Automatic segmentation of acoustic musical signals using hidden Markov models," IEEE transactions on pattern analysis and machine intelligence, 21 (4): 360-370.
[20] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., & Mitsufuji, Y. (2017, March), "Improving music source separation based on deep neural networks through data augmentation and network blending," Acoustics, Speech and Signal Processing (ICASSP): 261-265), IEEE