JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A survey paper based on "MUSIC & SPEECH SEPARATION"

Roshan Borse Dipali Partale Prajwal Bhosale Swapanaja Kulkarni
SINHGAD ACADEMY OF ENGINEERING

Abstract:

One practical requirement of the music copyright management is the estimation of music relative loudness, which is mostly ignored in existing music detection works. To solve this problem, the paper studies the joint task of music detection and music relative loudness estimation. To be specific, it is observed that the joint task has two characteristics, i.e., temporally and hierarchy, which could facilitate to obtain the solution. For example, a tiny fragment of audio is temporally related to its neighbour fragments because they may all belong to the same event, and the event classes of the fragment in the two tasks have a hierarchical relationship. Based on the above observation, we reformulate the joint task as hierarchical event detection and localization problem. To solve this problem, we further propose Hierarchical Regulated Iterative Networks (HRIN), which includes two variants, termed as HRIN-r (recurrent) and HRIN-cr, (convolutional recurrent) which are based on

recurrent and convolutional recurrent modules. To enjoy the joint task's characteristics, our models employ an iterative framework to achieve encouraging capability in temporal modelling while designing three hierarchical violation penalties to regulate hierarchy. Extensive experiments on the currently largest dataset (i.e., OpenBMAT) show that the promising performance of our HRIN in the segment-level and event-level evaluations. Index Terms—music detection, music relative loudness estimation, event detection, event localization, neural networks,
hierarchical classification.

Keywords- Music Separation, Convolutional Neural Network, Neural Network, Deep Learning, Recurrent Neural Network.

I. INTRODUCTION

Music detachment is an exceptional instance of sound source partition, which means to recuperate the performing voice and perhaps other instrumental sounds from a melodic polyphonic blend. Ongoing examinations have shown that profound brain organizations Deep Neural Networks (DNNs) can display complex capabilities and perform well on different errands. Many investigations have resolved the issue of single-channel source division with DNNs. The DNNs
commonly work on size or log-extent spectra in

the Mel space or the brief time frame Fourier change space. The assessed source signal is then acquired as the result of the info

combination signal and the assessed TF veil. Just not many of the studies consider the issue of music detachment while the others center around discourse partition. MUSIC detection (MD) refers to the task of finding out whether a music event happens in an audio file and what time it starts and ends, i.e., splitting an audio recording and annotating each fragment as music or non-music. MD not only has the basic application in automatic retrieving and localizing audio data based on the type of content but also has a more practical application of monitoring music for copyright management. The practical application in the music industry is the royalty collection in broadcasting. As elaborated in the Austrian National Broadcasting Corporation (ORF) requires knowing where exactly the music appears in the soundtrack of TV production, and detecting the music is in the foreground or the background. ORF posed this requirement.

1.1 Problem Formulation

As discussed before, the MD task and MRLE task are both event detection and localization problem, and the event categories of the two tasks form a hierarchy of two-level. We propose the Hierarchical Regulated Iterative Network (HRIN), a two-output deep neural network specifically designed to solve the Hierarchical Event Detection and Localization (HEDL) problem. HRIN propagates gradients from the two network outputs–each one corresponds to each hierarchical level. A corresponding loss function to each output is used for back-propagating the gradients from the event classes in the corresponding level. We use three penalties to regulate the hierarchy. In this section, we first present two variants of HRIN: a recurrent-only (HRIN-r) architecture and a convolutional-recurrent (HRIN-cr) architecture. Then we give a detailed description of the loss function.

1.2 Project Objective:

1. We will perform an extensive review of speech and music separation methods with a special focus on those applied to music signals.
2. We will propose a new deep learning technique based on MFCC features which is simple intuitive and computationally, less expensive, making it especially interesting.
3. We will propose a set of enhancements to state of the art, speech and music separation technique in semi-supervised scenarios.

II. Literature Survey

Music Detection Many works have been proposed for the single task of music detection. Seyerlehner et al proposed the manually-designed feature called Continuous Frequency Activation (CFA) for music/non-music detection. Benito-Gorron et al. explored different neural networks and trained them to solve speech detection and music detection separately and simultaneously. For speech detection only, they classified audio fragment into two classes of no-speech and speech; for music detection only, they categorized audio fragment into non-music and music; for simultaneous speech and music detection, they classified audio fragment into four classes which are no-speech, speech, non-music, and music. Lemaire et al.

Music Relative Loudness Estimation Music relative loud-ness estimation is a sub-task derived from the traditional music detection task. This task is normally combined with the music detection task as a joint task. Melendez-Catal ´ an´ et al. proposed a CNN based method called MMG (named by the initials of authors' last name) for the joint task in the 2018 MIREX competition. Melendez-Catal ´ an´ et al. also used MMG as the benchmark model to test the dataset they proposed called Open BMAT

Multi-Task Learning The general multi-task learning problem has been studied for a long time, and many works have been done in different research areas such as music information retrieval computer vision , natural language

processing and so on. Here the most related research is music information retrieval. Bock ¨ et al. proposed to use recurrent neural network for predicting probabilities of beats/downbeats and use dynamic Bayesian network to align the predicted beat and downbeat positions to the global best solution. Vogl proposed a system to detect drum instrument onsets along with the corresponding beats and downbeats using different neural networks, taking into consideration the additional meta-information like bar boundaries, tempo, and meter. Bittner et al.

Fig1. Proposed Architecture

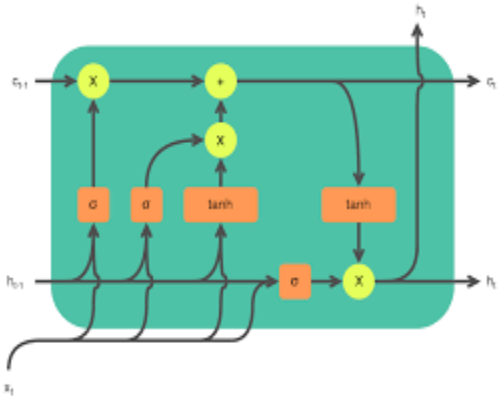


Fig2. RNN Architecture

III Proposed Method and Algorithm:

A. Proposed Methodology:

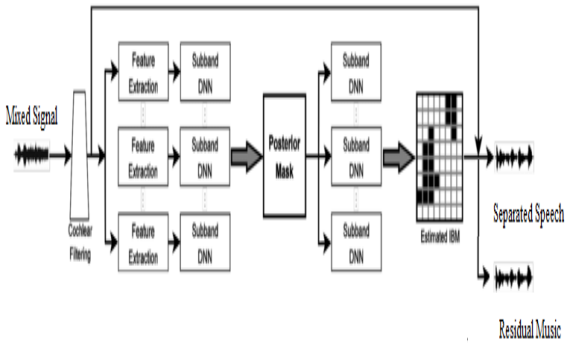
In a proposed system, we are proposing a music separation using deep learning techniques with limited set of supervised data as shown in figure1.

We are proposing a Recurrent neural network for feature extraction and classification. We are going to solve accuracy issue in diagnosis of speech with accurate stage predictions.

B. Algorithms

1. RNN

In this proposed research paper Deep Neural Network (RNN) will be used for music seperation. The exacted features from the speech data, rather than taking the features one by one. Generated weights are extracted from the different layers of NN such as hidden layers, activation layer and fully connected layers. Neural network layer is the key role of this network, which does the extraction of the features from the training data.



IV. Conclusion

In this paper, we will presented a DNN-based multichannel source separation framework where the multichannel filter is derived using the source spectra, which are estimated by DNNs, and the spatial covariance matrices, which are updated iteratively in an EM fashion. The weighted spatial parameter updates effectively handle bad estimation of spectral parameters by the DNN. The use of additional DNNs might improve the overall performance as long as overfitting can be avoided.

REFERENCES

1] S. Benito-Gorron et al. “speech detection and music detection” 2018 2nd International Conference on Micro-Electronics and Telecommunication Engineering.

2] G. R. Naik and W. Wang, Eds., Blind Source Separation: Advances in Theory, Algorithms and Applications. Berlin, Germany: Springer, 2014.

[3] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, “From blind to guided audio source separation: How models and side information can improve the separation of sound,” IEEE SPM, vol. 31, no. 3, pp. 107– 115, 2014.

[4] L. Deng and D. Yu, Deep Learning: Methods and Applications. Hanover, USA: Now Publishers Inc., 2014, vol. 7, no. 3-4.

[5] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, “Speech separation based on improved deep neural networks with dual outputs of speech features for both target and

interfering speakers," in Proc. ISCSLP, Singapore, 2014, pp. 250–254.

[6] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE/ACM Trans. ASLP, vol. 23, no. 12, pp. 2136– 2147, 2015.