MUSIC-NOISE SEGMENTATION IN SPECTROTEMPORAL DOMAIN USING CONVOLUTIONAL NEURAL NETWORKS

Taejin Park **Taejin Lee**
Electronics and Telecommunications Research Institute, Republic of Korea
{inctrl,tjlee}@etri.re.kr

ABSTRACT

We present a music-noise discrimination algorithm for the spectro-temporal domain based on the use of convolutional neural networks (CNNs). Unlike traditional handcrafted features that describe the characteristics of an audio signal, we classify an input signal using the learned features from the CNNs, which shows a high classification capability for various pattern recognition tasks. We investigated the performance of the music/noise discrimination algorithm based on CNNs trained using various noise and music signals. Finally, we segmented a music-noise mixture within the spectro-temporal domain based on the prediction results from the CNNs.

1. INTRODUCTION

Before we apply a music information retrieval technique to a variety of applications, such as music genre classification, audio fingerprinting, or music transcription, the use of processing and pre-classification algorithms is essential for obtaining a clear signal source before analyzing the music signal, particularly for real-life situations where noise can dominate an input signal. In this respect, preprocessing systems such as a speech/music discrimination algorithm [1] have been presented to pre-process a raw input signal. Such systems are based on the descriptive features such as the zero-crossing rate, linear prediction coefficient, and root mean square value. However, the recent developments of convolutional neural networks (CNNs) have shown remarkable performance improvements for both image- [2] and speech-classification [3] tasks without the use of any handcrafted feature sets. Therefore, we employed CNNs to classify a music/noise signal and analyzed the classification capability. In addition, unlike previous works that only discriminate the temporal frame, we segmented the time-frequency area within a spectrogram image containing music signals.

2. CLASSIFICATION SYSTEM

2.1 Classification task

The classification task defined in this paper is classifying an arbitrary area of a spectrogram to determine whether it contains a music or noise signal. To achieve this, we used a 2-D observation window, which has a bandwidth of 5.5 kHz and a temporal length of 0.74 s. The sample length of the window is 512 and the hop-length is 128. As a result, the pixel dimensions of the observed image are 64 by 256, as shown in figure 1. This 2-D observation window is shifted in both the temporal and frequency axes by 16 pixels for every classification event.

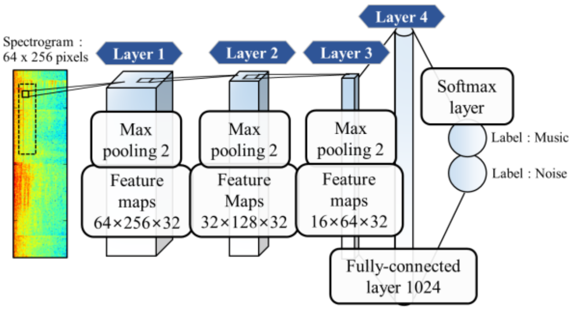

Figure 1. Structure of the neural network for classification.

2.2 Structure of deep neural network

The overall structure of the CNNs employed in our proposed classification system is shown in figure 1. We used 32 feature maps for every layer with a kernel size of 3 x 3, and max-pooled each convolutional layer. We also used a dropout technique to regularize the network and employed a rectified linear unit for activation. The entire network was optimized using a stochastic gradient descent algorithm.

3. SPECTROGRAM SEGMENTATION

3.1 Performance using an observation window

A music-noise classification task cannot be strictly defined as a binary classification problem because noise can be mixed at various signal-to-noise ratios (SNRs) under real-life conditions. Therefore, we tested the performance of our system at various SNRs and introduced a probability concept for the signal classification. To train the CNNs using training data, we used spectrogram images from both music and noise signals. For the music signals, we conducted the training using 450 popular music tracks from a US billboard chart. For the noise signals, we used environmental sounds (originally designed to test an audio scene classification algorithm) such as vehicle sounds and babble noises [4], which were recorded in public places including on the street, in a supermarket, and in a restaurant. We trained five noise-recording samples for each ten locations, for a total of 50 different noise samples. For the evaluation, another 50 songs and 50 noise sets were employed for the test. We tested the classifier at five different SNR

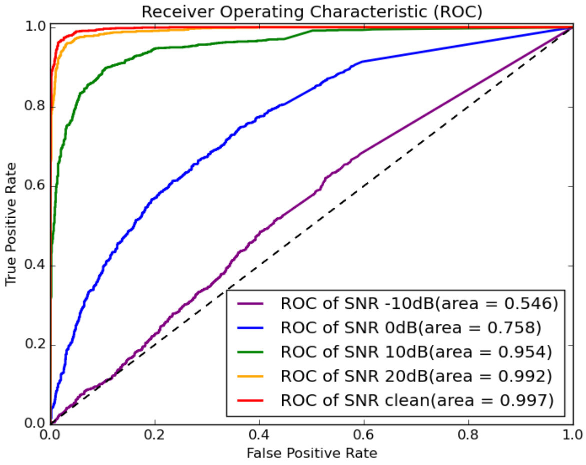

Figure 2. Receiver operating characteristics of proposed spectro-temporal music/noise classification system using a single observation window.

levels, as illustrated in figure 2. We evaluated the performance of the classifier using the area under curve (AUC) of the receiver-operating characteristic (ROC), which is based on the probability of a successfully classified window being a music signal. In an ideal case, which means the test dataset contains no music-noise mixtures and contains only clean signals of either music or pure noise, the proposed classifier showed an AUC of 0.997, which is a nearly perfect classification performance. However, as the SNR level decreases, the AUC of the classifier also decreases to 0.546. The results shown in figure 2 were derived from a node labeled as “music” in the CNNs shown in figure 1. In addition, the results from a softmax node, which was labelled with noise, showed a nearly identical performance.

3.2 Segmentation of music area from a spectrogram

We averaged the probability of all overlapping observation windows based on a segmentation of the spectrogram, as shown in figure 3. Thus, the unit area of the probability map for music existence is 16 x 16 pixels. The probability of a unit spectrogram segment can be computed as

$$p_{pm} = \frac{1}{M \cdot N} \prod_{f}^{M} \prod_{t}^{N} p(t, f) , \qquad (1)$$

where M is the number of overlapping observation windows along the frequency axis, and N is the number of overlapping observation windows along the temporal axis. As shown in figure 3, these parameters were set to $M = 9$ and $N = 16$. Here, $p(t,f)$ is the output of the softmax layer of the CNNs, which is in the form of the probability.

Figure 3 shows the segmentation results of a music and noise signal (bus sound) mixture with an SNR level of -10 dB. Whereas the noise signal dominates the spectrogram image, the proposed system segmented the music signal, which was not totally masked by the noise signal. As we can see in the top- and middle-right images in figure 3, the probability map shows a distinct difference in probability for each unit segmentation area. By comparing the magnitude of the two values from the probability maps, we segmented the spectrogram, as shown in the bottom-right of figure 3.

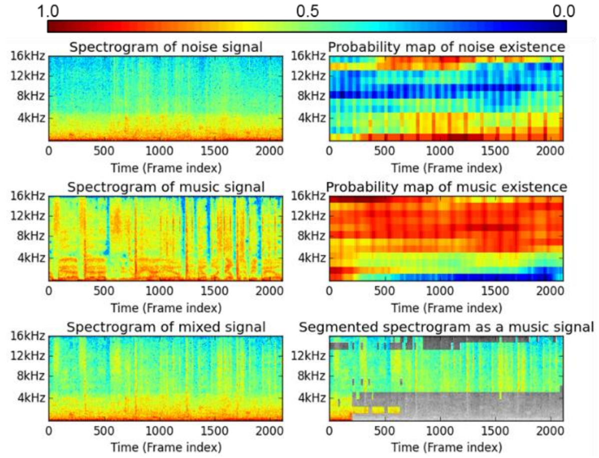

Figure 3. Segmented spectrogram of music and noise mixture with an SNR level of -10dB.

4. CONCLUSIONS

For the given test dataset, the rate of classification for music and noise was nearly perfect under clean conditions. This means that, within the given test set, according to the AUC value, the proposed classifier can discriminate with a nearly perfect success rate whether an input audio signal is music or noise using only a temporal length of 0.8 s. Finally, we segmented a spectrogram by combining the output probability of multiple observation windows. Therefore, the segmentation of a music signal in the spectro-temporal domain shows a reliable performance. Further investigation will include pixel-wise segmentation by training the CNNs using a more detailed ground-truth label and a deconvolution technique.

5. REFERENCES

[1] M. F. McKinney and J. Breebaart, “Features for audio and music classification,” In *ISMIR*, vol. 3, pp. 151-158. 2003.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NIPS)* 25, 2012.

[3] H. Lee, P. Pham, Y. Largman, and A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in Neural Information Processing Systems (NIPS)* 22. 2009.

[4] D. Giannoulis, E. Benetos, D. Stowell, and M. D. Plumbley, “IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events - Training Dataset for Event Detection Task, subtasks 1 - OL and 2 – OS,” Queen Mary University of London, 2012.