An Improved Approach to Audio Segmentation and Classification in Broadcasting Industries

Jingzhou Sun, School of Computer Science and Cybersecurity, Communication of China, Beijing, China

Yongbin Wang, School of Computer Science and Cybersecurity, Communication of China, Beijing, China

ABSTRACT

Audio segmentation and classification are the basis of audio processing in broadcasting industries. A Dual-CNN (Dual-Convolutional Neural Network) method is proposed in this article in which it is possible to pre-train a CNN with unlabeled audio data so as to deal with the scarcity of labeled data. Auto-encoders (including an encoder and a decoder) are utilized, thus the name "Dual." In the first place, audio sampling points and the derived STFT (Short-Time Fourier Transform) spectrograms pass through their own CNNs. Fusion of the extracted features is then performed. Finally, the merged features are sent to a fully connected network and the classification results are produced via Softmax. Being one of the segmentation-by-classification approaches, our solution also presents a novel smoothing method (SEG-smoothing) in order to deliver the best result of segmentation. A series of experiments have been conducted and their result verifies that the proposed approach for segmentation and classification outperforms alternative solutions.

KEYWORDS

Classification-Oriented Audio Processing, Dual-CNN, Smoothing, Sparse Autoencoder

INTRODUCTION

Over the years, broadcasting stations have accumulated a large amount of unlabeled audio content for programs. These valuable resources can be saved, indexed, and retrieved for later use by means of information technology. Efficient information retrieval calls for the help of labels that attach meaning to the data. According to an insider from China Radio International (CRI), as of July 2018, the total amount of all the audio content from CRI has exceeded 55 Terabytes, which corresponds to a 530,000 hours of audio playback. It is often impossible for human to accomplish such a tedious, time-consuming annotation task without the assistance from automatic or semi-automatic labeling techniques. Therefore, the potential of the audio content cannot be fully utilized.

Audio segmentation and classification are key techniques to successful completion of audio (data) labeling, or audio annotation, in that the classification results provide a starting point for efficient audio annotation. This topic has attracted researchers mainly from the AI (artificial intelligence) and signal processing communities (Castán et al., 2015).

The primary goal of automatic audio segmentation is to provide boundaries to delimit portions of audio with homogeneous acoustic content (Shin, Chang, & Kim, 2010). In the meantime, audio classification aims to help identify the semantic meaning of each portion derived from audio segmentation, whereas, it is, by no means, an easy task for audio streams such as broadcast news,

DOI: 10.4018/JDM.2019040103

containing single type classes and mixed type of classes (e.g., speech with music and speech with noise) (Xie, Fu, Feng, & Luo, 2011; Cheong, Oh, & Lee, 2004).

This article presents an audio classification solution to news broadcasting, which features coupling of audio segmentation and classification, namely segmentation-by-classification, see Background. A Dual-CNN (Dual-Convolutional Neural Network) was introduced to perform classification on clips with a fixed length. Unlike others, it can make use of both (a small amount of) labeled and (a large number of) unlabeled data for the training of CNNs. A novel smoothing method, SEG-smoothing, was then applied to the classification result, thus yielding portions of audio with homogeneous acoustic content. For performance evaluation of our proposed approach, a series of experiments involving Dual-CNN and other alternatives using datasets from Beijing People's Broadcasting Station and GTZAN, have been conducted. The results verify that, in terms of classification accuracy and segmentation error rate, Dual-CNN outperforms alternative solutions.

The remainder of the article is organized as follows. We presented related work in Section Background. This is followed by a detailed introduction to the Dual-CNN in Section the Dual-CNN Approach. The following section describes a smoothing method for audio segmentation. An array of experiments and related analysis for performance evaluation of the Dual-CNN is given in Section Evaluation. We conclude our work and identify future research in Section Conclusions and Future Work.

Background

We present, in this section, related work by which our proposed research was most inspired. In the beginning, we present two predominant categories for segmentation, i.e. segmentation-and-classification and segmentation-by-classification, and then review the state-of-the-art in related fields. This is followed by a discussion on the application of deep learning techniques, CNN and autoencoders, to deal with audio classification. We further this research by facilitating audio segmentation and classification in the broadcasting domain by a combination of both techniques. For a comprehensive and fair comparison, we investigate five approaches that are either classical in the field or share most features in common with our proposed work.

SEGMENTATION AND CLASSIFICATION

Audio segmentation/classification systems can be divided into two classes depending on how segmentation is performed (Castán, Ortega, Miguel, & Lleida, 2014).

Segmentation-and-Classification Approaches

This class of approaches detect the boundaries in the first place and then perform classification on each delimited segment (Huang & Hansen, 2006; Gallardo-Antolin, & Montero, 2010). For example, an approach using a temporally weighted fuzzy C-means algorithm was proposed in (Nguyen, Haque, Kim, & Kim, 2011). The Bayesian information criterion (BIC) is widely employed in many studies. Chen and Gopalakrishnan (1998) generated a breakpoint for every speaker change and environment/channel condition change in the broadcasting news domain. Moreover, Wu, Chiu, Shia, & Lin (2006) and Kotti, Benetos, & Kotropoulos (2008) utilize BIC to identify mixed-language speech and speaker change, respectively. Wu and Hsieh (2006) proposed a minimum description length (MDL) approach that allows multiple breakpoints for any generic data. Ali and Talha (2018) suggested that VAD (voice activity detection) be used to distinguish speech-absence and speech-presence segments in an audio file using the fractal dimension estimation algorithm. For each audio, a threshold for identifying speech-presence and speech-absence is computed automatically, with which segments are then categorized in response to their fractal dimensions.