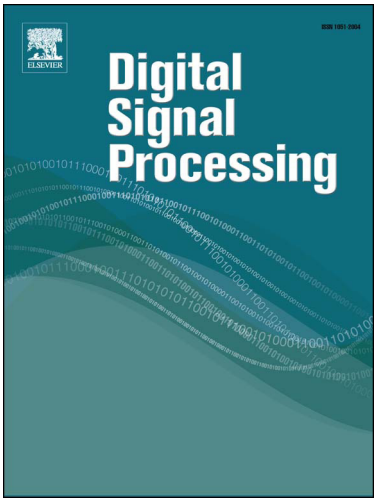
Accepted Manuscript



An adapted data selection for deep learning-based audio segmentation in multi-genre broadcast channel

Xu-Kui Yang, Dan Qu, Wen-Lin Zhang, Wei-Qiang Zhang

PII: S1051-2004(18)30074-5
DOI: https://doi.org/10.1016/j.dsp.2018.03.004
Reference: YDSPR 2299

To appear in: *Digital Signal Processing*

Please cite this article in press as: X.-K. Yang et al., An adapted data selection for deep learning-based audio segmentation in multi-genre broadcast channel, *Digit. Signal Process.* (2018), https://doi.org/10.1016/j.dsp.2018.03.004

An Adapted Data Selection for Deep Learning-Based Audio Segmentation in Multi-Genre Broadcast Channel

Xu-Kui Yang, Dan Qu, Wen-Lin Zhang, Wei-Qiang Zhang

***Abstract*—Broadcast audio transcription is still a challenging problem because of the complexity of diverse speech and audio signals. Audio segmentation, which is an essential module in a broadcast audio transcription system, has benefited greatly from the development of deep learning theory. However, the need of large amounts of labeled training data becomes a bottleneck of deep learning-based audio segmentation methods. To tackle this problem, an adapted segmentation method is proposed to select speech/non-speech segments with high confidence from unlabeled training data as complements to the labeled training data. The new method relies on GMM-based speech/non-speech models trained on an utterance-by-utterance basis. The long-term information is used to choose reliable training data for speech/non-speech models from the utterances at hand. Experimental results show that this data selection method is a powerful audio segmentation algorithm of its own. We also observed that the deep neural networks trained using data selected by this method are superior to those trained with data chosen by two comparing methods. Moreover, better performance could be obtained by combining the deep learning-based audio segmentation method with the adapted data selection method.**

***Index Terms*—deep learning, audio segmentation, voice activity detection, long-term information, multi-genre broadcast**

I. Introduction

Automatic transcription and retrieval for broadcast channel [1][2] has become one of the most attractive applications in the fields of audio signal processing and recognition. However, processing general broadcast audios is still a challenging task because of the varieties in terms of the data content, channel, and environment. Currently, many evaluations, such as multi-genre broadcast (MGB) challenge [3] and Albayzin evaluation [4], focus on audio data processing or speech recognition under broadcast channels and have attracted wide attentions. The content of broadcast audio is quite rich, including speech, music, and different types of noise or sound effects. Moreover, the speech data are very complex because of various speaking styles, different accents, mixed dialect, or with different types of background music or noise. Hence, automatic audio segmentation is a necessary front-end procedure for broadcast audio processing.

The purpose of audio segmentation is to split an audio record into segments of homogeneous content. Depending on the application, the term 'homogeneous' can be defined in terms of speaker, channel, or audio type. Generally, the first stage of audio segmentation is speech/non-speech detection to locate regions containing speech signals, which is also referred to as voice activity detection (VAD). There may be a further step of speaker segmentation/clustering to partition the speech regions into speaker-homogeneous segments. In this paper, we focus on voice activity detection.

Voice activity detection is an indispensable module for most speech signal related applications, and has a great influence on system performances. With the development of the deep learning theory, lots of deep neural network (DNN)-based VAD methods [5][6][7][8] have been proposed. Due to the success of modeling long-term dependences of input signals, recurrent neural networks (RNN) [9] and long short-term memory (LSTM) [10] recurrent neural networks have also been adopted. Convolutional neural network (CNN), known as time-delay neural network (TDNN) [11] in speech research, is also a widely used model *for* its advantages of learning spatial-temporal connectivity and reducing the number of free parameters.

Comparing with traditional VAD algorithms, deep learning-based VAD obtains much higher classification accuracies which benefits from not only the non-linear discriminative characteristics of algorithm, but also the enormous amounts of precisely-labeled training data (at least hundred hours of audio data). However, it is still a difficult task to collect such a large amount of audio data, not to mention labeling them exactly. And this problem partly restricts the applicability of deep learning- based VAD.

In the 2015 MGB challenge task [12], some data selection methods had been proposed, for example data selection based on light supervised alignments [13] and on phone-level force alignments[14][15]. These methods need a pre-trained automatic speech recognizer (ASR) which is difficult to be

This work was supported in part by the National Natural Science Foundation of China under Grants No. 61673395, No. 61403415, and Henan Provincial Natural Science Foundation under Grants No. 162300410331. The associate editor coordinating the review of this manuscript and approving it for publication was xxxx.

X.-K. Yang, D. Qu, and W.-L Zhang are with the National Digital Switching System Engineering and Technological R&D Center, Zhengzhou, 45001, China (e-mail: gzyangxk@gmail.com; qudanqudan@sina.com; zwlin_2004@163.com). The corresponding author is D. Qu.

W.-Q Zhang is with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wqzhang@mail.tsinghua.edu.cn).