Digital Signal Processing 81 (2018) 8–15

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

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

Xu-Kui Yang [a], Dan Qu [a,*], Wen-Lin Zhang [a], Wei-Qiang Zhang [b]

[a] *National Digital Switching System Engineering and Technological R&D Center, Zhengzhou, 45001, China*
[b] *Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

Article history:
Available online 9 March 2018

Keywords:
Deep learning
Audio segmentation
Voice activity detection
Long-term information
Multi-genre broadcast

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.

1. 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 [5–8] 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.

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

[*] Corresponding author.
E-mail addresses: gzyangxk@gmail.com (X.-K. Yang), qudanqudan@sina.com (D. Qu), zwlin_2004@163.com (W.-L. Zhang), wqzhang@mail.tsinghua.edu.cn (W.-Q. Zhang).

https://doi.org/10.1016/j.dsp.2018.03.004
1051-2004/© 2018 Published by Elsevier Inc.

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 trained in many situations. Also it is difficult to ensure the reliability of selected data since the accuracy of the label largely depends on the accuracy of ASR outputs.

In this paper, we propose an adapted training data selection method for multi-genre broadcast channel. Without requiring any alignments, each audio file in the unlabeled dataset is labeled using an audio segmentation method. The main steps are as follows: firstly, the long-term Mel spectral divergence (LTMD) [16] of each frame is used to classify frames into speech or non-speech class, and frames with highest confidences are selected; then speech/non-speech models are trained using features extracted from the selected frames; finally, all frames in the same audio are classified by the speech/non-speech models, and the speech segments are fine-tuned by a threshold detection based on long-term pitch divergence (LTPD) [17]. After the above processing, the reliable segments are chosen for DNN training according to some selection strategies. Experimental results show that this data selection method itself is a powerful audio segmentation algorithm. And the DNN models trained on the data chosen by it are more discriminative than those trained from two comparing methods using light supervised alignments or phone-level force alignments. Moreover, the performance can be improved further by fusing the outputs of adapted data selection method with those of DNN models.

The outline of this paper is as follows. Section 2 describes the DNN-based VAD procedure. A detailed description of the data selection methods for DNN training is given in Section 3. Section 4 presents the experimental data, setup, and results. And conclusions are drawn in Section 5.

2. VAD based on deep leaning

With higher accuracies in classification tasks, DNN models have been widely used in pattern recognition fields, besides VAD. The DNN model can be seen as a type of non-linear classifier which is able to learn complex pattern with its deep structure. As shown in Fig. 1, the procedure of DNN-based voice activity detector is as follows. Firstly, a set of feature vectors are extracted from audio data. And then, these feature vectors are fed into a pre-trained DNN models and transformed into speech or non-speech posterior probabilities. Some post-processing methods can be adopted to smooth the posterior probabilities before obtaining the VAD labels.

Besides standard full-connected forward DNN structure, RNN and TDNN have also been applied in voice activity detections for their advantages obtained from their special structures. Compared with standard DNNs, RNNs have a powerful advantage for their long contextual information representations because there are cyclic connections in the hidden layers of RNNs to model temporal correlations. However, the traditional RNNs suffer the gradient exploding or vanishing problems when being trained by the stochastic gradient descent (SGD) algorithm [18]. Thus, the LSTM RNN is proposed to alleviate this problem. The primary difference between LSTM RNN and the traditional RNN is that linear recurrent connections are used in LSTM RNN instead of non-linear ones in the conventional RNN, leading to more smooth back propagation of gradients.

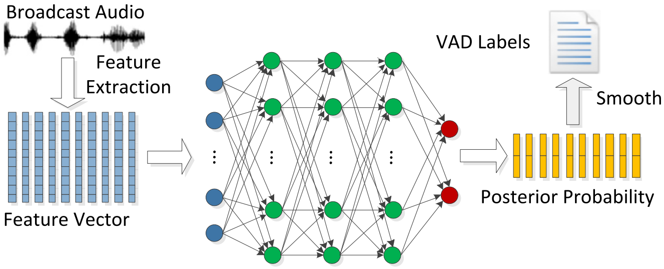

Fig. 1. DNN-based VAD procedure.

TDNN, seen as a precursor to convolutional neural networks (CNNs) [19,20], still has the ability to model long-term temporal dependencies from short-term input speech features without the affine transform in the initial layer as a standard DNN, since the temporal resolution that TDNN operates at increases from layer to layer. Moreover, under the assumption that neighboring activations are correlated, the sub-sampling processing is done to speed up the TDNN training and reduce the model size [21].

3. Data selection for DNN training

One of essential factors that make DNN models achieve much better performance is the large amount of labeled training data. As known to all, manual labeling is very costly, making labeled training data a bottleneck of DNN-based methods. To tackle this problem, some researches focus on how to enable DNN models to automatically learn from unlabeled data, for example transfer learning [22,23], or more recently dual-learning mechanism [24] in machine translation task. However, for VAD related applications, finding a similar or dual task is not easy. Here, we proposed an adapted data selection method, which automatically chooses data with high confidence from unlabeled data. In this section, we first introduce two data selection methods based on light supervised alignments or phone-level force alignments, and then present our proposed method.

3.1. Data selection based on light supervised alignments

MGB challenge mainly concentrates on an evaluation task of speech-to-text transcription of broadcast televisions. In this challenge task, the manual transcription of training data has no time information. To make the training dataset suitable for building speech recognition system, an ASR system is used to recognize the audio signal, and then the recognized transcriptions were aligned with the subtitles of TV programs to generate small speech segments with time information. However, these light supervised alignments may include some errors since both the automatic transcriptions from ASR system and subtitles were not precise. And phone matched error rate (PMER), word matched error rate (WMER), and average word duration (AWD) in seconds were computed to reject unreliable segments. Moreover, the evaluation organizers made rules that only audio data supplied by them can be used, but the light supervised alignments has no labels for non-speech segments. Thus, data selection can be done based on the assumptions [13] that segments without labels in light supervised alignments were likely to be non-speech segments. In other words, segments with labels were treated as speech segments, and non-speech data was extracted from the gaps between speech segments.

3.2. Data selection based on phone-level force alignments

More complex selection strategies based on phone-level force alignments was designed in [25,26]. In these methods, a monophone acoustic model was trained using data selected from speech segments according to WMER and AWD. Then, the monophone state sequence was obtained for each speech segment by force alignment from a pre-trained monophone acoustic model. After that, all phoneme frames were considered as speech data, while intra-segment non-speech portions were used as non-speech data. The speech data selected in this way are more precise than those chosen by previous method, because the silence or short pause embedded in speech segments were eliminated. However, there are still inaccuracies in non-speech data. If non-speech data contained the gaps between transcribed segments, there may be some speech data mixed into non-speech data because of the imprecision of light supervised alignment. If only intra-segment non-speech portions (silence and short pause) were used as non-speech data, it would lack of other types of non-speech segments such as music.

3.3. Adapted data selection method

The common shortcoming of the two data selection methods described above is the requirement of a pre-trained ASR system. This condition cannot be satisfied in most applications. VQVAD [27] is an unsupervised and self-adaptive VAD algorithm. It doesn't rely on any pre-trained models but uses a GMM-based classifier trained only from the under-detection recording, in which reliable training data is selected by short-term energy-based VAD algorithm. Similar to VQVAD, we propose an adapted VAD algorithm to select data for DNN training. In our method, VAD relies on speech/non-speech models trained on an utterance-by-utterance basis using short-term feature vectors. The training data of the speech/non-speech models consist of speech/non-speech frames of the utterance being processed with high confidence according to their long-term information instead of short-term energy in VQ-VAD. The proposed method is introduced as follows.

3.3.1. Long-term information

Because of its simplicity, robustness, and superior performance, long-term information [16,17] of speech signals has been studied deeply and widely in voice activity detection. Among these methods, long-term information based on auditory filter banks is more discriminative and robust. Auditory filter banks, such as Mel filter banks and Pitch filter banks are a set of filters designed according to the function of the human cochlea. The output of auditory filter banks can be regarded as a type of spectral decomposition in logarithmic forms. And the non-linear decomposition can represent some important acoustic cues like formants more explicitly.

The long-term spectrum divergence based on auditory filter banks between speech and noise can be defined as the deviation of long-term spectral envelops with respect to the average noise spectrums, which is given in Eq. (1).

$$D_*(l) = 10\log_{10}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{E_*^2(k,l)}{\bar{N}_*^2(k,l)}\right) \tag{1}$$

where K is the number of filter banks and R_d-order long-term spectral envelop is defined as follow:

$$E_*(k,l) = \max\left\{X_*(k,l-R_d+j)\,|\,j=0,1,\ldots,2R_d\right\} \tag{2}$$

where, $\boldsymbol{X}_*$ represents pitch features or Mel spectrum, and $X_*(k,l)$ is the kth band amplitude of $\boldsymbol{X}_*$ at frame l.

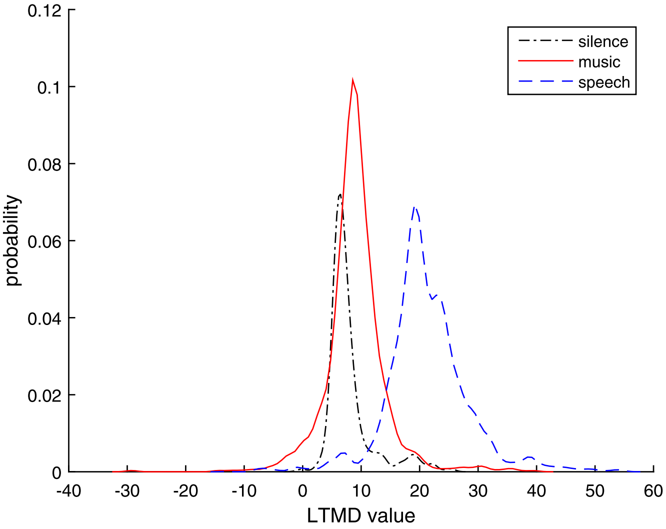

Fig. 2. The LTMD distribution of different audio types.

The noise spectrum N_* is estimated from X_* by the MMSE-based estimator [28]. And the average noise spectrum $\bar{N}_*(k,l)$ for the kth band at frame l is defined as:

$$\bar{N}_*(k,l) = \frac{1}{l}\big((1-\alpha)(l-1)\bar{N}_*(k,l-1) + \alpha N_*(k,l)\big) \tag{3}$$

where, $N_*(k,l)$ is the noise feature value of the kth band at frame l and $\bar{N}_*(k,1) = N_*(k,1)$. $0 \le \alpha < 1$ is a weight coefficient.

Two kinds of long-term information proposed by us [16,17] are used in adapted data selection method. The long-term information based on Mel filters, named long-term Mel spectral divergence (LTMD), is represented by $D_M(l)$. And the other one based on pitch filters, called long-term pitch divergence, is represented by $D_P(l)$. Fig. 2 shows the distributions of LTMD for different types of audios (speech, music, and silence) on MGB 2016 development dataset. The frame length and shift are 20 ms and 10 ms respectively, and $R_d = 20$. It can be seen that the distributions for music and silence almost overlap, which does not affect audio segmentation because we treat both silence and music as non-speech. Though there are also overlap intervals between speech and non-speech, some assumptions can be made with a quite high probability that frames with lowest LTMD values are non-speech frames while frames with highest values are speech frames. Thus, the speech and non-speech frames can be obtained with fixed percentages of the corresponding highest and lowest LTMD frames, respectively.

The distributions of LTPD for silence and speech data are shown in Fig. 3, where the frame length and shift are the same as Fig. 2, but the order of long-term spectral envelop is set to be 10. As shown in Fig. 3, the distribution of speech data has a long left tail and there is a small peak when the LTPD value is about −30. This is due to that there are some silence segments with short duration embedded in speech segments which aren't marked by the manual transcripts. Nevertheless, the discrimination ability of LTPD is excellent. We use LTPD to remove silence frames embedded in speech segments.

3.3.2. Algorithm framework

VAD can be done simply by setting up a threshold or more complexly using a discriminative model. The threshold detection is hard to adjust along with the changes of signal conditions, while the discriminative model suffers the problem of performance degradation in the mismatching environments between training and testing data. To overcome these problems, an adapted VAD method is proposed as follow. Firstly, clearly labeled frames are selected according to an assumption that frames with lowest LTMD

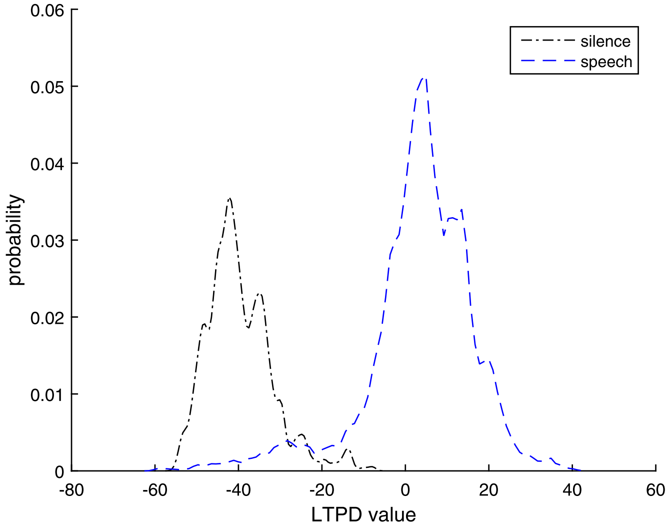

Fig. 3. The LTPD distribution of silence and speech data.

Table 1
The procedure of adapted data selection method.

```
Algorithm: Adapted data selection method.
Input: audio signal $s(n)$, set the parameters of frame length and shift;
1. Split $s(n)$ into frames, and extract short-term features
2. Calculate the LTMD $D_M(l)$ and LTPD $D_P(l)$ for each frame;
3. Select frames with clear labels based on LTMD;
4. Train the speech/non-speech model using the selected frames;
5. Classify all of the frames into speech and non-speech classes;
6. remove silence frames embedded in speech segments based on LTPD;
7. Smooth the results;
Output: the VAD labels for each frame.
```

values are non-speech frames while frames with highest values are speech frames. Then, a speech/non-speech model is trained based on these selected frames. Finally, all frames are classified by the model. The detailed procedure is summarized in Table 1.

As shown in Table 1, the short-term features and long-term information (LTMD $D_M(l)$ and LTPD $D_P(l)$) are extracted from the l-th frame. Then, the frames are sorted by the LTMD value, and the top partial frames with largest LTMD values are chosen as speech frames while the bottom partial frames with lowest LTMD values are chosen as non-speech frames. It should make sure that there are at least 1000 frames for each Gaussian component [29]. The corresponding short-term features of selected frames are used to train the speech/non-speech models.

The Gaussian mixture model (GMM) $\lambda^S : (\omega_m^S, \mu_m^S, \Sigma_m^S)$ and $\lambda^N : (\omega_m^N, \mu_m^N, \Sigma_m^N)$ are used as speech/non-speech models, where λ^S is the speech GMM and λ^N is the non-speech model, ω_m^*, μ_m^*, and Σ_m^* are the weight, mean, and variance of the m-th Gaussian respectively. The number of Gaussian components for λ^S and λ^N are the same, and denoted as M. The likelihoods of the l-th frame on λ^S and λ^N are shown as Eq. (4) and Eq. (5), respectively.

$$p\big(\mathcal{F}(l)|\lambda^S\big) = \sum_{m=1}^{M} \omega_m^S \mathcal{N}\big(\mathcal{F}(l)\big|\mu_m^S, \Sigma_m^S\big) \tag{4}$$

$$p\big(\mathcal{F}(l)|\lambda^N\big) = \sum_{m=1}^{M} \omega_m^N \mathcal{N}\big(\mathcal{F}(l)\big|\mu_m^N, \Sigma_m^N\big) \tag{5}$$

where, $\mathcal{F}(l)$ is the short-term feature vector of the l-th frame.

Thus, the speech/non-speech detection is done by comparing the corresponding log-likelihoods: the l-th frame is determined as speech frame if $\log(p(\mathcal{F}(l)|\lambda^S)) \geq \log(p(\mathcal{F}(l)|\lambda^N))$, otherwise determined as non-speech frame.

To simplify, the K-means algorithm is used to train the speech/non-speech models instead of expectation maximization (EM) to reduce the algorithm computation complexity. As a result, the maximum log-likelihood estimation can be simplified using nearest neighbor principle:

$$\theta_l = \min_m \big\| \mathcal{F}(l) - \mu_m^S \big\| - \min_n \big\| \mathcal{F}(l) - \mu_m^N \big\| \tag{6}$$

If $\theta_l \leq 0$, the frame is determined as speech frame, otherwise non-speech frame.

After the initial segmentation, more precise speech/silence detection will be performed on speech segments to remove the embedded silence frames, which is done by setting a threshold of −20 on LTPD values. In the end, the hang-over scheme [30] is applied to the output of step 6 to obtain final VAD decisions.

3.3.3. *Selection strategies*

Though post-processing techniques can help reducing errors, there are still some inevitable misclassifications in the obtained segments. To avoid the negative impact of misclassifications on DNN model training, some heuristic strategies for data selection are designed as follows:

1. Since the longer segments are more reliable, the speech and non-speech data used for DNN training are from the speech and non-speech segments with enough long duration, respectively. The minimum duration of both non-speech and speech segments is 3 seconds.
2. Since the misclassification often occurs in the margins of segments, a fixed number of frames at the beginning and end of each segment will be abandoned. In our experiments, the number is set to be 25.

4. Experiments

4.1. *Experimental data*

The 2016 MGB challenge is a task for state-of-the-art transcription systems of Arabic TV programs. Three datasets, termed as training, development, and evaluation, are released for model training, parameter setting, and performance evaluation, respectively. Since the labels of evaluation dataset are not available, we only use training and development datasets in our experiments.

The dataset used for DNN training was selected from the training set of the 2016 MGB challenge, which contains audios from more than 3000 episodes spanning over 19 programs with a total duration of 1200 hours. As described above, the audios were recognized using the QCRI Arabic LVCSR system, and then aligned with the human transcription to generate light supervised alignments with WMER, PMER, and AWD computed.

The VAD experiments were done on development dataset of the 2016 MGB challenge. To investigate the performance of DNN-based VAD models in the environment different from training set, we also tested them on the LDC Arabic broadcast news speech dataset (LDC2006S46).[1]

The development dataset of the 2016 MGB challenge include about 13 hours of audios from 17 different program episodes with manual segmentations and transcriptions. The show titles and genre labels were supplied. And all titles would have appeared in the training data.

Arabic broadcast news speech dataset consists of eight audio files recorded by the LDC from Voice of America (VOA) satellite radio news broadcasts in Arabic. The recordings were made at time

[1] https://catalog.ldc.upenn.edu.

Table 2
Information of two testing datasets.

Dataset	Speech duration	Non-speech duration
Development	9.94 hours	0.26 hours
LDC2006S46	8.82 hours	1.08 hours

of transmission between June 2000 and January 2001. The duration of each recording is either 60 minutes or 120 minutes, depending on the VOA broadcast schedule. Transcripts for these recordings are available as a separate corpus from the LDC: Arabic Broadcast News Transcripts (LDC2006T20).

The statistics of speech duration and non-speech duration for the two testing datasets are shown in Table 2. In this table, 'Development' means the development dataset of the 2016 MGB challenge.

4.2. Segmentation scoring

As in the Albayzin evaluation [32], the segmentation error rate (SER) would be used to measure the performance of DNN-based VAD models. This score was defined as the ratio of the overall segmentation error time to the sum of the durations of the segments that are assigned to each class in the file.

Given a test dataset Ω, each document was divided into contiguous segments at all class change points[2] and the segmentation error time for each segment n was defined as

$$\Xi(n) = T^n_{\text{duration}}\left[\max\left(N^n_{\text{ref}}, N^n_{\text{sys}}\right) - N^n_{\text{correct}}\right] \tag{7}$$

and segmentation error rate was defined as

$$SER = \frac{\sum_{n\in\Omega} \Xi(n)}{\sum_{n\in\Omega} T^n_{\text{duration}} N^n_{\text{ref}}} \tag{8}$$

where, T^n_{duration} is the duration of segment n, N^n_{ref} is the number of reference classes that are present in segment n, N^n_{sys} is the number of system classes that are present in segment n and N^n_{correct} is the number of reference classes in segment n correctly assigned by the segmentation system.

A forgiveness collar of one second, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a class begins or ends.

The tool used for evaluating the segmentation system is the one developed for the RT Diarization evaluations by NIST "md-eval-v21.pl", available in the web site of the NIST RT evaluations.[3]

4.3. Experimental setup

The frame length and shift used in our experiments are 20 ms and 10 ms, respectively. 20-dimensional Mel frequency cepstrum coefficients (MFCCs) with corresponding first and second-order delta coefficients and 12-dimensional Chroma vectors, resulting 72-dimensional short-term feature vectors, are used in adapted data selection method ('Adapt' for short). Without specification, this type of short-term feature is used for DNN-based VAD system.

The number of Gaussian components M is set to be 24. This value is automatically set by silhouette coefficient [31]. We randomly choose frames from the development dataset, cluster them using K-means algorithm with the number of clusters (K) ranging from 2 to 30, and compute the silhouette coefficient for each clustering. Fig. 4 shows the change of silhouette coefficients. It can be seen that the maximum is obtained when $K = 24$. Thus, the value of M is 24.

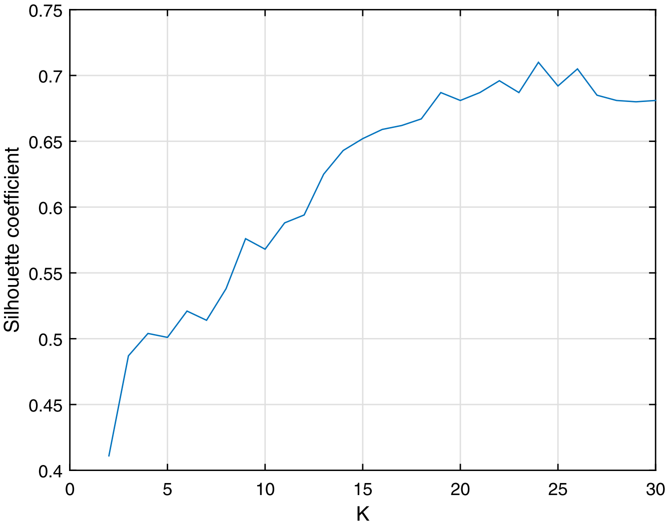

Fig. 4. The plot of silhouette coefficients.

Table 3
The amount of training data for DNN-based VAD selected from MGB training dataset by different data selection methods.

Methods	Speech duration	Non-speech duration
Light	77.15 hours	66.38 hours
Force	75.55 hours	74.21 hours
Adapt	74.47 hours	62.97 hours

For data selection based on light supervised alignments ('Light' for short), speech segments were randomly chosen from those with AWD ranging from 0.3 s to 0.7 s and WMER of equal to 0, and non-speech segments were selected from the gaps between speech segments with AWD ranging from 0.3 s to 0.7 s and WMER of less than 70.

For data selection based on phone-level force alignments ('Force' for short), an HMM–GMM acoustic model, which has 160 monophone states, was trained with segments chosen from those AWD ranging from 0.3 s to 0.7 s and WMER of equal to 0. For each segment in this training set, the monophone state sequence was obtained using force alignment. Frames corresponding to speech phones were used as speech data to train DNN models, otherwise were used as non-speech. Some non-speech segments randomly selected by data selection based on light supervised alignments were appended as non-speech data.

The amount of training data selected by different methods from MGB training dataset can be seen in Table 3. The three selected datasets will be used for DNN-based VAD systems training, and the performances of these systems are lists in Table 5.

4.4. Experimental results

4.4.1. Performance of adapted data selection method

The performance of the adapted data selection method as an audio segmentation method was evaluated on both testing datasets. The results were shown in Table 4. It can be seen that for the audio segmentation task the adapted method (see the column named 'Adapt') achieved excellent performance of its own. However, one shortcoming for this method that the audios under segmenting should contain both speech and non-speech data, and both types of data should not be less than 10% of the whole audio

[2] A "class change point" occurs each time any reference class or system class starts or ends. Thus, the set of active reference classes and/or system classes does not change during any segment.

[3] http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl.

Table 4
Performance comparison among adapted data selection method, energy-based VAD, and ASR-based VAD.

Dataset	Eval.	Adapt	Energy	ASR
Development	SER	2.40	4.42	0.73
	Time (s)	101	94	30549
LDC2006S46	SER	2.80	6.49	1.12
	Time (s)	45	42	14225

Table 5
Evaluation of DNN-based VAD systems with training data selected by different data selection methods.

Method	Development	LDC2006S46
Light	3.08	4.86
Force	2.87	4.22
Adapt	1.72	2.61

data. This condition sometimes cannot be met in real applications. Thus, we use this method to choose data for DNN-based VAD training.

In Table 4, we also compared the proposed adapted method with energy-based VAD [33] (‘Energy’ for short in the table) and decoder-based VAD [34,35] algorithms (ASR for short in the table). In the energy-based VAD algorithm, thresholds were carefully chosen for each testing dataset to obtain best performance. The ASR system used in the decoder-based VAD is the TDNN system, whose configuration and training stages are detailed in [16]. It can be seen that the energy-based achieves the worst performance since it fails to detect non-speech types like music, songs, and noises with strong energies. The decoder-based VAD is the best one, but there are still some errors, for example speech under low signal-to-noise ratio will be recognized as non-speech. Comparing with decode-based VAD algorithms, our method is still inferior to ASR. But the major advantages lie in two aspects.

Firstly, our “Adapt” method is simpler and much faster than “ASR” method as shown in Table 4. It is about more than 300 times faster.

Secondly, the decode-based VAD algorithms require various resources, multiple training stages, and significant expertise [36].

1) With respect to resources, it requires not only speech data with detail content labels, but also other resources such as dictionaries and phonetic questions.
2) Moreover, in the DNN–HMM hybrid approach, training of DNNs still relies on GMM models to obtain (initial) frame-level labels. Building GMM models normally goes through multiple stages (e.g., CI phone, CD states, etc.), and every stage involves different feature processing techniques (e.g., LDA, fMLLR, etc.).
3) In addition, the development of ASR systems highly relies on ASR experts to determine the optimal configurations of a multitude of hyper-parameters, for instance, the number of senones and Gaussians in the GMM models.

4.4.2. Comparison of different data selection methods for DNN-based VAD

The standard DNN-based VAD systems trained with different data selection methods were built. The configuration of these standard DNNs are: none splice context of input feature, 6 hidden layers with 100 units in each layer, and a final output layer of two units to represent speech and non-speech. Their performances on the two testing datasets were listed in Table 5. In this table, ‘Light’, ‘Force’, and ‘Adapt’ represent the data selection methods based on light supervised alignments and phone-level force alignments and the proposed adapted data selection method, respectively.

Table 6
Comparison of different types of input features.

Feature	Architecture	Development	LDC2006S46
FBK	DNN 1	2.51	4.33
	DNN 2	1.94	3.51
	DNN 3	1.72	2.84
MFCC+Chroma	DNN 1	1.72	2.61
	DNN 2	1.69	2.55
	DNN 3	1.70	2.60

It can be seen that the DNN model trained with ‘Force’ method performs better than that of ‘Light’ method. It is due to the fact that the training data selected by ‘Force’ method is more precise than that selected by ‘Light’ method. And the DNN model trained from data selected by ‘Adapt’ method achieves the best performance on both testing datasets, which means the ‘Adapt’ method is superior to the other two comparing methods.

4.4.3. Comparison of different types of short-term feature vectors for DNN-based VAD

In order to compare the performance of different input features, the 40-dimensional filterbank (FBK) features were also used. Three architectures of standard DNNs were built with different sizes of input context windows (0, 11, and 21) and different numbers of hidden units (100, 1000, and 2000).

Table 6 shows the performance comparison of different input features. In this table, the column named ‘Architecture’ means different architectures of standard DNN models. “DNN1” has a same configuration as those in previous section. Similarly, the architectures of “DNN2” has 11 frames of input feature context and 1000 units in each of 6 hidden layers, while that of “DNN3” has 21 frames of input feature context and 2000 hidden units.

From Table 6, it can be seen that DNN models trained with MFCCs + Chroma features are superior to those trained with FBK features. This is because the Chroma feature is applicable to music-related applications while the MFCC feature is suitable for speech-related applications, and a combined usage of both could improve the discrimination of DNN models. For DNN models trained with FBK features, the performance always improves as the complexities of the model grow. But for DNN models trained with MFCCs +Chroma, the DNN2 model obtains the best performance.

4.4.4. Evaluation on different types of neural networks

The performances of different types of neural networks were also investigated. We also trained LSTM RNN model and TDNN model. The LSTM RNN model has 3 LSTM hidden layers, where each LSTM layer contains 1024 cells, and 3 projection layers with 256 units to reduce the number of parameters. The TDNN model is composed of 6 layers, with connections of $[\{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}; \{0\}; \{-2, 2\}; \{0\}; \{-4, 4\}; \{0\}]$.

Table 7 lists the performance of all DNN models on the two testing datasets. In this table, the term ‘Fuse’ means whether the deep learning system is combined with the proposed adapted method of Section 3. A tick for this term implies the combination is done for the corresponding experiment while a dash means the results are obtained only by the DNN model.

From Table 7, it can be seen that the TDNN model achieve the best performance while the LSTM model performs worse than the other two models. And the detection speed of LSTM is much slower than DNN and TDNN. The table also tells us that all models combining with the adapted method improves the performances.

5. Conclusions

In this paper, we proposed an adapted data selection method for deep learning-based audio segmentation system training. As an

Table 7
Evaluation on different structure of DNN.

Type	Fuse	Development	LDC2006S46
DNN2	–	1.69	2.55
	✓	1.36	1.69
LSTM	–	2.21	3.06
	✓	1.97	2.59
TDNN	–	1.66	2.52
	✓	1.33	1.67

audio segmentation algorithm, this adapted data selection method could obtain good performance, but it suffers a shortcoming that the audios segmented by this algorithm have to contain both speech and non-speech data, and both types of data should not be less than 10% of the whole audio data. Thus, DNN models were trained with data selected from large amounts of unlabeled audios using this data selection method. Experimental results showed that these DNN models were superior to those trained with data selected by other comparing methods. Moreover, the SER was further reduced by combining the adapted data selection method with the DNN model.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dsp.2018.03.004.

References

[1] P. Lopez-Otero, L. Fernández, C. García-Mateo, Ensemble audio segmentation for radio and television programmes, Multimed. Tools Appl. 76 (5) (2017) 7421–7444, https://doi.org/10.1007/s11042-016-3386-2.

[2] S. Khurana, A. Ali, QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge, in: Proceedings of the IEEE Workshop on Spoken Language Technology, SLT, San Diego, California, USA, 2016, pp. 292–298.

[3] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, Y. Zhang, The MGB-2 challenge: Arabic multi-dialect broadcast media recognition, in: Proceedings of the IEEE Workshop on Spoken Language Technology, SLT, San Diego, California, USA, 2016, pp. 279–284.

[4] A. Ortega, I. Viñals, A. Miguel, E. Lleida, The Albayzin 2016 speaker diarization evaluation, in: Proceedings of 17th Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, USA, 2016.

[5] Neville Ryant, Mark Liberman, Jiahong Yuan, Speech activity detection on YouTube using deep neural networks, in: Proceedings of 14th Annual Conference of the International Speech Communication Association, INTERSPEECH, Lyon, France, 2013, pp. 728–731.

[6] Xiao-Lei Zhang, Ji Wu, Deep belief networks based voice activity detection, IEEE Trans. Audio Speech Lang. Process. 21 (4) (2013) 697–710.

[7] Xiao-Lei Zhang, De-Liang Wang, Boosting contextual information for deep neural network based voice activity detection, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (2) (2016) 252–264.

[8] Z. Koldovský, J. Malek, M. Boháč, J. Janský, CHiME4: multichannel enhancement using beamforming driven by DNN-based voice activity detection, in: Proceedings of the 4th International Workshop on Speech Processing in Everyday Environments, CHiME 2016, San Francisco, USA, Sept. 2016.

[9] Thad Hughes, Keir Mierle, Recurrent neural networks for voice activity detection, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2013.

[10] Florian Eyben, Felix Weninger, Stefano Squartini, Björn Schuller, Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2013.

[11] Samuel Thomas, Sriram Ganapathy, George Saon, Hagen Soltau, Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2014.

[12] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, P.C. Woodland, The MGB challenge: evaluating multi-genre broadcast media recognition, in: Proceedings of Automatic Speech Recognition and Understanding, ASRU, 2015, pp. 687–693.

[13] Quoc Do Truong, Michael Heck, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura, The NAIST ASR system for the 2015 multi-genre broadcast challenge: on combination of deep learning systems using a rank-score function, in: Proceedings of Automatic Speech Recognition and Understanding, ASRU, 2015, pp. 654–659.

[14] O. Saz, M. Doulaty, S. Deena, R. Milner, R. Ng, M. Hasan, Y. Liu, T. Hain, The 2015 Sheffield system for transcription of multi-genre broadcast media, in: Proceedings of Automatic Speech Recognition and Understanding, ASRU, 2015.

[15] P.C. Woodland, X. Liu, Y. Qian, C. Zhang, M.J.F. Gales, P. Karanasou, P. Lanchantin, L. Wang, Cambridge university transcription systems for the multi-genre broadcast challenge, in: Proceedings of Automatic Speech Recognition and Understanding, ASRU, 2015.

[16] X. Yang, D. Qu, W. Zhang, W. Zhang, The NDSC transcription system for the 2016 multi-genre broadcast challenge, in: Proceedings of the IEEE Workshop on Spoken Language Technology, SLT, San Diego, California, USA, 2016, pp. 273–278.

[17] X. Yang, L. He, D. Qu, W.Q. Zhang, Voice activity detection algorithm based on long-term pitch information, EURASIP J. Audio Speech Music Process. 2016 (2016) 14.

[18] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-forward neural networks, J. Mach. Learn. Res. 9 (2010) 249–256.

[19] O. Abdel-Hamid, A. Mohamed, H. Jiang, G. Penn, Applying convolutional neural networks concepts to hybrid NN–HMM model for speech recognition, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, Kyoto, Japan, 2012, pp. 4277–4280.

[20] T.N. Sainath, B. Kingsbury, A. Mohamed, B. Ramabhadran, Learning filter banks within a deep neural network framework, in: Proceedings of Automatic Speech Recognition and Understanding, ASRU, Olomouc, Czech Republic, 2013, pp. 297–302.

[21] V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, in: Proceedings of 16th Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 2015.

[22] Amit Das, Mark Hasegawa-Johnson, Cross-lingual transfer learning during supervised training in low resource scenarios, in: Proceedings of 14th Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 2015.

[23] Ritwik Giri, Michael Seltzer, Jasha Droppo, Dong Yu, Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2015.

[24] Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma, Dual learning for machine translation, in: Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems, NIPS 2016, Barcelona, Spain, 2016.

[25] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, The MGB challenge: evaluating multi-genre broadcast media recognition, in: Automatic Speech Recognition and Understanding, ASRU, 2015, pp. 687–693.

[26] L. Wang, C. Zhang, P. Woodland, M. Gales, P. Karanasou, P. Lanchantin, X. Liu, Y. Qian, Improved DNN-based segmentation for multi-genre broadcast audio, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2016, pp. 5700–5704.

[27] Tomi Kinnunen, Padmanabhan Rajan, A practical, self-adaptive voice activity detector for speaker verification with noise telephone and microphone data, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2013.

[28] Timo Gerkmann, Richard C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay, IEEE Trans. Audio Speech Lang. Process. 20 (4) (2012) 1383–1393.

[29] Guo Wu, The Session Variability Speaker Recognition, University of Science and Technology of China, China, 2007.

[30] Recommendation ITU-T G.720.1, Generic sound activity detector, 2010.

[31] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons Ltd., New York, 1990, pp. 72–91.

[32] D. Castán, D. Tavarez, P. Lopez-Otero, J. Franco-Pedroso, H. Delgado, E. Navas, L. Docio-Fernández, D. Ramos, J. Serrano, A. Ortega, E. Lleida, Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains, EURASIP J. Audio Speech Music Process. 2015 (2015) 33.

[33] H. Sun, B. Ma, H. Li, Frame selection of interview channel for NIST speaker recognition evaluation, in: Proceedings of 7th Int. Symposium on Chinese Spoken Language Processing, ISCSLP 2010, Nantou, Taiwan, December 2010, pp. 305–308.

[34] H. Sakai, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano, A. Lee, Voice activity detection applied to hands-free spoken dialogue robot based on decoding using acoustic and language model, in: Proceedings of ROBOCOMM, ICST/ACM, 2007, pp. 180–187.

[35] A. Lee, T. Kawahara, Recent development of open-source speech recognition engine Julius, in: Proceedings of Asia–Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, 2009.

[36] Y. Miao, M. Gowayyed, F. Metze, EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding, The Computing Research Repository (CoRR), arXiv:1507.08240v3, 2015 [2018-02-28], http://arxiv.org/abs/1507.08240.

Further reading

[37] L. Wang, C. Zhang, P.C. Woodland, et al., Improved DNN-based segmentation for multi-genre broadcast audio, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2016.

Xu-Kui Yang was born in Fujian, China, in 1988. He received the B.S. and M.S. degrees in information and communication from the Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2011 and 2014, respectively. He is currently working towards the Ph.D. degree on speech recognition at the National Digital Switching System Engineering and Technological R&D Center.

His research interests are in speech signal processing, continuous speech recognition, and machine learning.

Dan Qu was born in Jilin, China, in 1974. She received the B.S., M.S. and Ph.D. degrees in information and communication engineering from the Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2004, 2007 and 2013, respectively. From 2016 to 2017, she was a visiting scholar in Computer Science Institute of Carnegie Mellon University.

She is an Associate Professor in the National Digital Switching System Engineering and Technological R&D Center. Her research interests are in speech signal processing and pattern recognition & machine learning, and natural language processing.

Wen-Lin Zhang was born in Hubei, China, in 1982. He received the B.S., M.S. and Ph.D. degrees in information and communication engineering from the Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2004, 2007 and 2013, respectively.

He is an Assistant Professor in the National Digital Switching System Engineering and Technological R&D Center. His research interests are in speech signal processing, speech recognition, and machine learning.

Wei-Qiang Zhang was born in Hebei, China, in 1979. He received the B.S. degree in applied physics from University of Petroleum, Shandong, in 2002, the M.S. degree in communication and information systems from Beijing Institute of Technology, Beijing, in 2005, and the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, in 2009. From 2016 to 2017, he was a visiting scholar at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University.

He is an Associate Professor at the Department of Electronic Engineering, Tsinghua University, Beijing. His research interests are in the area of radar signal processing, acoustic signal processing, speech signal processing, machine learning and statistical pattern recognition.