INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Enhanced Feature Extraction for Speech Detection in Media Audio

Inseon Jang[1], *ChungHyun Ahn*[1], *Jeongil Seo*[1], *Younseon Jang*[2]

[1] Media Research Division, Electronics and Telecommunications Research Institute, Korea
[2] Department of Electronics Engineering, Chungnam National University, Korea
{jinsn, hyun, seoji}@etri.re.kr, jangys@cnu.ac.kr

Abstract

Speech detection is an important first step for audio analysis on media contents, whose goal is to discriminate the presence of speech from non-speech. It remains a challenge owing to various sound sources included in media audio. In this work, we present a novel audio feature extraction method to reflect the acoustic characteristic of the media audio in the time-frequency domain. Since the degree of combination of harmonic and percussive components varies depending on the type of sound source, the audio features which further distinguish between speech and non-speech can be obtained by decomposing the signal into both components. For the evaluation, we use over 20 hours of drama which manually annotated for speech detection as well as 4 full-length movies with annotations released for a research community, whose total length is over 8 hours. Experimental results with deep neural network show superior performance of the proposed in media audio condition.

Index Terms: Speech Detection, Voice Activity Detection

1. Introduction

Speech detection has traditionally been developed as a pre-processing of speech recognition and speech transcription. Due to diversification and development of media service technology, its application range is expanding and the speech detection has become an indispensable technology.

This paper addresses the speech detection aimed at discriminating the speech from non-speech in media contents. Media audio includes not only speech but also various sound sources, where they may be overlapped with each other. In addition, it involves speech that plays diverse situations such as emotional, whispering and singing voice. Hence, detecting the speech from media contents remains a challenge.

Among the media audio, the research on speech detection in broadcast news has been started early because the broadcast news domain has rich audio types and several application areas such as news transcription [1]. Both the hidden Markov model (HMM) trained on Mel-frequency cepstral coefficients (MFCC) and the support vector machine (SVM) trained on a combination of various audio features such as MFCC and zero crossing rate have been well-studied [2], [3]. There was also a study on acoustic features based on spectral properties and harmonic enhancement [4]. In this study, broadcast news audio was classified into speech, commercials, environmental sound, physical violence and silence using multi-model HMM.

Recent advances in machine learning have resulted in significant improvements in both speech and audio processing technology and a lot of researches on deep learning based speech detection have actively being reported. It has been presented the recurrent neural network (RNN) based voice activity detection (VAD) system trained only perceptual linear prediction (PLP) features and performance improvements by incorporating it into the automatic speech recognition (ASR) system [5]. There have been several studies to improve the VAD performance using convolutional neural network (CNN) alone and in combination with deep neural network (DNN) for noisy environment such as DARPA RATS program [6], [7], [8], [9], [10]. Furthermore, a comparison of the robustness of DNN, RNN and CNN for VAD under various noise condition was studied in [11]. For speech detection on web video such as Youtube, it has been investigated a combination of different audio features and discriminative classifiers [12]. Also, a DNN-based speech detection system trained on only MFCC has been proposed and it showed that the speech detection performance is further improved by using the extended context window as the input of the network [13].

Previous studies have shown that DNN-based speech detection has superior performance with a simple structure. However, there is a limitation to apply the existing research results to speech detection in media audio because audio features are extracted without deeply considering the acoustic properties of media audio.

For speech detection in movies, the use of long short-term memory (LSTM) RNN has been studied, which is able to model long range of temporal contexts [14]. Relative spectral analysis (RASTA)-PLP and their first order derivatives are used as audio features in this study. Lehner et al. investigated the audio feature set to represent the temporal characteristics and harmonicity of the signal and proposed a VAD system based on SVM [15]. In their work, four hours radio broadcasting was gathered and manually annotated for experiments, as well as 4 full-length Hollywood films.

In this paper, we present a novel audio feature extraction method to reflect the acoustic characteristics of the media audio in the time-frequency (TF) domain. Decomposing the spectrogram of media audio, it can be better represented the characteristics of the sound source contained in the media audio. In this work, we improve speech detection performance by using audio features extracted from the harmonic and percussive components of the media audio. We evaluate the suggested approach with DNN using two kinds of media audio datasets. One is over 20 hours of the drama dataset for speech detection and the other is 4 full-length movie audio with annotations released by Lehner et al., whose total length is over 8 hours [15]. Through the cross validation based on DNN, we confirm that the proposed method shows superior performance in media audio environment.

This paper is organized as follows. Section 2 introduces the background of decomposing media audio and describes the proposed audio feature extraction method. The description of datasets and experimental results are shown in Section 3 and 4, respectively. Finally, Section 5 concludes the works.

http://dx.doi.org/10.21437/Interspeech.2017-792

2. Enhanced Feature Extraction

2.1. Why to Decompose the Media Audio for Speech Detection

Media contents include a variety of sound sources such as music, sound effects, noise and speech, whose acoustic characteristics on the time-frequency domain depend on the type of sound source. For music signals, the frequency components of the sounds from the harmonic instruments such as violin are very smooth along the temporal direction in the spectrogram, while those from the percussive instruments such as drum are smooth along the spectral direction in the spectrogram [16], [17]. In the case of sound effects, the acoustic characteristics in the spectrogram are different for each individual sound effect. For example, fireworks, explosions, door-closing and horse-running sounds contain much of percussive components similar to percussion sounds. Meanwhile, wind sounds carry tonal information and are perceived as a harmonic part of a sound. Noise signals have an isotropic structure that is not biased to horizontal or vertical components in the spectrogram.

Speech can be divided into voiced and unvoiced. In contrast to the voiced, unvoiced sounds have no harmonic structure and are often acoustically similar to white noise. Singing voice constantly varies between voiced and unvoiced depending on the singing words, duration of the individual voiced and unvoiced parts of the words and the characteristics of the vocalist [18]. On the other hand, acoustic characteristics of laughing, breathing and crying sound from human are more similar to those of sound effects, than those of the speech.

As described above, the degree of combination of harmonic and percussive components differs depending on the type of sound source. Therefore it is necessary to decompose the spectrogram in order to better represent the acoustics characteristics, and the audio features obtained from the decomposed spectrogram are capable of improving the speech/non-speech classification performance. In this work, we adopt harmonic-percussive source separation (HPSS) to decompose the media audio. The aim of the HPSS is to separate audio mixture into harmonic and percussive components and it has been originally developed as an effective preprocessing for rhythm/harmonic instrument transcription and code detection as well as remixing purposes in the music processing area. Recently it has been considered as an elemental technology to improve the performance of singing voice separation for automatic lyrics recognition, automatic singer identification and automatic subtitle alignment [17], [18].

An example of HPSS for the media audio is shown in Figure 1. It was carried out using median filtering-based HPSS algorithm presented in [16] with same parameter setting as described in next section. As shown in the figure, string music and voiced speech are mainly composed of harmonic components while table knocking sound, actor's laughter sound and unvoiced speech are mainly composed of the percussive components. Also, as aforementioned, it can be seen that the tapping sound has similar spectrogram characteristics with that of the laughing sound.

In order to verify that the audio features extracted from the decomposed signals can more clearly distinguish between the speech and non-speech, we compared the 13-MFCCs extracted before and after applying HPSS in the drama dataset whose total length is over 20 hours. Figure 2 illustrates examples of the probability density function comparison of MFCCs for speech and non-speech. After applying the HPSS, it can be seen that the distinction between the probability density functions of each class becomes clearer, especially for percussion components.

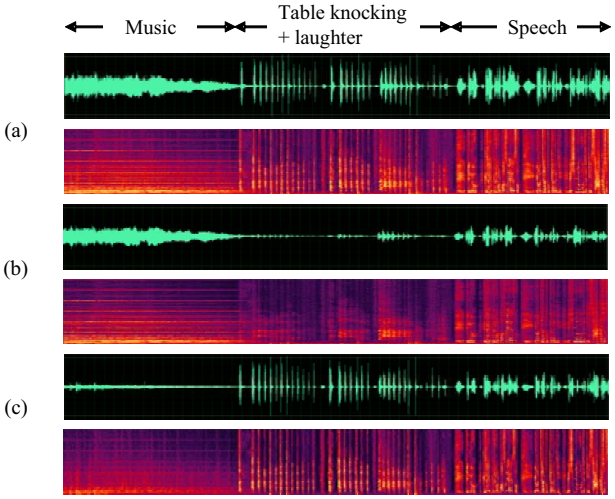

Figure 1: *A example of HPSS in drama audio: (a) original media audio; (b) harmonic and (c) percussive components after applying HPSS to media audio shown in (a).*

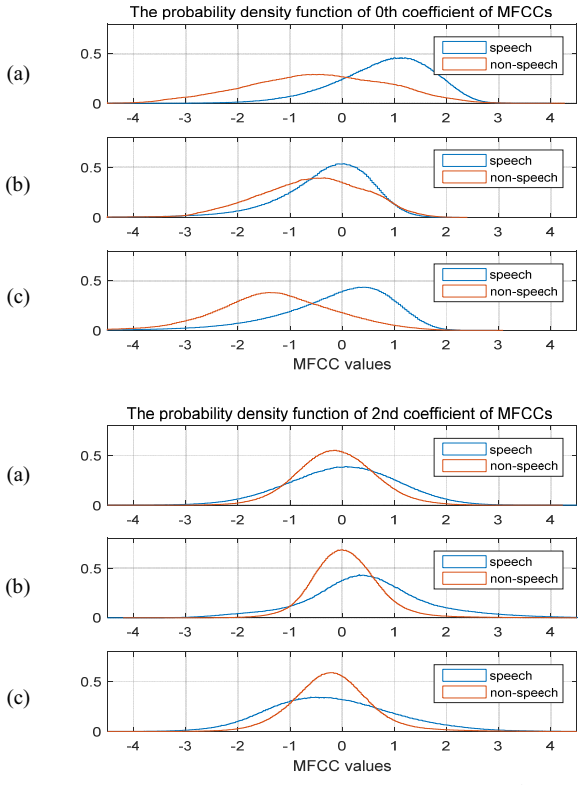

Figure 2: *Examples of the PDF comparison of 0th and 2nd coefficient of 13-MFCCs for speech and non-speech: extracted from (a) the original media audio; (b) harmonic and (c) percussive components after applying HPSS.*

2.2. Proposed Audio Feature Extraction

In order to decompose the media audio, we apply the HPSS using median filtering presented by FitzGerald et al [16]. It is based on the assumption that harmonic and percussive components exhibit horizontal and vertical lines on the spectrogram respectively, so that enhanced spectrograms for each component were calculated by using median filtering in perpendicular directions.

For HPSS of media audio down-sampled to 16 kHz, a spectrogram is carried out using an FFT size of 1024, Hann analysis window whose length is the same as the FFT size, and a hop size of 256. The length of median filter is 31 and soft masking is applied. Then, 13-MFCCs are extracted from the harmonic and percussive components, respectively. They are normalized on file-by-file basis to have zero-mean and unit variance. Finally, the feature vector is concatenated with that of the each 5 immediately preceding and following frames in order to include the contextual information [13].

2.3. DNN based Classifier

We trained a DNN with the following architecture: a 286-unit input layer, 3 hidden layers and 2-unit output layers. Each hidden layer contains the units whose number is the same as that of the input layer. The nonlinear activation functions of the hidden and output layers are a *sigmoid* and a *softmax*, respectively. The network was trained by backpropagation algorithm for 200 epochs using mini-batch gradient descent with a mini-batch size of 100 and a learning rate of 5e-3. Training was accelerated by use of a momentum of 0.5. To avoid over-adaptation, DNN training was stopped if there was no loss improvement over 5 epochs. The loss value between the targets and the network predictions for the validation data set is used as an evaluation criterion.

3. Data Sets

3.1. Drama Dataset

There are few databases which are publicly available for speech detection on media contents, except the Lehner's movie dataset which will be described in next section. The reason it is not actively released is copyright issues and time-consuming works.

In order to construct a large-scale database for media audio having various kinds of sounds, we first obtained 30 kinds soap opera contents from 3 different terrestrial broadcasters in Korea. The dataset consists of historical drama and modern drama since the types of sound effects and background sounds are different according to the genre of drama. Then, the ground truth is manually annotated on the dataset by labelling speech and non-speech segments as precisely as possible. As mentioned in [15], the annotation result depends on what the annotator considers speech, so it should be preceded by establishing a consistent rule before annotating. We have annotated the drama audio with the ASR system in mind, like Lehner's movie dataset so that laughing, breathing and singing sounds were annotated as non-speech in the dataset. Annotating the ground truth was a very time consuming, in our case, it took over 3 hours to annotate one hour of drama contents.

The total length of the drama dataset is 20.23 hours, in which it retains about 7.20 hours (35.97%) for speech. Table 1 gives detailed statistics for the drama dataset.

Table 1: *Statistics of the Drama dataset.*

Type	Number	Length	Speech [%]
Modern	18	10:50:20	35.62
Historical	12	9:32:23	36.36
Total	30	20:22:43	35.97

Table 2: *Statistics of the Movie dataset.*

Title	Length	Speech [%]
Bourne Identity	1:58:24	26.75
I Am Legend	1:40:22	18.35
Kill Bill 1	1:46:08	19.15
Saving Private Ryan	2:42:27	32.12
Total	8:07:21	25.16

3.2. Movie Dataset

Lehner et al. have announced to contribute the ground truth annotations of four Hollywood movies for a research community in [15]. Thanks for their contribution and cooperation, their ground truth was used for experiments in this paper.

The annotation set provided by them is divided into 30-minute chunks and consists of label data as speech or non-speech with 10ms, 20ms, and 200ms units, respectively, in each chunk. Also, the time and label information of the speech and non-speech segment boundaries in each chunk are included. We compared and aligned the speech segment boundaries of the annotations to the audio after extracting the audio tracks from movies. Table 2 gives detailed statistics for the movie dataset.

4. Experiments

We present three experimental results. First, we perform 5-fold cross validation on the drama dataset for the proposed method as well as the baseline method, where the MFCCs extracted from the mono down-mixed media audio is used as an audio feature. Both methods perform the speech detection using DNN having identical parameters and settings except for the number of nodes in the input layer. Second, we train both systems with complete drama dataset, then measure the performance of each resulting classifier for the movie dataset. We additionally perform a leave-one-out cross validation on the movie dataset because the performance of DNN-based classifier depends on the characteristics of training database.

Table 3: *5-folds CV results on drama dataset.*

	PREC	REC	F1	ACC	FRP	FNR
BASE	.9298	.8991	.9142	.9393	.0381	.1009
PROP	.9418	.9286	.9351	.9537	.0323	.0714

4.1. Cross Validation Results on Drama Dataset

The results of the 5-fold cross validation on drama dataset are listed in Table 3. The PREC, REC, F1 and ACC in column indicate the precision, recall, F1-score and accuracy, respectively. Also, we added the results regarding false positive rate and false negative rate in the columns FPR and FNR, respectively. The BASE and PROP in row contains the results from the baseline method and our proposed method, respectively.

Table 4: *Results on movie dataset using classifier trained with complete drama dataset.*

Title	PREC		REC		F1		ACC		FPR		FNR	
	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP
Bourne Id.	.8271	.8548	.5974	.6454	.6937	.7355	.8589	.8758	.0456	.0400	.4026	.3546
I Am Leg.	.8181	.8424	.4444	.5373	.5759	.6561	.8799	.8966	.0222	.0226	.5556	.4627
Kill Bill 1	.6120	.6947	.6007	.6875	.6063	.6911	.8506	.8823	.0902	.0716	.3993	.3125
Saving P.	.8888	.9157	.3902	.4647	.5423	.6166	.7884	.8143	.0231	.0202	.6098	.5353
Average	.7990	.8377	.4975	.5721	.5999	.6698	.8379	.8610	.0430	.0367	.5025	.4279

Table 5: *Leave-one-out CV results on movie dataset.*

Title	PREC		REC		F1		ACC		FPR		FNR	
	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP
Bourne Id.	.7068	.7629	.7151	.7444	.7109	.7535	.8444	.8697	.1083	.0845	.2849	.2556
I Am Leg.	.7119	.7876	.6420	.6768	.6751	.7280	.8866	.9072	.0584	.0410	.3580	.3232
Kill Bill 1	.5302	.6486	.7576	.7916	.6238	.7130	.8251	.8780	.1590	.1016	.2424	.2084
Saving P.	.8264	.8745	.5527	.6028	.6624	.7137	.8190	.8446	.0549	.0410	.4473	.3972
Average	.7093	.7803	.6552	.6936	.6684	.7262	.8404	.8709	.0913	.0648	.3448	.3064

The proposed approach improves accuracy by about 1.44% (93.93% to 95.37%). Also, a lower false negative rate 2.95% (10.09% to 7.14%) is observed and it means that the proposed audio features have better capability in detecting the non-speech such as sound effect and string music in media audio.

4.2. Results on Lehner's Movie Dataset

Table 4 shows the speech detection results on the movie dataset using DNN classifiers which are trained on complete drama dataset. The experimental results of the four movies and their average result weighted by the length of the movies are listed in the rows.

The proposed approach outperforms the baseline method, increasing the accuracy by 2.31% (83.79% to 86.10%) and decreasing 7.46% false error rate (50.25% to 42.79%) on average. This also shows that the audio features based on the proposed provide better performance in discriminating between the speech and non-speech. It should be noted that the performance of both methods are significantly worse than the cross validation results for the drama dataset shown in Table 3. This seems to be due to partly out-of-domain training, as reported in previous studies [14], [15].

Despite the similar performance, it is not appropriate to directly compare those results with the speech detection performance presented in [15] because the type and size of the training dataset are different in both experimental environments. To complement this, we provide additional cross validation results for publicly available movie dataset in the next section.

4.3. Cross Validation Results on Movie Dataset

The results of the leave-one-out cross validation on movie dataset are listed in Table 5. The experimental results for each movie's validation are listed in the rows as well as an average result weighted by the length of the movies.

The proposed approach outperforms the baseline, improved the accuracy by 3.05% (84.04% to 87.09%) as well as the false positive rate and the false negative rate by 2.65% (9.13% to 6.48%) and 3.84% (34.48% to 30.64%), respectively. As can be seen from the cross validation results, the audio features based on the proposed method shows consistently improved performance for both drama dataset and movie dataset.

5. Conclusions

A novel feature extraction method for speech detection in media audio was presented. Taking into account the acoustic characteristics in the TF domain, the proposed method extracts the audio features from the harmonic and percussive components of the media audio. Hence it provides robust speech detection performance even in the media audio where various sound sources are overlapped.

We performed the cross validation using DNN on the drama dataset which consists of over 20 hours drama audio with annotations for speech/non-speech and evaluated the performance on movie dataset from Lehner et al. using the DNN classifier trained with the complete drama dataset. Both experiments showed that the proposed approach outperforms conventional audio features, particularly improving accuracy and reducing false negative rate. In order to demonstrate the performance of the proposed method using opened dataset, additional cross validation was performed on the movie dataset. As a result, accuracy improvement was 3.05%, the false positive error rate and the false negative rate were improved by 2.65% and 3.84%, respectively.

Future research will examine the performance of DNN trained with the proposed audio features for various sound sources and compare it to the CNN of the TF domain.

6. Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (2015-0-00860, Development of assistive broadcasting technology for invisible and deaf people's media accessibility)

7. References

[1] DARPA Broadcast News Transcription and Understanding, http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa98/index.htm

[2] T. Hain and P. C. Woodland, "Segmentation and classification of broadcast news audio," *Proceeding of International Conference on Spoken Language Processing (ICSLP)*, pp. 2727–2730, 1998.

[3] L. Lu, H.J. Zhang, and S.Z. Li., "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482-492, 2003.

[4] T.L. Nwe and H. Li, "Broadcast news segmentation by audio type analysis," *Proceeding of 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[5] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," *Proceeding of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7378-7382, 2013.

[6] T. Ng, et al. "Developing a speech activity detection system for the DARPA RATS Program," in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 9-13, Portland, Oregon, USA, Proceedings*, 2012, pp. 1-4.

[7] G. Saon, et al. "The IBM speech activity detection system for the DARPA RATS Program," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25-29, Lyon, France, Proceedings*, 2013, pp. 3497-3501.

[8] S. Thomas et al. "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," *Proceeding of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2519-2523, 2014.

[9] S. Thomas, et al. "Improvements to the IBM speech activity detection system for the DAPRA RATS program," *Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4500-4504, 2015.

[10] L. Ferrer, G. Martin and V. Mitra, "A phonetically aware system for speech activity detection." *Proceeding of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5710-5714, 2016.

[11] S. Tong, G. Hao, and Y. Kai "A comparative study of robustness of deep learning approaches for VAD" *Proceeding of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5695-5699, 2016.

[12] A. Misra, "Speech/nonspeech segmentation in web video," in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 9-13, Portland, Oregon, USA, Proceedings*, 2012, pp. 1977-1980.

[13] N. Ryant, M. Libeman and J. Yuan, "Speech activity detection on YouTube using deep neural network," in INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25-29, Lyon, France, Proceedings, 2013, pp. 728-731.

[14] F. Eyben, F. Weninger, S. Squartini and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 483-487, 2013.

[15] B. Lehner, G. Widmer and R. Sonnleitner, "Improving voice activity detection in movies," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 2942-2946.

[16] D. FitzGerald, "Harmonic/percussive separation using median filtering," *Proceeding of the 13th International Conference on Digital Audio Effects (DAFx-10)*, 2010.

[17] C. Hsu, D "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1482-1491, 2012.

[18] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. 4, no. 1, pp. 62-73, 2010.