CSLT TECHNICAL REPORT-20150030 [Monday 19th October, 2015]

DNN-based Voice Activity Detection for Speaker Recognition

Fanhu Bie[1,3*], Zhiyong Zhang[1], Dong Wang[1] and Thomas Fang Zheng[1]

*Correspondence:
biefh@cslt.riit.tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
[3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Voice activity detection (VAD) plays an important role in speaker recognition. This paper proposes to use a novel DNN-based VAD which harnesses the power of deep neural networks (DNN) in learning speech patterns from a large labelled database designed for speech recognition, and thus deliberately optimizes the discrimination between speech and non-speech signals. More interestingly, the output of the DNN offers a noise prior, which may lend itself to a Bayesian treatment for the uncertainty of noise in speaker recognition.

The experiments were conducted on the mismatched-microphone condition (C3) of the SRE08 core test. It was found that the DNN-based VAD offered a relative reduction of 22.0% in equal error rate (EER) when compared to a fine-tuned energy-based VAD. When the Bayesian approach was employed, additional gains were obtained, particularly in noise conditions.

Keywords: speaker recognition; GMM-UBM; deep neural network

1 Introduction

Speaker recognition, also known as voiceprint recognition, has been broadly used to verify speaker identities. An important component of a practical speaker recognition system is the voice activity detection (VAD), which plays the role to select the speech segments which are the most effective for speaker discrimination.

VAD has been extensively studied in signal and speech processing communities. Early VAD algorithms are most based on features that reflect voice properties, e.g., the linear prediction coding (LPC) parameters [1], energy and formant [2], zero crossing rate (ZCR) [3], cepstrum [4], periodicity [5], pitch [6]. Another class of VAD algorithms is based on probabilistic models. Probably the most popular approach is the one based on likelihood ratio test (LRT), first proposed in [7]. Various extensions have been presented, by using more powerful LR estimations [8], better temporal smoothing [9], more discriminatively trained model [10] and more appropriate assumptions on the speech signal distribution [11]. The third category employs simple acoustic features (e.g., MFCCs) but relies on more powerful models and classifiers. The widely used models include Gaussian mixture model (GMM) [12, 13], hidden Markov model (HMM) [14], multi-layer perceptron (MLP) [15] and support vector machine (SVM) [16]. The MLP-based and GMM-based approaches were compared in [17], and the GMM and the maximum entropy (ME) model were studied in [13]. [18] compared the three popular models (GMM, SVM and MLP) and found the MLP produced better performance.

Recently, deep neural networks (DNN) gained popularity and many DNN-based approaches were reported. For example, [19] proposed to use DNN (called DBN

in their paper) to integrate heterogeneous primary and high-level features to gain a strong VAD. The novel approach based on recurrent neural networks (RNN) was proposed in [20], which follows the early work in [21], but involves a deep structure and uses a different activation function in a quadratic form. The RNN-based approach was also proposed in [22] but in a special form, i.e., the long short term memory (LSTM) RNN.

In spite of the rich research, the VAD approaches used in speaker recognition are rather simple. Probably the most popular one is still the simple energy-based approach, possibly with an adaptive threshold [23, 24]. Some researchers used the periodicity of speech frames or the power of noise-removed speech frames as the criterion to make speech/non-speech decisions [25, 26, 27, 28]. Another popular VAD approach is based on GMMs [29, 30]. The MLP was also used in [15], and was found to be effective in low-SNR conditions. A possible reason that most speaker recognition systems use simple VAD is that the dominant GMM-based speaker recognition framework is able to classify noise frames to the appropriate Gaussian components that represent noise, which reduces the impact of noise frames even if they are not well removed. Additionally, GMM-based speaker models do not rely on the temporal structure of speech signals, and so the complex smoothing methods and temporal constrains for VAD are not required.

Nevertheless, if the SNR is high, most of the simple VAD approaches tend to cause unacceptable errors, which in turn leads to serious performance reduction for speaker recognition. This is because true speech signals, even corrupted by noise, are still valuable for identifying speakers, and true noise signals simply degrade performance if they are not removed. In this situation, it is highly important to employ a strong VAD that is insensitive to noise so that real speech can be selected. This has been demonstrated by the experiments in [15], where the powerful MLP-based VAD offers more contribution in conditions with a low SNR.

This paper proposes a DNN-based VAD for speaker recognition. The DNN model has attained remarkable success recently in multiple research fields, particularly in speech recognition [31]. A valuable property of DNN is that it can learn high-level representations from raw features, and the learning can be performed within a complex feature space, for example, complicated noise conditions. These advantages can be harnessed to construct a strong VAD can deal with complex noise conditions, by learning from just raw features (e.g., the Fbank feature in this study).

An important by-product of the DNN-based VAD is the posterior probabilities read from the DNN output. Traditionally, these posteriors are used to make the speech/non-speech decision by comparing them with a pre-defined threshold. It works fine with clean speech, however in noisy conditions, this 'hard decision' is too abrupt as the decision boundary tends to be unreliable with noisy signals, and so any decision involves a large uncertainty. A better solution is treating the posteriors as prior confidence when the frames are pooled to perform recognition. This leads to a Bayesian treatment that we will present in Section 4.

The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 presents the DNN-based VAD, which is followed by the Bayesian approach described in Section 4. Section 5 presents the experiments and Section 6 concludes the paper.

2 Related work

The VAD method based on DNNs has been proposed in several papers, e.g. [19, 20, 22], though our focus is speaker recognition instead of VAD itself. Probably the most relevant work is [15] where the power of MLP was demonstrated, especially in noise conditions. The difference of our work is that we use a DNN instead of an MLP so that can leverage the power of deep structures to learn features that are important to discriminate speech and non-speech signals. Additionally, this work focuses more on the Bayesian treatment.

Besides frame selection, VAD has also been used to enhance speech signals for speaker recognition, combined with techniques such as Wiener filtering [29] and spectral subtraction [30]. Reasonable performance improvements have been reported by the enhancement methods. This paper does not try to remove noise, but address the uncertainty it caused. Nonetheless, noise removal and uncertainty compensation are complementary and can be combined to deliver a better treatment for noisy speech. We leave this combination as future work.

3 DNN-based VAD

DNN is a general non-linear classifier and can be trained to distinguish speech and non-speech signals, simply by collecting speech and non-speech frames and setting the training objective as the binary classification task. A potential difficulty is that training the VAD-DNN model requires speech/non-speech labels for each frame, which is not a trivial task although with forced alignment employed. An alternative approach is used in this paper: instead of training the VAD-DNN from scratch, we borrow it from a full-fledge automatic speech recognition (ASR) system.

DNN models have been widely used in ASR. These models are trained on large speech corpora that are labelled by words or phones, and so are quite powerful in phone discrimination. This power can be used to construct a strong VAD. The only change from an ASR-DNN to an VAD-DNN is that the output of the units corresponding to all speech (noise) phones need to be merged to produce the posterior that the input frame is speech (noise).

Figure 1 shows the structure of a VAD-DNN that is constructed based on an ASR-DNN. The ASR-DNN part involves 4 hidden layers and a phone-posterior layer (the output layer in the original ASR-DNN). The output layer merges the posteriors of speech and non-speech phones respectively to produce speech/non-speech posteriors.

The ASR-DNN was trained with about 100 hours of speech data and 12 minutes of noise data. The input layer consists of 200 units, corresponding to the dimension of the input features Each hidden layer consists of $1,200$ units, and the output layer (the phone posterior layer in Figure 1) consists of $3,440$ units, corresponding to the $3,440$ tri-phone states of the ASR system. Among these units, 16 units correspond to silence and noise, and the rest correspond to true speech. The output of these units are merged according to their corresponding phones, which has been represented by the connections from the PP layer to the output layer in Figure 1. At present all the weights of these connections are set to 1 without re-training. Note that the idea of reusing neural networks trained for ASR to perform VAD has been proposed in [15].

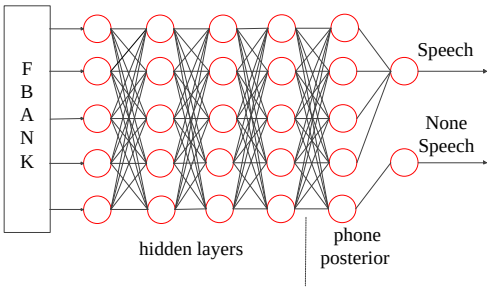

Figure 1 The DNN structure borrowed from ASR and used for VAD.

4 Bayesian treatment

This work is based on the gaussian mixture model-universal background model (GMM-UBM) framework. In this framework, the UBM is trained with speech features (MFCCs in this work) of a large number of speakers, and for each registered speaker, a speaker GMM is adapted from the UBM via the maximum a posterior (MAP) algorithm. In test, an utterance in the form of features X is tested against the model of the claimed speaker s. The confidence score that utterance X is spoken by s is computed as the log likelihood ratio (LLR), written by $log\Lambda(s;X)$. Normally, the LLR is computed as the average of the frame-based LLRs $\{\Lambda(s;x_t)\}$ where t indexes the speech frames. This is formulated as follows:

$$log\Lambda(s;X) = \sum_{t=1}^{|X|} log\Lambda(s;x_t) \quad (1)$$

$$= \sum_{t=1}^{|X|} log\frac{p(x_t|\mathcal{M}_s)}{p(x_t|\mathcal{M}_u)} \quad (2)$$

where $|X|$ denotes the number of frames in X, and $\mathcal{M}_s$ and $\mathcal{M}_u$ are the speaker GMM and the UBM respectively. $p(x_t|\mathcal{M})$ is the probability function of model $\mathcal{M}$, which is essentially a mixture of Gaussians. With the LLR score, the decision that X is spoken by s is achieved simply by comparing the LLR to a pre-defined threshold θ.

A potential problem of the LLR test in Eq. (1) is that all frames are treated equally in the average. This may lead to unreliable LLR estimation when some frames are not reliable. For example if some frames are seriously corrupted by noise, involving them in Eq. (1) just decreases performance.

This problem is generally addressed by VAD. VAD can be regarded as a frame-selection process that determines which frames are retained in the LLR test. This can be simply represented by a binary indicator variable $c(t)$ which is 1 when frame t is identified as speech by VAD, otherwise 0. This leads to the modified LLR score as follows:

$$log\Lambda(s;X) = \sum_{t=1}^{|X|} c(t) log\Lambda(s;x_t). \quad (3)$$

The VAD-based approach generally works well in conditions with a high SNR. However, for low-SNR conditions, simple VAD methods (e.g., the one based on simple energy threshold) tend to fail. A stronger VAD, for example the one proposed in this paper, can help to some extent, however the performance degradation is still significant compared to the case with an ideal VAD. This can be largely attributed to the intrinsic uncertainty in the speech/non-speech decision when the signal is noisy: it is really difficult to tell if a noisy segment is speech or not, even for people. This uncertainty means that the 'hard decision' based on VAD is over abrupt, and it is not appropriate to deal with noisy conditions.

A possible solution is a 'soft decision' that places a confidence that measures the probability that a frame is speech, rather than classifying it to speech or non-speech deterministically. This confidence can be regarded as a prior knowledge that a frame is speech, since it is derived from an exotic model. The modified LLR formula is given by:

$$log\Lambda(s;X) = \sum_{t=1}^{|X|} P(v|x_t) log\Lambda(s;x_t) \quad (4)$$

where v denotes the event that a frame is speech, and $P(v|x_t)$ is the probability that v occurs. This formulation leads to a probabilistic approach to deal with the decision uncertainty associated with VAD on noisy speech signals, and can be regarded as a Bayesian treatment in a general sense.

The confidence, or the prior probability $P(v|x_t)$ can be derived from any model that can produce posterior probabilities of v given x_t, for example, using GMMs with the Bayesian rule. This work chooses the DNN model described in the previous section, due to a multitude of advantages it possesses. First, it is a discriminative model and naturally produces posteriors. Second, the DNN is trained with a large amount of labelled data so can leverage rich information for phone discriminant. Third, the deep structure is powerful to learn multiple noise types which makes it very powerful in addressing complex noise conditions.

5 Experiment

This section presents the experiments. We start by describing the data and configurations, and then report the results on clean speech and noisy speech sequentially.

5.1 Data and configurations

The experiments are conducted on the C3 condition of the NIST SRE08 core test, which involves mismatched-microphone channels. The speech data are interviews recorded by different microphones, at an 8k Hz sampling rate with 16-bit precision. The data set involves 853 female speakers and 18,780 test utterances in total.

Each of the enrollment and test utterances lasts about 3 minutes. All the trials are divided into 2 groups, one is the development set and the other is the test set. The development set involves 200 speakers and $4,376$ trials, and the evaluation set involves 653 speakers and $14,404$ trials.

The recognition system is constructed based on the GMM-UBM architecture. The acoustic feature consists of 16-dimensional MFCCs and their first order derivatives. The UBM is trained with the Fisher English database. The training data consists of $4,000$ utterances that are randomly selected from the Fisher database. The UBM comprises $2,028$ Gaussian components, and the speaker GMMs are derived from the UBM by MAP adaptation.

5.2 Clean speech test

The first experiment evaluates the DNN-VAD and the Bayesian approach on clean speech. The baseline is based on the simple energy-based VAD that has been carefully tuned with the development set.

To have an intuitive idea, the posterior probabilities $P(v|x_t)$ produced by the DNN (prior probabilities for speaker recognition) for all the frames x_t of the development set are collected, and the distribution of the values is drawn in Figure 2.

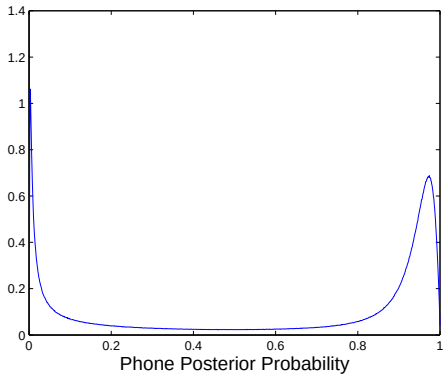

Figure 2 Distribution of the values $P(v|x_t)$ on the development set.

From the distribution, it is clear to see that $P(v|x_t)$ is rather discriminative: it assigns speech frames a value close to 1 and non-speech frames a value close to 0. This indicates that there is little uncertainty for VAD with the clean speech data.

Table 1 reports the results on the development set, in terms of equal error rate (EER). Three systems are reported: the baseline system that employs an energy-based VAD, the DNN-VAD system which employs the DNN-based VAD, and the Bayesian system that employs the Bayesian treatment given by Eq. (4). Performance with various thresholds θ are presented for the DNN-VAD system. For the baseline system, only the best result with the optimal threshold is reported since it is not the focus of the paper.

From Table 1, we see clear advantage of the DNN-based VAD when compared to the energy-based VAD employed in the baseline system. The Bayesian system also outperforms the baseline, but slightly worse than the best DNN-VAD system. This

System	Threshold(θ)	EER%
Baseline	-	32.56
DNN-VAD	0.6	24.35
DNN-VAD	0.7	23.88
DNN-VAD	0.8	**23.60**
DNN-VAD	0.9	23.67
DNN-VAD	0.95	24.56
Bayesian	-	24.13

Table 1 The EER results on the development set, tested with clean speech. The best result is in bold face.

System	Threshold(θ)	EER%
Baseline	-	31.49
DNN-VAD	0.8	**24.57**
Bayesian	-	25.00

Table 2 The EER results on the evaluation set, tested with clean speech. The best result is in bold face.

can be explained by the fact that the uncertainty in the VAD decision on clean speech is insignificant (see Figure 2), which means that the Bayesian approach is not very necessary. Anyway, the Bayesian approach enjoys a big advantage that there is not a threshold θ, which makes the method more robust against the change of acoustic conditions. This will be further verified by the noisy data test presented shortly.

The performance on the evaluation set is shown in Table 2, where the best threshold obtained from the development set is applied directly, for both the baseline and the VAD-DNN system. The observations are similar the results in Table 1: both the DNN-VAD and the Bayesian system outperform the baseline in a significant way, and the Bayesian treatment is slightly worse than the DNN-based VAD.

5.3 Noisy speech test

In order to test the DNN-based approach on noisy data, white noise is mixed to the speech signals in both the development set and the evaluation set. We test three noise levels, which are SNR=30, SNR=21 and SNR=12 respectively. Again, the distribution of the frame-level posteriors produced by the DNN on the development set is drawn in Figure 3. This figure clearly show that with a low SNR, the posteriors are much crowding in low values, indicating the significant reduction of speech/no-speech discrimination and the largely increased uncertainty with the DNN-based VAD.

The results on the development are reported in Table 5.3. It can be seen that mixing white noise leads to significant performance reduction for speaker recognition, in spite of which VAD approach is used. Nevertheless, in all the test conditions, the DNN-VAD system outperforms the baseline in a significant way. Interestingly, the Bayesian approach outperforms the DNN-VAD system in all the conditions. This indicates that with the high uncertainty associated with the VAD on noisy data, the Bayesian treatment can contribute due to its nature as a 'soft VAD decision'. Finally, we observe that with a very low SNR, all the three methods approach to a random decision, which means that the spectral structure has been seriously corrupted and so the performance is not related to VAD any more.

The results on the evaluation set is reported in Table 5.3, where the best threshold $\theta = 0.3$ has been employed. The same conclusions are drawn as from Table 5.3, that

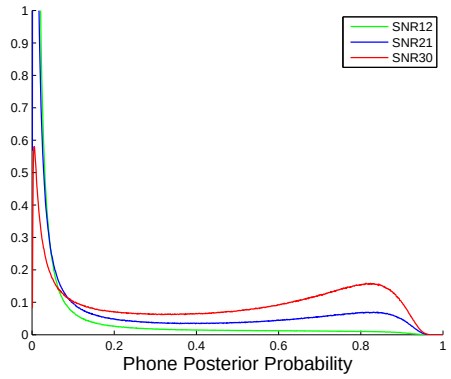

Figure 3 Distribution of the values $P(v|x_t)$ on the development set, with white noise mixed at different SNRs.

		EER%		
System	Threshold(θ)	SNR12	SNR18	SNR21
Baseline	-	48.30	44.54	39.33
DNN-VAD	0.5	47.24	35.54	27.47
DNN-VAD	0.3	47.49	35.48	27.26
DNN-VAD	0.1	49.65	40.07	29.25
Bayesian	-	**46.41**	**35.14**	**26.85**

Table 3 The EER results on the development set, test at various SNR levels. The best results are in bold face.

the DNN-VAD system outperforms the baseline consistently and significantly, and the Bayesian approach contributes in the noisy conditions. Again, we emphasize that the Bayesian approach possesses a big advantage in generalizability: it is parameter-free and can be easily migrated to different acoustic conditions. This is impossible for traditional VAD approaches, sine the threshold tuned in one condition is generally unacceptable when migrating to other conditions.

6 Conclusions

This paper presented a DNN-based VAD for speaker recognition. The DNN is borrowed from speech recognition and has been trained with a large amount of labelled data, and therefore is highly powerful for speech/non-speech discrimination. Additionally, a Bayesian treatment has been proposed to deal with the uncertainty in VAD decision, particularly with noisy speech signals. The results show that the DNN-based VAD is significantly better than the energy-based VAD, and the Bayesian treatment contributes to scenarios with noisy speech. Particularly, the Bayesian approach does not need a threshold as traditional VAD approaches do, and therefore is easy to be migrated to different acoustic conditions.

		EER%		
System	Threshold(θ)	SNR12	SNR18	SNR21
Baseline	-	48.59	42.07	37.14
DNN-VAD	0.3	45.90	36.54	27.95
Bayesian	-	**45.75**	**36.23**	**27.61**

Table 4 The EER results on the evaluation set, test at various SNR levels. The best results are in bold face.

In spite of the promising results, the present research is still in a preliminary stage. Particularly, we have not yet investigated the performance bound with a perfect VAD, due to the lack of speech/no-speech labels of the data in hand. Furthermore, our experiments show that in a low-SNR condition, VAD is not the deterministic factor for performance. DNN-based noise removal is an interesting work that we are focusing on.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61371136 and No. 61271389, it was also supported by the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302. It was also supported by Sinovoice and Huilan Ltd.

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. L. Rabiner and M. Sambur, "Voiced-unvoiced-silence detection using the itakura lpc distance measure," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'77.*, May 1977, vol. 2, pp. 323–326.
2. John D Hoyt and Harry Wechsler, "Detection of human speech in structured noise," in *Acoustics, Speech, and Signal Processing,IEEE International Conference on ICASSP'94., 1994*. IEEE, 1994, vol. 2, pp. II–237.
3. J. C. Junqua, B. Reaves, and B. Mark, "A study of endpoint detection algorithms in adverse conditions: Incidence on a dtw and hmm recognize," in *Eurospeech, 1991*, 1991, pp. 1371–1374.
4. J.A. Haigh and J.S. Mason, "Robust voice activity detection using cepstral features," in *TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on*, Oct 1993, vol. 3, pp. 321–324.
5. R Tucker, "Voice activity detection using a periodicity measure," *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, no. 4, 1992.
6. Fu-hua Liu and Michael A. Picheny, "Model-based voice activity detection system and method using a log-likelihood ratio and pitch," *US*, 2003.
7. Wonyong Sung Jongseo Sohn, Nam Soo Kim, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett*, vol. 6, no. 1, pp. 1 – 3, 1999.
8. Sang-Ick Kang, Q-Haing Jo, and Joon-Hyuk Chang, "Discriminative weight training for a statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 15, pp. 170 – 173, 2008.
9. Yu Tao and J.H.L. Hansen, "Discriminative training for multiple observation likelihood ratio based voice activity detection," *Signal Processing Letters, IEEE*, vol. 17, no. 11, pp. 897 – 900, 2010.
10. Q.-H. Jo, J.-H. Chang, J.W. Shin, and N.S. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205 – 210, 2009.
11. Jong Won Shin, Joon-Hyuk Chang, and Nam Soo Kim, "Voice activity detection based on statistical models and machine learning approaches," *COMPUTER SPEECH AND LANGUAGE*, vol. 24, no. 3, pp. 515–530, 2010.
12. D. Vlaj, Z. Kacic, and M. Kos, "Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria," *Computers and Electrical Engineering*, vol. 38, no. 6, pp. 1820–1836, 2012.
13. Ananya Misra, "Speech/nonspeech segmentation in web videos.," in *Interspeech'12*, 2012.
14. L. Hernandez O. Varela, R. San-Segundo, "Combining pulse-based features for rejecting far-field speech in a hmm-based voice activity detector," *Computers and Electrical Engineering*, vol. 37, no. 4, pp. 589–600, 2011.
15. Sriram Ganapathy, Padmanabhan Rajan, and Hynek Hermansky, "Multi-layer perceptron based speech activity detection for speaker verification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.
16. Lie Lu, Stan Z Li, and Hong-Jiang Zhang, "Content-based audio segmentation using support vector machines," in *ICME2001*, 2001.
17. Tim Ng, Bing Zhang, Long Nguyen, Spyros Matsoukas, Xinhui Zhou, Nima Mesgarani, Karel Veselỳ, and Pavel Matejka, "Developing a speech activity detection system for the darpa rats program.," in *Interspeech'12*, 2012.
18. B. Elizalde and G. Friedland, "Lost in segmentation: Three approaches for speech/non-speech detection in consumer-produced videos," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, July 2013, pp. 1–6.
19. Ji Wu Xiao-Lei Zhang, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697 – 710, 2013.
20. Thad Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection," in *ICASSP'13*, 2013.
21. Roberto Gemello, Franco Mana, and Renato De Mori, "Non-linear estimation of voice activity to improve automatic recognition of noisy speech," in *Interspeech'05*, 2005.
22. 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," in *Acoustics, Speech and Signal Processing,IEEE International Conference on ICASSP'13*, May 2013, pp. 483–487.
23. D.A. Reynolds, W. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami, "The 2004 mit lincoln laboratory speaker recognition system," in *Acoustics, Speech, and Signal Processing,IEEE International Conference on ICASSP'05*, March 2005, vol. 1, pp. 177–180.
24. T. Kinnunen, J. Saastamoinen, V. Hautamaki, M. Vinni, and P. Franti, "Comparing maximum a posteriori vector quantization and gaussian mixture models in speaker verification," in *ICASSP'09*, 2009.
25. V. Hautamaki, M. Tuononen, T. Niemi-Laitinen, and P. Franti, "Improving speaker verification by periodicity based voice activity detection," in *Proceedings of the 12th International Conference on Speech and Computer*, 2007, p. 645–650.

26. H. Sun, B. Ma, and H. Li, "An efficient feature selection method for speaker recognition," in *ISCSLP08*, 2008.
27. M.W. Mak and H.B. Yu, "Robust voice activity detection for interview speech in nist speaker recognition evaluation," in *Proceedings of the APSIPA ASC 2010*, 2010.
28. H. Yu and M. Mak, "Comparison of voice activity detectors for interview speech in nist speaker recognition evaluation," in *Interspeech'11*, 2011.
29. H. Sun, T. Nwe, B. Ma, and H. Li, "Speaker diarization for meeting room audio," in *Interspeech'09*, 2009.
30. Man-Wai Mak and Hon-Bill Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Computer Speech and Language*, vol. 28, no. 1, pp. 295–313, 2014.
31. George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.