SPEECH ACTIVITY DETECTION IN ONLINE BROADCAST TRANSCRIPTION USING DEEP NEURAL NETWORKS AND WEIGHTED FINITE STATE TRANSDUCERS

Lukas Mateju, Petr Cerva, Jindrich Zdansky and Jiri Malek

Technical University of Liberec,
Faculty of Mechatronics, Informatics, and Interdisciplinary Studies,
Studentska 2, 461 17 Liberec, Czech Republic

ABSTRACT

In this paper, a new approach to online Speech Activity Detection (SAD) is proposed. This approach is designed for the use in a system that carries out 24/7 transcription of radio/TV broadcasts containing a large amount of non-speech segments, such as advertisements or music. To improve the robustness of detection, we adopt Deep Neural Networks (DNNs) trained on artificially created mixtures of speech and non-speech signals at desired levels of signal-to-noise ratio (SNR). An integral part of our approach is an online decoder based on Weighted Finite State Transducers (WFSTs); this decoder smooths the output from DNN. The employed transduction model is context-based, i.e., both speech and non-speech events are modeled using sequences of states. The presented experimental results show that our approach yields state-of-the-art results on standardized QUT-NOISE-TIMIT data set for SAD and, at the same time, it is capable of a) operating with low latency and b) reducing the computational demands and error rate of the target transcription system.

Index Terms— deep neural networks, speech activity detection, weighted finite state transducers, speech recognition

1. INTRODUCTION

Speech activity detection is a problem of identifying both speech and non-speech segments in a sound recording. Over the years, various SAD approaches have been proposed and an SAD module has usually formed an integral component of a signal pre-processing algorithm in a wide range of tasks including, e.g., speech enhancement, speaker identification and, of course, speech transcription. Most of the existing SAD approaches are carried out in two subsequent stages: feature extraction, and speech/non-speech classification.

In the former phase, the classic approaches for feature extraction utilize energy [1], zero crossing rate [2] or auto-correlation function [3]. The family of more complex features, which have also been successfully applied, include MFCCs [4, 5], multi-resolution cochleagram features [6], multi-band long-term signal variability features [7] or channel bottleneck features [8]. Note that in [9], features based on the use of Deep Belief Networks (DBN) have also been proposed. In practice, various combinations of individual features are usually used to achieve the best possible results.

In the latter phase, various classification algorithms can be used, such as Support Vector Machines (SVM) [10] or Gaussian Mixture Models (GMMs) [11, 12]. In recent years, various DNN architectures started to be employed more and more frequently including fully connected feed-forward DNNs [5], Convolutional Neural Networks (CNNs) [13] or Recurrent Neural Networks (RNNs) [14, 15]. More complex approaches such as jointly trained DNNs [16] or boosted DNNs [6] have also been proposed. Moreover, in [17] a combination of DNN and CNN is used. The output from a given classifier can also be smoothed to further improve the accuracy of the detection. Recently, various techniques such as the Viterbi decoder [5, 18] or WFSTs [19] have been applied for this purpose.

Most of the aforementioned works aim primarily at offline applications, because applying SAD in an online environment brings further restrictions on the system, such as low computational demands and latency. The approaches developed namely for the online task include, for example, conditional random fields [18] or accurate endpointing with expected pause duration [20]. Another approach in [21] utilizes short-term features.

The goal of our efforts was to develop a SAD approach suitable for a system that is deployed for online 24/7 transcription of more than 80 TV and radio stations in several Slavic languages. From a speech recognition point of view, this type of input data is specific by containing a large amount of non-speech parts such as songs or advertisements. This applies namely to some local radio stations, whose broadcasts may contain only a few percent of speech segments. In this case, the utilization of an SAD module should reduce the computation demands on the transcription system dramatically.

An SAD module suitable for our target task should a) operate at a low level of Real-Time Factor (RTF), b) have a low latency, and c) reduce the Word Error Rate (WER) of transcription. To meet all these requirements at once, a new approach is proposed in the present paper. It adopts a DNN classifier that is trained on a data set created by mixing clean speech utterances with non-speech recordings at various desired levels of SNR. The output from DNN is then smoothed using a decoder based on WFSTs. To ensure high quality and accuracy of the detection, the employed transduction model is context-based, i.e., both speech and non-speech events are modeled as sequences of three consecutive states.

2. METRICS USED FOR EVALUATION

In this paper, three different overall accuracy metrics were used for evaluation including Frame Error Rate (FER), Miss Rate (MR) and False Alarm Rate (FAR) [5].

Moreover, the F-measure was utilized to evaluate the quality of the change-point detection between speech/non-speech events given the alignment between detected and reference boundaries [22]. Given the correctly detected boundaries (hits), it is also possible to calculate an error value for each hit (in seconds) and sort all the hits according to these values in ascending order. In this paper, the measure $\delta_{2/3}$ is utilized, which expresses (in seconds) the maximal error of the alignment for first two-thirds of the sorted (best) hits.

3. THE PROPOSED SAD APPROACH

The SAD approach presented in this paper was developed in a series of experiments described in the following subsections.

3.1. Data Used for Development

The data used for development and evaluation consisted of 6 hours of TV and radio recordings in several Slavic languages, e.g., Czech, Slovak, Polish or Russian. It contained not only clean speech segments but also segments with music, jingles and/or advertisements. Annotation of this data was created within two consecutive phases: speech/non-speech labels were produced automatically using the baseline DNN-based SAD approach (see the next section) at first, and then corrected by human annotators. In total, 70% of all frames were marked as containing speech.

3.2. Baseline DNN-Based Approach

The baseline approach employed a deep neural network with a binary output (i.e., without any smoothing) which was trained using the torch library[1]. The data for DNN training was composed of 7 hours of various non-speech events and/or noises, 30 hours of music recordings and 30 hours of clean speech utterances belonging to several Slavic languages and English.

The DNN had 5 hidden layers, each consisting of 128 neurons. The ReLU activation function and mini-batches of size 1024 were used within 10 epochs of training. The learning rate was set to 0.08. 39-dimensional log filter banks were used as features. The input vector for DNN had a length of 51 and was formed by concatenating 25 previous frames, the current frame, and 25 following frames. Local normalization was performed within one-second windows.

The accuracy of the baseline approach is summarized in Table 1 (see its first row). It is evident that it missed approximately 4% of speech segments. This fact affects the accuracy of the speech transcription system negatively, as the segments incorrectly marked as non-speech are not transcribed. Another problem of the baseline detector is the time precision of the change-point detection: the achieved value of $\delta_{2/3}$ is 0.42 s. This is also due to the fact that it is sometimes hard even for human annotators to determine the exact frame where a state change occurs. The baseline detector also produced a high number of false non-speech segments with a very short duration of one or two frames.

3.3. Smoothing the Output from DNN

As mentioned in the previous section, the baseline detector classified every input frame independently. On the other hand, every speech or non-speech segment usually lasts for at least several frames. Therefore, our next efforts were focused on smoothing the output from DNN. For this purpose, weighted finite state transducers were utilized using the OpenFst library[2].

The resulting scheme consists of two transducers (see Fig. 1). The first models the input signal. The other one is the transduction model and represents the smoothing algorithm. It consists of three states. The first state, denoted by 0, is the initial state. The transitions between states 1 and 2 emit the speech/non-speech labels and are penalized by penalty factors P1 a P2, respectively. Their values (500 and 500) were determined in several experiments not presented in this paper. Note that these values were tuned on different data set.

[1] http://torch.ch

[2] http://www.openfst.org/twiki/bin/view/FST/WebHome

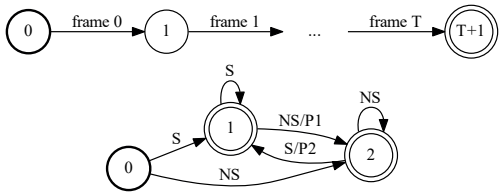

Fig. 1. The transducers representing the input signal (upper) and the basic smoothing model without any context (lower).

Given the two described transducers, the decoding process is performed using on-the-fly composition of the transduction and the input model of unknown size. This is possible since the input is considered to be a linear-topology, unweighted, epsilon-free acceptor. After each composition step, the shortest-path (considering tropical semi-ring) determined in the resulting model is compared with all other alternative hypotheses. When a common path is found among these hypotheses (i.e., with the same output label), the corresponding concatenated output labels are marked as the final fixed output. Since the rest of the best path is not known with certainty, it is denoted as a temporary output (i.e., it can be further refined).

The results obtained with the aid of the DNN-based approach with smoothing are summarized in the second row of Table 1. They show an overall significant boost in accuracy. For example, MR was reduced from 3.7% to 2.2% and the value of $\delta_{2/3}$ from 0.42 s to 0.27 s.

3.4. Using Artificial Training Data

The level of MR yielded so far, i.e., around 2%, still leads to a small increase in WER of a transcription system (e.g., from 13% to 14%), as the speech frames incorrectly classified as non-speech are omitted from transcription. The analysis we performed showed that the detector specifically misclassified the speech segments with background noise. The reason for this behavior is that the speech data used for DNN training so far were recorded in clean conditions (they served originally for training of an acoustic model for speech recognition systems).

Hence in the next step, the goal was to employ training data containing non-speech events, such as music or jingles in the background. The lack of such annotated data forced us to create an artificial source by mixing 30 hours of clean speech with non-speech recordings. For this purpose, a larger set of non-speech recordings of a total length of 100 hours was prepared first. After that, every speech recording was mixed with a randomly selected non-speech recording from the prepared set. Note that every non-speech recording used for mixing had to have the same or longer duration than the given input speech recording (the selected non-speech recording was trimmed to match the length of the speech recording) and its volume was increased or decreased to match the desired level of SNR (which was also selected randomly from an interval between -30 dB and 50 dB).

The labeling of this artificial data was carried out automatically: when SNR of the recording was higher than a defined threshold of 0 dB, the recording was marked as containing speech. In the opposite case, the recording was labeled as non-speech.

The results after using only these 30 hours of mixed training data are shown in the third row of Table 1. It is evident that this approach led to an increase in F-measure and a significant reduction in MR from 2.2% to 0.3%. Unfortunately, these improvements are

Table 1. Summarized results of individual SAD approaches described in Sect. 3.

Approach	FER	MR	FAR	F-measure	$\delta_{2/3}$
Baseline DNN-based	4.7%	3.7%	7.1%	0.3%	0.42 s
+ Basic smoothing	2.9%	2.2%	4.7%	28.5%	0.27 s
+ Artificial training data with noise	3.1%	0.3%	10.1%	41.3%	0.34 s
Modified artificial training data + Context-based smoothing	2.4%	0.5%	7.2%	52.7%	0.26 s

all accompanied by an increase in FAR and, even more importantly, an increase in $\delta_{2/3}$ from 0.27 s to 0.34 s. This negative fact motivated us to further improve the smoothing algorithm.

3.5. Improved Context-Based Smoothing

The scheme of the improved smoothing transducer that utilizes context is depicted in Fig. 2. In this case, both the speech and non-speech events are represented as sequences of three states, where the first and third states (the outer states) model the context. Similarly to smoothing without any context, the penalties are defined just for transitions between the speech and non-speech events, i.e., for transition a) from the end state of speech (*end_S*) to the start state of non-speech (*start_NS*), and b) from the end state of non-speech (*end_NS*) to the start state of speech (*start_S*).

To prepare training data containing transitions between speech and non-speech events, the data set from Sect. 3.4 was modified. At first, two recordings were chosen randomly from the artificial training set; one speech and one non-speech. After that, these two recordings were joined in a random order. The resulting recording then contained one of the two possible transitions (i.e., from speech to non-speech or from non-speech to speech) and was annotated automatically as follows:

1. The number of transition frames was derived from the input feature context window (25-1-25).
2. Only the 50 frames at the inner boundary of the two joined recordings were annotated as transitional, i.e., using 25 labels *stop_S* followed by 25 labels *start_NS* or 25 labels *stop_NS* followed by 25 labels *start_S*.
3. All other frames were marked as either speech or non-speech.

The results of the experiment with the context-based smoothing (see the fourth row of Table 1) show that this approach addresses the issue of an increase in $\delta_{2/3}$, which has emerged due to the use of the artificial training data (see the third row of Table 1). The value of $\delta_{2/3}$ was reduced from 0.34 s to 0.27 s. At the same time, a significant decrease in FAR, an increase in F-measure, and only a slight decrease in MR by 0.2% was achieved when compared to the previous experiment.

3.6. Evaluation on QUT-NOISE-TIMIT Corpus

The evaluation on QUT-NOISE-TIMIT corpus [23] shows the performance of the proposed approach in comparison with five approaches already presented in [23] and two techniques reaching the state-of-the-art results [24, 12]. The five approaches were: standardized VAD system ITU-T G.729 Annex B [25], standardized advanced front-end ETSI [26], Sohn's likelihood ratio test [27], Ramirez's long-term spectral divergence (LTSD) [28] and GMM based approach with use of MFCC features [23]. The latter two techniques were voice activity detection using subband noncircularity (SNC) [24] and complete-linkage clustering (CLC) for VAD [12].

The QUT-NOISE-TIMIT corpus was designed for training and evaluation of SAD systems in various noise scenarios and SNR levels. The data set combines clean speech from TIMIT corpus [29] with background noise recordings from QUT-NOISE data set [23]. The QUT-NOISE data set contains five types of background noises (scenarios: cafe, home, street, car, reverb) each from two different locations. Total amount of 600 hours were compiled and divided into two groups (A, B). Each group contains recordings from all scenarios in various SNR levels.

The training and testing protocols recommended for QUT-NOISE-TIMIT corpus presented in [23] were followed. The training was done with prior knowledge of target environment SNR; low noise (10, 15 dB), medium noise (0, 5 dB) and high noise (-10, -5 dB). However no prior knowledge of target environment scenario was utilized during the training phase. For each target SNR, group A was used for training and group B for testing and vice-versa. The proposed SAD module was trained as described in Sect. 3 with the exception of the use of artificial training data.

Figure 3 presents the comparison of the proposed SAD module and above introduced SAD approaches at three different noise levels: low, medium and high. In addition to MR and FAR, Half-Total Error Rate (HTER) was also evaluated. It is defined as equal-weighted average of MR and FAR. The obtained results show that our solution outperforms other SAD systems in low and medium noise conditions. The absolute reduction in HTER is over 2% over the previously best complete-linkage clustering approach. However, the HTER is approximately 2% worse in high noise conditions. The rest of the techniques are still being outperformed by a fair margin.

Our solution thus achieve state-of-the-art results in both low and medium noise conditions while staying competitive in high noise conditions on QUT-NOISE-TIMIT corpus.

4. RESULTS OF THE PROPOSED SAD APPROACH IN A REAL SPEECH TRANSCRIPTION SYSTEM

Given the findings and results from all previous experiments, the resulting SAD approach with the context-based smoothing was evaluated in a real speech-transcription system.

For this purpose, two test sets of Czech broadcasts were utilized. The first set represents 4 hours (22204 words) recorded from a Czech live news TV channel. Approximately 60% of its content consisted of speech segments. The length of the other set was 8 hours, it contained 7212 words, and speech frames formed only 10% of its content. This set represents broadcast of a Czech local radio station.

The transcription system employed an acoustic model based on a Hidden Markov Model - Deep Neural Network (HMM-DNN) hybrid architecture [30], where the baseline Gaussian Mixture Model (GMM) is trained as context-dependent, speaker-independent and contains 3886 physical states. The data for training of this model contained 270 hours of speech recordings. The parameters used for the DNN training were as follows: 5 hidden layers with a decreasing number of neurons per hidden layer (1024-1024-768-768-512),

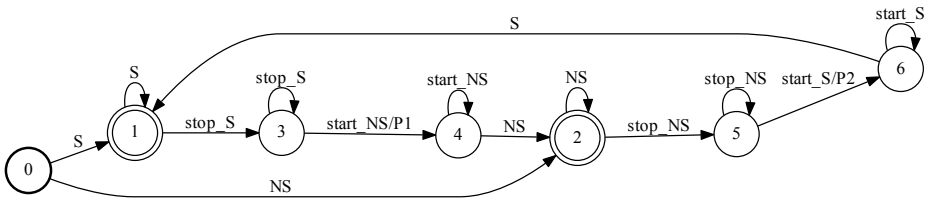

Fig. 2. The scheme of the WFST representing the context-based smoothing model.

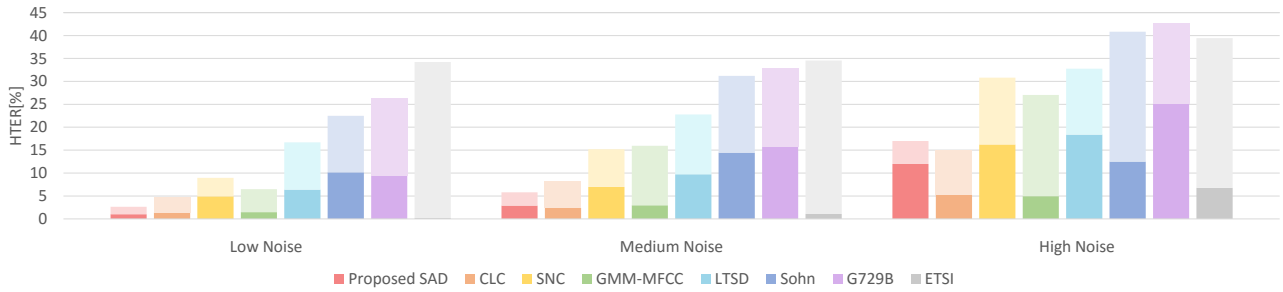

Fig. 3. Comparison of various SAD systems across QUT-NOISE-TIMIT corpus. HTER is defined as equal-weighted average of MR and FAR. The percentage contribution of MR and FAR to HTER bars is displayed by darker and lighter shades, respectively.

ReLU activation function, mini-batches of size 1024, 35 training epochs, learning rate 0.08. For signal parameterization, log filter banks were used with context windows of 5-1-5 and local normalization was employed within one-second windows.

The linguistic part of the system was composed of a lexicon and a language model. The lexicon contained 550k entries with multiple pronunciation variants and the language model was based on N-grams. For practical reasons (mainly with respect to the very large vocabulary size), the system used bigrams. However, 20 percent of all word-pairs actually include sequences containing three or more words, as the lexicon contains 4k multi-word collocations. The unseen bigrams are backed-off by Kneser-Ney smoothing.

4.1. Experimental Results

Within the performed experiments, both test sets were transcribed a) with and b) without the use of the SAD module. The obtained results are presented in Table 2, which contains values of Word Error Rate (WER) and Correctness (Corr) to show the transcription accuracy of the system. To measure computational demands with and without SAD, values of RTF (the ratio of the processing time to recording length) are also presented.

Table 2. Evaluation of the proposed SAD approach in a real speech transcription system.

Test set	live news TV channel		local radio station	
SAD module	Yes	No	Yes	No
WER [%]	12.4	12.7	14.0	17.9
Corr [%]	89.7	89.7	88.5	88.4
RTF	0.42	0.77	0.08	0.83

The obtained results indicate that the utilization of the proposed SAD approach was advantageous on both test sets. The yielded Corr and WER show that the SAD module limited the insertions coming from the non-speech parts and omitted hardly any speech parts.

The SAD module allowed the transcription system to operate with improved accuracy and, at the same time, RTF was almost two times, and more than ten times lower for the first and second test set, respectively. Of course, the reason for this difference is that the data in the second set contained fewer speech segments. Note that RTF of the SAD module itself is around 0.01 and all presented RTF values were measured using processor Intel Core i7-3770K @ 3.50GHz.

The transcription system complemented with SAD can also be utilized for online transcription without any major delay, because its latency is around 2 seconds.

5. CONCLUSIONS

In this paper, a new SAD approach suitable in offline as well as online speech transcription systems is proposed. The approach utilizes

- a DNN-based classifier;
- training data created artificially by mixing speech and non-speech recordings at various levels of SNR;
- a WFST-based decoder that smooths the output from DNN using a context-based model in which both the speech and non-speech events are represented as sequences of states.

The application of this approach to a real speech-recognition system leads to a) a slight decrease in WER, and b) significant reduction in RTF of the whole transcription process. The latter advantage is namely important for 24/7 monitoring of streams containing a large proportion of music (e.g., local radio stations), where the computational demands on the transcription system can be reduced dramatically.

6. ACKNOWLEDGEMENTS

This work was supported by the Technology Agency of the Czech Republic (Project No. TA04010199) and partly by the Student Grant Scheme 2017 of the Technical University in Liberec.

7. REFERENCES

[1] Georgios Evangelopoulos and Petros Maragos, "Speech event detection using multiband modulation energy.," in *INTERSPEECH*. 2005, pp. 685–688, ISCA.

[2] Bojan Kotnik, Zdravko Kacic, and Bogomir Horvat, "A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm.," in *INTERSPEECH*. 2001, pp. 197–200, ISCA.

[3] Houman Ghaemmaghami, Brendan Baker, Robert Vogt, and Sridha Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function.," in *INTERSPEECH*. 2010, pp. 3118–3121, ISCA.

[4] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah, "A model based voice activity detector for noisy environments," in *INTERSPEECH*, 2015, pp. 2297–2301.

[5] Neville Ryant, Mark Liberman, and Jiahong Yuan, "Speech activity detection on youtube using deep neural networks.," in *INTERSPEECH*. 2013, pp. 728–731, ISCA.

[6] Xiao-Lei Zhang and DeLiang Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection.," in *INTERSPEECH*. 2014, pp. 1534–1538, ISCA.

[7] Andreas Tsiartas, Theodora Chaspari, Nassos Katsamanis, Prasanta Kumar Ghosh, Ming Li, Maarten Van Segbroeck, Alexandros Potamianos, and Shrikanth Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *INTERSPEECH*, 2013, pp. 718–722.

[8] Jeff Ma, "Improving the speech activity detection for the darpa rats phase-3 evaluation.," in *INTERSPEECH*. 2014, pp. 1558–1562, ISCA.

[9] Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, April 2013.

[10] Jong Won Shin, Joon-Hyuk Chang, and Nam Soo Kim, "Voice activity detection based on statistical models and machine learning approaches," *Comput. Speech Lang.*, vol. 24, no. 3, pp. 515–530, July 2010.

[11] Tim Ng, Bing Zhang 0004, Long Nguyen, Spyros Matsoukas, Xinhui Zhou, Nima Mesgarani, Karel Vesely, and Pavel Matejka, "Developing a speech activity detection system for the darpa rats program.," in *INTERSPEECH*. 2012, pp. 1969–1972, ISCA.

[12] Houman Ghaemmaghami, David Dean, Shahram Kalantari, Sridha Sridharan, and Clinton Fookes, "Complete-linkage clustering for voice activity detection in audio and visual speech," in *INTERSPEECH*, 2015, pp. 2292–2296.

[13] George Saon, Samuel Thomas, Hagen Soltau, Sriram Ganapathy, and Brian Kingsbury, "The ibm speech activity detection system for the darpa rats program.," in *INTERSPEECH*. 2013, pp. 3497–3501, ISCA.

[14] Thad Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection.," in *ICASSP*. 2013, pp. 7378–7382, IEEE.

[15] 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 *ICASSP*, May 2013, pp. 483–487.

[16] Qing Wang, Jun Du, Xiao Bao, Zi-Rui Wang, Li-Rong Dai, and Chin-Hui Lee, "A universal vad based on jointly trained deep neural networks.," in *INTERSPEECH*. 2015, pp. 2282–2286, ISCA.

[17] Samuel Thomas, George Saon, Maarten Van Segbroeck, and Shrikanth S. Narayanan, "Improvements to the ibm speech activity detection system for the darpa rats program.," in *ICASSP*. 2015, pp. 4500–4504, IEEE.

[18] Chao Gao, Guruprasad Saikumar, Saurabh Khanwalkar, Avi Herscovici, Anoop Kumar, Amit Srivastava, and Premkumar Natarajan, "Online speech activity detection in broadcast news.," in *INTERSPEECH*. 2011, pp. 2637–2640, ISCA.

[19] Hoon Chung, Sung Joo Lee, and Yunkeun Lee, "Endpoint detection using weighted finite state transducer.," in *INTERSPEECH*. 2013, pp. 700–703, ISCA.

[20] Baiyang Liu, Björn Hoffmeister, and Ariya Rastrow, "Accurate endpointing with expected pause duration," in *INTERSPEECH*. 2015, pp. 2912–2916, ISCA.

[21] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *Signal Processing Conference, 2009 17th European*, Aug 2009, pp. 2549–2553.

[22] Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar, "An improved speech segmentation quality measure: the r-value.," in *INTERSPEECH*, 2009, pp. 1851–1854.

[23] David Dean, Sridha Sridharan, Robert Vogt, and Michael Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms.," in *INTERSPEECH*. 2010, pp. 3110–3113, ISCA.

[24] S. Wisdom, G. Okopal, L. Atlas, and J. Pitton, "Voice activity detection using subband noncircularity," in *ICASSP*, April 2015, pp. 4505–4509.

[25] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "Itu-t recommendation g.729 annex b: a silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, Sep 1997.

[26] Jin-Yu Li, Bo Liu, Ren-Hua Wang, and Li-Rong Dai, "A complexity reduction of etsi advanced front-end for dsr," in *ICASSP*, May 2004, vol. 1, pp. I–61–4 vol.1.

[27] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[28] Javier Ramrez, Jose C. Segura, Carmen Bentez, Angel De La Torre, and Antonio Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 3–4, 2004.

[29] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[30] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30 –42, jan. 2012.