INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech Activity Detection Based on Multilingual Speech Recognition System

Seyyed Saeed Sarfjoo, Srikanth Madikeri, and Petr Motlicek

Idiap Research Institute, Martigny, Switzerland
{ssarfjoo, msrikanth, petr.motlicek}@idiap.ch

Abstract

To better model the contextual information and increase the generalization ability of the Speech Activity Detection (SAD) system, this paper leverages a multilingual Automatic Speech Recognition (ASR) system to perform SAD. Sequence-discriminative training of Acoustic Model (AM) using Lattice-Free Maximum Mutual Information (LF-MMI) loss function, effectively extracts the contextual information of the input acoustic frame. Multilingual AM training causes the robustness to noise and language variabilities. The index of maximum output posterior is considered as a frame-level speech/non-speech decision function. Majority voting and logistic regression are applied to fuse the language-dependent decisions. The multilingual ASR is trained on 18 languages of BABEL datasets and the built SAD is evaluated on 3 different languages. On out-of-domain datasets, the proposed SAD model shows significantly better performance with respect to baseline models. On the Ester2 dataset, without using any in-domain data, this model outperforms the WebRTC, phoneme recognizer based VAD (Phn_Rec), and Pyannote baselines (respectively by 7.1, 1.7, and 2.7% absolute) in Detection Error Rate (DetER) metrics. Similarly, on the LiveATC dataset, this model outperforms the WebRTC, Phn_Rec, and Pyannote baselines (respectively by 6.4, 10.0, and 3.7% absolutely) in DetER metrics.
Index Terms: speech activity detection, multilingual automatic speech recognition, logistic regression, multilingual SAD

1. Introduction

Speech Activity Detection (SAD), a process of identifying the speech segments in an audio utterance [1], is a critical part of Automatic Speech Recognition (ASR), speaker recognition, speaker diarization, and other speech-based applications. Developing an accurate SAD system, operating in the noisy environment is an active research field in speech processing [2–6].

This paper explores SAD built around multilingual ASR systems, as we hypothesize it can offer better *generalization ability* by leveraging the contextual information extracted by ASR [7]. Generally, this paper employs a conventional multi-task network as a multilingual Acoustic Model (AM) trained using the Lattice-Free Maximum Mutual Information (LF-MMI) framework, capable of extracting the language-dependent contextual information. Using a multilingual dataset for the AM training was investigated in several studies [8–13]. Unlike applying a simple block-softmax loss on stacked input data with added language indicator for phoneme names, we apply LF-MMI loss on multi-task architecture, which provides a scalable approach to develop multilingual AM. Practically, we use PKWRAP, a PyTorch based Kaldi [14] wrapper for LF-MMI training of acoustic models [15][1]. The proposed multilingual acoustic model was trained on 18 languages of the BABEL datasets[2]. The original motivation for using this dataset is to train a SAD system robust to noise and language variabilities. Within each language-dependent part of AM, speech and non-speech acoustic frames are mapped to a different set of output context-dependent phones (i.e. posteriors, cf. Section 4). For each language, we use the index of maximum output posterior as a frame-level speech/non-speech decision function. In order to fuse the decisions from different languages, conventional logistic regression [16] and majority voting techniques are employed.

To investigate the generalization ability of the proposed SAD, experiments presented in the paper were performed on both in-domain and out-of-domain data. For out-of-domain experiments, two specific conditions are considered: (i) access to a small development set is available, or (ii) no in-domain data is available at all. Results with logistic regression and majority voting fusion are reported for these conditions. Concretely, the development part of the BABEL Kurdish dataset is used as an in-domain evaluation set. Eval parts of Ester2[3] and LiveATC[4] datasets are used as out-of-domain sets. BABEL Kurdish contains conversational telephony speech (CTS) in Kurdish. Ester2 is a broadcast news dataset in French. LiveATC comprises a large number of conversations between Air Traffic Controllers (ATCo) and pilots with a large variety of accents in English. To investigate the generalization ability of our SAD model, we consider different real-life scenarios with high variability in channel, background noise, and language.

We show that the proposed multilingual architecture offers comparable results on the in-domain set and significantly outperforms the baselines on the out-of-domain Ester2 and LiveATC datasets. For a fair comparison with the Google WebRTC and the popular BUT pre-trained phoneme recognizer based SAD (Phn_Rec)[5] in out-of-domain evaluation, we also assumed that no in-domain data is available during training. In addition, using a small development set in the logistic regression method further improves the performance of the proposed SAD system.

The rest of this paper is organized as follows: related works are discussed in Section 2. Multilingual acoustic model training is briefly explained in Section 3. The proposed multilingual ASR-based SAD is described in Section 4. Experiment setup and results are shown in Section 5. Conclusions are discussed in Section 6.

[1] Multitask acoustic modeling code will be made available as a part of PKWRAP

[2] One language (Somali) is part of MATERIAL project https://www.iarpa.gov/index.php/research-programs/material. Here we call the total dataset BABEL.

[3] http://catalog.elra.info/en-us/repository/browse/ELRA-S0338/

[4] https://www.liveatc.net/

[5] https://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

http://dx.doi.org/10.21437/Interspeech.2021-1058

2. Related works

Large effort was invested in the past to find the optimal features [17–20], or classifier [21–23] for the SAD task. We can mention Gaussian Mixture Model (GMM) [21], Hidden Markov Model (HMM) [22], or Support Vector Machines (SVM) [23] as the often used classifiers for the SAD task. With the advent of Deep Neural Networks (DNNs), several DNN-based architectures were proposed for the SAD task [24, 25] including Convolutional Neural Network (CNN) [26] and Recurrent Neural Network (RNN) [27] architectures. Recently, for training the SAD model in a noisy environment, DNN models with attention mechanism in temporal domain [6] and a combination of temporal and spectral domains [5] were investigated.

Contextual Information (CI) is important for training a robust SAD system, especially at low Signal-to-Noise Ratios (SNR) [28]. Several methods for boosting contextual information have been proposed. In [29], by boosting CI, Zhang and Wang proposed to generate multiple different predictions from a single DNN and reported a significant improvement over the standard DNN in challenging noise scenarios with low SNR levels. In [28], a boosted DNN (bDNN)-based SAD was proposed. Zhang and Wang exploited the input/output CI by adopting multiple input/output units for the DNN. In addition, to aggregate long-short term CI, they proposed an ensemble model that contains bDNNs of various sizes. However, the computational cost of the ensembled method is significantly higher than that of a single-bDNN-based SAD.

Capturing sequential contextual information using RNN architecture was investigated in [27], nevertheless, the improvement in the results was observed when the models were trained as *noise-dependent*. Using multilingual BABEL or Public Safety Communications (PSC) datasets for training the DNN based SAD with simple feed-forward architecture was investigated in [13]. PSC corpus that contains simulated first-responder type background noises and speech effects, was introduced in NIST OpenSAT 2019 challenge [30]. Similar to LiveATC, this dataset is challenging for ASR and SAD tasks.

3. Multilingual acoustic model training

Training a multilingual ASR system is an effective way to compensate for data shortages in low-resourced languages. DNN based acoustic models can be considered as a feature extractor to train a monolingual acoustic model for the specific target language. The multilingual models can either share the output layer or have separate output layers, which are called single- and multi-task models, respectively. Without any loss in performance, multi-task ASR training provides a much more scalable approach to develop multilingual AM [7]. LF-MMI significantly outperformed the conventional cross-entropy (CE) for training the multilingual AM [31]. The MMI cost function uses a numerator and a denominator graph to model the observed feature sequence based on the ground truth and compute the probability over all possible sequences, respectively. Sequence-discriminative training of multilingual AM using the LF-MMI loss function effectively extracts the contextual information of the input acoustic frame. In this paper, for training the multilingual AM, time-delayed neural network (TDNN) architecture with LF-MMI loss was applied. In order to obtain alignments to train all the TDNN models, HMM/GMM models were first trained for each language.

In multi-task training of AM, we have L objective functions where L is the number of training languages, computed independently of each other based on the language of the input utterance:

$$\mathcal{F}_{\text{MMI}}^{(l)} = \sum_{u=1}^{U_l} \log \frac{p\big(\mathbf{x}^{(u)}|\mathcal{M}_{\mathbf{w}(u)}^{l}, \theta\big)p(\mathbf{w}(u))}{p\big(\mathbf{x}^{(u)}|\mathcal{M}_{\text{den}}^{l}, \theta\big)}, \tag{1}$$

where U_l is the number of utterances in the current minibatch for language l, θ contains the shared and language-dependent parameters, $\mathcal{M}_{\mathbf{w}(u)}^{l}$ and $\mathcal{M}_{\text{den}}^{l}$ are language-specific numerator and denominator graphs, respectively. The overall cost function is the weighted sum of all language-dependent cost functions:

$$\mathcal{F}_{\text{MMI}} = \sum_{l=1}^{L} \alpha_l \mathcal{F}_{\text{MMI}}^{(l)}, \tag{2}$$

where α_l is the language-dependent weight for computing the total loss. Gradients for language-dependent layers are computed and updated for each minibatch. Using backpropagation, the shared parameters are then updated.

4. Multilingual ASR based SAD

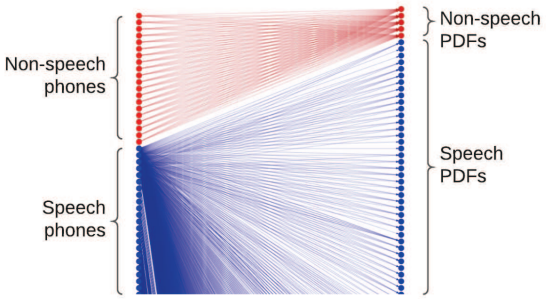

Figure 1: *Mapping between input phones and output PDFs in the HMM/GMM ASR model of Assamese language. Non-speech phones are mapped to the first five initial PDFs.*

In the trained HMM/GMM model for each language, we can observe the mapping between input phones and the output Probability Density Functions (PDFs). Figure 1, shows the mapping between input phones and output PDFs in the HMM/GMM ASR model of BABEL Assamese language. We can observe that the non-speech phones are mapped to the specific non-speech PDFs. As a result, the output posteriors of the language-dependent AM model are separated for input speech and non-speech frames. For training the AM model, the sequential discriminative LF-MMI loss function was applied. As a result, these output posteriors can be effective for discriminating the speech and non-speech frames.

The structure of multilingual ASR-based SAD is shown in Figure 2. After training the multilingual AM model, for each language we considered the speech/non-speech (SP/NSP) block, to detect the speech frames based on the PDF index of frame-level maximum output posterior. If the maximum output posterior belongs to one of the non-speech PDFs, we considered the current frame as a non-speech frame and vice versa. For fusing the decisions from different languages, we performed logistic regression and majority voting techniques. In logistic regression, we concatenated the frame-level decisions of SP/NSP block of each language as input features to predict the final decision. In the majority voting, we considered a frame as speech if

SP/NSP block of majority of the languages consider it as speech frame.

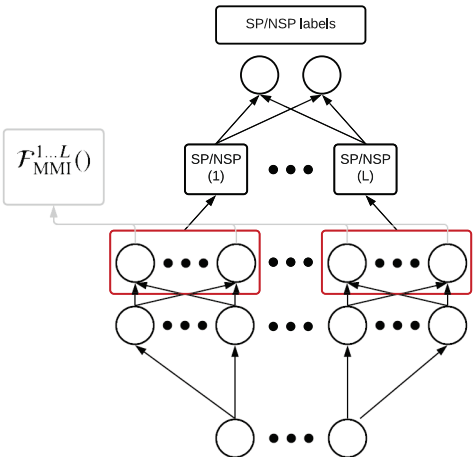

Figure 2: *Structure of the multilingual ASR based SAD. SP/NSP block detects speech frames based on the PDF index of frame-level maximum output posteriors. Logistic regression was used for fusing the language dependent speech/non-speech labels. In other words, fusion blocks use binary SP/NSP labels which are extracted from index of the maximum DNN posteriors.*

5. Experimental setup and results

5.1. Dataset and DNN configuration

To demonstrate the scalability of the multi-task system, we considered training the multilingual acoustic model with 18 languages from BABEL datasets with approximately 1000 hours of data. All languages are available at Linguistic Data Consorcium[6] (LDC). The name of BABEL languages used for training is shown in Table 1.

Table 1: *BABEL languages used for training.*

Languages
Assamese, Bengali, Cantonese, Haitian, Kazhak, Kurmanji_kurdish, Lao, Lithuanian, Pashto, Somali, Swahili, Tagalog, Tamil, Telugu, Tok_pisin, Turkish, Vietnamese, Zulu

For training the AM, we used 40-dimensional MFCCs as acoustic features, derived from 25 ms frames with a 10 ms frameshift. In addition, an online i-vector extractor of 100 dimensions was trained. For speeding up the training, we used a frame sub-sampling factor of 3. We also augmented the data with 2-fold speed perturbation in all the experiments. The network consists of 8 layers of TDNN with 1024 nodes in each layer. The pre-final layer has only 200 units. For training the AM model, PKWRAP, a PyTorch package for LF-MMI training of acoustic models was used [15]. Real-time factor for extracting the MFCC, i-vector, and computing the DNN posteriors on 4.20GHz Intel(R) Core(TM) i7-7700K CPU are 0.0076, 0.0032, and 0.18, respectively. The real-time factor for computing the DNN posteriors on GeForce GTX 1080 Ti GPU is 0.0066. The time for computing the logistic regression scores w.r.t. forward pass of the DNN is negligible. Using CPU and GPU for processing, this model is roughly five and sixty-two times faster than in real-time. Because of this reason, this approach is convenient as a pre-processing step for audio-based interactive systems.

For investigating the generalization ability of the proposed SAD, we performed experiments on in-domain and out-of-domain scenarios. The development part of the BABEL Kurdish dataset was used as an in-domain evaluation set. Eval parts of Ester2 and LiveATC datasets were used as out-of-domain sets. LiveATC was collected in the automatic collection and processing of voice data from the air-traffic communications (ATCO2) project.[7] For all evaluations, we considered the conditions when we have access to the in-domain development set, which is used for training the logistic regression (ASR_Mul_LR) and the condition that we do not have a development set, which is the case for majority voting (ASR_Mul_MV) and SP/NSP blocks of single best language. For ASR_Mul_LR models, the threshold for SP/NSP detection was set based on Half Total Error Rate (HTER). The duration and number of segments in the selected datasets are shown in the Table2. For investigating the *generalization ability* of our SAD model, we considered different real-life scenarios with high variation in channel, background noise, and language.

Table 2: *Duration and number of segments in the selected datasets.*

Dataset	Duration (hour)	# Segments
LiveATC_dev	2.7	1.0k
LiveATC_eval	6.8	0.9k
Ester2_dev	7.4	1.2k
Ester2_eval	7.2	1.7k
BabelKurdish_dev	20.6	11.0k
BabelKurdish_eval	20.0	11.3k

In this experiment, False Alarm (FA), Miss detection (Miss), and Detection Error Rate (DetER) were used as performance measures. DetER is defined as:

$$\text{DetER} = \frac{\text{False alarm} + \text{Miss detection}}{\text{Total duration of speech in the reference file}}. \quad (3)$$

FA and Miss, are performance measures with just considering the False alarm and Miss detection in the numerator of DetER, respectively. In this paper, we considered Brno University's phoneme recognizer-based VAD (Phn_Rec), Google webRTC[8], and Bi-directional LSTM (BLSTM) based SAD from Pyannote [32] as baseline models. Phn_Rec is a Hungarian phoneme recognizer, with all the phoneme classes linked to the 'speech' class [33]. Hungarian data which was collected in SpeechDat-E project[9], was found as the best for generic phoneme recognition working over the different languages [34]. For Pyannote SAD, we trained the model using the same 18 BABEL languages. For a fair comparison with WebRTC and Phn_Rec, we considered the out-of-domain scenarios when we don't have access to the in-domain data. Based on the SAD result on the second DIHARD challenge[10], the aggressiveness mode of WebRTC SAD was set to 3.

[6] https://www.ldc.upenn.edu

[7] https://www.atco2.org/

[8] https://github.com/wiseman/py-webrtcSAD

[9] http://www.fee.vutbr.cz/SPEECHDAT-E

[10] https://dihardchallenge.github.io/dihard2

5.2. In-domain evaluation

Comparison of SAD results on in-domain experiment for BabelKurdish_eval set is shown in the Table 3. To reduce the noise in the classifier's output, in each ASR-based SAD, we applied temporal smoothing for detecting the start and end of each speech segment. In all experiments output of Tok_pisin language showed a single best result which is ASR_Single_Best in Table 3. The majority voting and logistic regression fusion multi-language results are called ASR_Mul_MV, and ASR_Mul_LR, respectively. For investigating the result of trainable ASR-based and Pyannote models in the in-domain scenario, the result of pre-trained Phn_Rec and WebRTC models are not shown in Table 3.

For the in-domain experiment, temporal smoothing parameters are tuned using the in-domain development set. Here, w.r.t. ASR_SingleBest model, ASR_Mul_LR improved the DetER by 1.2 %. This LR fusion caused to decrease in the miss detection with increasing the false alarm. ASR-based SAD showed comparable performance w.r.t. the Pyannote model. Using different DNN architectures and temporal smoothing methods are the main reasons for observing the difference in the performance of these two systems.

Table 3: *Comparison of SAD results on in-domain BabelKurdish_eval set. ASR_SingleBest, ASR_Mul_LR, and ASR_Mul_MV are multilingual ASR based SAD systems when single best system, logistic regression based, or majority voting based fusion is considered, respectively.*

SAD Model	DetER (%)	FA (%)	Miss (%)
ASR_SingleBest	19.9	**4.0**	15.9
ASR_Mul_LR	18.7	5.2	13.5
ASR_Mul_MV	19.3	5.6	13.7
Pyannote	**18.1**	5.9	**12.2**

5.3. Out-of-domain evaluation

Comparison of SAD results on out-of-domain LiveATC evaluation set is shown in the Table 4. Here ASR_SingleBest and ASR_Mul_MV models are not using any in-domain data. ASR_Mul_LR model was trained using the in-domain development set. Without considering the ASR_Mul_LR model, ASR_SingleBest and ASR_Mul_MV models significantly outperformed the baseline models based on DetER performance measure. Training the multilingual AM model is one of the reasons for observing good results in the ASR_SingleBest model. The ASR_Mul_LR model outperformed the ASR_SingleBest model with a relative improvement of 4.0% on DetER performance measure. Comparison of SAD results on out-of-domain Ester2 evaluation set is shown in the Table 5. In this out-of-domain set, we observed the same pattern, and based on DetER performance measure, the proposed model significantly outperformed the baselines. The ASR_Mul_LR model outperformed the ASR_SingleBest model with a relative improvement of 38.4% on DetER performance measure. Based on the observed results, the proposed multilingual ASR-based SAD showed strong *generalization ability*. We believe that training procedure as a multi-task learning system has the main effect on achieving this *generalization ability*. In addition, having a small in-domain dataset improves the performance of the proposed method.

Table 4: *Comparison of SAD results on out-of-domain LiveATC evaluation set.*

SAD Model	DetER (%)	FA (%)	Miss (%)
ASR_SingleBest	10.1	4.9	5.2
ASR_Mul_LR	**9.7**	6.1	**3.6**
ASR_Mul_MV	11.1	**4.3**	6.8
Phn_Rec	20.1	4.6	15.5
WebRTC	16.5	9.4	7.1
Pyannote	13.8	10.1	3.7

Table 5: *Comparison of SAD results on out-of-domain Ester2 evaluation set.*

SAD Model	DetER (%)	FA (%)	Miss (%)
ASR_SingleBest	5.2	4.7	0.5
ASR_Mul_LR	**3.2**	**2.3**	0.9
ASR_Mul_MV	4.7	4.2	0.5
Phn_Rec	6.4	3.9	2.5
WebRTC	11.8	6.5	5.3
Pyannote	7.4	7.3	**0.1**

6. Conclusions

Contextual information is important for training a robust SAD system, especially at noisy sets. In this paper, we trained the SAD system using the multilingual ASR model. This ASR model was trained with LF-MMI loss on multi-task architecture which provides a much more scalable approach to develop AM. The decision for detecting speech/non-speech frames is based on the index of maximum output posterior. Majority voting and logistic regression were applied to fuse the language-dependent decisions. We observed the significant improvement w.r.t. baselines on out-of-domain Ester2 and LiveATC evaluation sets. More specifically, for the Ester2 dataset, the proposed SAD method outperformed the WebRTC, Phn_Rec, and Pyannote BLSTM SAD models by absolute 7.1%, 1.7%, and 2.7% in DetER respectively. Similarly, w.r.t. WebRTC, Phn_Rec, and Pyannote BLSTM SAD models, respectively, we obtained an absolute improvement of 6.4%, 10.0%, and 3.7% in DetER on LiveATC dataset. In addition, using small development set in the logistic regression method, further improved the performance of the proposed SAD system. In in-domain experiments, with tuning the temporal smoothing parameters we observed comparable results w.r.t. the Pyannote model.

7. Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 833635 (ROXANNE: Real time network, text, and speaker analytics for combating organised crime). The research is also partially based upon the work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract #FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

8. References

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[2] B. Sharma, R. K. Das, and H. Li, "Multi-level adaptive speech activity detector for speech in naturalistic environments." in *Twentieth annual conference of the international speech communication association (INTERSPEECH)*, 2019, pp. 2015–2019.

[3] F. Martinelli, G. Dellaferrera, P. Mainar, and M. Cernak, "Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8544–8548.

[4] G. Dellaferrera, F. Martinelli, and M. Cernak, "A bin encoding training of a spiking neural network based voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3207–3211.

[5] J. Lee, Y. Jung, and H. Kim, "Dual attention in time and frequency domain for voice activity detection," *arXiv preprint arXiv:2003.12266*, 2020.

[6] Z. Zheng, J. Wang, N. Cheng, J. Luo, and J. Xiao, "Mlnet: An adaptive multiple receptive-field attention neural network for voice activity detection," *arXiv preprint arXiv:2008.05650*, 2020.

[7] S. Madikeri, B. Khonglah, S. Tong, P. Motlicek, H. Bourlard, and D. Povey, "Lattice-free maximum mutual information training of multilingual speech recognition systems," *Twenty first annual conference of the international speech communication association (INTERSPEECH)*, 2020.

[8] D. Imseng, P. Motlicek, P. N. Garner, and H. Bourlard, "Impact of deep mlp architecture on different acoustic modeling techniques for under-resourced speech recognition," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 332–337.

[9] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, and H. Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2322–2326.

[10] P. Motlicek, D. Imseng, B. Potard, P. N. Garner, and I. Himawan, "Exploiting foreign resources for dnn-based asr," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–10, 2015.

[11] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of multilingual asr using end-to-end lf-mmi," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6061–6065.

[12] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, May 2014, pp. 7639–7643.

[13] M. Karafiát, M. K. Baskar, I. Szöke, H. K. Vydana, K. Veselỳ, J. Černockỳ *et al.*, "But opensat 2019 speech recognition system," *arXiv preprint arXiv:2001.11360*, 2020.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[15] S. Madikeri, S. Tong, J. Zuluaga-Gomez, A. Vyas, P. Motlicek, and H. Bourlard, "Pkwrap: a pytorch package for lf-mmi training of acoustic models," *arXiv preprint arXiv:2010.03466*, 2020.

[16] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

[17] H. K. Maganti, P. Motlicek, and D. Gatica-Perez, "Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–1037.

[18] E. Chuangsuwanich and J. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," in *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[19] A. Misra, "Speech/nonspeech segmentation in web videos," in *Thirteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[20] R. Zazo Candil, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," 2016.

[21] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matějka, "Developing a speech activity detection system for the darpa rats program," in *Thirteenth annual conference of the international speech communication association (INTERSPEECH)*, 2012.

[22] H. Veisi and H. Sameti, "Hidden-markov-model-based voice activity detector with high speech detection rate for speech enhancement," *IET signal processing*, vol. 6, no. 1, pp. 54–63, 2012.

[23] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *6th International Conference on Signal Processing, 2002.*, vol. 2. IEEE, 2002, pp. 1124–1127.

[24] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2012.

[25] A. Ivry, B. Berdugo, and I. Cohen, "Voice activity detection for transient noisy environment based on diffusion nets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 254–264, 2019.

[26] S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5549–5553.

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

[28] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2015.

[29] X. L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Fifteenth annual conference of the international speech communication association (INTERSPEECH)*, 2014.

[30] F. R. Byers, J. G. Fiscus, S. O. Sadjadi, G. A. Sanders, and M. A. Przybocki, "Open speech analytic technologies pilot evaluation opensat pilot," Tech. Rep., 2019.

[31] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi." in *Nineteenth annual conference of the international speech communication association (INTERSPEECH)*, 2018, pp. 12–16.

[32] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020.

[33] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2006, pp. I–I.

[34] P. Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapso, T. Mikolov, and O. Plchot, "But system description for nist lre 2007," in *Proc. 2007 NIST Language Recognition Evaluation Workshop*, 2007, pp. 1–5.