Two-Dimensional Convolutional Recurrent Neural Networks for Speech Activity Detection

Anastasios Vafeiadis[1], Eleftherios Fanioudakis[2], Ilyas Potamitis[2], Konstantinos Votis[1], Dimitrios Giakoumis[1], Dimitrios Tzovaras[1], Liming Chen[3], Raouf Hamzaoui[3]

[1]Information Technologies Institute, Center for Research & Technology Hellas, Thessaloniki, Greece
[2]Technological Educational Institute of Crete, Department of Music Technology and Acoustics, Crete, Greece
[3]Faculty of Computing, Engineering and Media, De Montfort University, Leicester, UK

{ta1829@edu, potamitis@staff}.teicrete.gr, {anasvaf, kvotis, dgiakoum, tzovaras}@iti.gr, {liming.chen, rhamzaoui}@dmu.ac.uk

Abstract

Speech Activity Detection (SAD) plays an important role in mobile communications and automatic speech recognition (ASR). Developing efficient SAD systems for real-world applications is a challenging task due to the presence of noise. We propose a new approach to SAD where we treat it as a two-dimensional multilabel image classification problem. To classify the audio segments, we compute their Short-time Fourier Transform spectrograms and classify them with a Convolutional Recurrent Neural Network (CRNN), traditionally used in image recognition. Our CRNN uses a sigmoid activation function, max-pooling in the frequency domain, and a convolutional operation as a moving average filter to remove misclassified spikes. On the development set of Task 1 of the 2019 Fearless Steps Challenge, our system achieved a decision cost function (DCF) of 2.89%, a 66.4% improvement over the baseline. Moreover, it achieved a DCF score of 3.318% on the evaluation dataset of the challenge, ranking first among all submissions.

Index Terms: speech activity detection, voice activity detection, convolutional recurrent neural networks

1. Introduction

One of the most important problems in the area of speech signal processing is distinguishing speech from non-speech periods in an audio signal [1]. SAD is part of many applications (e.g., ASR [2] and speaker diarization [3]). Recently, SAD has received attention especially in research projects [4] and challenges [5, 6]. The main reason is that real-world speech recordings, such as the Apollo audio data [7], are characterized by multiple noise types and several overlap instances over most channels. Most audio channels are degraded due to high channel noise, system noise, attenuated signal bandwidth, transmission noise, cosmic noise, analog tape static noise, and noise due to tape aging.

Voice activity detection (VAD) algorithms have been extensively researched [8, 9, 10, 11]. These algorithms are mainly probabilistic models, use temporal and power spectral characteristics of sound and do not require training. Because of their low complexity, the majority of VAD algorithms work well for real-time applications. However, they require extensive fine-tuning of their hyper-parameters and have lower performance, to a certain extent, in low signal-to-noise-ratio (SNR) environments.

A large number of SAD methods and models have been proposed for highly degraded acoustic conditions. Most of them are supervised and exploit the time and frequency properties of speech and noise to effectively separate speech from non-speech [12, 13]. Some of them use energy operators and multi-band modulations [14], autocorrelation coefficients [15], as well as time and frequency feature-level fusion [16].

The problem with most supervised methods is that the data has to be well annotated (in milliseconds), which is a time-consuming task requiring specialists to hand-label the audio data. To address this problem, semi-supervised and unsupervised methods have been proposed. In particular, Gaussian Mixture Models (GMMs) have been extensively used [17, 18]. GMM-based SAD systems are typically composed of two GMMs: one trained on speech frames and one on non-speech frames.

More recently, deep learning (DL) based methods have been proposed to solve the SAD problem. The ability of deep neural networks to automatically extract low-level features from a given signal segment, has made them popular in various scientific fields. Various DL methods have been explored, compared and fused with VAD algorithms [19], in order to select the ones best suited for the SAD problem [20]. Among these DL based methods, Recurrent Neural Networks (RNNs) have several properties that make them a popular choice for SAD [21, 22].

In this paper, we consider SAD as a multilabel classification problem and use a 2D CRNN to address it. The ability of the Convolutional Neural Network (CNN) layers to capture the temporal and frequential information of the audio signal and the ability of the recurrent layers to identify the time intervals, for long sequences, of the classified events (speech, non-speech), make them suitable for this problem. Furthermore, we use the stratified k-fold cross-validation method, to preserve the percentage of samples for each class in different folds and use majority voting to calculate the DCF. Finally, we perform convolutions on the k-fold majority voting results to smooth the output.

The remainder of the paper is organized as follows. Section 2 describes our methodology, including raw audio signal preprocessing, feature extraction and network architecture. The evaluation of the networks for the dataset is presented in Section 3. Finally, Section 4 concludes the paper.

2. Methodology

This section describes the steps of our proposed approach, starting from the extraction of the features of the audio signal that are used as input to the 2D network architectures, to describing the neural networks architectures used in our experiments.

As an augmentation method, we used random time shifts for each 1 s of recording (-8000 samples, +8000 samples), creating an array with elements that roll between the last position and the re-introduced at first in order to keep all the signal information. Finally, we did not apply any denoising method, since we wanted the network to learn how to distinguish between noisy and ambient recordings.

2.1. Feature Extraction

As a first step, we split the recordings into segments of 1 s and assign to each of the samples the corresponding label (0: non-speech, 1: speech). The main advantage of deep neural networks is their ability to extract features from raw data. We calculated the Short-time Fourier Transform (STFT) spectrogram for each 1 s - recording and extracted the corresponding grayscale spectrogram image. The length of the Fast Fourier Transform (FFT) was 256, with a hop length of 64. We selected the Hanning window for the FFT, since it is commonly used for speech signals [23]. This resulted in a 129x126 spectrogram used as input to the 2D CRNN.

2.2. Network Description

As a network architecture, we used a modified 2D CRNN, where we permute the dimensions of the CNN output and then reshape them to feed the Gated Recurrent Unit (GRU) of the RNN. Additionally, we apply max-pooling on the frequency domain only, when calculating convolutions, allowing the entire time information to be processed by the RNN. The network was trained in Keras [24] with TensorFlow [25] backend, using a batch size of 32 for 25 epochs.

Our 2D CRNN architecture is shown in Figure 1. The architecture was inspired by Bartz *et al.* [26], who applied 2D CRNNs for language identification in text documents. We applied a similar architecture for SAD. The CNN part of our 2D CRNN architecture consists of five convolutional layers. The

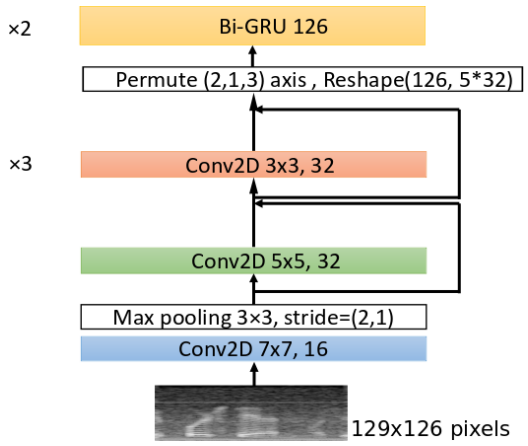

Figure 1: *2D CRNN architecture with STFT spectrogram image as input*

first one has 16 filters, the second 32, the third 32, the fourth 32 and the fifth one 32. The first layer computes convolutions over the time and frequency domain, using 7x7 kernels. A 3x3 max pooling operation follows each convolutional layer and the subsampled feature maps are fed to the next convolutional layer. We used 2x1 strides for each max pooling operation since we wanted to sub-sample the frequency domain and leave the time domain as is to be processed by the RNN part. The size of the kernels was decreased to 5x5 in the second convolution and to 3x3 in the third, fourth and fifth. Each convolution was followed by batch normalization [27] of its outputs, before the element-wise application of the rectified linear unit (ReLU) [28] activation function. Finally, the resulting feature maps of the consecutive convolution-max pooling operations were permuted and reshaped to be used as input to two bi-directional GRUs (RNN part), each one having a filter size of 126. We used the Adam [29] optimizer with an initial learning rate l_r=0.001 which was reduced by a factor of 0.01, when there was no DCF improvement for 5 consecutive epochs.

3. Evaluation and Analysis

We conducted experiments on the development and evaluation datasets of the Fearless Steps Challenge (Apollo 11) [30]. These datasets consist of 39 and 40 recordings, respectively, each recording containing a total of approximately 30 min audio in wav format and sampled at 8000 Hz.

3.1. Systems

We compared the proposed 2D CRNN model with STFT spectrograms as input to a 1D CRNN that uses raw waveforms as inputs, our 2D CRNN where we replaced the STFT spectrograms by MFCC images, a GRU-RNN [21] using MFCCs as input, a state-of-the-art VAD system [31] and the baseline system results [7], provided by the organizers. MFCCs are one of the most popular features for voice recognition [32]. For our experiments, we calculated 20 double-delta coefficients (including the $0th$ energy coefficient) using an FFT with a Hanning window size of 2048 and a hop length of 512, which resulted in a 20x16 vector. This 2D vector was used as an input to the 2D CRNN. The main reason for selecting the double-delta coefficients is that they convey richer information about the frames context [33].

The GRU-RNN is a basic network consisting of two bi-directional GRU units, each one having a filter size of 126. All the networks were trained using the same parameters as described in Section 2.2.

3.1.1. 1D CRNN

The 1D CRNN architecture is shown in Figure 2. The CNN part of it consists of five convolutional layers. The filter size at each layer increases as a power of two. Specifically the first one has 16, the second 32, the third 64, the fourth 128 and the fifth one 256. The first layer performs convolutions over the time domain (raw waveform), using 1x3 kernels. A 1x2 max pooling operation follows on each convolutional layer and the subsampled feature maps are fed to the next convolutional layer. Each convolution is followed by batch normalization of its outputs, before an element-wise application of the ReLU activation function. We selected this activation function for each layer, as it is the most commonly used. Finally, the resulting feature maps of the consecutive convolution-max pooling operations are fed as input to two bi-directional GRUs (RNN part), each one having a filter size of 126. The main reason for using bi-directional GRUs over the original long short term memory (LSTM) units proposed for CRNNs, is that they train faster and we can achieve comparable performance with the LSTMs. Furthermore the bi-directional unit could learn the context based on future and past values (e.g., speech followed by non-speech, large periods of silence or noise). We used the Adam optimizer

with an initial learning rate lr=0.001. We reduced the lr by a factor of 0.01, when there was no DCF improvement for five consecutive epochs, which boosted our DCF score. The 1D and 2D CRNNs use the same number of convolutional layers, but with different kernel sizes and number of filters, since the nature of the input data is different.

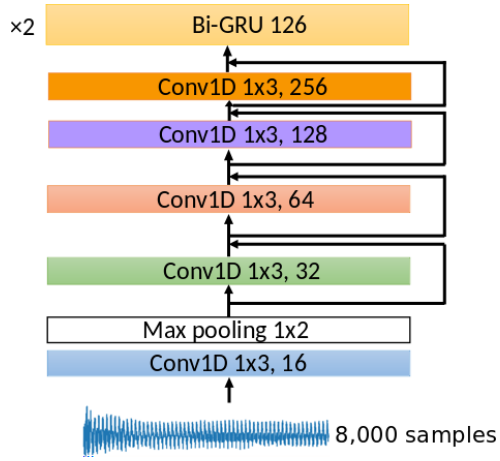

Figure 2: *1D CRNN architecture with raw waveform as input*

3.2. Network Training

In order to avoid bias over a single class, we trained our networks using the stratified k-fold method ($k = 5$). This helped us preserve the percentage of samples for each class. The metric that was selected as a performance measurement was the DCF score, which is defined as follows:

$$DCF(\theta) = 0.75 \times P_{FN}(\theta) + 0.25 \times P_{FP}(\theta)$$

where θ denotes a given system decision-threshold setting. P_{FP} is the probability of a false positive (FP), which is equal to the $total\ FP\ time$ divided by the $annotated\ total\ non-speech\ time$ and P_{FN} is the probability of a false negative (FN), which is equal to the $total\ FN\ time$ divided by the $annotated\ total\ speech\ time$. To optimize parameter θ we tested all values from 0 to 1 with a step size of 0.01.

3.3. Results

The results are summarized in Table 1. Our approaches significantly outperformed the unsupervised baseline algorithm and the VAD system [31] for the development set (ground truth given). VAD algorithms can predict continuous speech segments in some noisy channels but they require a lot of fine-tuning based on the recorded channel.

Figure 3 depicts the SAD performance for the different architectures. We notice that the 2D CRNN with STFT spectrograms as input is able not only to accurately detect the speech and non-speech segments, but also to correct the labeling of the ground truth.

Since we evaluated the CRNNs using a 5-fold stratified cross validation, it was also necessary to compare the performance of each fold. Figure 4 shows that the average of the 5-folds achieved the best performance amongst them, and the standard deviation of each fold was very small, justifying the robustness of the CRNN architectures.

Table 1: *Performance of different architectures using DCF as a metric on the development dataset. No collars are used*

Systems	DCF (%) without filtering
1D CRNN	3.02
2D CRNN (STFT spec. image)	**2.89**
2D CRNN (MFCC image)	4.02
MFCC RNN	4.08
Google VAD [31] (mode 0)	13.99
Baseline [7]	8.6

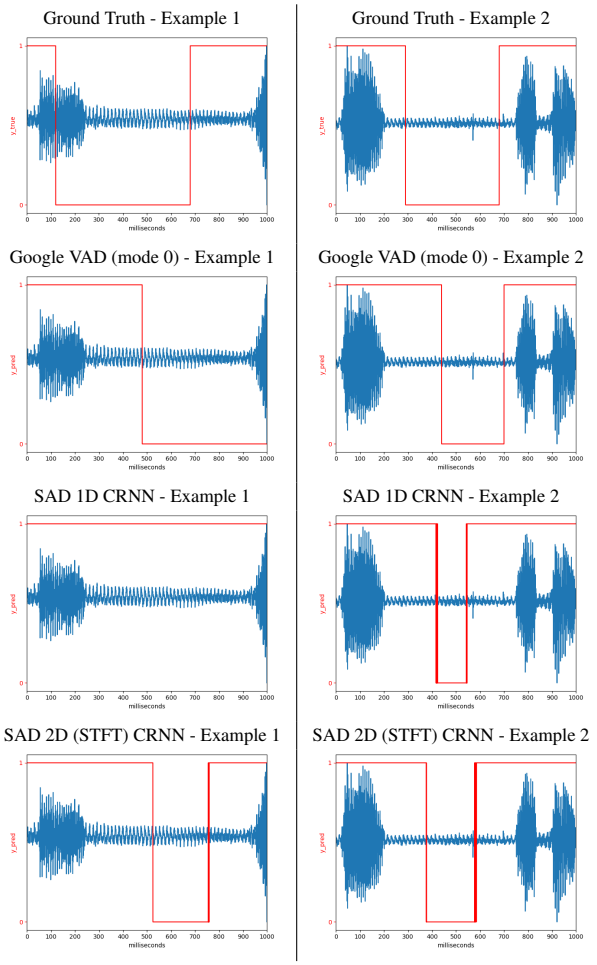

Figure 3: *Examples of speech and non-speech activity detection of 1D and 2D (STFT) CRNN architectures with the ground truth of the development dataset*

Figure 5 shows the advantage of our moving average (temporal smoothing) post-processing filter. The CRNN architectures output many spikes in the waveform as speech predictions. These spikes usually range from 0.01 to 0.5 s (red line). By calculating convolutions of 10 ms windows (80 samples) average, we were able to correct the predictions (black line). Additionally, the average filter can also work as a confidence score for each predicted segment. The main problem that we

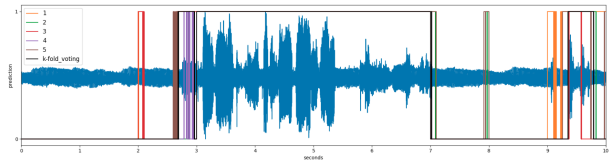

Figure 4: *SAD results for the 5 folds and the ensembled majority on an example of the evaluation dataset (no ground truth given) using 2D (STFT) CRNN*

are trying to solve is the misclassification of spikes (either detected as speech or non-speech). As another post-processing step, segments whose duration was shorter than 150 ms (1200 samples) and which were predicted as speech were relabelled as non-speech if their preceding and following segment was predicted as non-speech. This is because speech segments cannot be too short due to the inertia of human articulators.

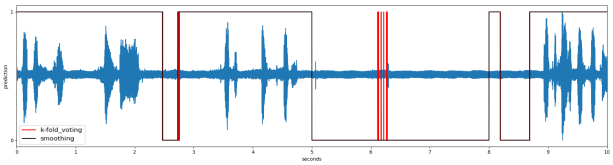

Figure 5: *SAD results using the moving average filter of the convolutions (temporal smoothing) after averaging the 5 folds*

Finally, our 2D CRNN architecture, using STFT spectrograms as input, achieved a DCF score of 3.318% on the evaluation dataset of the 2019 Fearless Steps SAD Challenge [7], ranking first among the 27 submissions.

4. Conclusions

We proposed a system that exploits a 2D CRNN for SAD. On Task 1 of the 2019 Fearless Steps Challenge, our system outperformed a well-known VAD algorithm [31] and achieved the first place among the 27 submissions. The novelty of our approach lies in treating SAD as a 2D image multilabel classification problem, where the input is an STFT spectrogram of the audio recording. The operational simplicity of our system makes it also power efficient. As future work, we will compute the DCF for each channel, since each channel may have a different SNR, and average the scores. We will also try other deep learning architectures, pseudo-labeling, test time shift augmentation techniques, consider other signal lengths and, finally, embed our approach in mobile and ASR devices.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157, project ACROSSING.

6. References

[1] A. Sholokhov, M. Sahidullah, and T. Kinnunen, "Semi-supervised speech activity detection with an application to automatic speaker verification," *Computer Speech & Language*, vol. 47, pp. 132–156, 2018.

[2] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[3] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[4] 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*, 2012.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[6] H. Dubey, A. Sangwan, and J. H. Hansen, "Robust feature clustering for unsupervised speech activity detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2726–2730.

[7] J. H. Hansen, A. Joglekar, M. Chandra Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," *Proc. Interspeech 2019*, 2019.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[9] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, vol. 2. IEEE, 1999, pp. 789–792.

[10] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, 2008.

[11] R. Martin and I. Cohen, "Single-channel speech presence probability estimation and noise tracking," in *Audio Source Separation and Speech Enhancement*. Wiley, 2018, ch. 6, pp. 87–106.

[12] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[13] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[14] G. Evangelopoulos and P. Maragos, "Speech event detection using multiband modulation energy," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[15] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[16] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, "Voice activity detection based on conditional random fields using multiple features," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra *et al.*, "All for one: feature combination for highly channel-degraded speech activity detection." in *Interspeech*, 2013, pp. 709–713.

[18] X. Wu, M. Zhu, R. Wu, and X. Zhu, "A self-adapting gmm based voice activity detection," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. IEEE, 2018, pp. 1–5.

[19] B. Liu, Z. Wang, S. Guo, H. Yu, Y. Gong, J. Yang, and L. Shi, "An energy-efficient voice activity detector using deep neural networks and approximate computing," *Microelectronics Journal*, 2019.

[20] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3391–3398.

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

[22] G. Gelly and J.-L. Gauvain, "Optimization of RNN-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, 2018.

[23] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[24] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[25] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[26] C. Bartz, T. Herold, H. Yang, and C. Meinel, "Language identification using deep convolutional recurrent neural networks," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 880–889.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations (ICLR-15)*, 2014.

[30] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," *Proc. Interspeech 2018*, pp. 2758–2762, 2018.

[31] "Google WebRTC," 2016. [Online]. Available: https://webrtc.org/

[32] B. Milner, J. Darch, I. Almajai, and S. Vaseghi, "Comparing noise compensation methods for robust prediction of acoustic speech features from mfcc vectors in noise," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[33] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with lombard and noisy speech," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990, pp. 857–860.