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

The Application of Learnable STRF Kernels to the 2021 Fearless Steps Phase-03 SAD Challenge

Tyler Vuong[1], *Yangyang Xia*[1], *Richard M. Stern*[1,2]

[1] Department of Electrical and Computer Engineering, Carnegie Mellon University, USA
[2] Language Technologies Institute, Carnegie Mellon University, USA

tvuong@andrew.cmu.edu, raymondxia@cmu.edu, rms@cs.cmu.edu

Abstract

We describe a deep-learning-based system developed for the Fearless Steps Phase-03 Speech Activity Detection (SAD) challenge. The system includes both learnable spectro-temporal receptive fields (STRFs) and unconstrained 2-dimensional convolutional kernels in the first layer. Experiments show that the inclusion of learnable STRFs in the first layer increases the system's robustness to additive noise. Additionally, we found that utilizing SpecAugment during training improves generalization on unseen data. By incorporating these enhancements and others our system achieved the best score in the official SAD challenge.

Index Terms: speech activity detection, spectro-temporal receptive field, gabor filters

1. Introduction

Speech activity detection (SAD) is fundamental to speech-related technologies such as automatic speech recognition and speaker identification. The design of robust SAD algorithms involves strategies that separate speech signals from common or novel environmental sounds, particularly at low signal-to-noise ratios (SNRs).

Organized by the University of Texas at Dallas's Center for Robust Speech Systems (UTD-CRSS), the Fearless Step Challenge Phase-03 (FSC P3) is the third phase of a competition over multiple speech technologies on the recovered and digitized recordings from the NASA Apollo-11 Mission [1, 2, 3]. The FSC data contains 19,000 hours of naturalistic speech recordings degraded by a combination of additive noise, degradations due to linear and nonlinear channel, bandwidth restriction, and many other effects that pose challenges to speech processing algorithms [1]. For the task of speech activity detection, the FSC data are especially challenging for their low sampling rate (8 KHz), the high variation of SNRs (between 0 and 20 dB), and the variety of degradation mentioned above [1]. The CMU Robust Speech Group participated in the SAD task of the FSC P3 Challenge during spring 2021. This paper describes our efforts to apply the STRFNet system that has enjoyed recent success in the tasks of voice type discrimination [4] and spoofing detection [5] to the SAD task, evaluating on the challenging data from the FSC.

Classical SAD algorithms based on statistical signal processing principles typically rely on the non-stationary nature of speech and make assumptions about the energy distribution of the short-time Fourier transform of the signal [6]. Such energy-based SADs tend to fail if the distractors violate the statistical assumptions such as stationarity. More robust SAD could be achieved in the modulation domain [7, 8] because the modulation spectra of speech are highly localized [9]. Modulation-based methods are therefore more robust to additive noise and reverberation compared to other energy-based methods [7]. Our development of this system is informed by our experiences with voice-type discrimination (VTD) [5], which attempts to discriminate between live speech in a room and speech from other sources that are not live. In essence, VTD combines elements of SAD and detection of human presence, and we found the contributions of modulation features and spectral features to be complementary in this task.

More recent data-driven methods using deep neural networks (DNNs) have advanced SAD performance [10]. These systems typically make use of popular deep learning devices such as convolutional neural networks (CNNs) to extract discriminative patterns of speech in the time-frequency domain [11] or the learning of long-term temporal patterns using recurrent structures such as long short-term memory (LSTM) [10], or both of the above [12].

The development of spectro-temporal receptive fields (STRFs) has been ongoing for decades. Motivated by structures that have been observed in the brainstem and the central auditory system [13], STRFs respond to a range of patterns of temporal modulation and spectral modulation [14]. Multiple research groups have introduced differing mathematical formulations of the STRFs [15, 16, 17] and have applied them to speech processing tasks such as automatic speech recognition and speech activity detection (*e.g.* [7]). In the SAD task, modulation features have shown to be especially robust to environmental degradation (*e.g.* [7, 8]). Nevertheless, since the range of potential two-dimensional spectro-temporal modulation patterns that are helpful for various tasks could be quite large, substantial effort has been devoted to finding the best STRF parameters through dimensionality-reduction techniques (*e.g.* [7]). These dimensionality-reduction processes are typically independent of the application task and therefore they could be sub-optimal. The STRFNet system that we developed for VTD is one of the few deep learning systems with adaptable parameters that correspond to spectral-temporal filter shapes (similar models include [18], [19], etc.).

In this paper, we describe the SAD system that we used to participate in the SAD task of FSC P3. Our system is a variation of the STRFNet that is constrained to extract spectro-temporal modulation features in the first layer of the network [5]. We show that an unconstrained network with a similar formulation was able to achieve first place in the Challenge. Although not reflected in the Challenge results, we show that the hybrid STRFNet system is more robust in low SNR conditions than our baseline. From the analysis of the learned STRF parameters, we confirmed the importance of temporal modulation around the nominal speaking rate of English speakers.

Organization of this paper. In the next section, we introduce the Hybrid-STRFNet SAD system that placed first in the official FSC P3 SAD Challenge. We then describe the evalua-

http://dx.doi.org/10.21437/Interspeech.2021-651

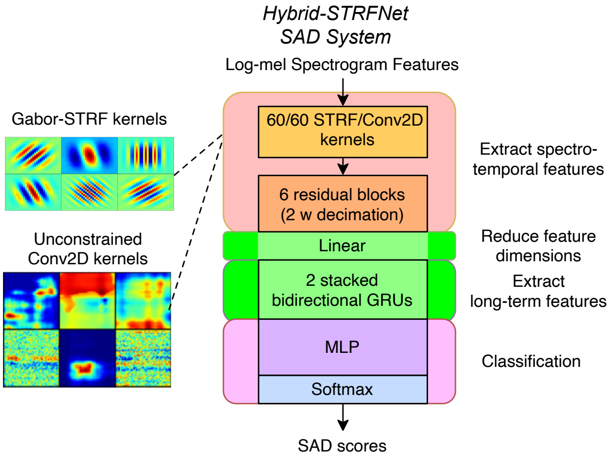
Hybrid-STRFNet
SAD System
Log-mel Spectrogram Features
Gabor-STRF kernels
Unconstrained
Conv2D kernels
60/60 STRF/Conv2D kernels
6 residual blocks (2 w decimation)
Linear
2 stacked bidirectional GRUs
MLP
Softmax
SAD scores
Extract spectro-temporal features
Reduce feature dimensions
Extract long-term features
Classification

Figure 1: *Block diagram of the Hybrid-STRFNet SAD system that includes both learnable STRFs and unconstrained 2D convolutional kernels (Conv2D) in the first layer.*

tion procedure using the provided Challenge data. Finally, we discuss our experimental results.

2. Hybrid-STRFNet SAD System

This section describes the Hybrid-STRFNet SAD system that we developed for the FSC P3 SAD Challenge. We first extract spectro-temporal features using both STRF and standard convolutional kernels. Then the extracted features are used as input to recurrent layers to learn long temporal patterns and finally a multi-layer perceptron layer predicts speech activity probabilities. Figure 1 illustrates the overall organization of the Hybrid-STRFNet SAD system.

2.1. Feature extraction and data augmentation

We used the standard log-mel spectrogram as input features to our system. The spectrogram was obtained using a 20-ms Hamming window with 50% overlap between frames and 512-point discrete Fourier transform. 80 mel channels were calculated, and power compression using the natural logarithm was applied subsequently. To increase robustness to unseen testing conditions and to prevent overfitting on the training set, we applied the SpecAugment method [20] to the input features during training. Although SpecAugment was initially developed to improve robustness in automatic speech recognition, it has improved robustness in other tasks such as spoofing detection [5]. Since SAD requires output predictions at a fine time resolution, we randomly masked out frequency channels only and did not mask time frames or apply time warping.

2.2. Speech activity detection system

The Hybrid-STRFNet system we developed for SAD closely resembles the STRFNet system we recently developed for voice type discrimination and for spoof detection (*i.e.* distinguishing live speech from machine-generated speech) [5]. The SAD system consists of three stages based on deep learning principles and learns to predict speech activity probabilities from the input features.

The first stage consists of processing the input log-mel spectrogram features with 2-D convolutional kernels to extract spectro-temporal patterns. Previous studies have shown that processing the input features with convolutional layers preceding recurrent layers provides substantial improvements in SAD [12]. Rather than using only unconstrained 2-D convolutional kernels, we constrain half of the kernels in the first layer to be in the shape of STRFs that are parameterized by modulation frequencies in the time and frequency domain, respectively [16]. The use of hybrid CNN and STRF kernels was motivated by our previous study which showed that including both learnable STRF kernels and convolutional kernels improved robustness in spoofing detection [5].

We follow the implementation similar to [16] and define the Gabor-based STRF kernel as the product between a complex sinusoidal function, $s[n,k;\omega_n,\omega_k]$, and a hanning envelope, $h[n,k]$, as follows.

$$h[n,k] = (.5 - .5\cos(\tfrac{2\pi(n+1)}{W_n+1}))(.5 - .5\cos(\tfrac{2\pi(k+1)}{W_k+1})) \quad (1)$$

$$s[n,k;\omega_n,\omega_k] = \exp[j\omega_n(n-n_0) + j\omega_k(k-k_0)] \quad (2)$$

$$\mathrm{STRF}[n,k;\omega_n,\omega_k] = s[n,k;\omega_n,\omega_k]\cdot h[n,k] \quad (3)$$

where n_0 and k_0 is the center index of the kernel in time and frequency, respectively, and W_n and W_k is the window length in time and frequency, respectively.

Each Gabor-based STRF kernel has two parameters that allow the kernel to be tuned to specific spectro-temporal modulation frequencies. The two modulation parameters, ω_n and ω_k, that control the temporal modulation (rate) and spectral modulation (scale), respectively, are adapted with gradient backpropagation during training. Compared to the STRF [15] that we used in the original STRFNet system [5], the Gabor-based STRF is computationally more efficient and has half as many learnable parameters. We found empirically that the Gabor-based STRF is sufficient for the SAD task. The other half of the kernels in the first stage are unconstrained 2-D convolutional kernels. In total, there are 60 2-D convolutional kernels in the first layer. Half of the kernels are unconstrained 2-D kernels size (5, 5) and the other half are learnable STRF kernels. The learnable STRF kernels span 300 milliseconds in time and 30 mel bins, resulting in a kernel size of (30, 30). Following the initial convolutional layer are six residual blocks [21] to further enhance the spectro-temporal features. The residual blocks downsample the initial frame rate of the input features by 4, resulting in a 40-ms output resolution.

Following the convolutional stage that extracts spectro-temporal features, we flattened each time frame's features by reshaping the features to a single vector before passing them through a fully connected layer to reduce the feature dimensionality. Next, we passed those features through two stacked bidirectional gated recurrent units (GRUs) [22]. Recurrent neural networks have shown tremendous sequential modeling capabilities and have been widely used for SAD [10, 11, 12]. The main difference between our SAD system and our recently-developed STRFNet system is that we omit the self-attentive pooling layer originally used to compress the time dimension after the GRU layer. In this specific Challenge, the system must estimate speech activity at a much finer time resolution than the VTD task; therefore, we do not reduce the time dimension.

Finally, each GRU's output is passed through a one-hidden-layer multi-layer perception (MLP) with two dimensions for the output. The output scores are normalized using the softmax function to obtain estimated posterior probabilities of speech for every 40-ms segment of input. In total, the network contains roughly 280 thousand learnable parameters. Specific hyperparameters and the model implementation can be found at `http://www.cs.cmu.edu/~robust/code.html`.

3. Experimental Setup

3.1. Datasets

We evaluated our system using the Fearless Steps Challenge Phase-03 dataset hosted by the University of Texas at Dallas Center for Robust Speech Systems [1]. The dataset contains audio from both the Apollo 11 and Apollo 13 mission. Each file is about 30 minutes long and is sampled at 8 kHz. This dataset is challenging due to the channel variability and varying SNR estimated to be between 0-20 dB throughout each file's duration. Additionally, some speech regions only last 0.5 seconds, with as many as 15 speakers talking within 10 seconds [3]. We were provided with approximately 63 hours of training data and 15 hours of development data from the Apollo 11 mission. The systems were evaluated on 34 hours of blind test data which includes data from additional unseen channels and unseen Apollo 13 mission that was not in previous challenges. To further study our system's robustness to additive noise, we performed an internal evaluation where we added additional noise at various SNRs to the development data. The additional noise recordings were obtained from the Deep Noise Suppression (DNS) dataset [23].

3.2. Training and evaluation procedure

To train our SAD systems, we randomly selected 30-second audio segments from the training data and computed the log-mel spectrogram. As mentioned in the previous section, we applied SpecAugment to the input spectrogram to improve generalization. Specifically, we randomly masked out different amounts of frequency channels for each training sample. The system was optimized using the negative log-likelihood loss function between the output predictions and target distribution. Each STRF kernel's rates and scales were randomly initialized from uniform distributions of $[0, 25)$ Hz and $[-.25, 0.25)$ cycles per channel, respectively, and has a time support of 300 milliseconds and spans 30 mel channels. All SAD systems were trained using the AdamW optimizer [24] with a learning rate of $5e^{-4}$, a batch size of 20, and a weight decay of .01 in PyTorch. We trained each system for 50 epochs using a single GeForce GTX 1080 Ti GPU card with 12 GB memory and used the best performing system on the development set for the official evaluation.

The Fearless Steps Phase-03 Challenge uses the Decision Cost Function (DCF) as the primary metric. The DCF is a weighted average of the probability of a false alarm and probability of a miss, with misses weighted three times as much as false alarms. Since our system predicts a speech probability for every 40-millisecond segment of audio, we threshold the probabilities to make a binary decision. We use a threshold of .05, which we found to be the optimal value on the development set. We did not apply any additional postprocessing to generate the predicted speech activity timestamps for the challenge evaluation.

In addition to the official challenge evaluation, we carried out a supplementary study to evaluate the robustness to additional additive noise at various SNRs. A variety of noise signals were added from the DNS dataset at SNRs ranging from -6 to 30 dB. Recognizing that the original data already contains noise, the true SNR of the corrupted speech will be lower. Since the evaluation labels were not provided, testing the robustness to additive noise was carried out using the development data. We used the Equal Error Rate (EER) metric for these studies since the EER is a reliable indicator of raw performance. These experiments did not require any additional training because we were mainly interested in the robustness of the systems trained on the original training set. It takes on average 287 seconds to process a 30-minute audio file on a single CPU, resulting in a real-time factor of .16.

Table 1: *Comparison of the DCF scores on the development set without and with SpecAugment.*

System	SpecAugment? No/Yes
CNN-RNN Baseline	.0104 / .0072
Hybrid-STRFNet	.0100 / .0074

Table 2: *Comparison of the DCF scores on the Fearless Steps Phase-03 SAD Challenge development and evaluation sets.*

System	Dev DCF	Eval DCF
Challenge Baseline	.125	.156
Second Place Team	—	.0192
CNN-RNN Baseline	.0072	.0151
Hybrid-STRFNet	.0074	**.0147**

3.3. Baseline systems

The organizers provided an official Challenge baseline system based on the Combo-SAD system developed for the DARPA RATS evaluation [25, 26]. The baseline system consists of Gaussian Mixture Models (GMMs) and uses four different voicing measures and spectral flux as features. We developed a second baseline in addition to the provided baseline. To evaluate the benefits of using both learnable Gabor-based STRF kernels and unconstrained 2D kernels in the first layer, we developed a CNN-RNN system that contains only unconstrained 2-D kernels in the first layer. To make the number of kernels equal, the CNN-RNN baseline system doubled the amount of unconstrained 2-D kernels in the first layer. The rest of the network architecture is identical to the Hybrid-STRFNet SAD system and was trained using the same procedure.

4. Experimental Results and Discussion

In this section, we discuss the results of the official Fearless Steps Phase-03 SAD Challenge and our internal evaluation to study the system's robustness to additive noise. Additionally, we analyze the final distributions of the learned STRF parameters.

4.1. Fearless Steps Phase-03 Challenge Evaluation Results

Table 2 summarizes the results on the development and blind evaluation set for the Challenge. As shown in Table 1, applying SpecAugment during training provides improvements of up to 30% relative DCF on the development set. This simple data augmentation method helped reduce overfitting on the training set and was used in all our further experiments. Our Hybrid-STRFNet system achieved first place in the official challenge, outperforming the second-place system and the official challenge baseline on the evaluation set by 23% and 90% relative

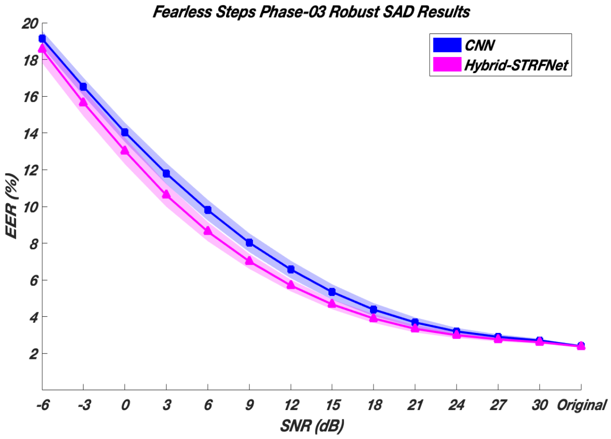

Figure 2: *Equal error rates when additional noise is added to the development set*

DCF, respectively. The CNN-RNN baseline system that we developed achieved similar performance to our Hybrid-STRFNet system; however, in the following section, we illustrate the advantages of the Gabor-based STRF kernel when additive noise is present.

4.2. Robustness to additive noise

The performance of our CNN-RNN baseline system and Hybrid-STRFNet system when noise is added to the development set is shown in Figure 2. Each system was trained and tested seven times. The solid line depicts the average EER, and the shaded regions depict the standard deviation. When testing on the original development data, the two systems' performance is almost identical, similar to the results in the previous section. However, when noise is added, the Hybrid-STRFNet system outperforms the baseline CNN-RNN system at all SNRs and achieved up to 13% relative improvement in EER. This result is consistent with the results of our previous study [5] and further suggests that STRF kernels provide additional robustness compared to standard convolutional kernels.

We believe that the additional robustness can be attributed to STRF kernels being constrained to analyze the spectro-temporal modulations, while standard convolutional kernels are unconstrained and free to pick up any cues from the input. Extracting modulation information is useful because speech and noise are more separable in the modulation domain due to the unique characteristics of the two [7]. Nevertheless, based on our initial experiments, the use of STRF kernels alone degrades the performance. This indicates that standard convolutional kernels are still necessary. The use of STRF kernels by themselves is not sufficient since the kernels span 300 milliseconds to analyze medium-time modulation patterns and the time resolution required for SAD is much smaller.

4.3. Learned STRF parameter distributions

A benefit of having learnable STRF kernels is that they are more interpretable compared to standard convolutional kernels. Additionally, the learnable component allows us to not have to determine the optimal rates and scales for each specific task beforehand since the parameters are adapted during training. Figure 3 depicts the rates and scales that were learned by our Hybrid-STRFNet SAD system after training. We display the distribution of the absolute value of the learned rates and scales because the sign only determines the drifting direction of each STRF. Although the rates and scales were randomly initialized from uniform distributions of $[0, 25)$ Hz and $[-.25, 0.25)$ cycles per channel, respectively, they converged to a non-uniform distribution. Specifically, the rates distribution has a peak around 4-5 Hz, which is known to be the rate corresponding to the syllabic rate of English [27, 28]. Previous studies have shown that temporal modulations around 2 Hz to 8 Hz are important for robust speech recognition [29] and filtering a signal with a narrow bandpass filter centered at 4 Hz results in a more noise-robust representation [30]. We found that the learned scale distribution converges to a distribution with a peak at slow spectral modulations. Analyzing slow spectral modulation is similar to keeping the lower order coefficients after computing Mel Frequency Cepstral Coefficients, which is often done for feature extraction. Interestingly, learned STRFs were hardly ever in the form of purely temporal or purely spectral filters.

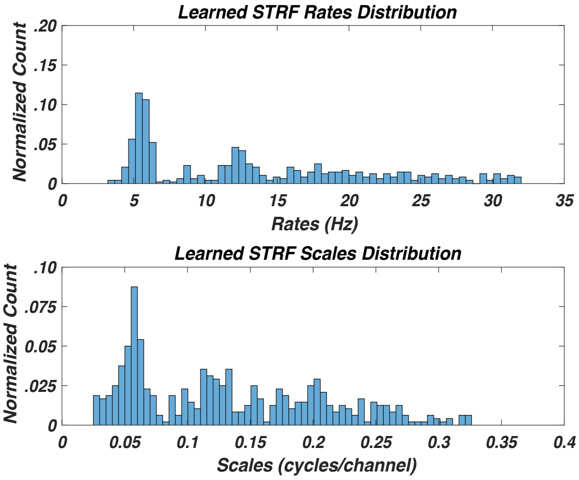

Figure 3: *Distribution of the learned rates(top) and learned scales(bottom) after training*

5. Conclusions

In this paper we describe the state-of-the-art SAD system that CMU developed for the UTD-CRSS Fearless Steps Phase-03 Challenge. While the SAD system follows a popular CNN-RNN architecture, a key distinction is that the initial convolutional layer includes both learnable Gabor-based STRF kernels and traditional unconstrained convolutional kernels. In addition to achieving the best performing score in the Challenge, we demonstrated that the inclusion of both STRF and CNN kernels in the first layer makes the SAD system more robust to additive noise. Similar to the results of other studies, we found that utilizing SpecAugment during training improves the generalization capabilities for the SAD task. In the future we intend to develop a real-time SAD system based on spectro-temporal receptive fields. Additionally, we will develop a learnable front-end representation that is more suitable for STRFs.

6. Acknowledgement

This work was funded by the Johns Hopkins University Applied Physics Laboratory.

7. References

[1] A. Joglekar, J. Hansen, M. Shekar, and A. Sangwan, "FEARLESS STEPS Challenge (FS-2): Supervised Learning with Massive Naturalistic Apollo Data," in *INTERSPEECH*. ISCA, Oct. 2020, pp. 2617–2621.

[2] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio," in *INTERSPEECH*. ISCA, sept 2019, pp. 1851–1855.

[3] 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," in *Proc. Interspeech 2018*, 2018, pp. 2758–2762.

[4] C. Richey, Z. Armstrong, and A. Lawson, "Distant microphone conversational speech in noisy environments," SRI International, Tech. Rep., 2019.

[5] T. Vuong, Y. Xia, and R. M. Stern, "Learnable spectro-temporal receptive fields for robust voice type discrimination," in *INTERSPEECH*. ISCA, Oct. 2020, pp. 1957–1961.

[6] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 1998, pp. 365–368.

[7] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[8] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, 2015.

[9] M. Elhilali, *Modulation Representations for Speech and Music*. Cham: Springer International Publishing, 2019, pp. 335–359.

[10] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 483–487.

[11] J. Heitkaemper, J. Schmalenstroeer, and R. Haeb-Umbach, "Statistical and neural network based speech activity detection in non-stationary acoustic environments," in *INTERSPEECH*. ISCA, Oct. 2020, pp. 2597–2601.

[12] Q. Lin and M. L. Tingle Li, "The DKU speech activity detection and speaker identification systems for fearless steps challenge phase-02," in *INTERSPEECH*. ISCA, Oct. 2020, pp. 2607–2611.

[13] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 382–395, 1995.

[14] D. L. Wang, "Auditory stream segregation based on oscillatory correlation," in *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing*, Ermioni, Greece, September 1994, pp. 624–632.

[15] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[16] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

[17] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 416–422, 2002.

[18] E. Bavu, A. Ramamonjy, H. Pujol, and A. Garcia, "TimeScaleNet: A multiresolution approach for raw audio recognition using learnable biquadratic IIR filters and residual networks of depthwise-separable one-dimensional atrous convolutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 220–235, 2019.

[19] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *IEEE Workshop on Spoken Language Technology*, 2017, pp. 1021–1028.

[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple augmentation method for automatic speech recognition," in *INTERSPEECH*. ISCA, Sep. 2019, pp. 2613–2617.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[23] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *INTERSPEECH*. ISCA, Oct. 2020, pp. 2492–2496.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, May 2019.

[25] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[26] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, and D. W. Oard, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions," in *INTERSPEECH*. ISCA, Sep. 2014, pp. 1544–1548.

[27] T. Houtgast, H. J. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," *Acta Acustica united with Acustica*, vol. 46, no. 1, pp. 60–72, 1980.

[28] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *International Conference on Spoken Language Processing (ICSLP)*, vol. 96. Citeseer, 1996, pp. 24–27.

[29] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[30] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, 1998.