DeepComboSAD: Spectro-Temporal Correlation Based Speech Activity Detection for Naturalistic Audio Streams

Aditya Joglekar, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Speech activity detection (SAD) serves as a crucial front-end system to several downstream Speech and Language Technology (SLT) tasks such as speaker diarization, speaker identification, and speech recognition. Recent years have seen deep learning (DL)-based SAD systems designed to improve robustness against static background noise and interfering speakers. However, SAD performance can be severely limited for conversations recorded in naturalistic environments due to dynamic acoustic scenarios and previously unseen non-speech artifacts. In this letter, we propose an end-to-end deep learning framework designed to be robust to time-varying noise profiles observed in naturalistic audio. We develop a novel SAD solution for the UTDallas Fearless Steps Apollo corpus based on NASA's Apollo missions. The proposed system leverages spectro-temporal correlations with a threshold optimization mechanism to adjust to acoustic variabilities across multiple channels and missions. This system is trained and evaluated on the Fearless Steps Challenge (FSC) corpus (a subset of the Apollo corpus). Experimental results indicate a high degree of adaptability to out-of-domain data, achieving a relative Detection Cost Function (DCF) performance improvement of over 50% compared to the previous FSC baselines and state-of-the-art (SOTA) SAD systems. The proposed model also outperforms the most recent DL-based SOTA systems from FSC Phase-4. Ablation analysis is conducted to confirm the efficacy of the proposed spectro-temporal features.**

***Index Terms*—Fearless steps challenge (FSC), NASA Apollo missions audio, spectro-temporal correlations, speech activity detection (SAD).**

I. Introduction

In recent years, there has been an increased use of spoken language in human-machine interactive systems, encompassing a broad spectrum of speech and language technology (SLT) tasks. Speech recorded in naturalistic scenarios often includes environmental distortions, overlapping speech from other speakers, and extended periods of silence. To optimize the performance of deep learning (DL)-driven SLT solutions, it is imperative that these systems accurately identify active speech regions across entire audio streams. Thus, most speech processing pipelines use speech activity detection (SAD) as a core front-end scheme to extract speech regions.

SAD is a two-step process; i) feature extraction, followed by ii) binary classification employed to differentiate between speech and non-speech segments. Several supervised and unsupervised SAD solutions have been proposed over the years. Initial SAD mechanisms employed threshold-based classification on time-domain acoustic features such as energy, zero-crossing rate, and frame-wise pitch estimations [1], [2]. Robustness to environmental distortions was later pursued by leveraging statistical modeling techniques to temporal, spectral, and cepstral speech features [3], [4]. Subsequently, gaussian mixture models (GMM) and similar cluster schemes emerged as effective strategies to model log-energy distributions in speech features [5], [6], [7]. In alignment with this, Combo-SAD [8] was designed to linearly transform spectro-temporal features into a 1-dim feature space through principal component analysis (PCA), followed by GMM classification. This letter was extended by TO-Combo-SAD [9] which employed a threshold optimization strategy to mitigate speech density variability issues in Apollo audio streams. To address noisy speech profiles, unsupervised strategies including rVAD and GammatoneSAD [10], [11] also incorporated speech enhancement mechanisms prior to classification.

Recent advancements in DL have introduced convolutional neural networks (CNN) and recurrent neural networks (RNN) for spatial and temporal modeling in SAD design [12], [13], [14], [15]. Spatio-temporal modeling capabilities of time-distributed feed-forward (FFN) CNNs have demonstrated promising outcomes [12], [13]. These approaches, although showcasing superior performance under in-domain (IN) noisy conditions, have exhibited limitations when applied to naturalistic data with unseen acoustic characteristics [16], [17], [18], [19], [20], [21]. The CRSS-UTDallas Fearless Steps (FS) Apollo audio is such a +150k-hr collection of time-synchronized multi-speaker communications recorded in varying noise types. The FS Challenge (FSC) corpora [22], [23], [24] utilize this data, incorporating audio with out-of-domain (OOD) characteristics in FSC Phases 3 & 4 (referred to as "FSC-P#" throughout this letter for brevity). Consequently, datasets released for FSC-P3 & FSC-P4 have motivated generalizable system development. Drawing inspiration from 'Combo-SAD' & 'TO-Combo-SAD' system designs [8], [9], this letter introduces the 'DeepComboSAD' model. This framework seeks to attain generalizability through learnable spectro-temporal filters and threshold optimization.

Manuscript received 14 June 2023; revised 8 September 2023; accepted 8 September 2023. Date of publication 26 September 2023; date of current version 24 October 2023. This work was supported by NSF-CISE CCRI Community Resource Program under Grant 2016725, and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Keshet. *(Corresponding author: John H. L. Hansen.)*

The authors are with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: aditya.joglekar@utdallas.edu; john.hansen@utdallas.edu).

Digital Object Identifier 10.1109/LSP.2023.3319229

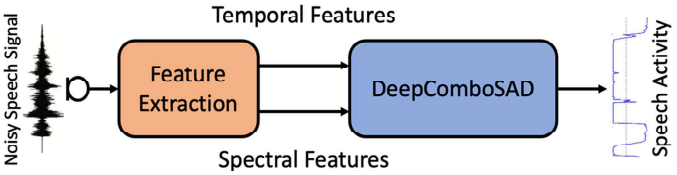

Fig. 1. Overview diagram of the DeepComboSAD end-to-end process.

II. System Formulation

The proposed system overview is depicted in Fig. 1. Similar to Combo-SAD, the proposed framework extracts voicing measures and spectral features from time-domain (Section II-A1) and STFT-domain (Section II-A2) for a given utterance. These inputs are processed through convolutional attention layers to derive "Deep-Combo" features which then aid in estimating speech activity. The goal is to achieve robustness against adverse distortions induced during speech capture by learning from temporal, complex-spectral, and time-frequency (TF) context using self-attention mechanisms.

A. Deep-Combo Feature Extraction

We initially processes a mono audio segment $x(t) \in \mathbb{R}^{1\times M}$ by splitting into $x_f(n) \in \mathbb{R}^{N\times T}$ frames to extract spectro-temporal features. Here, 'M', 'T', 'n', 'N' represent segment length (4.82 sec), no. of frames in a segment, frame index, & window size (32 ms with 12 ms skip rate) respectively.

1) Temporal Features: Voicing measures such as energy, harmonicity, clarity, and linear predictive coding (LPC) are usually obtained from normalized auto-correlations in time [25], [26], [27]. We compute such correlations across time samples within each frame. $r_{xx}(n,k) \in \mathbb{R}^{L\times T}$. The normalized auto-correlation matrix is computed in (1) as:

$$r_{xx}(n,k) = \frac{\sum_{j=0}^{N-1} (x_f(j)w(j)) \cdot (x_f(j+k)w(j+k))}{\sum_{j=0}^{N-1} w(j)w(j+k)} \quad (1)$$

where '$w(j)$' is square root of the Hanning window, & 'k' is auto-correlation lag index. For each frame in $x_f(n)$, the first 'L' positive and negative auto-correlation lags are considered since normalizing with a window function mitigates the impact of strong correlation peaks, thus obviating need for low-pass filtering. [28]. In addition, the frame-wise auditory features i) Log-energies $e(n) \in \mathbb{R}^{1\times T}$ (to measure signal loudness), ii) peak-to-valley ratio $p(n) \in \mathbb{R}^{1\times T}$ (to detect abrupt peaks/drops in power), and iii) first-order difference $d(n) \in \mathbb{R}^{1\times T}$ (to assist with detecting unvoiced speech frames by preserving high-frequency content), are also computed.

$$x_{\text{temporal}}(n) = [r_{xx}(n,k),\ e(n),\ p(n),\ d(n)] \quad (2)$$

The temporal input feature set $x_{\text{temporal}}(n) \in \mathbb{R}^{(L+3)\times T}$ is formed by concatenating above mentioned auditory features with the auto-correlation matrix (detailed in (2)).

2) Spectral Features: The short-time fourier transform (STFT) of input '$x(t)$' is initially computed to produce a complex-valued spectrogram. Here, 1-dim convolutional layers perform a 256-point DFT operation using square root of Hanning window on frames after zero padding. Most SAD systems overlook phase information in spectral feature extraction, which is suitable for utterances degraded by static noise. However, recent speech enhancement efforts have emphasized the importance of phase-sensitive information for speech severely degraded by dynamic noise [29]. Hence, we employ complex-valued spectrogram as a core spectral feature input.

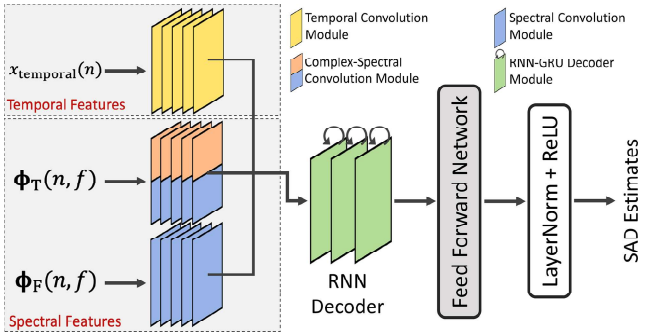

Fig. 2. Schematic representation of the proposed DeepComboSAD network. Three input features with dedicated convolutional processing modules *(left)* are: temporal *(yellow)*, spectral *(blue)*, and complex-spectral *(blue-orange)*. RNN-GRU decoder *(green)* followed by Linear Projection layers generate SAD estimates from stacked spectro-temporal feature embedding inputs.

$$\mathbf{X_t}(n,f) = [X(n,f), X(n-1,f), ..X(n-N_d+1,f)]^T \quad (3)$$

Inter-frame & intra-frame correlations similar to (1) are computed in spectral domain by defining an N_t-dim speech vector $\mathbf{X_t}(n,f)$ (depicted in (3)) with freq. bin index 'f'.

$$\mathbf{\Phi_T}(n,f) = \text{LayerNorm}(\mathbf{X_t}(n,f)\mathbf{X_t}(n,f)^H). \quad (4)$$

$\mathbf{X_t}(n,f)$ is used to compute a '$N_t \times N_t$'-dim complex correlation matrix $\mathbf{\Phi_T}(n,f)$ as shown in (4).

Unlike conventional correlation operations, layer normalization with learnable affine transformations normalize the inter & intra-frame correlations, thereby mitigating computational problems during prolonged silences in recordings. These complex-valued correlations are converted to real-valued vectors by concatenating real and imaginary values along the feature dimension, and for efficiency, we only retain the lower half of the frequencies, reducing memory and computational demands. Similar correlation computations across frequency dimensions yield the '$N_f \times N_f$' time-varying correlation matrix $\mathbf{\Phi_F}(n,f)$, capturing impact of dominant speech formant regions, as described in (5) below,

$$\mathbf{\Phi_F}(n,f) = \text{LayerNorm}(\mathbf{X_f}(n,f)\mathbf{X_f}(n,f)^H) \quad (5)$$

where $\mathbf{X_f}(n,f)$ is a N_f-dim speech vector which consists of the current and past N_f-1 frequency bin information.

The DeepComboSAD network (detailed in Fig. 2) individually processes inputs from the "Deep-Combo" feature set $[x_{\text{temporal}}(n), \mathbf{\Phi_T}(n,f), \mathbf{\Phi_F}(n,f)]$. This spectro-temporal feature extraction approach paired with nonlinear modeling capabilities of DL systems seeks to enhance SAD performance for adverse noise distortions in naturalistic settings.

B. DeepComboSAD Network

The DeepComboSAD network is formulated using recurrent (RNN) and feed-forward networks (FFN) that incorporate dedicated feature-processing convolutional modules as detailed in the schematic representation (Fig. 2). Though the RNN & FFN layers are shared by temporal, complex-spectral, & spectral modules, separate convolutional layers are used to consolidate information across the Deep-Combo features.

The temporal convolution module utilizes 1-dim convolutional layers with (kernel, stride, padding) set to (5,1,2), followed by a 1-dim batch normalization & leaky rectified linear unit (ReLU) non-linearity. Three temporal convolution modules are stacked sequentially to extract salient temporal information

TABLE I
DCF PERFORMANCE (%) OF SAD SYSTEMS ON THE FSC-P4 DEV, EVAL SETS, AND EVAL SUB-SETS

SAD Systems	Dev % DCF ↓	Eval % DCF ↓	A11 (IN)	A11 (OOD-C)	A08 (OOD-M)	A13 (OOD-M)
rVAD	13.19	16.68	15.99	17.49	11.99	24.03
Combo-SAD	*8.25*	*11.24*	*4.32*	*19.69*	*11.00*	*17.11*
TO-Combo-SAD	13.77	15.63	11.16	17.71	21.17	18.52
CL-CNN	10.13	11.89	8.68	15.24	10.85	16.71
U-Net-SpecAug	*3.77*	*7.89*	*4.86*	*9.19*	*10.38*	*11.68*
ACAM-logMel	5.82	10.49	8.87	13.92	12.04	7.35
ACAM-MRCG	4.69	9.21	7.55	12.48	10.14	*7.2*
STAM-logMel	2.78	5.71	3.84	7.94	4.33	9.0
TFMR-MelFB	2.01	4.28	2.30	*2.72*	5.09	12.09
CRNN-fusion	-	4.12	2.34	*3.34*	6.74	7.69
STRF-logMel	-	3.72	2.07	*2.39*	6.97	*7.22*
DeepComboSAD	***1.77***	***3.42***	***1.96***	***3.43***	***3.79***	***7.34***

Systems above the dashed line are unsupervised, whereas those below are supervised. For this study, COMBO-SAD and U-Net-SpecAug serve as the unsupervised and supervised baseline models, respectively. Abbreviations of domain types for each eval sub-set: **IN** (in-domain), **OOD** (out-of-domain), **OOD-C** (sourced from an unseen apollo-11 channel), and **OOD-M** (originating from a different apollo mission).
Underlined values indicate baseline system results, and boldface indicates proposed system results."

across frames. Output filters are set to 256 except for the last module configured to 128. The spectral & complex-spectral convolution module architecture is identical, comprising of 2-dim convolutional layers with (kernel, stride, padding) configured to $((5 \times 5), (2 \times 1), (0 \times 2))$, followed by a time-frequency self-attention (TF-SA) network [30], [31], [32], batch normalization, & leaky ReLU. TF-SA layer computes a 2-dim attention map, enabling the network to model speech energy distributions along time and frequency dimensions. Correlation features formulated in (4) and (5) are transformed through five sequential complex-spectral/spectral convolutional modules. Output filters for the complex-spectral modules with $\mathbf{\Phi_T}$ input are set to $\{56, 84, 84, 112, 112, 128\}$, while filters for spectral modules processing $\mathbf{\Phi_F}$ are configured to $\{30, 45, 45, 60, 60, 128\}$.

Encoded context from the spectro-temporal modules is then concatenated feature-wise to produce a 384-dim speech embedding per frame. This embedding is processed through a three-layered uni-directional gated recurrent unit (GRU) with a 256 hidden size. RNN-GRU analyzes frame-wise similarities to produce 256-dim contextual embedding outputs. The FFN block transforms RNN-GRU outputs to frame-wise speech presence probabilities for each successive segment (as shown in Fig. 2). Two linear layers followed by ReLU and Dropout *($p = 0.3$)* form the initial FFN block. To achieve training stability, layer normalization and ReLU are employed in the last FFN layer. Finally, the proposed network is trained in an end-to-end manner on 400-frame segments. Unlike traditional cross-entropy loss, weighted mean squared loss ($\mathcal{L}_{\text{WMSE}}$) addresses speech density variations in the data, & offsets the imbalance between speech/non-speech samples [24]. The proposed weighted mean squared loss is computed as follows:

$$\mathcal{L}_{\text{WMSE}} = \sum_{i=1}^{D} \alpha_i * (x_i - y_i)^2 \tag{6}$$

where $\alpha_i = 1.5$ if y_i is speech, else $\alpha_i = 1$, using which, we generate frame-wise {0,1} decisions. The last three modules, in tandem with the loss function, seek to discern a relationship between frame-wise speech density and contextual embeddings. This process learns an optimal scaling factor for the speech presence probability outputs against a static threshold, culminating in the threshold optimization strategy.

III. EXPERIMENTAL SETUP

A. Dataset

FSC-P4 data was curated to evaluate system robustness for seen and unseen acoustic conditions. To achieve this, audio from 5 main Apollo-11 *(A11)* channels [16] were reserved for Train (70-hr) & Dev (20-hr) sets, as well as the Eval in-domain (IN) sub-set. out-of-domain (OOD) Eval sub-set audio was sourced from an unseen *A11* channel, Apollo-8 *(A08)*, & Apollo-13 *(A13)* missions audio [22], [23], [24]. FSC-P4 comprises a 35-hr Eval set with equally weighted IN & OOD speech.

B. Baseline Comparison Systems

The proposed system is benchmarked against ten unsupervised & supervised systems (See Table I). The unsupervised systems i) *Combo-SAD* [8], ii) *TO-Combo-SAD* [9], & iii) *rVAD* [10] have previously exhibited strong performance on *A11* audio [16], [22]. Supervised DL-based systems iv) *CL-CNN*, a convolutional network trained with curriculum learning strategy [33], & v) *U-Net-SpecAug* [18], employing a U-Net architecture with Specaugment [34], were specifically designed for prior FSC datasets [23], [24]. To contrast the efficacy of our TF attention modules, SAD DL architectures utilizing attention mechanisms are also considered. These include vi) *ACAM* [35], utilizing temporal attention with *MRCG* (multiresolution cochleagram) features, & vii) *STAM* [36], processing log-Mel-Spectrogram *(logMel)* features through sequential spectral & temporal attention modules. Finally, the top-3 performers from FSC-P4, namely viii) *TFMR-MelFB*, with Mel-filterbank features and four-layered Transformer [30], ix) *CRNN-fusion* [20], a convolutional recurrent model with fused inputs from 1-dim & 2-dim convolutional modules, & x) *STRF-logMel* [21], a CNN decoder with adaptable spectro-temporal filters, are also included. DCF scores for ix) and x) are derived from FSC-P4 system submissions data [37], [38].

C. Training and Evaluation Procedure

Unsupervised system thresholds were determined using a greedy search. Except for *ACAM* & *STAM*, DL systems were trained on 400-frame segments with a 200-frame overlap for 50 epochs, utilizing the ADAM optimizer, 32 batch size, and a scheduler reducing the learning rate of $1e-4$ after three epochs

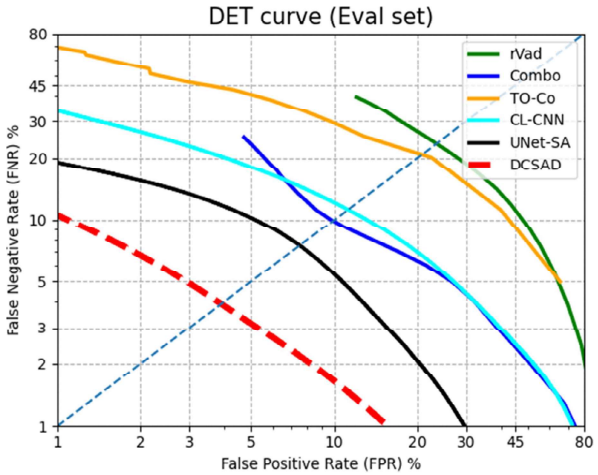

Fig. 3. DET curve computed over FSC-P4 Eval (IN & OOD) set includes rVad *(green)*, TO-Combo-SAD *(gold)*, Combo-SAD *(blue)*, CL-CNN *(light blue)*, U-Net-SpecAug *(black)*, and DeepComboSAD *(dashed red)*.

without improvement. Attention-frame selection and training procedure for *ACAM* & *STAM* adhered to the implementation detailed in [35], [36]. The optimal layer count & kernel sizes were determined using Dev set DCF scores. DL systems typically converged around the 32$^{\text{rd}}$ epoch. Best threshold values for systems were ascertained by analyzing the ROC curve on Dev set. DeepComboSAD decisions were calculated with threshold $\theta = 0.123$. All SAD systems were evaluated with the NIST-defined DCF measure [39] with 0.25 sec collar.

IV. Results and Discussion

Overall DCF performance for proposed and comparative systems is summarized in Table I. We report an absolute 7.8% reduction in DCF *(Eval)* over the unsupervised, & 4.47% over the supervised baseline system, with far superior performance demonstrated over *rVad*, *TO-Combo-SAD*, & *CL-CNN*. *ACAM* and *STAM*, designed for noise resilience, excel for OOD data but under perform significantly for in-domain (IN) audio, lagging behind DeepComboSAD by an absolute 6.8 and 2.3% DCF. *ACAM* fares better with MRCG than with logMel features, while *STAM-logMel*'s superior results can likely be attributed to the added spectral attention module. *TFMR-MelFB* and *CRNN-fusion* models perform relatively better, with 20% and 17% relative DCF degradation compared to DeepComboSAD. *STRF-logMel* exhibits similar performance for both IN & OOD data, akin to DeepComboSAD, given their use of learnable filters. Nonetheless, the *logMel* features in STRF don't explicitly capture temporal or phase-sensitive contexts, as confirmed by the ablation analysis (Section IV-A). Temporal & complex-spectral context incorporated in an end-to-end manner results in DeepComboSAD's 8% relative improvement over *STRF-logMel*. In the OOD Eval sub-sets, DeepComboSAD delivers SOTA results for *A08 (OOD-M)* & *A11 (IN)*. Although some systems show slight enhancements for *A11 (OOD-C)* & *A13 (OOD-M)*, their consistency across other sub-sets is lacking. A DET curve analysis for the Eval set (Fig. 3) highlights this DeepComboSAD robustness, with consistent performance across all thresholds. A relatively constant slope for Eval DET curves is also observed for DeepComboSAD. Additionally, predictions from DeepComboSAD trained with WMSE-loss skew towards a lower FNR as compared to other systems. This suggests a preference towards speech detection as opposed to noise suppression.

TABLE II
Ablation Analysis for DeepComboSAD Network, Including Comparisons With Log-Mel-Spectrogram (*logMel*) Features

Sub-Systems	Dev % DCF ↓	Eval % DCF ↓
TF (logMel)	*2.61*	*7.22*
T	1.81	6.59
F	1.51	5.43
TF	1.84	5.59
T-F	1.47	6.05
T-TF	1.52	4.24
F-TF	1.41	5.13
T-F-TF (logMel)	*1.81*	*6.08*
T-F-TF (*Ours*)	***1.77***	***3.42***

Abbreviation T denotes time-domain feature & their associated temporal convolutional (conv) module. F signifies complex-spectral feature & their conv module. TF represents spectral feature & their conv module. TF (*logMel*) implies the use of *logMel* features with spectral conv module. Any abbreviation containing a '-' suggests the employment of multiple features or modules. All models adopt a consistent decoder framework (GRU + FFN + layernorm) & WMSE loss.

Indented values indicate baseline ablation result, while boldface indicates proposed system result. Underlined values depict the best performing sub-systems for Dev & Eval sets."

A. Ablation Analysis

An ablation study was performed to evaluate the impact of different feature combinations within the DeepComboSAD network. This also includes a comparison using log-Mel-Spectrograms (*logMel*). All DCF results are shown in Table II. Notably, the (Dev & Eval) performance dropped considerably using only the *TF (logMel)* compared to other combinations. When comparing *T-F-TF* features (*logMel* & *Ours*), the proposed network benefits significantly from learnable *TF* filters, resulting in better generalization. Although the learnable features perform similarly on Dev (IN) set, *T-TF* domain combination yields vastly improved Eval results. While *T-F*, *T-TF*, and *F-TF* display variations in IN & OOD results, they collectively achieve Eval set performance that exceeds their individual contributions. This showcases the high degree of non-mutual information captured by the Deep-Combo features.

V. Conclusion

In this letter, we proposed a novel spectro-temporal correlation-based speech activity detector that leverages temporal features extracted from inter-frame & intra-frame correlations, combined with real and complex-valued spectral features to achieve robustness in dynamically varying naturalistic speech environments. We benchmarked our system's efficacy over the FSC-P4 dataset, demonstrating superior performance compared to all previously established state-of-the-art SAD solutions. The DCF, DET curve, & ablation analyses detailed in this research demonstrate that DeepComboSAD i) effectively extracts robust features from time-domain signals through its learnable filters, ii) is highly adaptable to time-varying acoustic characteristics, iii) sustains a notably low false negative rate under unseen noise profiles & iv) demonstrates consistent performance across both in-domain & out-of-domain scenarios. For our future work, we aim to extend this proposed letter by developing a jointly modeled SAD and speaker diarization system that employs temporal and complex-spectral schemes. This system is intended to be deployed in a pipeline to generate diarized transcripts for the +150k-hr audio archive of the twelve NASA Apollo Missions.

REFERENCES

[1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T recommendation G.729 AnnexB: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice & data applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 64–73, Sep. 1997.

[2] S. G. Tanyer and H. Ozer, "Voice activity detection in non-stationary noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000.

[3] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[4] I.-C. Yoo, H. Lim, and D. Yook, "Formant-based robust voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2238–2245, Dec. 2015.

[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[6] J. Wu and X.-L. Zhang, "Maximum margin clustering based statistical VAD with multiple observation compound feature," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 283–286, May 2011.

[7] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 3, pp. 217–231, Mar. 2001.

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

[9] A. Ziaei, L. Kaushik, A. Sangwan, J. H. L. Hansen, and D. W. Oard, "Speech activity detection for NASA Apollo space missions: Challenges and solutions," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1544–1548.

[10] Z.-H. Tan, A. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, 2020.

[11] V. Kothapally and J. H. Hansen, "Speech detection and enhancement using single microphone for distant speech applications in reverberant environments," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1948–1952.

[12] H. Soltau, G. Saon, and T. N. Sainath, "Joint training of convolutional and non-convolutional neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2014, pp. 5572–5576.

[13] S.-Y. Chang et al., "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2018, pp. 5549–5553.

[14] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7378–7382.

[15] G. Gelly and J.-L. Gauvain, "Optimization of RNN-Based speech activity detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 646–656, Mar. 2018.

[16] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2758–2762.

[17] J. H. L. Hansen, K. Hickman, N. Jones, H. Dubey, A. Sangwan, and V. Kothapally, "UTDallas-PLTL: Leveraging spoken language technology for assessment of communication based learning behavior in peer-led team learning," in *Proc. 6th Annu. Conf. Peer-Lead Team Learn.*, Chicago, IL, 2017, pp. 5–10.

[18] W. Wang et al., "The DKU-Duke-Lenovo system description for the fearless steps challenge phase III," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 1044–1048.

[19] T. Vuong, Y. Xia, and R. M. Stern, "The application of learnable STRF kernels to the 2021 fearless steps Phase-03 SAD challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4364–4368.

[20] P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Unsupervised representation learning for speech activity detection in the fearless steps challenge 2021," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4359–4363.

[21] T. Vuong, N. Madaan, R. Panda, and R. M. Stern, "Investigating the important temporal modulations for deep-learning-based speech activity detection," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2023, pp. 525–531.

[22] J. H. Hansen et al., "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1851–1855.

[23] A. Joglekar, J. H. Hansen, M. Chandra-Shekar, and A. Sangwan, "FEARLESS STEPS challenge (FS-2): Supervised learning with massive naturalistic Apollo data," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2617–2621.

[24] A. Joglekar, S. O. Sadjadi, M. C. Shekar, C. Cieri, and J. H. L. Hansen, "Fearless steps challenge Phase-3 (FSC P3): Advancing SLT for unseen channel and mission data across NASA Apollo audio," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 986–990.

[25] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. 11th Annu. Conf. Int. Speech Comm. Assoc.*, 2010, pp. 3118–3121.

[26] N. Esfandian, F. Jahani Bahnamiri, and S. Mavaddati, "Voice activity detection using clustering-based method in spectro-temporal features space," *J. AI Data Mining*, vol. 10, pp. 401–409, 2022.

[27] Q. Lin and Y. Shao, "A novel normalization method for autocorrelation function for pitch detection and for speech activity detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2097–2101.

[28] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 2010.

[29] V. Kothapally and J. H. Hansen, "SkipConvGAN: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1600–1613, 2022.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[31] V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang, "SkipConvNet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3935–3939.

[32] Q. Zhang, Q. Song, Z. Ni, A. Nicolson, and H. Li, "Time-frequency attention for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7852–7856.

[33] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Speech activity detection in naturalistic audio environments: Fearless steps Apollo corpus," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1290–1294, Sep. 2018.

[34] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2613–2617.

[35] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.

[36] Y. Lee, J. Min, D. K. Han, and H. Ko, "Spectro-temporal attention-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 27, pp. 131–135, 2020.

[37] "NIST fearless steps challenge Phase-03 2021 web-platform," 2021. Accessed: Mar. 01, 2021. [Online]. Available: https://sat.nist.gov/fsc3

[38] "The fearless steps challenge Phase-04 2022 website," 2022. Accessed: Mar. 01, 2022. [Online]. Available: https://fearless-steps.github.io/ChallengePhase4/

[39] F. R. Byers, J. G. Fiscus, S. O. Sadjadi, G. A. Sanders, and M. A. Przybocki, "Open speech analytic technologies pilot evaluation OpenSAT pilot," US Department of Commerce, National Institute of Standards and Technology, Multimodal Information Group, NIST, 2019.