Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection

Midia Yousefi, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Speech technology systems such as Automatic Speech Recognition (ASR), speaker diarization, speaker recognition, and speech synthesis have advanced significantly by the emergence of deep learning techniques. However, none of these voice-enabled systems perform well in natural environmental circumstances, specifically in situations where one or more potential interfering talkers are involved. Therefore, overlapping speech detection has become an important front-end triage step for speech technology applications. This is crucial for large-scale datsets where manual labeling in not possible. A block-based CNN architecture is proposed to address modeling overlapping speech in audio streams with frames as short as 25 ms. The proposed architecture is robust to both: (i) shifts in distribution of network activations due to the change in network parameters during training, (ii) local variations from the input features caused by feature extraction, environmental noise, or room interference. We also investigate the effect of alternate input features including spectral magnitude, MFCC, MFB, and pyknogram on both computational time and classification performance. Evaluation is performed on simulated overlapping speech signals based on the GRID corpus. The experimental results highlight the capability of the proposed system in detecting overlapping speech frames with 90.5% accuracy, 93.5% precision, 92.7% recall, and 92.8% Fscore on same gender overlapped speech. For opposite gender cases, the network scores exceed 95% in all the classification metrics.**

***Index Terms*—1-D CNN, binary classifier, co-channel speech detection, cocktail party problem, convolutional neural network, overlapping speech detection, residual learning, simultaneous speaker detection, speech modeling, speech separation.**

I. Introduction

THE cocktail party problem was first introduced by Colin Cherry in 1953 [1], which has triggered research in a range of areas that after almost 70 years are still active [2], [3]. A core part of the cocktail party phenomena are spontaneous conversations, where multiple speakers are talking. In [1], Cherry studied the human auditory system ability to selectively focus on one talker/conversation at a time, while ignoring interfering sources. This contributed to a theory for selective attention and early "filter" models for multi-speaker cocktail party communications [4]–[6]. The cocktail party problem as depicted in Fig. 1 is a psychoacoustic phenomena, which refers to the remarkable ability of the human auditory system to selectively attend, recognize and extract meaningful information from a complex auditory signal, where interfering sounds are produced by competing talkers [7]–[9].

Over the past decade, there has been increasing interest in formulating solutions with the same auditory capabilities as humans by employing both engineered signal processing disciplines and machine learning techniques. Almost all state-of-the-art speech technologies such as Automatic Speech Recognition (ASR), speaker diarization, speaker identification and speech synthesis systems operate effectively when the input is a clean single speaker signal. In contrast, these systems degrade rapidly in real world naturalistic scenarios, especially in existence of an interfering talker [10]–[12]. Therefore, detecting overlapping speech segments and extracting meaningful information from them remains a challenging task, and is an active field of research in both signal processing and machine learning communities.

Speech is a highly non-stationary signal consisting of a sequence of sounds called phonemes. The time variations of these sounds result in a very wide dynamic range of multiple frequency components as shown in Fig. 1. Therefore, time-frequency analysis techniques have been developed to study the frequency components of speech signal as a function of time. Short Time Fourier Transform (STFT) has historically been the most popular time-frequency analysis technique, which reveals important features/structures of the speech signal. Fundamental frequency, formants, Mel Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral flux, spectral density, spectral envelope, etc. are among important STFT-based spectral features, which characterize different aspects of speech, and could help advance overlapping speech research.

It is well known that the structure of a speech signal can be adversely degraded by the presence of a simultaneous talker. Therefore, the presence of a co-current talker can be identified by observing potential deviations in speech structure/features. Many studies have used manually designed features such as: Spectral Autocorrelation Peak Valley Ratio (SAPVR) [13], harmonicity [14], zero crossing rate, kurtosis [15], Non-negative Matrix Factorization (NMF)-based speaker-specific energy estimation [16], [17], Spectral analysis of frequency modulated

Manuscript received June 4, 2020; revised September 17, 2020; accepted October 10, 2020. Date of publication November 6, 2020; date of current version December 7, 2020. This work was supported in part by the National Science Foundation (NSF) under Grants # 1918032 and # 2016725, (PI: Hansen), and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. *(Corresponding author: John Hansen.)*

The authors are with the Center for Robust Speech Systems, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: Midia.Yousefi@utdallas.edu; John.Hansen@utdallas.edu).

Digital Object Identifier 10.1109/TASLP.2020.3036237

2329-9290 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

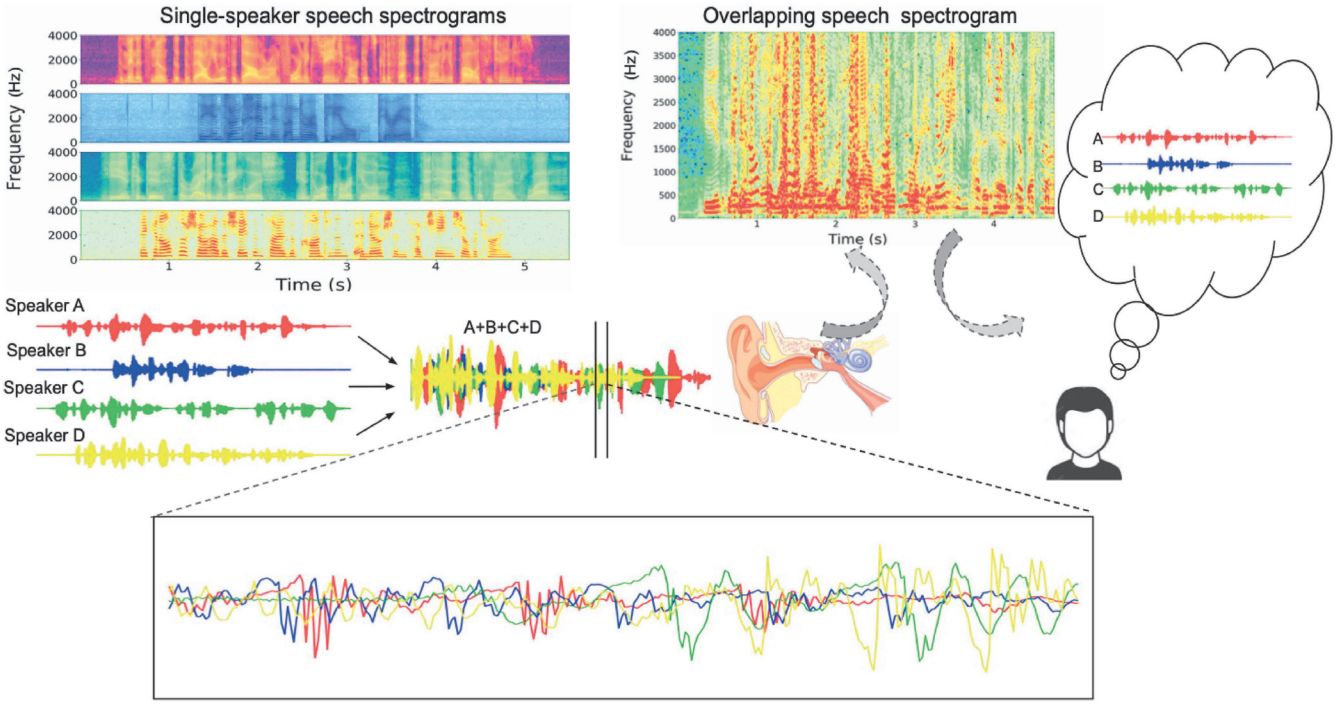

Fig. 1. Single speaker speech waveforms are shown in red, blue, green and yellow colors. Their corresponding spectrograms are also shown with the same color code. As depicted in this figure both the temporal and spectral structure of the speech signals are degraded by the presence of competing talkers.

sub-bands (SFM) [18], and pyknogram [19], [20] to detect overlapping speech segments within an audio stream. However, these features are hand crafted and may not necessarily be the optimal representation for the degraded speech. Also, these features are directly estimated from the speech waveform and can not learn from the data. Therefore, they have the potential to be very fragile in noisy, competing speaker conditions. Thus, incorporating Deep Neural Networks (DNNs) for modeling a single speaker speech could be a more effective way for classifying single speaker versus overlapping speech segments.

Some studies such as [21]–[23] have recently applied DNN architectures to address the classification of overlapping speech segments. In [21], the authors used a Long Short-Term Memory (LSTM) network to address overlapping speech detection for the AMI corpus [24]. The network was trained based on several features such as: kurtosis, spectral flux, harmonicity and MFCC, which results in 76% accuracy in detecting overlapping speech segments. However, since AMI is a real meeting scenario corpus, it is not balanced in terms of the ratio between the number of overlapping and non-overlapping speech samples. As reported in [23], AMI only contains approximately 5–10% overlapping speech, which may not be sufficient for training a neural network model without overfitting the data.

Consequently, [22] used artificially generated overlapping speech to train their CNN network. The CNN was trained based on FFT, MFCC and spectral envelope extracted from 25, 100, and 500 ms audio segments. Their system achieved 74–80% accuracy on different frame size features. The authors also reported F-score, which is the harmonic mean of the precision and recall to be 72% for a 25 ms frame length and 80% for longer duration frames.

In this study, we build on our previous work [25] and use a CNN-based architecture to resolve the overlapping speech detection problem on frame level segments as short as 25 ms. We also explore the effect of alternate features such as spectral magnitude, pyknogram, Mel Filter-Banks (MFB) and MFCC on the performance of overlapping speech detection, considering both computation time and classification measures. The contributions of this study are threefold:

- Proposing a block-based high performance 1-D CNN architecture for frame-level overlapping speech detection. Compared to our previous work [25], the convolutional layers are replaced by processing blocks performing max pooling and normalization steps in addition to the convolutional layers.
- Evaluating the performance of the proposed overlapping speech detection on different input spectral features considering both classification measures and processing time.
- Analyzing the effect of increasing network depth on performance of the introduced model by employing skip connections and residual learning.

The remainder of the paper is organized as follows. We present the problem formulation, generating overlapping speech mixtures, and extracting spectral features in Section II. Details of the proposed CNN architectures are explained in Section III. We report on the experimental procedures and results in Section IV. The results are discussed in Section V, and finally the conclusion is presented in the last section.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

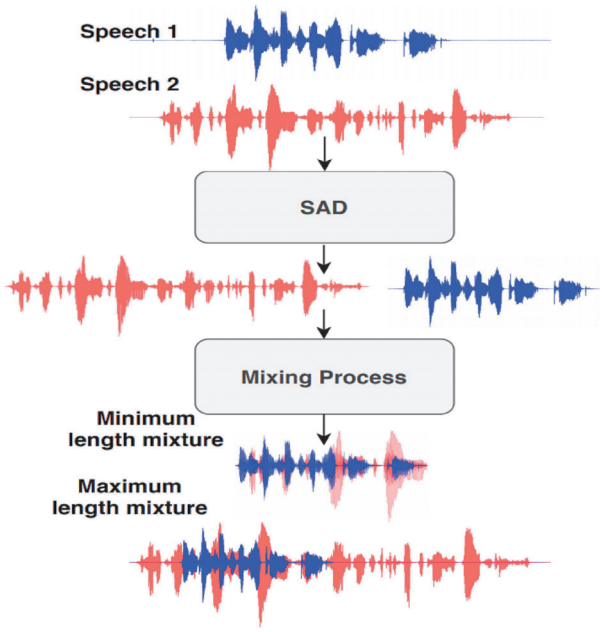

Fig. 2. Overlapping speech dataset generation pipeline.

II. Overlapping Speech Detection

A. Dataset Design

In this study, we generate the overlapping speech utterances based on the GRID corpus, which is a multi-speaker, sentence-based corpus used in monaural speech separation and recognition challenge [26]. This corpus contains 34 speakers, 16 female and 18 male speakers, each providing 1000 sentences, which have been frequently used in several overlapping speech detection and separation studies [16], [19], [27], [28].

For generating overlapping speech, random utterances from random speakers are selected. Next, Speech Activity Detection (SAD) is performed on the train/test utterances to remove all silent portions. In order to generate an audio mixture consistent with naturalist data, two scenarios are considered for simulating the mixtures. These scenarios are depicted in Fig. 2. In the first scenario, the longer utterance is cut so that its length matches the shorter utterance, then they are summed with a random Signal to Interference Ratio (SIR) uniformly distributed between 0 to 5 dB. In the second scenario, the shorter utterance is zero padded to match the length of the longer utterance before summing with a random SIR. Thus, each generated mixture file is either entirely overlapping speech or contains segments of both single-speaker and overlapping speech. We also include utterances that are entirely single-speaker speech to balance the final dataset. Also, we consider same-gender and opposite-gender speech mixtures. Therefore, we create three datasets for Male-Male, Male-Female, and Female-Female mixed speech signals. For each dataset, we have generated 20 h of mixed data for the training set, 3 h for development, and 2 h mixtures for the test set.[1] Also, speakers used for generating the test set are separate from those used in training and development sets.

[1]The corpus generated here will be shared with the speech community.

B. Acoustic Features

Speech is a non-stationary signal and its statistics changes over time. However, it is assumed that on a short-time scale, speech can be considered stationary. This is the reason for framing the speech signal into 20–30 ms segments. There is a trade-off between the length of each frame and its quality. In the longer duration frames, the stationary assumption may be violated, while in short frames, the number of samples may not be sufficient for estimating a reliable spectrum. Also, a time domain speech signal is not the most efficient choice to train the network due to the data sequence length. Accordingly, we use a set of spectral-based features that are more suitable in analyzing differences between single-speaker and overlapping speech segments. Prior to extracting spectral features, a pre-emphasis filter with its coefficient set to 0.97 is applied to the speech to boost the magnitude of high frequencies resulting in a more overall balanced spectrum.

1) Spectral Magnitude: is a 256-dim feature vector calculated using a 512-dim Short Time Fourier Transform (STFT) computed over a 25 ms Hamming window with a 10 ms frame shift. The spectral magnitude is the most basic spectral feature, which estimates active frequencies in each frame.

2) Mel FilterBank (MFB): is a 40-dim feature vector calculated by applying filterbanks to the power spectrum of the speech signal. In contrast to a linear frequency scale (Hz), the Mel scale is more discriminative in lower frequencies and less discriminative in higher frequencies, resulting in an improved resolution at lower frequencies. This is depicted in Fig. 3(A)(B). The filterbank consists of 40 triangular filters with center frequencies distributed along the Mel scale, which helps capture energy at each critical frequency band and approximates the overall spectral shape.

3) Mel Frequency Cepstral Coefficient (MFCC): is a 39-dim feature vector calculated as MFCCs and their first and second derivatives. Extracting MFCC is the same as MFB features with three additional steps. First, the log of the filterbank output is derived. Second, a Discrete Cosine Transform (DCT) is applied to the log of filterbank energies in order to decorrelate them across the channels, and achieve a more compressed representation of the filterbank energies. Third, keeping the first 13 spectral envelope-type elements of the DCT coefficients and discarding higher order elements. It is well known that MFCC features are very successful in many speech processing technology systems compared to MFB features since their coefficients are uncorrelated.

4) Teager-Kaiser Energy-Based Pyknogram: is a 120-dim, spectral-based feature vector. Pyknogram enhances the speech spectrogram by performing an AM-FM (Amplitude-Frequency Modulation) analysis, which decomposes the speech spectral sub-bands into amplitude and frequency components. As shown in [29], a linear combination of several AM-FM signals can represent the speech signal as:

$$x[n] = A[n]\cos(\omega[n]), \tag{1}$$

where A and ω are time varying amplitude and frequency terms used to represent the overall speech signal x. By calculating

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

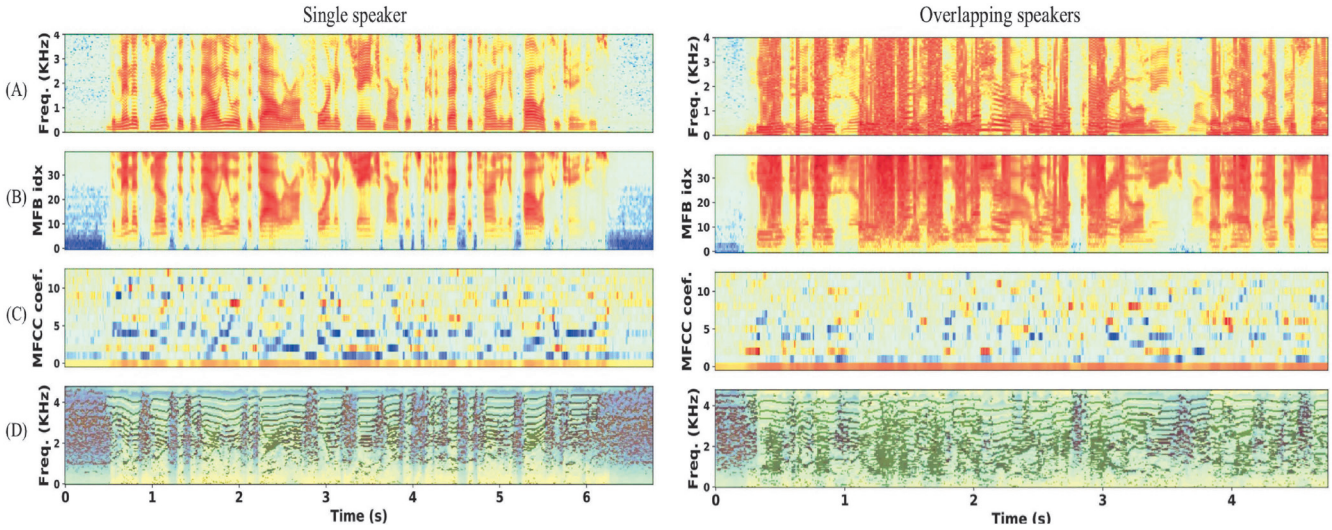

Fig. 3. Alternate feature representations, (A): 256-dim spectrogram, (B) 40-dim Mel-FilterBank, (C) 13-dim MFCC, and (D) 120-dim Pyknogram representations of single speaker speech and multiple speaker speech.

amplitude A and frequency ω as a function of time n, we can estimate the pyknogram representation of the speech signal. This procedure is carried out in two steps: first, disintegrating the speech signal into multiple components using a bandpass filterbank, where each component carries information in a subband frequency of the original speech signal. Second, decomposing each subband signal into frequency and amplitude partitions using an energy estimation operator.

The logarithmically spaced gammatone filterbank is chosen as the bandpass filter due to its similarity with the human auditory system e.g., cochlea. Next, the energy of each bandpass signal is estimated using the Teager-Kaiser Energy Operator (TEO) [19], [30]. The relation between the energy of each subband signal and its corresponding amplitude and frequency elements can be described by [30], [31]:

$$E_{teo}(x_i[n]) \propto A_i^2\omega_i^2 = x_i^2[n] - x_i[n-1] * x_i[n+1], \quad (2)$$

$$\omega_i[n] = \frac{1}{2\pi}\cos\left(1 - \frac{E_{teo}(x_i[n] - x_i[n-1])}{E_{teo}(x_i[n])}\right), \quad (3)$$

$$A[n] = \sqrt{\frac{E_{teo}(x_i[n])}{\sin^2(2\pi\omega[n])}}, \quad (4)$$

where E_{teo} is the estimated energy using TEO for the subband signal x_i, A_i, and ω_i are the amplitude and frequency belonging to subband i.

TEO is very effective in estimating the energy of the signal for two reasons [32], [33]: first, it uses only a window of three time-domain samples to estimate the energy, which implies excellent temporal resolution in capturing energy fluctuations in the signal. Second, the energy is estimated in a nonlinear manner, which makes it quite suitable for processing speech due to the non-linearity of speech energy distribution across the frequency domain (energy of voiced phonemes are more concentrated in lower frequencies, while constant phonemes of speech have more energy in high frequencies). After deriving $\omega[n]$ and $A[n]$, a short-time estimate value for the dominant frequency in each subband can be calculated as:

$$f_\omega(t,i) = \frac{\sum_t^{n+T}\omega_i[n]A_i^2[n]}{\sum_t^{n+T}A_i^2[n]}, \quad (5)$$

where t, i and T represent the t^{th} frame, the i^{th} filterbank, and the number of samples per frame. Finally, if the extracted frequencies are aligned with their corresponding filterbank bandwidth, they are selected as resonant peaks of pyknogram.

III. Network Architecture Design

In this section, we develop and configure three 1-D Convolutional Neural Network (CNN) architectures to perform overlapping speech detection. The fundamental elements of the designs including the number of layers, kernel size, channel size of the convolutional layers, and learning rate are taken into account for classification of multi-talker speech segments. The architectures of the proposed 1-D CNNs are depicted in Fig. 4.

Model-1: Fig. 4(a) shows the structure of the "1-D CNN" model. This structure is simply formed by stacking six convolutional layers followed by 2 Fully Connected (FC) layers. The 1-D CNN model was proposed in our previous work [25], which has proven to be very effective in achieving the state-of-the-art overlapping speech detection on frames as short as 25 ms. As discussed in [25], all hyperparameters are tuned using the development set resulting in an optimum choice of six 1-D convolutional layers. Each layer has a 128 input and output channel size except for the first layer which has 1 input channel to match the input data. Hyperbolic tangent (tanh) is used as the activation function between the convolutional layers. Subsequently, the extracted features in all channels are concatenated into one dimension and submitted to the first FC layer with 128 hidden neurons and the Rectified Linear Unit (ReLU) activation function. The second FC serves as the final layer of the network

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

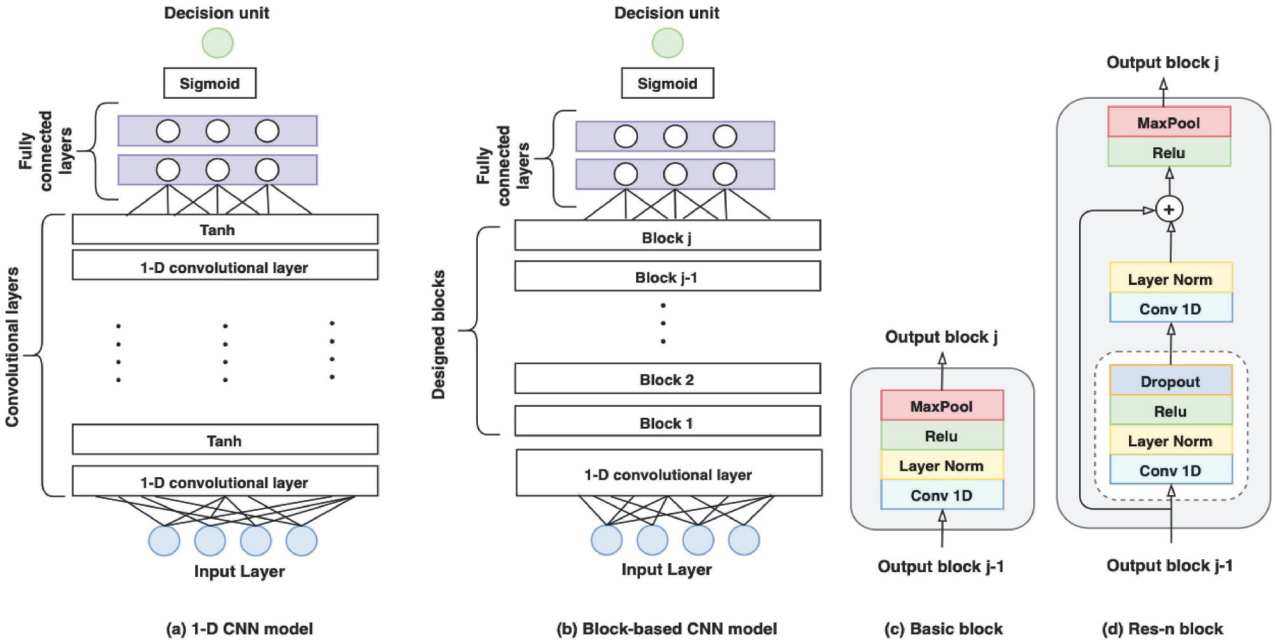

Fig. 4. Proposed architectures for overlapping speech detection. (a) 1-D CNN model consisting of several 1-D convolutional layers, with the output layer followed by two Fully Connected (FC) layers and a sigmiod activation function to derive the final decision. (b) the simple 1-D convolutional layers are replaced by designed building blocks; input is fed into a 1-D convolutional layer to adjust channel size; output of this layer is submitted to J blocks having the same structure. (c) and (d) are the building blocks used in the Block-based CNN architecture. Block (d) is a Res-n block with a skip connection and a dropout layer to avoid overfitting.

with an output size of 1 and a sigmoid activation function to yield the output probability of each class.

Model-2: For the second model, we design blocks consisting of a convolutional layer, layer normalization, ReLU activation function, and a max pooling layer. This means that the main difference between Model-1 and Model-2 is the addition of normalization and max pooling in each layer of the network. The details of the blocks used in Model-2 is depicted in Fig. 4(b) with block Fig. 4(c).

Considering that all blocks have the same design with the same input and output size, we apply a convolutional layer on the input to adjust for feature dimension so that it matches the dimensions of the first block. The first convolutional layer has 256 output channels with a kernel size of 3 and stride of 1 sample followed by a ReLU activation function. Padding is also employed on the input in order to maintain the same dimension size. In each block, the convolutional layer has an input and output channel size of 256, kernel size of 3, and stride of 1. The second step of each block is to normalize the output of the convolutional layer. There are two reasons for using layer normalization here: first, it speeds up the overall learning process. Second, layer normalization can help with reducing "internal covariate shift" defined as the change in distribution of network activations due to the change in network parameters during training [34]. Put differently, by employing layer normalization, the outputs of each layer are "whitened" - (i.e., linearly transformed to have zero means and unite variances). Therefore, layer normalization helps fix the distribution of the input data for each block, which alleviates the effect of internal covariate shift. Layer normalization solves this problem by forcing the deeper layers to learn the underlying structure of the data by themselves and more independently of the weights which are learned in the shallower layers. We normalize the features over both feature dimension and channels as follows:

$$LN(Y) = \frac{Y - E[Y]}{Var(Y) + \epsilon} * \gamma + \beta, \tag{6}$$

$$E[Y] = \frac{1}{MC} \sum_{MC} Y, \tag{7}$$

$$Var(Y) = \frac{1}{MC} \sum_{MC} (Y - E[Y])^2, \tag{8}$$

where Y is the output of the convolutional layer, γ and $\beta \in \mathbb{R}^{1\times C}$ are learned parameters, M is the size of the feature and C represents the number of channels.

The next step in the proposed processing block is to submit the normalized features into a max pooling layer to progressively reduce the dimension of the features. As a consequence, the amount of trainable parameters is decreased, which prevents overfitting the data. Finally, the output of the final block is passed to two FC layers with the same parameters as in Model 1 for the final decision (see Fig. 4(b)).

Model-3: Inspired by the skip connections used in ResNet [35], we modify the designed basic block shown in Fig. 4(c) by adding skip connections as depicted in Fig. 4(d). In ResNet usually one or two convolutional layers are used in each block. Thus, in our setting, if one convolutional layer is used the block is denoted as Res-1, and if two convolutional layers are used the block is named Res-2. The processing steps denoted by the dashed line in Fig. 4(d) is performed only in Res-2. In Res-1, we use only one convolutional layer and perform layer

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

normalization, where the normalized output is then summed with the input through a skip connection path. Next, ReLU activation and max pooling operations are employed.

The motivation behind using skip connections in this study lies behind the deterioration we witnessed in results from stacking more blocks in Model-2. Therefore, with using skip connections along with a deeper network (through stacking additional blocks), we want to explore the following question: Is the decrease in the performance caused by data and gradient vanishing in the deeper layers of the network?

In some studies [36], the success of deep neural network is accredited to the depth of the network. As explained in [37], front-end layers of the network generally learn basic features, while as we consider deeper layers, each layers learn more complex features. In applications where the underlying structure of the data is not very complicated, the performance of the shallow and deep network should be similar (i.e, their functions may be repetitive). In contrast, if the data has a high degree of complex structure, deeper networks produce a more balanced comprehensive feature or higher context structure that represent all details of the data.

Residual learning is an effective way to train deep models using skip connections. Assume that the general block in Fig. 4(c) learns the mapping function $H(x)$ in the convolutional layer right before the max pooling layer as:

$$H(x_{j-1}) = ReLU(LN(W_j x_{j-1} + B_j)), \tag{9}$$

where W_j and B_j are the weight and bias matrix of the convolution layer j, and LN represents a layer normalization step. With the introduction of skip connections, the network needs only learn $F(x)$ as:

$$F(x_{j-1}) = H(x_{j-1}) - x_{j-1} \tag{10}$$

and then the final mapping $H(x)$ is calculated as:

$$\begin{aligned} H(x_{j-1}) = F(x_{j-1}) + x_{j-1} = \\ ReLU(LN(W_j x_{j-1} + B_j) + x_{j-1}). \end{aligned} \tag{11}$$

Since $F(x)$ is a simpler function compared to $H(x)$, the network is able to more effectively characterize the model training structure. Also, the skip connection provides a path for the input data toward deeper blocks, thereby preventing data vanishing for deeper layers. Also, the vanishing gradients problem is solved by constructing ensembles of many short networks together, instead of preserving the gradient flow throughout the entire network.

IV. Experimental Results

In this section, we investigate the performance for different overlapping speech detection architectures shown in Fig. 4 using classification metrics. We also analyze the effect of alternate hyperparameters, as well as input features on overall performance.

Evaluation metrics: we evaluate performance of the proposed overlapping speech detection models on datasets generated from the GRID corpus. Accuracy is a widely used metric for classification problems, and is defined as the ratio of the correct predictions divided by the total number of samples. Nevertheless, accuracy is not sufficient for scenarios where the dataset is not balanced in terms of the number of data samples belonging to each class. As a consequence, for unbalanced datasets, it is more useful to consider how many samples are mis-classified within each class. This can be achieved by using other measures defined based on the confusion matrix such as precision, recall, and F-score. Precision and recall are frequently used to evaluate performance of the system for classification. Precision reflects the ratio of the correctly detected overlapping segments to the total number of detected overlapping segments and is defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{12}$$

where TP (True Positive) represents the correctly classified overlapping speech frames, and FP (False Positive) represents the single-speaker speech frames which are mis-classified as overlapping speech. On the other hand, recall is the ability of the model to find all the overlapping frames in the dataset, measured as the ratio of correctly detected overlapping segments to the total number of actual overlapping segments, defined as follows:

$$Recall = \frac{TP}{TP + FN}, \tag{13}$$

where FN (False Negative) represents the actual overlapping speech frames which are mis-classified as single-speaker speech. The harmonic mean of precision and recall is reflected by the F-score defined as:

$$F - score = \frac{2 \times precision \times recall}{recall + precision}, \tag{14}$$

A validation set is used for hyperparameter tuning. Models are trained with 100 epochs using a Binary Cross Entropy (BCE) loss and Stochastic Gradient Descent (SGD) optimizer [38]. The initial value of the learning rate is tuned to 0.001, with a 50% decay if no improvement is observed in the cross validation loss for three successive epochs.

1) Results for Model-1: the first model is based on 1-D CNN model in Fig. 4(a), which serves as the primary system in this study, and therefore considered as the baseline. The 1-D CNN model is trained with a range of alternate input features. The motivation behind this set of experiments is to determine which spectral features are more effective for training a high performing classifier.

The hyperparameters of the network are tuned to ensure that network has a viable initial setup for the overlapping speech detection task prior to experiments with the alternative input features. The hyperparameters are regulated based on the development set. In order to reduce the overall number of experiments and computational cost of hyperparameter optimization, we first tune the number of layers, then the channel size of each convolutional layer. Eventually, we tune the learning rate. Our experiments show that 6 layers of convolutional layers, with 128 output channels and a learning rate of 0.001 are the best choices for the hyperparameters. The training loss and cross validation loss of the network with the selected hyperparameters are shown in Fig. 5 (left). As shown, 100 epochs is sufficient for training the network. In addition, the validation loss, represented in red, is very close to the training loss as the epoch count

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

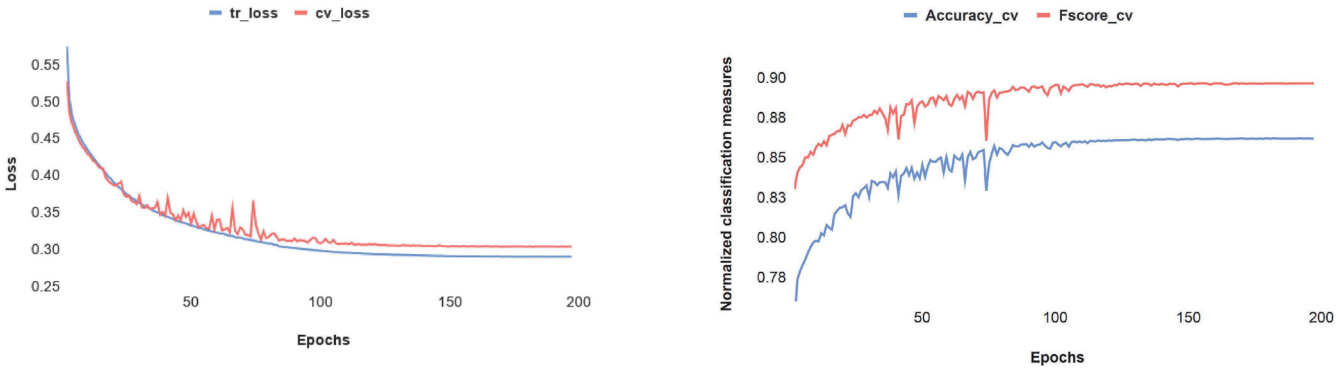

Fig. 5. Training loss and cross validation loss associated with the tuned hyperparameters of 1-D CNN model (Model-1). Also, Accuracy and F-score of Model-1 in development phase using cross validation dataset are plotted.

TABLE I
THE PERFORMANCE OF 1-D CNN MODEL (MODEL-1) ON OVERLAPPING SPEECH DETECTION. M-M, F-F, AND M-F STAND FOR OVERLAPPING SPEECH GENERATED BASED ON MALE-MALE, FEMALE-FEMALE, AND MALE-FEMALE SPEAKERS

1-D CNN model (model-1)					
Dataset	Metrics	MFCC	MFB	Pykno	Spec
M-M	Accuracy	81.3%	79.7%	81.3%	80.3%
	Precision	84.2%	81.7%	83.1%	82.7%
	Recall	89.1%	90.8%	91.5%	90.1%
	Fscore	86.2%	85.5%	86.5%	85.8%
F-F	Accuracy	82.9%	82.0%	83.4%	82.6%
	Precision	85.7%	84.1%	86.1%	85.8%
	Recall	90.7%	91.5%	91.2%	89.9%
	Fscore	87.6%	87.1%	87.9%	87.3%
M-F	Accuracy	88.7%	88.7%	88.9%	88.6%
	Precision	91.8%	92.4%	91.3%	92.3%
	Recall	91.7%	91.0%	90.8%	90.9%
	Fscore	91.4%	91.4%	91.4%	91.3%

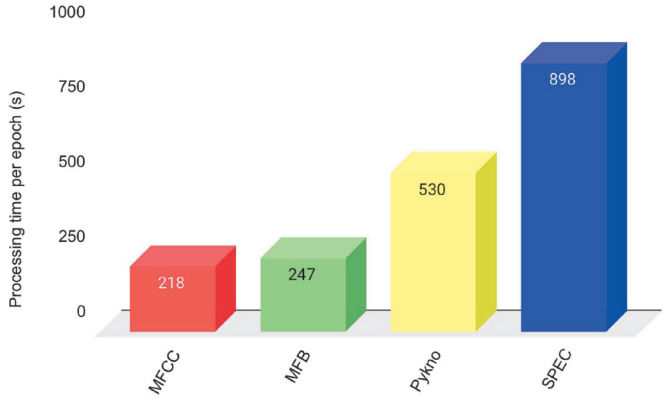

Fig. 6. Processing time per epoch (in second) for training 1-D CNN model (model-1) based on different input features. Spec stands for spectral magnitude, and Pykno stands for pyknogram. For example, the processing time for training 1-D CNN model with MFCC is 218 s per epoch.

increases, indicating the generalization power of the network to speech samples not seen in the training phase. The right plot in Fig. 5 shows the accuracy and F-score of Model-1 in the development phase, confirming that after 100 epochs, the model attains a stable performance in classifying overlapping speech frames.

Subsequent to hyperparameter tuning, four independent networks are trained, with the same selected hyperparameters based on four alternate input spectral features. The performance of each network is summarized in Table I.

For the experiments, three sets of overlapping speech data based on the GRID corpus are prepared (refer to Table I). The first dataset, tagged as M-M, contains mixed speech utterances in which both speakers are male. The second dataset, shown as F-F, consists of overlapping speech signals from two female speakers, and in the M-F dataset, one of the speakers is male while the other is female. As inferred from the table, classification of overlapping speech segments for opposite gender mixtures is easier than the same gender mixtures. This is due to the difference in the vocal tract physiology and fundamental frequency, and their harmonics, which exists in speech between male and female speakers. However, for all three datasets, the model trained on pyknogram achieves the best performance compared to other features in terms of both accuracy and F-score. MFCC achieves the second best performance compared to other input features.

Additionally, Fig. 6 illustrates the processing time per epoch for training the network based on alternative features. As noted earlier, spectral magnitude is a 256-dim feature vector which is highly redundant in information content. The experiments show that processing spectral magnitude requires significantly longer processing time compared to other features without improvement in classification. In contrast, the 120-dim pyknogram feature vector achieves the best overall performance at the cost of more processing time compared with either the 40-dim MFB or 39-dim MFCC feature sets.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

TABLE II
THE PERFORMANCE OF BLOCK-BASED CNN MODEL WITH BASIC BLOCK FOR OVERLAPPING SPEECH DETECTION WITH ALTERNATE HYPERPARAMETERS. ACCURACY-CV IS ACCURACY MEASURED ON CROSS VALIDATION DATASET. ACCURACY-TT IS ACCURACY MEASURED ON THE TEST DATASET

# Blocks	Kernel	Channel	Learning rate	Accuracy-cv	Fscore-cv	Accuracy-tt	Precision-tt	Recall-tt	Fscore-tt
2	3	256	0.001	87.8%	90.9%	87.8%	90.9%	91.0%	90.7%
3	3	256	0.001	89.2%	91.6%	89.3%	92.7%	91.2	91.7%
4	3	256	0.001	**90.3%**	**92.5%**	**90.2%**	**93.3%**	**92.1%**	**92.5%**
5	3	256	0.001	90.0%	92.5%	89.9%	92.5%	92.4%	92.3%
6	3	256	0.001	90.1%	92.5%	89.9%	92.8%	92.2%	92.3%
4	2	256	0.001	89.3%	91.9%	89.2%	92.0%	92.0%	91.8%
4	4	256	0.001	89.8%	92.3%	89.7%	92.5%	92.1%	92.1%
4	5	256	0.001	89.5%	92.0%	89.4%	92.6%	91.5%	91.8%
4	6	256	0.001	89.8%	92.2%	89.6%	92.3%	92.3%	92.0%
4	3	64	0.001	88.1%	91.1%	88.1%	90.6%	91.9%	91.0%
4	3	128	0.001	89.0%	91.7%	89.0%	91.8%	91.7%	91.6%
4	3	256	0.0001	90.5%	92.8%	86.9%	90.0%	90.7%	90.1%
4	3	256	0.01	86.8%	86.8%	90.4%	93.5%	92.1%	92.6%

Based on their results, pyknogram and MFCC achieved better overall classification performance, however, according to Fig. 6, the processing time per epoch for MFCC is 218 seconds, while it is 530 seconds for pyknogram. This indicates greater efficiency of MFCCs in detecting co-existing talker in speech segments, which comes from the compactness of the feature. MFCC carries all the required information for detecting overlapping speech in a 39-dim feature vector, whereas pyknogram requires a 120-dim vector. Deduced from Table I, in most occasions, MFCC either accomplishes the same performance as pyknogram, or lags by at most 1%, which is tolerable for frame-level overlapping speech detection. Deciding between using MFCC or pyknogram depends substantially on the end application. For real-time applications, MFCC is clearly the best choice, in price of lower accuracy compared to pyknogram. However, for off-line processing, pyknogram provides slightly better accuracy.

2) Results for Model-2: this architecture consists of the Block-based CNN model Fig. 4(b) with the Basic block in Fig. 4(c). Having added the normalization and max pooling layers, we choose to tune the hyperparameters of Model-2 independently from the first model as a means of finding the best parameter combinations for improved performance. Motivated by the efficiency of MFCCs in Model-1, we tune the second model by training the network on MFCC features derived from the Male-Male dataset. Classification results for alternate hyperparameters are reported in Table II.

The first column of Table II presents the number of basic blocks Fig. 4(c) used in the network. In the first five experiments, the number of the blocks are varied from two to six, while the rest of the hyperprameters are kept constant. Based on classification performance, the best number of blocks is selected for this architecture. Next, other hyperparameters are tuned. According to Table II, 4 blocks with 256 channels, and a kernel size of 3 is the best combination of hyperparameters for this structure. However, the choice between the learning rate is yet to be decided. Considering classification performance on both cross validation and test datasets, the network trained with a learning rate of 0.001 achieves the best overall performance compared with a learning rate 0.01 and 0.0001. A small learning rate increases the training time for convergence to an optimal local minimum. On the other side, a large learning rate causes larger steps in the training process which increases the probability of converging to a poor local minimum. Thus, we adjust the learning rate to 0.001 to sustain the converging process with the goal of an effective local minimum on both cross validation and test datasets.

Based on the chosen hyperparameters for Model-2, a set of experiments using all input spectral features derived from M-M, F-F, and M-F datasets are carried out and reported in Table III.

A comparison of the results from Table III and Table I, shows that block based 1-D CNN (Model-2) is superior compared to the 1-D CNN model (Model-1). With employing a more compact feature set such as MFCC, Model-2 achieves better classification performance compared to Model-1. This confirms the effectiveness of using layer normalization and max pooling layers. By standardizing the output of each block, layer normalization assures that the input distribution of the upcoming blocks are consistent during parameter updating. This stabilizes training and also fortifies the network against variations in the input features and their successive learned mappings. Alternatively, max pooling is a down-sampling step that replaces a filter-lengthed sub-array with its maximum value. The maximum value can generally be considered as a statistics summary of that sub-array. The main advantage of pooling layers is that the network depends more on an approximate of the mapped features, rather than focusing on their exact forms. As a consequence, the network is more robust to local variations from input features for each block. According to all the aforementioned facts, both layer normalization and pooling components boost performance of overlapping speech detection by 9% absolute improvement for the challenging same gender overlapping speech datasets.

3) Results for Model-3: the block shown in Fig. 4(d) is designed to examine if stacking more blocks while using skip connections is able to improve the performance of Model-2. As

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

TABLE III
THE PERFORMANCE OF MODEL-2 ON OVERLAPPING SPEECH DETECTION. M-M, F-F, AND M-F STAND FOR OVERLAPPING SPEECH FROM MALE-MALE, FEMALE-FEMALE, AND MALE-FEMALE SPEAKERS

Block-Based CNN model with basic block (model-2)					
Dataset	Metrics	MFCC	MFB	Pykno	Spec
M-M	Accuracy	90.3%	87.8%	90.4%	89.6%
	Precision	93.3%	90.9%	93.3%	93.1%
	Recall	92.2%	91.0%	92.5%	91.3%
	Fscore	92.5%	90.7%	92.6%	92.0%
F-F	Accuracy	90.8%	88.7%	91.2%	90.4%
	Precision	93.6%	92.2%	94.0%	93.8%
	Recall	92.9%	91.3%	93.2%	92.1%
	Fscore	93.1%	91.5%	93.4%	92.7%
M-F	Accuracy	95.6%	93.9%	95.4%	94.1%
	Precision	97.2%	96.2%	97.0%	96.3%
	Recall	96.2%	94.6%	96.2%	94.9%
	Fscore	96.7%	95.3%	96.5%	95.5%

stated in Section III, depending on the number of convolutional layers used in Res-n block, Res-1 and Res-2 are introduced. Res-1 is the exact same architecture used in Model 2 except for the additional skip connection (i.e., summing the input of each block with its output right after layer normalization step). Furthermore, in Res-2, some added steps such as one additional 1-D convolutional layer, layer normalization, ReLU activation function, and a dropout layer are performed. In order to study the effect of introducing the skip connection on classification performance, we use the same hyperparameters values for Model-3 as used in Model-2.

In the first set of experiments, we investigate the effect of stacking further blocks in Model-3 by training 12 networks independently. In the first six networks, a block-based CNN model with Res-1 block are trained by stacking 2, 4, 6, 8, 10, 12 blocks in the architecture. For the other six networks, a block-based CNN model with Res-2 block are employed to study the effect of increasing the depth of the network for the same overall number of blocks. Performance of overlapping speech detection is illustrated in Fig. 7. As inferred from the plots, Model-2 is more capable of classifying overlapping speech segments compared to both Res-1 and Res-2 blocks in Model-3. The only case where Model-3 outperforms Model-2 is when there are only two blocks of Res-1 in the network architecture. However, employing a larger number of layers (i.e., four to twelve layers), the skip connections begin to degrade overall performance.

One reason for the performance loss of Model-3 with Res-2 block in detecting overlapping speech frames could be due to the overfitting problem caused by adding extra convolutional

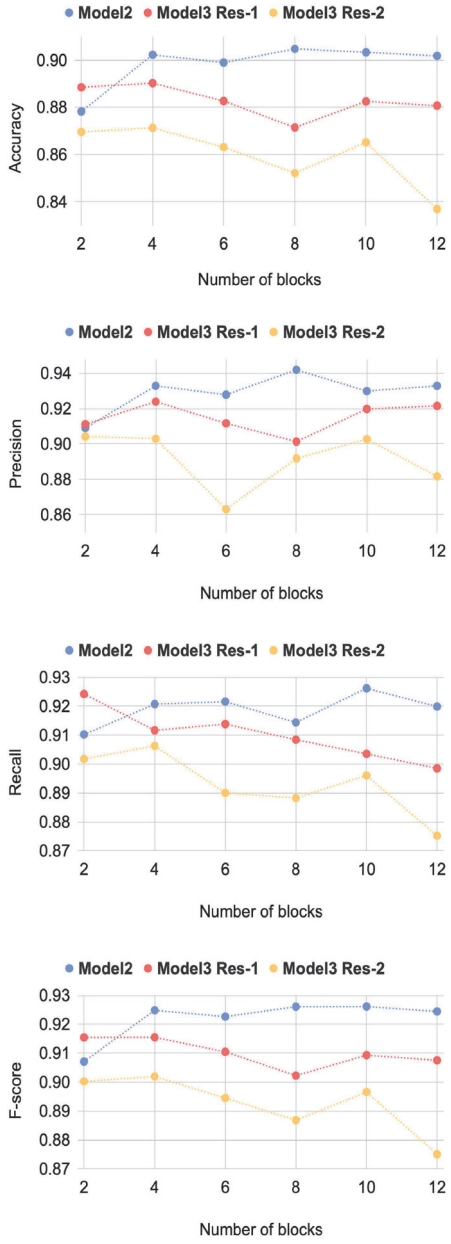

Fig. 7. Effect of increasing depth of network by stacking more blocks in the architecture for overlapping speech detection.

layer in each block. This increases the network's parameters and capacity, which results in poor generalization to the test dataset. Therefore, the network tends to "memorize" the samples in the training set instead of learning the underlying patterns belonging to each class. This also can be verified by comparing the results for Res-1 and Res-2 in Model-3. Res-1 achieves better accuracy and F-score compared to Res-2 due to the smaller model size.

Alternatively, Res-1 block also falls behind Model-2 for all classification measure. As noted in Section III, the motivation behind designing a Res-n block was to determine if a deeper network will boost performance. For this reason, we employed skip connections to insure the stability in optimization and also prevent data vanishing in deeper layers. The experiments

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
THE PERFORMANCE OF MODEL-3 WITH RES-1 BLOCK ON OVERLAPPING SPEECH DETECTION. M-M, F-F, AND M-F STAND FOR OVERLAPPING SPEECH FROM MALE-MALE, FEMALE-FEMALE, AND MALE-FEMALE SPEAKERS

Block-Based CNN model with Res-1 block (model-3)					
Dataset	Metrics	MFCC	MFB	Pykno	Spec
M-M	Accuracy	89.0%	86.9%	89.7%	89.6%
	Precision	92.4%	90.1%	92.7%	93.1%
	Recall	91.2%	90.6%	92.0%	91.3%
	Fscore	91.6%	90.0%	92.1%	92.0%
F-F	Accuracy	89.7%	87.6%	90.6%	89.7%
	Precision	92.8%	90.9%	93.6%	93.2%
	Recall	92.1%	91.2%	92.7%	91.6%
	Fscore	92.3%	90.7%	92.9%	92.2%
M-F	Accuracy	94.6%	93.3%	95.0%	94.1%
	Precision	96.5%	95.6%	96.7%	96.3%
	Recall	95.3%	94.3%	95.8%	94.9%
	Fscore	95.9%	94.9%	96.2%	95.5%

performed here proved that patterns needed for classifying a single-speaker speech versus multi-speaker frames may not be as complex, and can be learned by a small model with 4 to 6 layers. In this case, adding skip connections in each block might impact the quality of the learned patterns adversely.

We also report results of Model-3 with a Res-1 block on M-M, F-F, M-F datasets in Table IV for comparison with Model-2 and Model-1. Model-3 performs better than Model-1 in distinguishing multi-talker speech frames. For example, on the M-M dataset, Model-1 achieves 81.3% accuracy using pyknogram. Model-3 scores 89.7% accuracy using the same features resulting in an +8.4% absolute improvement. This is true for other scores as well, +9.6% improvement in precision, and +5.6% improvement in F-score. Comparing Tables I and IV, we observe that pyknogram features achieve recall of 91.5% for Model-1, and 92.0% for Model-3. Since, recall is the ability of the model to find all overlapping segments, this observation indicates that the first model is able to classify 91.5% of overlapping frames in the test dataset, however, the low precision makes the first model vulnerable to confusing single-speaker frames with overlapping speech. In other words, Model-1 has a high false alarm rate that can be problematic in many real world applications. For MFCC features, Model-3 improves recall by +2.1%, reinforcing the fact that a stronger classifier is able to compensate for a weak feature.

The pattern of results is repeated for the F-F dataset as well. Accuracy is improved by +6.8% for MFCC, +5.6% for MFB, +7.2% for pyknogram and +7.1% for spectral magnitude in Model-3 compared to Model-1. Precision is also boosted by +7.1% for MFCC, +6.8% for MFB, +7.5% for both pyknogram, and spectral magnitude. Model-3 improves the recall score of Model-1 by +1.4% for MFCC, -0.3% for MFB, +1.5% for pyknogram, and +1.7% for spectral magnitude.

The third dataset known as M-F, has the same advancement trend in the results. Model-3 achieves 95% accuracy with 96.7% precision and 95.8% recall for pyknogram indicating almost a +5% overal improvement compared to Model-1.

V. DISCUSSION

To the best of our knowledge, previous overlapping speech detection systems such as [19], [21], [22] were designed to perform well on long segments of speech. In [19], an unsupervised energy-based approach was proposed, which requires an input speech segment of no less than 2 seconds for reliable overlapping speech classification. In [21] a supervised LSTM-based network was introduced that achieved 76% accuracy on the AMI corpus, which is a real scenario meeting corpus with different duration overlapping speech segments. The most recent work [22] used a CNN-based network to address distinguishing multi-talker segments on 25, 100, and 500 ms segments. The reported accuracy for the 25 ms frames was 74% with 72% recall. The best accuracy of their proposed system was 80% on 500 ms input segments.

In contrast to these previously proposed techniques, we proposed a 1-D CNN-based network in [25] to classify 25 ms overlapping speech frames with an accuracy of 85%, recall of 91% and precision of 87%. Building on our previous work, in this study we extended the network architecture in order to boost the accuracy and precision of the overlapping speech classifier on 25 ms frames. We used our previous work [25] as the baseline in this study. This baseline was presented as Model-1 here.

The baseline was trained using MFCCs with final training and cross validation loss shown in Fig. 5. Model-1 achieves good generalization power for the development set with a cross validation loss close to that seen for the training loss in Fig. 5(left). It is also worth noting that the used Binary Cross Entropy (BCE) cost correlates well with the classification measures. This can be inferred from Fig. 5(right), where peaks in cross validation loss match valleys for the accuracy and F-score, suggesting that a higher BCE cost will result in lower accuracy and F-score. Model-1 is evaluated using alternate input features such as: spectral magnitude, MFB, MFCC, and pyknogram. Based on the evaluation metrics in Table I, pyknogram and MFCC achieved the best scores. However, based on Fig. 6, processing time per epoch for network training with MFCC was 218 sec, whereas it was 530 sec for pyknogram. Since MFCC is a compact 40-dim feature vector compared to the 120-dim pyknogram, it reduced the processing time by almost 60% in the training phase, while achieving competitive performance compared to pyknogram. Although, Model-1 was successful in retrieving most of the overlapping speech frames with a recall of 89.1% for Male-Male mixtures, 90.7% for Female-Female mixtures, and 91.7% for Male-Female mixtures based on MFCCs, it showed lower precision and accuracy for all noted datasets. Low precision suggests a high false alarm rate, which may be undesired in many applications. This motivated us to expand the capacity of Model-1 to improve both the precision score and accuracy of overlapping speech detection.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

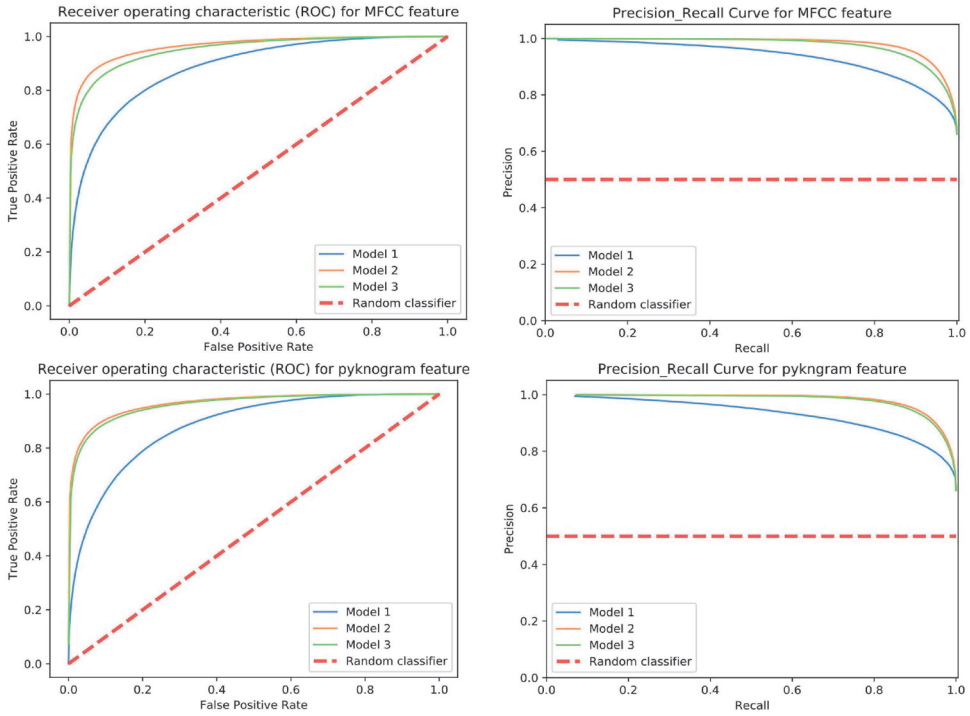

Fig. 8. The ROC and precision-recall curves for the proposed architectures. The networks are trained based on MFCC and pyknogram features extracted from male-male dataset.

Model-2 was introduced as a block-based CNN architecture Fig. 4(b) with the basic block in Fig. 4(c) to boost precision and accuracy of the first model. Model-2 consisted of one 1-D convolutional layer for adjusting the number of channels followed by several identical blocks each performing: a 1-D convolutional layer, layer normalization, ReLU activation function and max pooling layer. The number of used blocks is a very important hyperparameter affecting both processing time and the evaluation metrics. Finally, the output of the last block was submitted to two FC layers with a sigmoid activation for the final decision. The hyperparameters in Model-2 were tuned based on cross validation and test scores reported in Table II. For a fair comparison between Model-2 and Model-1, the same experiments were conducted for Model-2, and were presented in Table III. The results confirmed that Model-2 was effective in improving accuracy and precision of Model-1, achieving 90.3% accuracy, for MFCCs based on Male-Male dataset, whereas accuracy for Model-1 is 81.3% in the same feature set, indicating a +9% absolute improvement. Precision was also increased by +9.1% from 84.2% to 93.3% for MFCCs derived from M-M dataset, resulting in less false alarms while detecting multi-talker speech frames. Table III shows similar improvement for all three datasets. The best numbers for Model-2 belongs to Male-Female mixtures with over 95% accuracy, precision and recall. The success of Model-2 can be attributed to the layer normalization and max pooling layers, leading to a smoother optimization landscape, a more stable training process, and a model more robust to variations of parameters distribution.

Model-3 was proposed in an effort to explore impact of network depth on classification performance. In the block-based CNN architecture, a Res-n block as shown in Fig. 4(d) was employed. Parameter "n" in Res-n was determined by the number of the convolutional layers used in each block. Skip connections in Model-3 helped addressing vanishing input data problem for deeper layers, which are responsible for learning more complex features. We used the same hyperparameter values from Model-2 for Model-3 to investigate the effect of using skip connections on the final evaluation metrics. Fig. 7 presents the effect of increasing network depth on overlapping detection accuracy, precision, recall and F-score. In these figures, Model-2 and Model-3 with Res-1 blocks and Res-2 blocks were considered with different numbers of blocks in their architectures. As conveyed from Fig. 7, Model-3 with Res-2 block underperforms the Res-1 block, which could be the result of overfiting caused by increasing the capacity of the model. Also, Res-1 Model-3 lags behind Model-2 in classifying overlapping frames by 2% in accuracy, 0.9% precision, and 1% recall.

Finally, Fig. 8 presents the Receiver Operating Characteristics (ROC), and pecision-recall curve for MFCC feature (top) and pyknogram (bottom), extracted from Male-Male mixtures. In these two feature sets, both Model-3 and Model-2 outperform Model-1 while their difference is more perceptible for MFCCs compared to pyknogram.

VI. Conclusion

In this study, we proposed block-based CNN architectures for addressing the problem of overlapping speech detection. Our focus was to classify single versus multi-talker speech in segments as short as 25 ms. Furthermore, we explored the effects of

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

using alternate input features such as spectral magnitude, MFB, MFCC, and pyknogram on both final classification performance, and computational time. The proposed architecture was shown to be robust to local deviations of input features, and achieved the best classification results with 90.5% accuracy, 93.5% precision, 92.7% recall, and 92.8% Fscore for MFCC feature on the same gender overlapping speech mixtures. Additionally, using pyknogram features improved classification slightly in price of higher processing time. It is worth mentioning that an important aspect of choosing the right feature for overlapping speech detection is based on the application in hand. For those application where detecting all overlapping segments is crucial, while some false alarms can be tolerated, a feature with higher recall is desired even if precision and accuracy are low. For real-time applications, MFCC is the best option because of the required processing time and computational cost is lowest compared to other features. Therefore, the choice between these features is dictated by available online computing resources. We demonstrated that using pyknogram can provide better classification accuracy in applications where off-line processing is allowed. This proposed frame-based model architecture with such a high accuracy, precision and recall could eventually provide an effective solution for the overlapping speech detection and make this a pre-processing step for most of speech technology systems designed for real-world applications.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] C. Guan-Lin, W. Chan, and I. Lane, "Speaker-targeted audio-visual models for speech recognition in cocktail-party environments," in *Proc. ISCA Interspeech 2016*, 2016, pp. 2120–2124.

[3] D. Oberfeld and F. Kloeckner-Nowotny, "Individual differences in selective attention predict speech identification at a cocktail party," *Elife*, vol. 5, pp. e16747, 2016.

[4] D. E. Broadbent, *Perception and Communication*, Elsevier, 2013.

[5] A. M. Treisman, "Contextual cues in selective listening," *Quart. J. Exp. Psychol.*, vol. 12, no. 4, pp. 242–248, 1960.

[6] J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations.," *Psychol. Rev.*, vol. 70, no. 1, pp. 80, 1963.

[7] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[8] J. H. McDermott, "The cocktail party problem," *Current Biol.*, vol. 19, no. 22, pp. R1024–R1027, 2009.

[9] M. A. Bee and C. Micheyl, "The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it?" *J. Comp. Psychol.*, vol. 122, no. 3, pp. 235, 2008.

[10] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1670–1679, Oct. 2015.

[11] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, 2010.

[12] K.i Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4353–4356.

[13] R. E. Yantorno, K. R. Krishnamachari, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "The spectral autocorrelation peak valley ratio (SAPVR)-A usable speech measure employed as a co-channel detection system," in *Proc. IEEE Int. Workshop Intell. Signal Process.*, 2001, vol. 21.

[14] M. Zelenák, C. Segura, and J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Proc. Eleventh Annu. Conf. Int. Speech Commun. Assoc.*, 2010.

[15] K.i Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Proc. Twelfth Annu. Conf. Int. Speech Commun. Assoc.*, 2011.

[16] M. Yousefi, N. Shokouhi, and J. H. L. Hansen, "Assessing speaker engagement in 2-person debates: Overlap detection in United States Presidential debates.," in *Proc. Interspeech*, 2018, pp. 2117–2121.

[17] M. Yousefi and M. H. Savoji, "Supervised speech enhancement using online group-sparse convolutive NMF," in *Proc. 8th Int. Symp. Telecommun.*, 2016, pp. 494–499.

[18] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. HL Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 2834–2838.

[19] N. Shokouhi and J. H. L. Hansen, "Teager–Kaiser energy operators for overlapped speech detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1035–1047, May 2017.

[20] N. Shokouhi, A. Ziaei, A. Sangwan, and J. HL Hansen, "Robust overlapped speech detection and its application in word-count estimation for prof-life-log data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4724–4728.

[21] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proc. INTERSPEECH 2013*, 2013.

[22] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short time frames using deep learning," in *Proc. INTERSPEECH*, 2017, pp. 1198–1202.

[23] T. v. Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 91–95.

[24] J. Carletta *et al.*, "The ami meeting corpus: A pre-announcement," in *Proc. Int. workshop Mach. Learn. Multimodal Interact.* Springer, 2005, pp. 28–39.

[25] M. Yousefi and J. H. L. Hansen, "Frame-based overlapping speech detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 4–8, 2020, pp. 6744–6748.

[26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[27] Y.-H. Tu, J. Du, L.-R. Dai, and C.-H. Lee, "Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 61–65.

[28] M. Yousefi, S. Khorram, and J. H. L. Hansen, "Probabilistic permutation invariant training for speech separation," in *Proc. ISCA Interspeech 2019*, Graz, Austria, Sep. 15–19, 2019, Paper 1827, pp. 4604–4608.

[29] A. C. Bovik, P. Maragos, and T. F. Quatieri, "Am-fm energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3245–3265, Dec. 1993.

[30] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

[31] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Classification of speech under stress based on features derived from the nonlinear teager energy operator," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1998, vol. 1, pp. 549–552.

[32] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, Mar. 2001.

[33] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Methods for stress classification: Nonlinear teo and linear speech based features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 1999, vol. 4, pp. 2087–2090.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learn.*, 2015.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. Thirty-first AAAI Conf. Artif. Intell.*, 2017.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.

[37] L. C. Yan, B. Yoshua, and H. Geoffrey, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[38] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT'2010*, Springer, 2010, pp. 177–186.

Midia Yousefi (Student Member) received the B.S. degree in electrical engineering from the University of Kurdistan, Sanandaj, Iran, in 2014 and the M.S. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2016. She became the Ph.D. student with the University of Texas at Dallas (UTD) in 2017. Since then, She has been a Graduate Research Assistant with Center for Robust Speech Systems (CRSS) at UTD. She has published the results of her research in various authored or coauthored journal and conference publications in the field of speech processing and language technology. Her research interests focus on automatic speech recognition, speech separation, speech enhancement, and machine learning.

John H. L. Hansen (Fellow, IEEE) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, USA, and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, GA, USA. In 2005, he joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTDallas), where he is currently an Associate Dean for Research, as well as a Professor of Electrical and Computer Engineering, the Distinguished University Chair in Telecommunications Engineering, and holds a joint appointment as a Professor with the School of Behavioral and Brain Sciences (Speech & Hearing). From August 2005 to December 2012, he was the Department Head of Electrical Engineering with UTDallas, overseeing a five times increase in research ($4.5 M to 22.3 M) with a 20% increase in enrollment along with hiring 18 additional T/TT faculty, growing UTDallas to the 8th largest EE program from ASEE rankings in terms of degrees awarded. At UTDallas, he established the Center for Robust Speech Systems (CRSS). He was the Department Chairman and Professor of Speech, Language and Hearing Sciences (SLHS), and a Professor in Electrical & Computer Engineering with the University of Colorado - Boulder (1998–2005), where he cofounded and was the Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTDallas. He has supervised 92 Ph.D./M.S. thesis candidates (51 Ph.D., 41 M.S./M.A.). He is the author or coauthor of 762 journal and conference papers including 13 textbooks in the field of speech processing and language technology, signal processing for vehicle systems, coauthor of textbook *Discrete-Time Processing of Speech Signals* (IEEE Press, 2000), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and the lead author of the report "The Impact of Speech Under Stress on Military Speech Technology", (NATO RTO-TR-10, 2000). His research interests include digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, signal processing for hearing impaired/cochlear implants, robust speech recognition with emphasis on machine learning and knowledge extraction, and in-vehicle interactive systems for hands-free human–computer interaction. In April 2016, he was awarded the honorary degree Doctor Technices Honoris Causa from Aalborg University, Aalborg, Denmark, in recognition of his contributions to the field of speech signal processing and speech/language/hearing sciences. He was recognized as the IEEE Fellow in 2017 for his contributions in "Robust Speech Recognition in Stress and Noise," International Speech Communication Association (ISCA) Fellow in 2010 for his contributions on research for speech processing of signals under adverse conditions, and was the recipient of the Acoustical Society of America's 25 Year Award (2010) in recognition of his service, contributions, and membership to the Acoustical Society of America. He is currently the ISCA President (2017–2019; re-elected 2020–2022) and a member of the ISCA Board. He is also the Vice-Chair on U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain (2015–2017; 2018–2021). Previously, he was the IEEE Technical Committee (TC) Chair and a Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) (2005–2008; 2010–2014; elected IEEE SLTC Chairman for 2011–2013, Past-Chair for 2014), and elected ISCA Distinguished Lecturer (2011–2012). He was also a member of the IEEE Signal Processing Society Educational Technical Committee (2005–2008; 2008–2010); a Technical Advisor to the U.S. Delegate for NATO (IST/TG-01); the IEEE Signal Processing Society Distinguished Lecturer (2005–2006), an Associate Editor for the IEEE transactions on speech and audio processing (1992–1999), an Associate Editor for IEEE signal processing letters (1998–2000), an Editorial Board Member for IEEE *Signal Processing Magazine* (2001–2003); and a Guest Editor (October 1994) for the Special Issue on Robust Speech Recognition for the IEEE transactions on speech and audio processing. He is currently an Associate Editor for Journal of the Acoustical Society of America, and was on the Speech Communications Technical Committee for the *Acoustical Society of America* (2000–2003). He was the recipient of the 2020 Provost's Award for Excellence in Graduate Student Supervision - University of Texas-Dallas and the 2005 University of Colorado Teacher Recognition Award as voted on by the student body. He organized and served as the General Chair for ISCA Interspeech-2002, September 16–20, 2002, a Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX, USA, March 15–19, 2010, and the Co-Chair and Organizer for IEEE SLT-2014, December 7–10, 2014, Lake Tahoe, NV, USA. He will be the Technical Program Chair for IEEE ICASS-2024, and the Co-Chair and Organizer for ISCA INTERSPEECH-2022.

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 16,2021 at 03:51:03 UTC from IEEE Xplore. Restrictions apply.