Letters in High Energy Physics
ISSN: 2632-2714
Volume 2024

Implementation of Speaker Identification and Speaker Emotion Recognition System

[1]Ravi Shankar. D, [2]Manjula R.B.
[1]Research Scholar, School of ECE, REVA University
ravishankard@reva.edu.in
[2]Professor, School of ECE, REVA University
manjula.rb@reva.edu.in

Abstract:

Audio classification incurs unique difficulties in speaker recognition and human emotion detection, which have applicable relevance to the real world. This paper introduces a novel multimodal solution to the two challenges of speaker verification and sentiment detection in a customer service call centre setting. For speaker recognition, utilizing a small subset of the LibriSpeech Library, features are extracted via Mel-frequency cepstral coefficients (MFCCs). A three-layer Long Short-Term Memory (LSTM) architecture using triplet loss for training produces an Equal Error Rate (EER) of 6.89%, demonstrating efficacy and precision. Simultaneously, we also conduct emotion detection on the RAVDESS dataset via CNN to classify eight feelings the emotions proposed by Ekman, plus neutral and relaxed resulting in an F1 score of 0.85. This contribution demonstrates that such deep learning approaches can be applied in the real world for telephone speaker authentication and help centers, as speaker verification and emotion detection provide additional meaning to what is being conveyed.

Keywords: Convolutional Neural Network (CNN), Equal Error rate, Speaker Authentication, MFCC, LSTM.

1. Introduction

The significance of speaker recognition and emotion detection pertains to multiple use-case scenarios. These features are necessary where advancements in real-time and human-computer interaction occur[1][2]. From security access to empathetic response, the need for speaker identification and emotion detection extends across various sound processing functions. As society becomes more digitally advanced and voice interactive from customer service numbers to security access to voice-activated smart home devices these two developments have additional functional opportunities for machine learning[3][4]. In addition, the research compilation notes that prior research using MFCC features for speaker identification has positive outcomes and that LSTMs can appropriately learn time-variant speech features over time[5]. Other research notes that CNNs correctly identify when a person is attempting to express emotion through their voice[6]. However, with such progress on either side, much of the research delves into experimentation on either speaker identification or emotion classification without consideration of both and without acknowledging the conversion process in the real world, like hearing a known person speaking to you happily when they sound like themselves[7][8]. However, despite the advancements in speaker recognition and emotion detection, very few studies exist that attempt to utilize them simultaneously in one system. Either researchers independently attempt to increase the stand-alone recognition rates, or they attempt to create an extremely advanced emotion detection system[9][10]. In addition, while the speaker recognition systems hold up relatively well across most environments, there is still research lacking in multimodal and noisy environments where there are too many persons and where one person can exhibit multiple emotions. This research evaluates a deep learning multimodal system for speaker and emotion recognition to solve the issues discussed in varying emotional and acoustic environments. Therefore, the importance of this study comes from practical implementation, for a system that needs speaker and emotion recognition simultaneously, and from a field where such integration is goal-oriented and beneficial, as doing both speaker recognition and emotion recognition is feasible and practically useful. This research aims to develop and evaluate a multimodal system that combines speaker identification and emotion recognition. to solve the issues presented by varying emotional and acoustic environments. Therefore, the importance of this study comes from practical implementation, for a system that needs speaker and emotion recognition simultaneously, and from a field where such integration is goal-oriented and beneficial, as doing both speaker recognition and

emotion recognition is feasible and practically useful. The proposed system employs MFCCs as acoustic features for speaker recognition and a three-layer LSTM network for speaker identification trained with triplet loss to improve speaker differentiation accuracy. The system yielded a 6.89% EER score, which denotes acceptable reliability across different sound environments. Simultaneously, the author employed CNN for emotion recognition trained on RAVDESS, a dataset containing recordings of eight different emotional states. This model yielded an F1 score of 0.85, which signifies accurate emotional recognition that would greatly benefit customer service interactions. The structure of the paper is as follows. The Introduction section outlines the importance of both speaker and emotion recognition. The literature review includes related works as well as the omission of both recognition tasks. The methodology includes the datasets, models Used, and training process. The results and discussion section presents the performance of the system and how it compares to other approaches. Finally, the Conclusion highlights the study's key findings and suggests future research directions.

2. Literature review

The advancements in acoustic feature extraction and processing merely in the past year for text-independent speaker recognition and speechemotion recognition show that features created now are already cutting edge, and the accuracy of performance for systems over the past several years has shifted many systems. For speaker recognition, for instance, text-dependent systems have adjusted beyond mere access to where they've been restricted in the past. System accuracy has increased using MFCC with LSTM networks[11], 3DCNN networks with LSTM architecture, and LPC with Log-Mel spectrum to derive acoustic features that are processed through an LSTM architecture[12]. Where the architectures differ, however, are pre-trained weights, performance, and sequential pooling. For example, a wav2vec2-based architecture for speaker recognition, where either a single-utterance classifier or an utterance-pair classifier achieves better results than the traditional approaches[22]. Where many approaches leverage other networks, AutoSpeech takes VGG-Net and ResNet, operating from examined classical features[23], to determine the most effective operation of neural cells to establish a custom CNN architecture for speaker recognition. The Additive Margin MobileNet1D (AM-MobileNet1D) is based on portability, meaning its resource footprint is small[24]. For example, this architecture requires only 11.6MB of space, while SincNet and AM-SincNet need 91.2MB; it runs 7 times faster and with 1/8 of the parameters, advantageous for mobile applications where processing capabilities are limited. As for speaker recognition, we constructed an LSTM model with MFCCs through a three-layer structure with triplet loss and achieved a surprisingly low Equal Error Rate. Such a stable, effective model makes it suitable for deployment. The range of models utilized for speech emotion recognition tasks spans from parallel CNNs to Transformer encoders applied to different means of data entry and processing[13]. For instance, one group of researchers uses known findings about MFCCs, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast to input 1D CNNs and create emotion detection through audio alone without any additional visual aids[14]. However, these results demonstrate that non-specialized CNNs do not acquire large emotional features on a wide scale adequately. Thus, the Global-Aware Multi-scale (GLAM) neural network uses convolution kernels of scale to acquire multi-scale feature representation, while the global-aware fusion module acquires globally salient emotional features[25]. Where new advancements in Speech Emotion Recognition (SER) exist, they sidestep an analogous progression but rather position emotions as discrete occurrences with beginnings and endings addressing the predicament of "When does a certain emotion occur?" Ours is effective and efficient [26]. It takes in audio sound bytes with the intent of a CNN-based architecture replicating the appropriate higher-order temporal and frequency patterns for emotional detection. There exist 2 Conv1D layers with ReLU activation, dropout (0.2), and Softmax classification output these layers are known for highly accurate classification with relatively simplistic composition.

3. Methodology:

In this section, we present the methodology employed to conduct our multimodal addressing two critical tasks: First, for Speaker Recognition, we take the LibriSpeech dataset and extract MFCC for our audio features. Second, for Speech Emotion Recognition, we take the RAVDESS dataset and use the original audio files as our input. We use a three-layer LSTM architecture for our Speaker Recognition task, which was optimally trained via hyperparameter tuning using triplet loss. We use a CNN-based architecture for our emotion recognition

task. The following subsections explain our process for each task—from data acquisition to architecture to evaluation.

3.1 Feature Extraction:

To prepare the sound waves being input into the neural network, they would need to be processed and conditioned into a feature set. Thus, we would be looking for the extraction of Mel Frequency Cepstral Coefficients (MFCCs) features as this is the most common and their relevance is to speech signals more so than phonetic information. Figure 1 shows the extraction of MFCC features from a speechsignal and notates each block entry to the right of it for what it accomplishes in the degradation.

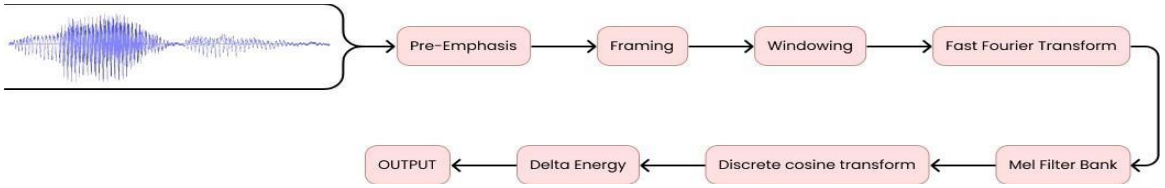


Fig. 1 MFCC block diagram

a) Pre-emphasis: The initial step involves applying pre-emphasis to the audio signal, achieved through a first-order high-pass filter. This filter is typically implemented using a straightforward difference equation:

$$y[n] = x[n] - \alpha \cdot x[n-1] \quad (1)$$

In this equation:

y[n] is the output signal after pre-emphasis.
x[n] is the input audio signal.
α is the pre-emphasis coefficient where $0.9 < \alpha < 1.0$.

This equation, in effect, raises the amplitude of the high-frequency parts of the signal. It does so by taking the current sample, x[n], and subtracting a percentage of the previous sample, x[n−1], from it so that higher frequency changes between samples are more pronounced.

b) Framing: This implies that the speech signal is contained in 20-30 ms windows with overlapping successive frames of N (N>M) where typical M=100 and N=256. This framing is necessary because speech is a time-varying signal, but over short time intervals, its properties remain fairly constant, allowing for short-time spectral analysis.

c) Windowing: To maintain signal continuity,each of the frames mentioned above is multiplied by a Hamming window. This windowing process minimizes

$$Y[n] = x[n] * w[n] \quad (2)$$

Here, w[n] represents the window function applied to the signal.

d) Fast Fourier Transform: FFT is a method used to transform a signal from the time domain into the frequency domain. This transformation allows us to obtain the magnitude frequency response of each frame. The output of this process is a frequency spectrum.

e) Triangular band pass filters: To obtain a smoother magnitude spectrum and reduce the dimensionality of features, we multiply the magnitude frequency response by a set of 20 triangular bandpass filters. These filters are typically based on the Mel scale, and the Mel frequency can be calculated using the formula:

$$Mel(f) = 1125 * \ln(1 + f/700) \quad (3)$$

spectral distortion by gradually reducing

f is the frequency in Hertz that you want to convert to the Mel scale.

f) Discrete cosine transform: We apply Discrete Cosine Transform (DCT) [16] to the 20 log energy values (E_k) obtained from the triangular bandpass filters, which results in L mel-scale cepstral coefficients. The DCT formula is asfollows:

the signal to zero at both the beginning and end of each frame. The operation is expressed as:

$$C_m = \sum_{k=1}^{N} \cos\left[m * (k - 0.5) * \pi/N\right] * E_k, \quad m=1,2\ldots\ldots L \tag{4}$$

In this formula, N represents the number of triangular bandpass filters (usually 20), and L is the number of mel-scale cepstral coefficients (typically 12). DCT transforms the frequency domain into a time-like domain known as the quefrency domain. These features are referred to as the mel-scale cepstral coefficients (MFCCs).

For speech recognition, MFCCs alone can be used, but to improve performance, the log energy can be added, and delta operations can be performed on these features. This enhanced feature set is commonly employed for more accurate speech recognition.

1. Log energy: This feature calculates the energy content of the audio signal within a frame.

2. Delta cepstrum: These derivatives, calculated as velocity and acceleration, provide information about how the energy and MFCC values change over time.

$$\Delta C_m(t) = [\sum^{i} = -M^{M} C_m (t+\tau)\, \tau] / \sum^{i} = -M^{M} \tau^2$$

$$c \qquad\qquad c \qquad (5)$$

Here, "M" typically has a value of 2. When we add velocity as a feature, the total feature dimension becomes 26. If we include both velocity and acceleration, the feature dimension increases to 39.

3.2 Speaker Recognition Model

The architecture of our speaker recognition model is LSTM-based. LSTMs are effective for this type of learning because speaker recognition is based on sequential information, such as audio waveforms and phoneme information. LSTMs function well with this type of sequential learning and have the potential to understand long-term dependencies inherently found in such time series data used for this audio-based, time-dependent task. Therefore, LSTMs excel at retaining information over extended time intervals, allowing them to better capture the contextual details within sequential data [17]. We implemented triplet loss for enhanced accuracy of the model, which gives feedback to the model to lower embeddings of the same speaker by one distance while raising the distance for any other embeddings found within the same feature space. While this is an extra step for the model to calculate, in the long run, it provides much better learned differentiation for speakers. LSTM does not have the vanishing gradient problem, allowing it to learn gradients for much longer sequences than a traditional RNN; therefore, it is a better trained model over longer sequences.

The architectural design of an LSTM cell as shown in figure 2, forms the basis of its four essential components, i.e., the input gate, the forget gate, the cell state, and the output gate. Among them, the input gate determines how much of the current input should be allowed to be added to the cell state and the forget gate [18] decides what portion of the old cell state will influence our new cell state. The cell state itself functions as long-term memory and stores information that has been repeatedly used in the LSTM circuit. The output gate decides what parts of the cell state will be passed to the next layer in the network. These gates are expressed using through sigmoid neural network layers, which have values between 0 and 1 that act as information flow regulators for each gate.

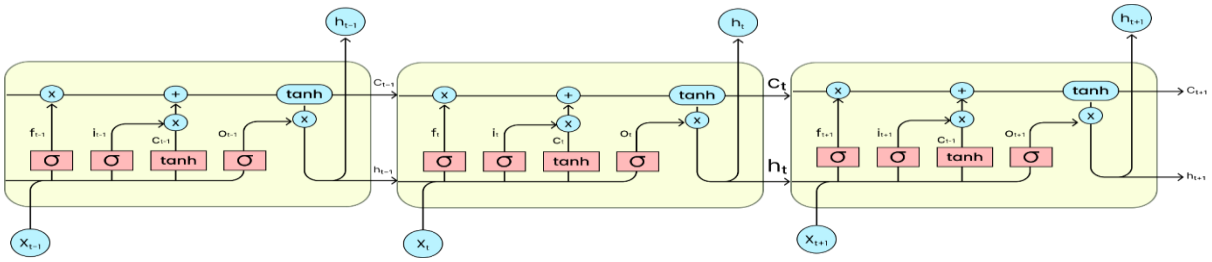


Fig. 2 LSTM cell architecture

An LSTM cell operates in such a way that, first, an input gate determines how much of the current input should be contributed to the cell state. Next, a forget gate determines how much of the last cell state should be kept in memory or, instead, intentionally "forgotten." Subsequently, the cell state is either added to or subtracted from based on the previously

made determinations of the input and forget gates. Finally, the output gate determines how much of thisupdated cell state will be passed on to the subsequentlayer.

The following equations [5] describe the architecture of an LSTM cell:

Input gate: $i_t = \sigma (W_i [x_t, h_{t-1}] + b_i)$ (6)

Forget gate: $f_t = \sigma (W_f [x_t, h_{t-1}] + b_f)$ (7)

Output gate: $o_t = \sigma (W_o [x_t, h_{t-1}] + b_o)$ (8)

Candidate cell state: $\tilde{c}_t = tanh (W_c [x_t, h_{t-1}] + b_c)$ (9)

Cell state: $c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$ (10)

Output: $h_t = o_t * tanh (c_t)$ (11)

i_t = represents input gate.
f_t = represents forget gate.
o_t = represents output gate.
σ = represents sigmoid function.
w_x = weight for the respective gate(x) neurons.
h_{t-1} = output of the previous lstm block (at timestamp t - 1).
x_t = input at current timestamp.
b_x = biases for the respective gates(x).
c_t = cell state (memory) at timestamp(t).
$\tilde{c}_t$ = represents candidate for cell state at timestamp(t).

But in addition to these main components, LSTM cells also feature what's called a hidden state, which is a vector of data concerning what has been learned about the sequence up until that point and, similar to the cell state, it is transmitted to subsequent timestamps. The power of LSTMs is that they can encode and assess such long-term dependencies across temporally separated sequences. We execute speaker recognition using LSTM cells as shown in Fig. 3. We apply Mel-frequency cepstral coefficients (MFCCs) to the audio streams. The MFCCs are applied to the audio sample as they best capture the spectral and temporal dynamics of the audio sample. They transform an audio signal into a frequency domain while simultaneously capturing the temporal domain by segmenting the audio signal into small, overlapping "windows." As for the spectral dynamics, these are observed as the shape of the MFCCs; they show where energy is present in certain frequency bands, which helps differentiate among certain speakers. As for the temporal

dynamics, these are observed over time as the person is speaking; they help differentiate different accents, differences in intonation, and speed of speech, which are all unique to certain speakers. Furthermore, MFCC features must acknowledge the temporal characteristics of speech. These features are a blend of spectral and temporal information, thus making it more likely for the model to distinguish between subtle differences in how each speaker may sound and, subsequently, learn their inherent speech patterns. Therefore, this creates a more effectively trained feature set for speaker recognition that operates in conjunction with its capacity to learn persons by voice despite variations in accents, quality, or presentation. Following this process, the MFCC characteristics are sent to a 3-layer LSTM network. LSTM networks are particularly effective for sequential data, which is essentially what time-dependent spoken data is. Three layers are effective due to accelerated training, and this layer arrangement gives excellent speaker identification results. These layers identify all the subtle distinctions required for accurate identification. We implement a triplet loss function to train our model, whereby intra-speaker variance reduces (distance between anchor and positive) while inter-speaker variance increases (distance between anchor and negative). Therefore, our LSTM-based solution is not only holistic but incredibly efficient for speaker recognition.

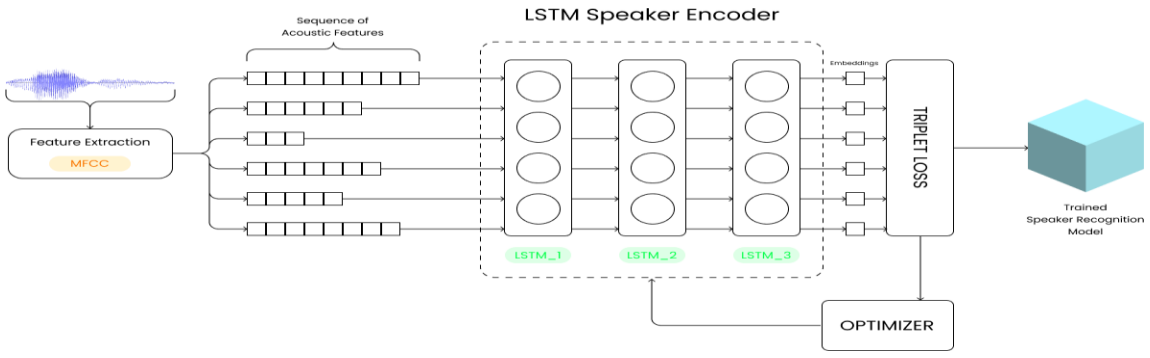


Fig. 3 Proposed Model for Speaker Recognition.

3.2.1 Speaker Authentication:

There are various components of authentication for security and accuracy in use across multiple dimensional applications. For example, one step involved with authentication is speaker recognition, which enables speaker verification and speaker identification based on distinct features of a voice. Our speaker authentication model implements a basic audio matching technique to assess whether the claim of the test sample speaker is that of the enrolled speaker's claim versus that of the test sample's claim. In short, we are trying to match audio from the enrolled speaker with that of the testsample. It all starts with enrollment. Enrollmenthappens with the user's specific voiceprint during thefirst use. The enrolled speaker reads a fixed prompt, which is MFCC feature extracted and LSTM modeled. The audio file generated from the enrollment process is the system's adjusted version of the user's voice at that point in time. From then on, the LSTM turns those features into a d-vector, which is stored for accurate recognition the next time. Yet at enrollment and for testing, a different audio is used for testing. This testing audio is, however, associated with the person trying to verify his identity to see if he is who he claims to be. Some MFCC features are taken from this testing audio and input into the same LSTM network to create a d-vector [19] for this person's identity. The intention is to analyze whether this d-vector from the testing audio is comparable to the d-vector from the enrolled speaker. We evaluate this using a cosine similarity metric [20]. We compute the cosine similarity of the d-vector for the test sample and thed-vector of the enrolled speaker. The output value indicates how similar two voices are. If the value is greater than our threshold of 0.8, we accept this speaker as the enrolled speaker. This threshold is setto yield the greatest accuracy without excessive falsereject and false acceptance rates. Should the threshold be higher, authentication is stricter, meaning only those who match up exactly will be letin; however, it will false reject those speakers who are slightly different, but still reputable. If the threshold is lower, excess people are let in (both good and bad) because the system gives them access with slight different allowances. 0.8 is the Goldilocks value of basically everything.

The results of the tests (1 to 3) are shown below in Fig. 4, Fig. 5 and Fig. 6:

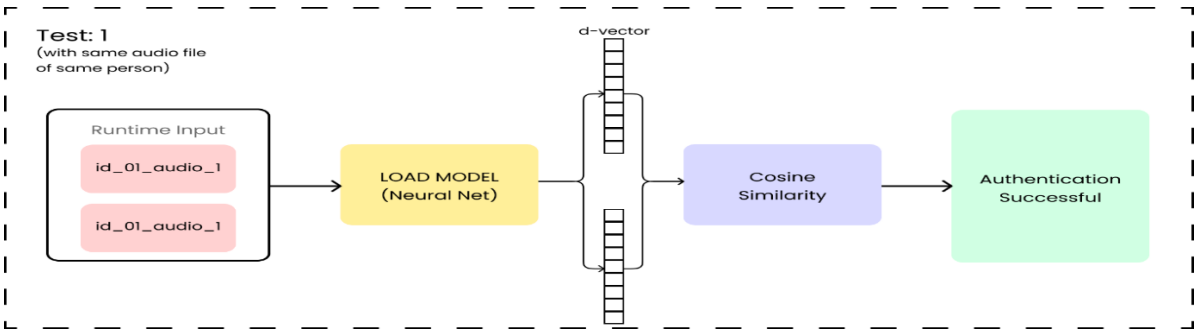


Fig. 4 Test: 1

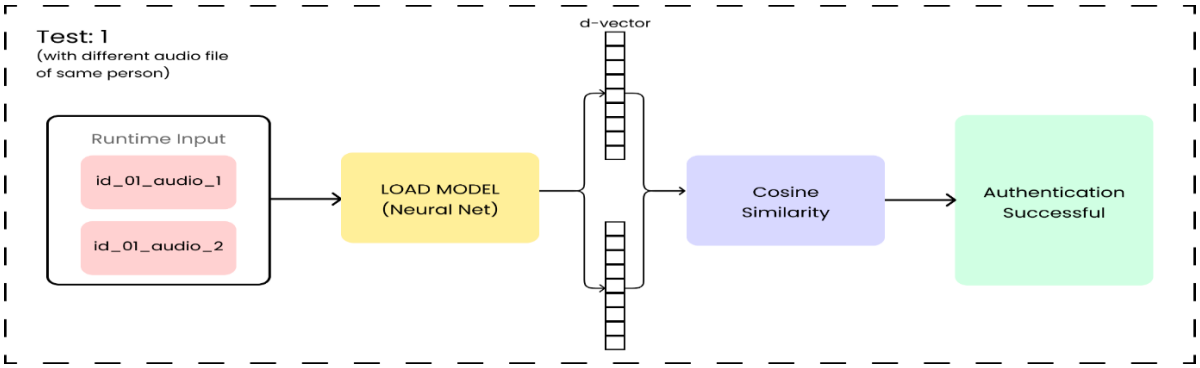


Fig. 5 Test: 2

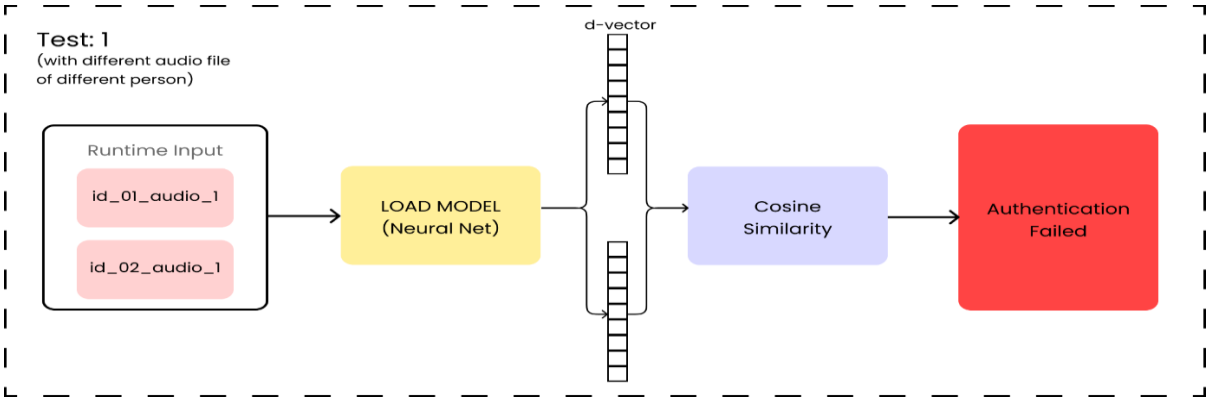


Fig. 6 Test 3

3.3 Speech Emotion Recognition

In our research framework, we employ a Convolutional Neural Network (CNN) as shown in Fig. 7 to tackle the task of emotion classification based on audio features. Our baseline model consists of one-dimensional convolutional layers, integrated with crucial components such as dropout layers,

batch normalization, and activation functions. The input layer of our CNN is designed to accept 40 × 1

arrays, corresponding to the audio feature representations extracted from sound files. Following this, the network initiates with an initial convolutional layer featuring 64 filters, each with a kernel size of 5 and 'same' padding. This layer employs the Rectified Linear Unit (ReLU) activation function and includes dropout with a rateof 0.2 to mitigate overfitting. The convolution operation can be mathematically represented as:

$$Z^{[1]} = X * W^{[1]} \quad (13)$$

where $Z^{[1]}$ is the output feature map at the first convolutional layer, X is the input feature map, and $W^{[1]}$ is the convolutional filter at the first layer. The ReLU activation is applied as:

$$A^{[1]} = \max(0, Z^{[1]}) \quad (14)$$

A subsequent convolutional layer follows, comprising 128 filters and mirroring the configurations of the preceding layer. It alsoemploys ReLU activation and dropout at the same rate, contributing to the model's resilience. Upon the convolutional layers, a flattening layer transforms the output into a one-dimensional tensor for further processing. Subsequently, a fully connected layer adapts its size according to the number of distinct emotion classes, serving as the output layer. This layer incorporates a softmax activation function to compute class probabilities:

$$\text{Softmax}(Z^{[L]})_i = e^{Z^{[L]}_i} / \sum_{j=1} e^{Z^{[L]}_j} \quad (15)$$

where $\text{Softmax}(Z^{[L]})_i$ is the probability of class *i*, and $Z_i^{[L]}$ is the logit (pre-activation) for class i. In terms of model training, we configure it with categorical cross entropy loss:

$$L(y, \hat{y}) = - \sum_{i=1}^{c} y_i \log(\hat{y}_i) \quad (16)$$

where C is the number of classes, y_i is the true label (one-hot encoded), and $\hat{y}_i$ is the predicted probability for class *i*. We use the Adam optimizer and accuracy as the evaluation metric.

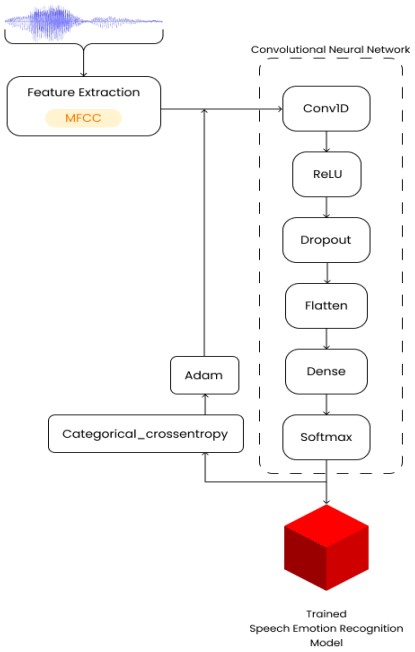
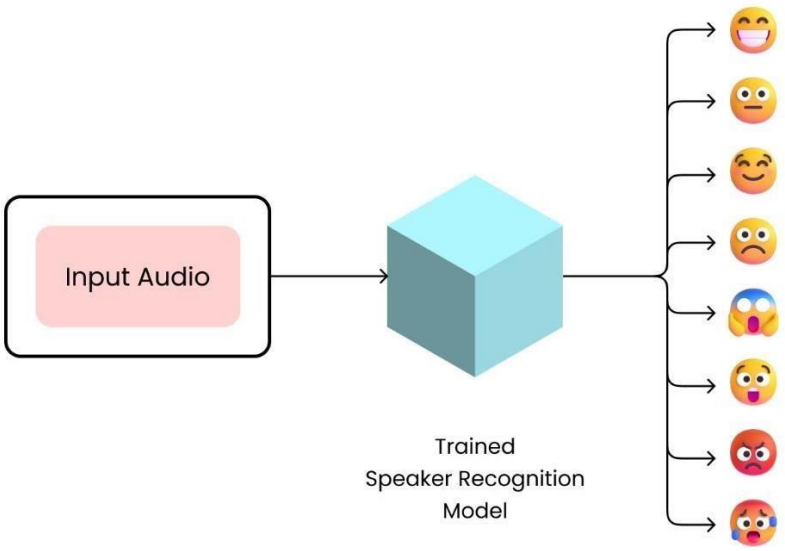


Fig. 7 Proposed Model for Speech Emotion Recognition

The accuracy is used as the evaluation metric and the Adam optimizer. Training occurs for 50 epochs with a batch size of 16 and real-time validation on the validation set. The metrics used to assess the model to see if it can predict emotions from audio are confusion matrices and classification reports.

3.3.1 Speech Emotion Recognition (SER) for Customer Service:

One of the most practical applications of these SER systems in the realm of customer service is customer engagement. Since these systems are based on the detection and classification of emotions expressed via vocalization, it stands to reason that they would work best assessing how customers feel and react emotionally over virtual communication. Thus, for instance, our system shows high levels of accuracy across the board for all emotional classifications from happy, joy, and satisfaction to frustration, anger, confusion, and surprise. A more holistic schema of emotion detects issues more effectively in online customer service situations. Take customer service, for example. The use of SER makes customer problems more visible, sooner, and with a more collaborative approach to resolution. An ongoing evaluation and check-in of feelings create better interpersonal relationships with customers and a more customer-oriented attitude of the big company since customer service is there for on-demand access and cross-collaborative integration of all—emotional and practical. Furthermore, as so much customer service is rendered online these days, an active customer facilitating their existence in an SER world is more likely to receive empathetic, on-the-spot, and successful customer service that simultaneously solves the emotional aspect of the purchase.



Fig, 8 Emotional States Detected by SER

3.4 Multimodal Architecture:

Our unique contribution is a robust multimodal architecture, detailed in Fig. 9, which combines Speaker Recognition and Speech Emotion Recognition for better processing of audio, as it identifies the speaker and, at the same time, identifies his/her emotional state. This is beneficial for call centers with privacy and operational needs and for customer service speaker recognition that accommodates real-time, nuanced emotion recognition during online meetings. The system includes Speech Emotion Recognition(SER) [21] along with Speaker Recognition in the sample audio assessment. Thus, it functions on a dual level of knowing who is speaking and at the same time, how they're feeling. Thus, it's a more comprehensiveview of client needs.

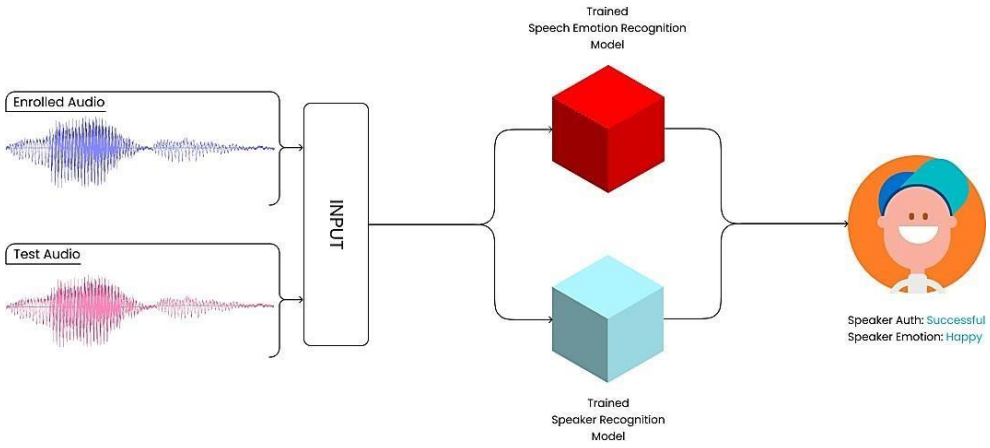


Fig. 9 Proposed multimodal approach

4. RESULTS & DISCUSSIONS:

This chapter presents and analyzes the results of the speaker recognition and speech emotion recognition models, respectively. It begins with a speaker recognition ablation study that demonstrates how the LSTM-based method developed in this text proves superior to all other architectural baselines set as control. The LibriSpeech corpus is utilized for all testing, which is a large, diverse speaker database for the most comprehensive testing atmosphere. Next are the results of the SER experiments using a CNN-based approach. These were done using the RAVDESS database to determine whether the model accurately predicted what the final emotional state would be for a given spoken word. This is a quantitative assessment of the two models for accuracy and feasibility of integration for future speaker recognition and realistic applications of emotion detection.

4.1 Experimental Setup

An NVIDIA GeForce RTX 3060 GPU, 6144 MiB (6 GB) VRAM was utilized for training and testing. The experimental platform was Ubuntu 22.04 LTS. Both models are developed under Python 3.10.12. The speaker recognition model was developed using PyTorch, while the speech emotion recognition model was developed using TensorFlow.

4.2 Datasets

The speaker recognition model was evaluated on the LibriSpeech corpus. LibriSpeech is an English corpus containing approximately 1,000 hours of audio spoken at a sampling rate of 16 kHz. It is derived from LibriVox audiobooks and contains naturally occurring segmentation and alignment. The proposed model was developed on chunks of LibriSpeech—train-clean-100, train-clean-360, and train-other-500—with SpecAugment during training to ensure model development efficacy and generalization. The evaluation of the model occurred on the LibriSpeech test-clean chunk. Table 2 illustrates the developed and evaluated corpuses. The RAVDESS database was used to train the speech emotion recognition task. The RAVDESS database contains 7,356 audio and video files created by 24 professional actors (12 female, 12 male). Each actor recorded his or her voice saying two lexically matched sentences and one neutral sentence. All actors use a neutral North American accent. The emotionally rendered stimuli include: calm (spoken and sung), happiness (spoken and sung), sadness (spoken and sung), anger (spoken and sung), fear (spoken and sung), surprise (spoken only), and disgust (spoken only)—the latter two alsohave two lexically matched songs. Each emotion hasa neutral face and is performed at a mean and stronglevel of intensity. In the experimental setup, we usedthree types of files: (1) audio-only files in .wav format (16-bit, 48kHz), (2) audio-video files in .mp4 format (720p H.264 video with AAC 48kHz audio),and (3) video-only files without sound. The dataset was split, with 65% used for training and 35% for testing.

Table 1: Splits of LibriSpeech used in experiments

LibriSpeech datasets	Hours	Per-speaker minutes	Female Speakers	Male Speakers	Total Speakers
test-clean	5.4	8	20	20	40
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-clean-500	496.7	30	564	602	1166
TOTAL	**966.3**	**88**	**1148**	**1230**	**2378**

4.3 Speaker Recognition Results:

This section provides an in-depth analysis of the speaker recognition experiments using the proposed LSTM-based model. Where accuracy is critical for speaker recognition, identifying a different speaker can compromise the intention and security of voice-based user verification systems. Therefore, refer to Table 1 for results of accuracy from various architectures via EER and Table 2 for speaker recognition architecture comparison within the field. The proposed architecture for LSTM with processing via MFCC features contains an EER of 6.89%, the lowest, which not only means it essentially recognizes speakers at a high rate but also makes it the preferred architecture for any such application with a need for speaker verification and security. In comparison, LPC EER with the LSTM model is 9.14%, and Log-Mel EER with the LSTM model is 7.89%. Thus, the features related to distinguishing between human-generated voices suggest how critical feature representation is for speaker recognition. The ability to outperform the others suggests that the effectiveness of the proposed model was due to MFCC features and LSTM processing for speaker recognition. The practical relevance of the LSTM-based model further demonstrates that this approach would work in practical, real-world settings where speaker recognition exists for increased security and authentication.

Table 2: Comparison of Speaker Recognition Performance

Architecture	EER in %
Proposed Model (LSTM + MFCC)	6.89
LSTM + LPC	9.14
LSTM + Log-Mel	7.89

4.4 Speech Emotion Recognition Results:

This study is pertinent to Speech Emotion Recognition (SER), which has been gaining popularity across various applications like customer service, sentiment analysis, and human-machine interaction. The subsequent sections presentfindings from the emotion recognition experiments utilizing a CNN approach. The confusion matrix (Table 3) demonstrates how well the model was ableto classify the eight overall emotion categories like Angry, Happy, Neutral, Unhappy, Relaxed, Fearful, Disgusted, Surprised. There is a way to measure such classification effectiveness from the confusion matrix. For instance, rage was assessed correctly in the first cell of the confusion matrix 173 times; it was assessed incorrectly, however, as Happy,Unhappy, and Neutral. It had a decent time assessingRelaxed and Fearful but an ineffective time assessing Happy and Surprised. Table 4 assesses theeffectiveness and accuracy of the model.

Table 3: Confusion Matrix

Angry	[[173	8	1	7	1	0	1	1]
Happy	[1	102	7	4	2	2	4	1]
Neutral	[0	18	212	3	11	7	7	6]
Unhappy	[3	7	9	227	5	14	2	8]
Relaxed	[1	1	9	5	216	10	4	6]

Fearful	[3	4	7	17	3	200	1	6]
Disgusted	[0	5	1	4	7	7	171	2]
Surprised	[0	0	7	1	3	4	10	165]]
	Angry	Happy	Neutral	Unhappy	Relaxed	Fearful	Disgusted	Surprised

Table 4: Performance of the model on the test set for each class.

Emotion	precision	recall	F1-score	support
Angry	0.96	0.9	0.93	192
Happy	0.7	0.83	0.76	123
Neutral	0.84	0.8	0.82	264
Unhappy	0.85	0.83	0.84	275
Relaxed	0.87	0.86	0.86	252
Fearful	0.82	0.83	0.82	241
Disgusted	0.85	0.87	0.86	197
Surprised	0.85	0.87	0.86	190
Accuracy			0.85	1734
Overall avg	0.84	0.85	0.84	1734
Weighted avg	0.85	0.85	0.85	1734

The total weighted average F1 score across the entire dataset was 0.85, providing a true representation of how the model functioned in the grand scheme. In the end, there was 85% efficiency across all emotion categories. These findings about the Speech Emotion Recognition (SER) model validate its performance and effectiveness in identifying different emotions via vocal intonation. Furthermore, the model scored high on F1 for "Angry," "Relaxed," and "Disgusted," which means that by knowing when people are angry and relaxed (more than some of the other categories), this sentiment analysis can be applied in professional settings like customer service, where emotional concerns need to be addressed immediately—and with quality service. The weighted average F1-score of 0.85 suggests that the model performs reliably across the various categories of emotion, which makes applicability and feasibility across various environments. Such an accurate real-world application would mean that nothing would be lost in translation down the line when people use their words motivated by emotion, and it needs to be understood precisely. The consistent decrease in the loss function over the 50 epochs of training, illustrated in Figure 10, indicates that predictionerror was decreasing consistently. Simultaneously, accuracy was increasing, as shown in Figure 11, meaning that relative to its defined parameters, the model was making more appropriate predictions. Therefore, the smoothing of the loss function depicted in Figure 10 and the increase in accuracy demonstrated in Figure 11 suggest that this model transfers successfully on a micro scale during training and successfully generalizes on a macro scale to accurately determine speaker identity through emotion recognition for real-world application in call center environments.

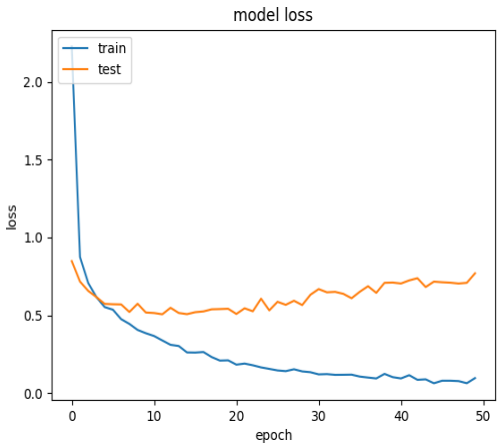


Fig 10. Model loss trajectory over the span of 50 epochs.

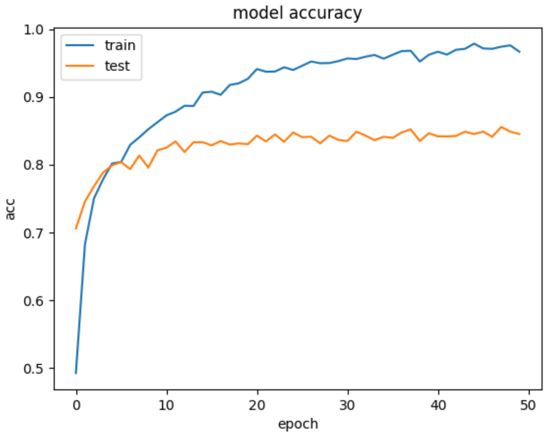


Fig 11. Accuracy trajectory over the span of 50 epochs.

The terminal output that confirms multimodal is shown in Figure 12. The program received two audio files—one TEST UTTERANCE and one UTTERANCE that enrolled since it was from the same speaker—and noted that TEST UTTERANCE sounded like ENROLLED UTTERANCE because it had access to both. It exceeded the 0.8 threshold for cosine similarity calculated beforehand and stated that speaker verification was successful. All of this occurred within a 0.10 second timeframe. In addition, it evaluated the emotion, as well, correctly identifying the emotion of the test audio as "happy" during a 0.19 second emotion detection challenge. Where the quasi-experiment had only one tester involved, though, the model was meant to havemany testers and can authenticate one tester and identify emotion while concurrently trying toauthenticate the speech of other authenticatedtesters.

```
python3 proposed_models.py
Similarity score between TEST_UTTERANCE & ENROLLED_UTTERANCE: 0.899976372718811
Authentication: SUCCESSFUL
Authentication Time: 0.10160565376281738 seconds

Emotions:
0: neutral
1: relaxed
2: happy
3: unhappy
4: angry
5: fearful
6: disgust
7: surprised
PREDICTED EMOTION: [[0.000197 0.00519  0.527822 0.000142 0.461361 0.000317 0.002314 0.002656]]
Index of the maximum value in PREDICTED EMOTION: 2
Predicted emotion of TEST_UTTERANCE is: happy
Emotion Recognition Time: 0.1902909278869629 seconds
```

Fig. 12 Authentication and Emotion Detection Results.

5. Conclusion

By combining multimodal speaker recognition and speech emotion recognition, the paper presents a novel framework capable of training for generalization for speaker verification in security and positive/negative emotion detection in customer service. The speaker recognition module employs a three-layer LSTM network trained by a triplet loss function and obtains an Equal Error Rate (EER) of 6.89% after training and validation on the LibriSpeech database; therefore, it has suitable trustworthiness and precision for voice authentication and security. The trained emotion recognition module employs CNNs for emotion detection and successfully recognizes eight different emotions in RAVDESS with an F1-Score of 0.85; therefore, it has excellent efficiency for telehealth, sentiment analysis, and customer service. Where this multimodal model could be expanded upon is speaker identification and sentiment analysis; however, the audio component is bolstered by more contextualized details. Where this will go for future research is in real-time applications, multilingual applications, applications with ethical/privacy concerns, and more generalizability of HRI beyond this study to other fields.

REFERENCES:

[1] Samia Abd El-Moneim, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy and Fathi E. Abd El-Samie, *"Text-independent speaker recognition using LSTM-RNN and speech enhancement,"* Multimed Tools Appl 79, 24013–24028, 2020.

[2] Seong-Hu Kim, Hyeonuk Nam and Yong-Hwa Park, *"Temporal Dynamic Convolutional Neural Network for Text-Independent Speaker Verification and Phonemetic Analysis,"* ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6742-6746, 2022.

[3] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, *"Librispeech: An ASR corpus based on public domain audio books,"* 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),5206-5210, 2015.

[4] Durairaj, Prabakaran and Sriuppili, S, *"Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation,"* Journal of Physics: Conference Series, 2021.

[5] Christian Bakke Vennerød, Adrian Kjærran and Erling Stray Bugge, *"Long Short-term Memory RNN*," arXiv, 2021.

[6] Emmanuel Maqueda, Javier Alvarez-Jimenez, Carlos Mena and Ivan Meza, *"Triplet loss-based embeddings for forensic speaker identification in Spanish,"* Springer Science and Business Media {LLC}, Volume 35, 18177-18186, 2021.

[7] Anant Singh and Akshat Gupta, *"Decoding Emotions: A comprehensive Multilingual Study of Speech Models for Speech Emotion Recognition,"* arXiv, 2023.

[8] M. G. de Pinto, M. Polignano, P. Lops and G. Semeraro, *"Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients,"* 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2020.

[9] Livingstone SR and Russo FA. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS One, 2018.

[10] EKMAN, P. *"Basic emotions, Handbook of cognition and emotion,"* 45-60, 1999.

[11] Hu, Z.F. & Si, X.T. & Luo, Y. & Tang, S.S. and Jian, F. *"Speaker recognition based on 3dcnn-lstm,"* Engineering Letters, 29, 463-470, 2021.

[12] Q. Xu, M. Wang, C. Xu and L. Xu, *"Speaker Recognition Based on Long Short-Term Memory Networks,"* 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), 318-322, 2020.

[13] Ullah, Rizwan & Asif, Engr. Dr. Muhammad & Ali Shah, Wahab & Anjam, Fakhar & Ullah, Ibrar & Khurshaid, Tahir & Wuttisittikulkij, L. & Shah, Shashi & Ali, Syedmansoor & Alibakhshikenari, Mohammad, *"Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer,"* Sensors, 23, 2023.

[14] Dias Issa, M. Fatih Demirci and Adnan Yazici, *"Speech emotion recognition with deep convolutional neural networks,"* Biomedical Signal Processing and Control, Volume 59, 2020.

[15] P. Manikandan, K. Shrimathi, M. Kiruthika and A. Mubeena, "Speech Recognition using Fast Fourier Transform Algorithm," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), Volume 10, 2022.

[16] Meghanani, Amit & Ramakrishnan, A.G, *"Pitch-synchronous Discrete Cosine Transform Features for Speaker Identification and Verification,"* 2020.

[17] J. Oruh, S. Viriri and A. Adegun, *"Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition,"* in IEEE Access, Volume 10, 30069-30079, 2022.

[18] C. -J. Peng, Y. -J. Chan, C. Yu, S. -S. Wang, Y. Tsao and T. -S. Chi, *"Attention-Based Multi-Task Learning for Speech-Enhancement and Speaker-Identification in Multi-Speaker Dialogue Scenario,"* 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, pp. 1-5, 2021.

[19] Jung, Jeeweon & Heo, Heesoo & Yang, Ilho & Yoon, Sunghyun & Shim, Hye-Jin & Yu, Hajin, *"D-vector based speaker verification system using Raw Waveform CNN,"* 2017 International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2017), 2018.

[20] Jason W. Pelecanos, Quan Wang, Yiling Huang and Ignacio Lopez-Moreno, *"Parameter-Free Attentive Scoring for Speaker Verification,"* The Speaker and Language Recognition Workshop, 2022.

[21] Stephen Obadinma, Faiza Khan Khattak, Shirley Wang, Tania Sidhorn, Elaine Lau, Sean

Robertson, Jingcheng Niu, Winnie Au, Alif Munim, and Karthik Raja Kalaiselvi Bhaskar, *"Bringing the State-of-the-Art to Customers: A Neural Agent Assistant Framework for Customer Service Support,"* Proceedings ofthe 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 440–450, 2022.

[22] N. Vaessen and D. A. Van Leeuwen, *"Fine-Tuning Wav2Vec2 for Speaker Recognition,"* ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7967-7971, 2022.

[23] Shaojin Ding, Tianlong Chen, Xinyu Gong, Weiwei Zha and Zhangyang Wang, *"AutoSpeech: Neural Architecture Search for Speaker Recognition,"* Interspeech, 2020.

[24] Nunes, João Antônio, Macêdo, David, Zanchettin, Cleber, *"AM-MobileNet1D: A Portable Model for Speaker Recognition,"* 2020.

[25] W. Zhu and X. Li, *"Speech Emotion Recognition with Global-Aware Fusion on Multi-Scale Feature Representation,"* ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6437-6441, 2022.

[26] Nunes, João Antônio, Macêdo, David, Zanchettin, Cleber, *"AM-MobileNet1D: A Portable Model for Speaker Recognition,"* 2020.