Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 18 No. 3 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijim.v18i03.43013

PAPER

Convolutional Neural Network Architectures for Gender, Emotional Detection from Speech and Speaker Diarization

Thaer Mufeed Taha[1](✉), Zaineb Ben Messaoud[2], Mondher Frikha[3]

[1]ATISP Research Unit, National School of Electronics and Telecommunications of Sfax (ENET'Com), University of Sfax, Sfax, Tunisia

[2]Higher Institute of Computer Science and Multimedia of Gabès (ISIMG), Teboulbou, Tunisia

[3]Director of 'Advanced Technologies for Image and Signal Processing' (ATISP) Research Unit, University of Sfax, Sfax, Tunisia

thaer.alamer85@gmail.com

ABSTRACT

This paper introduces three system architectures for speaker identification that aim to overcome the limitations of diarization and voice-based biometric systems. Diarization systems utilize unsupervised algorithms to segment audio data based on the time boundaries of utterances, but they do not distinguish individual speakers. On the other hand, voice-based biometric systems can only identify individuals in recordings with a single speaker. Identifying speakers in recordings of natural conversations can be challenging, especially when emotional shifts can alter voice characteristics, making gender identification difficult. To address this issue, the proposed architectures include techniques for gender, emotion, and diarization at either the segment or group level. The evaluation of these architectures utilized two speech databases, namely VoxCeleb and RAVDESS (Ryerson audio-visual database of emotional speech and song) datasets. The findings reveal that the proposed approach outperforms the strategy level in terms of recognition results, despite the real-time processing advantage of the latter. The challenge of identifying multiple speakers engaging in a conversation while considering emotional changes that impact speech is effectively addressed by the proposed architectures. The data indicates that the gender and emotion classification of diarization achieves an accuracy of over 98 percent. These results suggest that the proposed speech-based approach can achieve highly accurate speaker identification.

KEYWORDS

deep learning, gender recognition, speaker diarization, voice recognition, emotional speech

1 INTRODUCTION

Recent studies indicate an increasing demand for speaker recognition across various applications, including security systems, biometric verification, criminal investigations, and customer care [1] [8] [9]. The innate human ability to recognize

Taha, T.M., Messaoud, Z.B., Frikha, M. (2024). Convolutional Neural Network Architectures for Gender, Emotional Detection from Speech and Speaker Diarization. *International Journal of Interactive Mobile Technologies (iJIM)*, 18(3), pp. 88–103. https://doi.org/10.3991/ijim.v18i03.43013

Article submitted 2023-07-19. Revision uploaded 2023-11-27. Final acceptance 2023-11-27.

individuals by their voice, often overlooked, is a crucial aspect of human-computer interactions [2] [3].

Convolutional neural networks (CNNs) have found applications across multiple domains due to their versatile nature, particularly in speech analysis endeavors such as gender detection, emotional detection from speech, and speaker diarization [10].

Gender detection from speech, which is the process of discerning a speaker's gender based on their vocal attributes, has greatly benefited from CNN architectures. Specifically, the convolutional layers extract hierarchical features from speech signals, while the fully connected layers are responsible for gender classification based on these features [6] [7].

The time-delay neural network (TDNN) is another significant architecture, characterized by its dilated convolutions, which allow it to perceive temporal dependencies across multiple time scales. This approach is essential for gender detection, as it enables the capture of long-range contextual information in speech signals.

Emotional speech detection aims to understand the emotions conveyed by a speaker. CNNs have also demonstrated effectiveness in this area. One commonly used architecture for this purpose is the 1D-CNN. The system processes speech in a one-dimensional pattern, using convolutional layers to extract local acoustic attributes and pooling layers to identify the most important features.

Attention-based CNN (ACNN) has recently emerged as a prominent method for emotion detection. Equipped with attention mechanisms, the ACNN can identify crucial regions of an input speech by focusing on pertinent acoustic segments. This precision enhances the capture of emotional indicators, thereby improving classification accuracy.

Speaker diarization, the process of separating an audio stream based on speaker identities, has been another area where CNNs have shown proficiency. A prominent architecture in this context is the time-domain CNN (TCNN), which operates directly on raw waveforms to analyze speaker characteristics. By using 1D convolutions, TCNNs can differentiate between local and global patterns in audio, which enables precise speaker differentiation.

The x-vector system combines the strengths of CNNs and recurrent neural networks (RNNs) to improve speaker diarization. Here, CNNs analyze speech to infer frame-level attributes, which are then utilized by RNNs to identify temporal patterns and produce speaker embeddings. This system has achieved state-of-the-art results in speaker diarization tasks.

Supervised diarization using CNNs represents a new and rapidly evolving research area. While conventional diarization techniques rely on unsupervised clustering or speaker embedding methods, supervised diarization uses annotated data to guide the training of a CNN for diarization-specific tasks. This method includes various CNN architectures designed to meet specific requirements. Primarily, convolutional layers extract distinctive features from audio input, while fully connected layers classify or determine speaker labels for each segment. This network's core capability lies in its ability to identify speaker-specific attributes from inputs and link them to the corresponding speaker labels. The iterative learning process involves inputting audio segments into the CNN, assessing the disparity between predicted and true labels, and optimizing network parameters through backpropagation and gradient descent.

This paper is structured to provide a comprehensive understanding, starting with a detailed overview of CapsNets in Section 2, followed by descriptions of the datasets in Section 3. Section 4 explores the proposed CapsNets-based diarization model, while Section 5 deals with experimental investigation and results in Section 5, concluding with the findings in Section 6.

2 METHODS AND TECHNIQUES

Diarization algorithms, as indicated by scholarly literature, initially relied on methods that did not have formal endorsement.

1. Speaker diarization: The use of CNNs to extract audio data parameters for speaker diarization has gained traction in the field of deep learning [3] [11]. CNNs showcase their ability to identify subtle details such as background noise or variations in speaker accents, showing resilience to minor changes in input data. Another beneficial approach involves RNNs, which capture correlations within the data, thereby enhancing diarization accuracy. In addition to supervised methodologies, unsupervised speaker diarization has been extensively researched. Such unsupervised techniques focus on clustering speakers without prior knowledge of their identities, utilizing clustering and low-dimensional embeddings [12] [4]. The most recent studies have explored contrastive learning and other self-supervised techniques for generating speaker embeddings without relying on labeled data. Across both supervised and unsupervised domains, the application of deep learning methodologies has been crucial in improving speaker diarization accuracy [20] [21].

   A crucial technique in diarization is the i-vector model, which converts input data into low-dimensional speaker embeddings to identify speakers. However, this model requires a large amount of training data to achieve optimal accuracy [22] [23]. Deep neural networks (DNNs) have recently excelled at extracting more discriminative features and clustering in a shared embedding space, thereby significantly enhancing speaker diarization accuracy [24] [25]. A prominent DNN architecture in this field is the deep clustering network (DCN), which is skilled at mapping input audio data for clustering purposes [26] [27] [28].
2. Gender speaker: The intersection of language and gender has been a focal point of extensive research. Predominantly, studies have focused on gender-based linguistic differences across various contexts. Language expresses gender through constructs such as “gendered speech,” including pronouns, vocabulary, and syntax. This section provides a summary of recent literature on this topic. Vocabulary distinctions between genders indicate that women tend to use tag questions and hedging phrases to seek audience validation, while men exhibit more direct speech. Research suggests that men tend to use concise expressions, while women’s linguistic structures are more intricate, possibly reflecting nuanced communication needs [13].

   The complex interplay between language and gender has been the focus of numerous studies, with researchers consistently revealing the subtle ways in which gender influences and is influenced by linguistic patterns [36]. Historically, much linguistic research has focused on the differences in language usage by men and women across various contexts [37].

   The term “gender speaker” encapsulates this concept. It refers to linguistic constructs that encompass pronouns, vocabulary, and syntax used to portray or convey gender. For example, languages often have gender-specific pronouns, which, in turn, influence perceptions and biases related to gender roles [38]. Additionally, vocabulary choices often reveal gendered patterns. Some studies suggest that women might use tag questions such as “don’t you think?” Women tend to use hedging phrases like “kind of” more frequently than men [39]. These linguistic choices often serve to seek validation or soften assertions. In contrast, some studies indicate that male speakers tend to use a more direct and assertive speech pattern [40].

Syntax, another significant aspect of language, is also influenced by gender. Research shows that men tend to use shorter, more direct phrases, while women's linguistic constructions tend to be more elaborate, characterized by longer sentences and intricate structures [40]. This variation in syntactic choices may arise from societal expectations, communication priorities, or a multitude of other factors that intersect with gender [41].

3. Emotional speaker: Given the central role of emotions in human interactions, there has been a surge in linguistic and psychological studies on emotional expression. Communicating emotions involves voice modulation, facial expressions, and vocabulary. This segment reviews current research on speakers' emotions. The authors of [29] describe a semi-automated procedure for constructing a balanced diachronic voice corpus that spans age, gender, and recording periods, with the ultimate goal of creating a rich dataset from the French National Institute of Audiovisuals (INA). Figure 1 illustrates a semi-automated pipeline designed to efficiently extract target speaker segments from a large audiovisual dataset.

In contrast, [30] introduces a novel feature extraction technique using bag-of-audio-words (BoAW) for conversational audio data. They propose an advanced emotion detection model based on RNNs that comprehensively captures conversational context and individual states. Figure 2 illustrates the method, which utilizes the BoAW technique for representing audio features and then feeds these features to the RNN model. The low-level descriptors (LLDs) extracted from audio streams, which are essential for emotion classification tasks, undergo a sophisticated encoding strategy before being input into the RNN model. This emotion identification pipeline is validated using the IE-MOCAP and MELD datasets.

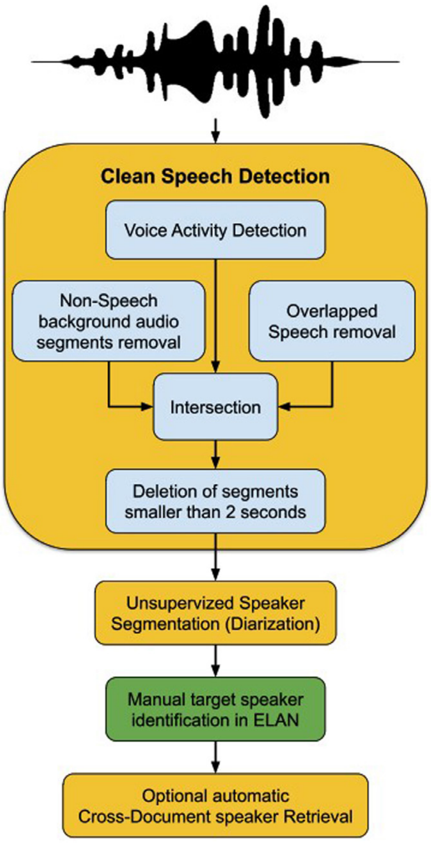

Clean Speech Detection
Voice Activity Detection
Non-Speech background audio segments removal
Overlapped Speech removal
Intersection
Deletion of segments smaller than 2 seconds
Unsupervized Speaker Segmentation (Diarization)
Manual target speaker identification in ELAN
Optional automatic Cross-Document speaker Retrieval

Fig. 1. Flowchart of gender and speaker diarization approach [29]

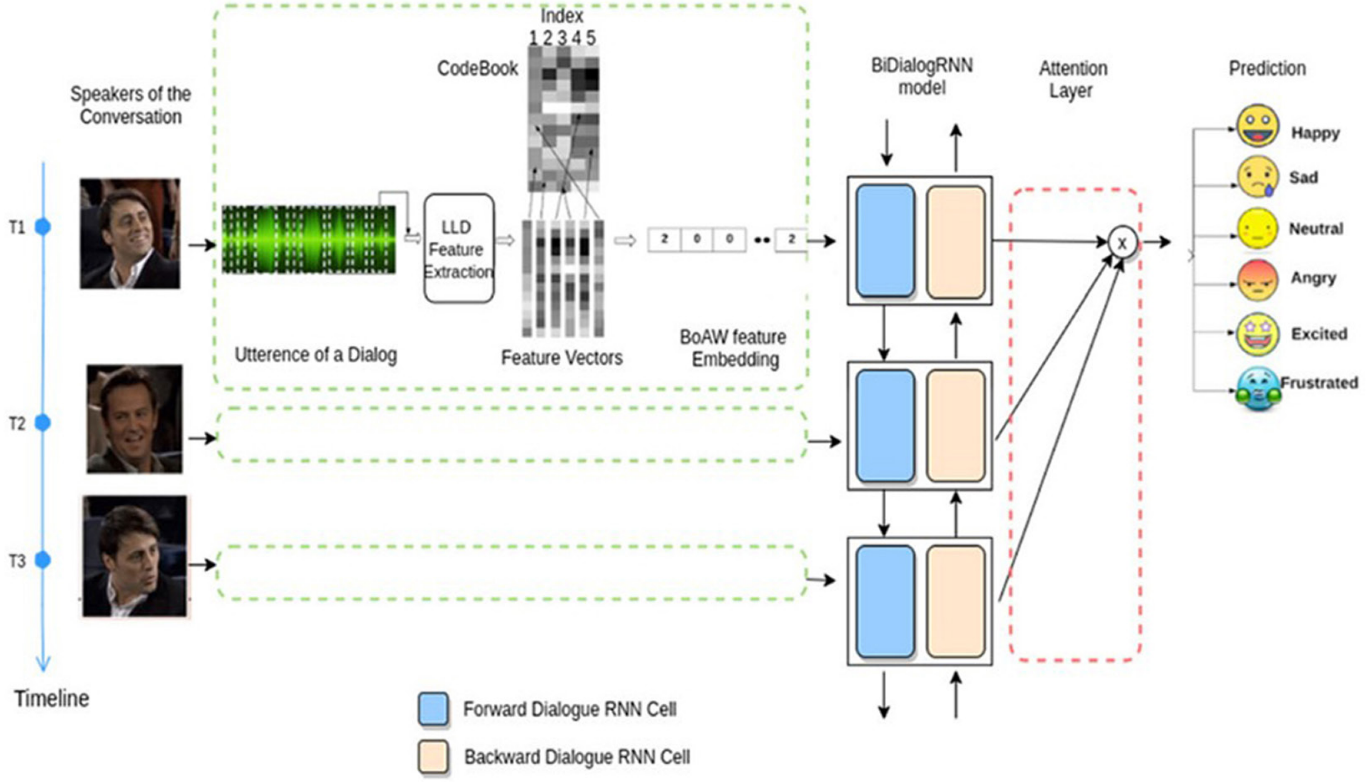


Fig. 2. Flowchart of emotion and speaker diarization approach [30]

3 EMOTIONAL GENDER SPEAKER WITH DIARIZATION MODEL

3.1 Proposed approach

The process of categorizing speech segments based on the speaker's gender and emotional state is referred to as emotional gender speaker diarization. This method considers both gender and emotion to provide a more comprehensive analysis of speech than approaches that treat them as separate diarization tasks [29] [30].

This strategy might lead to more accurate speaker diarization, which is a potential advantage. The clustering method might be more effective in discriminating between speakers if it considers both emotional state and gender rather than just one of them.

It may also provide more nuanced insights into the interaction between gender and emotion in speech, which is another potential advantage. For example, it might be possible to identify patterns in which speakers of a specific gender are more likely to express certain emotions.

However, there may be challenges associated with this strategy as well. For example, it might be more challenging to train a clustering algorithm that considers both gender and emotional state than to train separate algorithms for each aspect. If both aspects are considered simultaneously, the resulting clusters might also be more challenging to interpret.

Figure 3 demonstrates that emotional gender speaker diarization has the potential to be an effective technique for speech analysis, but further research is needed to fully understand its advantages and drawbacks.

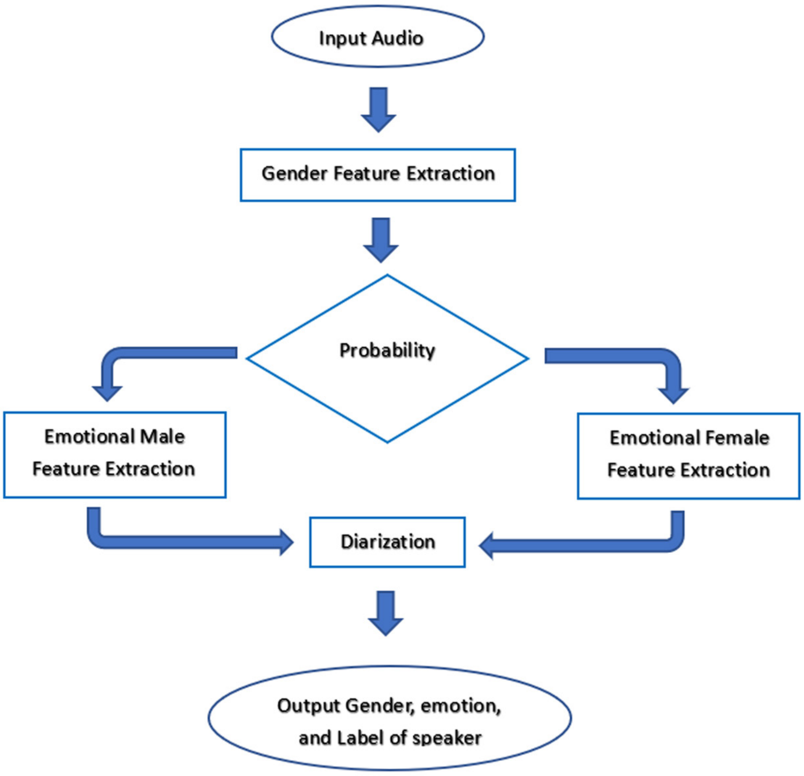

Input Audio
Gender Feature Extraction
Probability
Emotional Male Feature Extraction
Emotional Female Feature Extraction
Diarization
Output Gender, emotion, and Label of speaker

Fig. 3. Flowchart of proposed approach

During the clustering stage, similar speaker segments are grouped together using a clustering algorithm. Using the features identified during the CNN step as input, the clustering algorithm divides the speaker segments based on their emotional state and gender.

A potential method for emotional gender speaker diarization using CNN on the dataset is outlined below [31]:

The collection comprises several audio recordings that feature speeches from various speakers. The first step is to preprocess the dataset.

When segmenting the audio into speaker turns, a diarization model utilizes the CNN model to categorize the mood and gender of each speaker turn.

To distinguish between the different speakers in the audio, this diary model might utilize strategies such as speaker embeddings or clustering.

The proposed approach would involve using CNNs [32] to classify the gender and emotion of each speaker. Subsequently, this data would be fed into a diarization model to identify the different speakers and their corresponding gender and emotion labels. This technology has various applications, including speech analysis, speaker identification, and emotion recognition.

3.2 Datasets

The primary objective of our study is to assess the efficacy of our proposed architecture in utilizing diarization for speaker identification based on gender and emotional speech. To accomplish this, we will conduct a comparative analysis between our model and baseline models using various speech datasets. Detailed descriptions of these datasets are presented below.

1. The VoxCeleb datasets are extensive audio-visual collections commonly utilized in speaker recognition and diarization research. They are designed to facilitate the development and evaluation of tasks related to speakers using deep learning methods. There are two main versions of the VoxCeleb dataset [17]:

   VoxCeleb1: VoxCeleb1 is the first version of the dataset, which was released in 2016. It consists of approximately 100,000 statements from 1,251 celebrities. The audio recordings are collected from various sources, such as YouTube videos, interviews, and movie clips. The dataset covers a wide range of speakers with diverse accents, languages, and recording conditions [18].

   VoxCeleb2: VoxCeleb2 is an expanded version of the dataset that was released in 2017. It addresses some limitations of VoxCeleb1 by including more data and additional speakers. VoxCeleb2 contains over 1 million speech segments from 6,112 celebrities. Similar to VoxCeleb1, the audio recordings were collected from online sources. VoxCeleb2 aims to offer a more extensive and varied dataset for speaker recognition research [19].

   Both VoxCeleb1 and VoxCeleb2 offer audio recordings and accompanying metadata for each speaker, enabling researchers to investigate a range of speaker-related tasks, including speaker identification, verification, and diarization. The datasets are available for research purposes and have been widely utilized by the research community to develop and benchmark cutting-edge speaker recognition systems.
2. The RAVDESS (Ryerson audio-visual database of emotional speech and song) is a publicly accessible database that offers a compilation of recordings of emotional speech and song. It was developed by Ryerson University's Sensing, Imaging, and Signal Processing Laboratory (SISPLab) [5].

   The RAVDESS database contains recordings from 24 professional actors (12 male and 12 female) who were directed to produce various vocalizations to convey different emotions. The database covers emotions such as calm, happiness, sadness, anger, fear, surprise, and disgust. Each actor recorded their speech and song in a neutral tone.

   Audio features: The database contains a total of 1,440 audio files, each with a duration of approximately 3–5 seconds. The audio files are sampled at 48 kHz and stored as 16-bit WAV files. The speech recordings are primarily in English [14].

   Emotional labels: Each audio file in the RAVDESS database is labeled with the intended emotion and the gender of the actor. The emotional labels provide a baseline for emotion recognition tasks. The database also includes additional metadata, such as actor ID, intensity, and statement ID [15].

   Song recordings: In addition to emotional speech, the RAVDESS database also includes recordings of actors singing songs with a neutral emotion. These song recordings can be utilized for tasks related to analyzing singing voices or emotional content in singing.

   Availability: The RAVDESS database is freely available for research purposes. Researchers can access and download the dataset from the official RAVDESS website or other approved sources. The terms of use specified by the RAVDESS database must be adhered to when using the dataset [16].

3.3 Experiments

The dataset for this study was divided into three categories: training, validation, and testing. Specifically, the training subset contained 80% of the total files, while the validation and test subsets each comprised 10%. The partitioning was determined empirically to ensure sufficient training data and appropriate testing samples. It's worth noting that, due to privacy considerations, the dataset did not include speaker ID labels. While efforts were made to ensure that speakers were unique to each subset, it cannot be conclusively guaranteed.

The research's modeling approach was inspired by previous successful architectures, with intentional customizations. Existing literature showed consistent success using CNN techniques [33] for analyzing signals. We used spectrograms because they are compatible with tensor transformations and CNNs have been proven effective in analyzing tensor data.

1. Gender Speaker: Deep learning structures, notably CNNs, are commonly selected for speaker gender classification, as illustrated in Figure 4. A CNN model architecture for gender detection has been proposed [34] and is structured as follows:

   - Preprocessing: The audio undergoes a preprocessing stage to extract features such as mel frequency cepstral coefficients (MFCCs) or log-mel spectrograms. Subsequently, these features are normalized to ensure a mean of zero and a variance of one.
   - Convolutional layers: These normalized features undergo processing through successive convolutional layers. Here, a hierarchical representation of the audio features has been developed. After each layer, a non-linear activation function, typically the rectified linear unit (ReLU), is applied.
   - Pooling layers: After each convolutional operation, pooling layers (either max or average) are used to reduce the spatial dimensions and capture dominant information.
   - Fully connected layers: The output from the final pooling layer is flattened and then passed through one or multiple fully connected layers. This setup identifies a non-linear relationship between retrieved features and gender tags.
   - Output layer: The output of the final fully connected layer is passed through a softmax activation function, which generates probabilities for each gender category (male or female).

   To optimize the model, hyperparameters such as the number of layers, the number of filters per layer, and the learning rate are adjusted, often using methods like grid search. Data augmentation techniques, such as random cropping, noise injection, and time manipulation, further enhance the model's robustness.
2. Emotional gender speaker: CNN-based models enable the automatic detection of both the emotional state and gender of a speaker from an audio input [35]. Primarily used for image recognition, CNNs also effectively process audio data. The model's architecture includes:

   - Feature extraction with CNN: The audio data is processed through a series of convolutional and pooling layers to extract essential auditory features.
   - Analysis layer: Extracted features are transmitted to a fully connected layer, or an RNN, which is responsible for predicting gender and emotional tone.

   Popular CNN architectures such as VGG, ResNet, and Inception have demonstrated outstanding performance in audio classification and can be adapted for tasks related to predicting emotional gender.
3. Emotional gender diarization speaker: This technique identifies both the gender and emotional tone in audio streams. Initially, the audio is divided into

homogeneous segments, each corresponding to distinct speakers—a process known as speaker diarization. Post-diarization, features are extracted from each segment, forming the basis for predictions about gender and emotional state.

4 RESULTS AND DISCUSSION

The technique presented in the preceding section was utilized in a variety of optimization tests. This study examined the influence of supervised diarization parameters on the results.

A) Gender speaker

To assess the efficacy of a CNN-based gender speaker categorization model, the dataset is commonly divided into training and testing subsets. After being trained on the training set, the model is evaluated using a variety of performance metrics, such as accuracy and recall, on the testing set. These measurements show the system's accuracy in determining the gender of a speaker.

Various situations, such as the following, need to be evaluated to assess the proposed system:

The model can determine the genders of 80% of the speakers with an accuracy of 0.8. Precision: The model correctly classifies 0.85 of the male speakers, or 85% of them.

Recall: Since the model correctly identifies 75% of the speakers as men, the recall is 0.75.

B) Emotional gender speaker

In our proposed approach, we will graph the total number of emotions, of which the proportions are shown above. Then, using Librosa, waveforms related to each emotion will be generated.

During the training of a CNN model with a labeled dataset, the weights and biases of the network are iteratively adjusted. The objective of this adjustment procedure is to minimize a selected loss function, which quantifies the difference between the expected output labels and the actual labels.

The loss function steadily decreases throughout training, indicating an improvement in model accuracy. Training trends often involve graphs that display the training accuracy and loss as functions of training iterations or epochs. The training loss plot illustrates the changes in the loss function over time, while the training accuracy plot depicts the proportion of accurate predictions generated by the model using the training data. These graphs are valuable tools for evaluating the performance of the model during training and for identifying instances of overfitting. Overfitting happens when the model becomes overly complex and starts memorizing the training data instead of learning general patterns.

To assess the performance of a classification model, a confusion matrix is commonly used. The confusion matrix enables a comparison between the expected and actual results, offering insights into the model's accuracy in predicting emotions, especially in male and female emotion classification.

C) Emotional gender with diarization speaker

Our method, known as speaker diarization, is utilized to identify and classify different speakers in an audio recording. Simple-diarizer typically produces two main components as its output:

The incoming audio stream is divided into smaller, speaker-specific segments as part of our methodology. Each section is associated with a specific speaker, who is assumed to be speaking continuously throughout that segment.

Labeling: Each segment that is specific to a speaker should have a label or identification assigned to it. Different speakers in the audio data are identified using labels.

The results can be visualized using various techniques, such as a waveform plot or spectrogram. When a speaker is changed in a waveform plot, the output displays a unique color for each speaker. A spectrogram's output identifies speaker changes by displaying distinct frequency bands for each speaker.

Each record in the list, which corresponds to a specific speaker, contains the following details:

Start: The number of seconds since the beginning of the audio recording at which the speaker's turn begins.

End: The number of seconds since the beginning of the audio recording at which the speaker stops speaking.

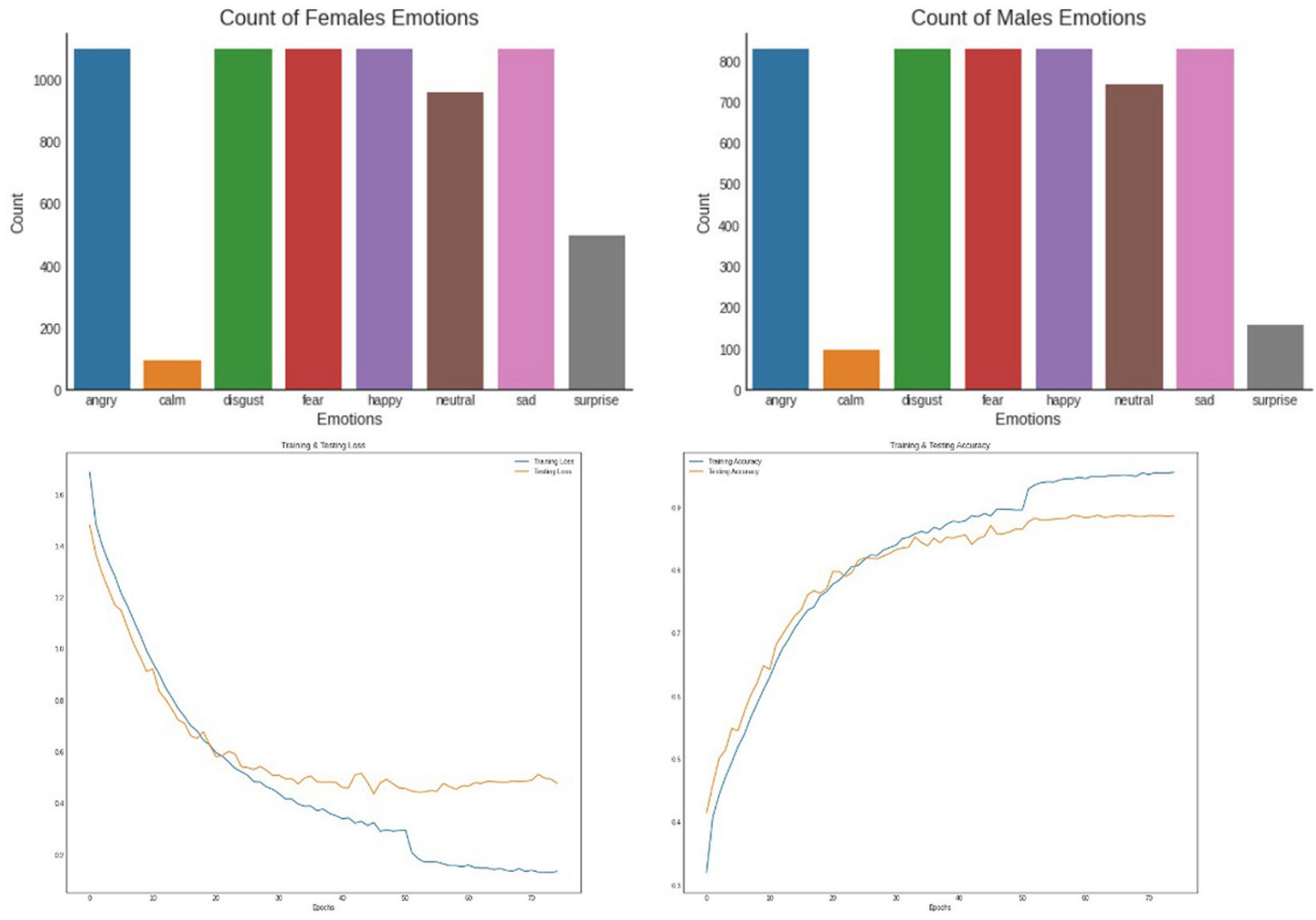


Fig. 4. Results of training and validation proposed model

Accuracy is a commonly cited metric that quantifies the percentage of correctly classified instances (speakers) in a dataset [36]. An alternate, yet pivotal, metric is the F1 score, which combines precision and recall to provide a single evaluative measure [37]. Recall measures the true positive predictions out of all potential positive predictions, while precision quantifies the true positives out of all the positive predictions made. In datasets with class imbalances, the F1 score is a balanced metric that effectively combines precision and recall [38].

The combination of emotion detection, gender classification, and speaker diarization provides a more comprehensive understanding of spoken interactions, whether in multi-party dialogues or standalone audio recordings, as outlined in Table 1 [39]. Deciphering the emotional tenors of speakers, which can range from joy and despair to anger and calmness, provides valuable insights into the emotional dynamics of a conversation. The classification of gender, in distinguishing between male and female timbres, further deepens our understanding of gender dynamics within dialogues. Lastly, speaker diarization, which separates auditory signals from different speakers, identifies speaking sequences, highlighting conversation dynamics [40].

This study on emotion detection, gender classification, and speaker diarization reveals significant insights into spoken language, strengthening fields such as voice recognition, natural language processing, and sentiment analysis [40]. Its relevance is particularly strong in sectors such as healthcare, where understanding patient emotions and gender nuances is crucial for customizing care [41]. While evaluating model performance, it is crucial to carefully select evaluation benchmarks and modalities to ensure a comprehensive and accurate assessment.

Our model's efficiency was tested using the well-known RAVDESS and VoxCeleb speech databases. Audio excerpts from these repositories were strategically divided to facilitate model training and testing.

Table 1. Our approach vs. others' approaches

Approach	Gender	Emotion	Speaker Diarization
[29] Rémi et al.	0.8	X	0.92
[30] Chamishka et al.	X	0.91	0.93
Our Approach	0.96	0.97	0.98

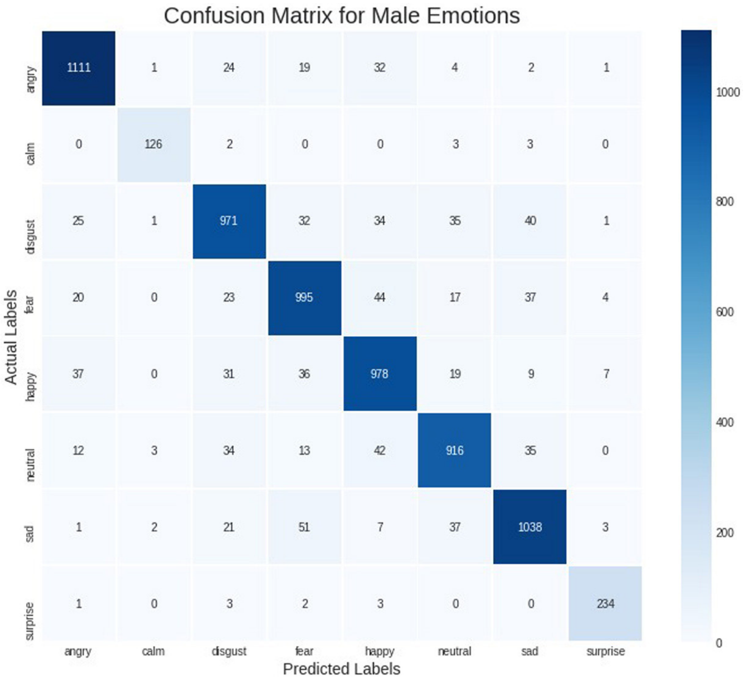
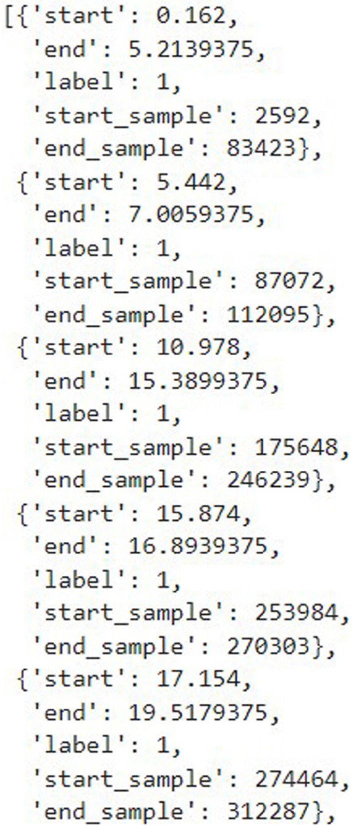


Fig. 5. Confusion matrix



```
[{'start': 0.162,
  'end': 5.2139375,
  'label': 1,
  'start_sample': 2592,
  'end_sample': 83423},
 {'start': 5.442,
  'end': 7.0059375,
  'label': 1,
  'start_sample': 87072,
  'end_sample': 112095},
 {'start': 10.978,
  'end': 15.3899375,
  'label': 1,
  'start_sample': 175648,
  'end_sample': 246239},
 {'start': 15.874,
  'end': 16.8939375,
  'label': 1,
  'start_sample': 253984,
  'end_sample': 270303},
 {'start': 17.154,
  'end': 19.5179375,
  'label': 1,
  'start_sample': 274464,
  'end_sample': 312287},
```

Fig. 6. The content of the segments with details

RAVDESS database: There are two distinct categories of utterances in the RAVDESS dataset. Only speech samples with a neutral emotional tone in the initial utterance were used to train the models during the training phase. The models were then tested on their ability to recognize speakers using neutral speech samples from the second utterance in order to maintain a text-independent approach, as shown in Table 2.

Table 2. Validation of our approach on the RAVDESS dataset

Approach	Emotion	Speaker Diarization
[30] Chamishka et al.	0.91	0.93
Our Approach	0.97	0.98

VoxCeleb database: The VoxCeleb dataset, which contains 35 phrases spoken in a variety of moods, was utilized to assess the model's ability to recognize speakers expressing specific emotions. However, only 15 neutral emotion examples were used to train the algorithms. The performance of the models was then evaluated by identifying the speakers from samples of emotional and neutral speech in the remaining 20 words. The tests were conducted five times to establish the validity of the findings. The results from each trial were recorded and averaged to provide the final findings.

Metrics such as speaker diarization error rate (DER), gender classification accuracy, and emotion classification accuracy will be used to evaluate the performance of the proposed method on the VoxCeleb dataset.

Table 3. Validation of our approach on VoxCeleb dataset

Approach	Gender	Speaker Diarization
[29] Rémi et al.	0.8	0.92
Our Approach	0.96	0.98

5 CONCLUSION

In conclusion, the paper introduces three innovative system architectures for speaker identification that effectively overcome the limitations of both diarization and voice-based biometric systems. Conversely, voice-based biometric systems are limited to identifying individuals in single-speaker recordings and struggle with the complexity of identifying speakers in conversational contexts. Emotional shifts can alter voice characteristics and make gender identification challenging in such scenarios. The proposed architectures integrate methods for gender classification, emotion recognition, and diarization at either the segment or group level, offering a comprehensive approach to speaker identification in challenging scenarios. The assessment of these architectures using two speech databases, VoxCeleb and Ryerson audio-visual database of emotional speech and song (RAVDESS), demonstrates their impressive performance in terms of recognition results. Despite the real-time processing advantages of other approaches, the proposed method outperforms them, achieving a speaker diarization accuracy of 0.98. These results demonstrate that the proposed speech-based approach is highly effective in accurately identifying speakers, particularly in conversational settings with emotional variations. This research introduces new possibilities for speaker identification in various applications, including security, forensics, and human-computer interaction. The integration of gender, emotion, and diarization in the proposed architectures substantially enhances the resilience and precision of speaker identification systems, thereby making them a valuable contribution to the field. Future perspectives may include refining these architectures and investigating their suitability in real-world, practical scenarios to further improve their usefulness and acceptance.

6 REFERENCES

[1] R. Jahangir, T. Y. Wah, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021. https://doi.org/10.1016/j.eswa.2021.114591

[2] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, "CASA-Based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," *Applied Soft Computing*, vol. 103, pp. 1–24, 2021. https://doi.org/10.1016/j.asoc.2021.107141

[3] S. Sun, J. Wang, W. Huang, W. Li, J. Li, H. Liu, and S. Li, "Two-stage end-to-end neural diarization for meeting speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1146–1158, 2020.

[4] R. Wang, X. Hu, K. Zhao, K. Chen, and Y. Qian, "Unsupervised speaker diarization using a clustering-based deep embedding approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1447–1462, 2021.

[5] X. Wu, M. Zhao, and P. Li, "Processing of emotional prosody in Mandarin: An ERP study with the RAVDESS," *Brain and Language*, vol. 207, p. 104832, 2020.

[6] T. M. Al-Hadithy, M. Frikha, and Z. K. Maseer, "Speaker diarization based on deep learning techniques: A Review," in *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, pp. 856–871, 2022. https://doi.org/10.1109/ISMSIT56059.2022.9932710

[7] T. M. Al-Hadithy, Z. B. Messaoud, and M. Frikha, "Improved speaker recognition system based on CNN algorithm," *J. Harbin Inst. Technol.*, vol. 54, no. 6, pp. 1–10, 2022.

[8] S. Moideen, T. V. Sreenivas, and M. R. Kaimal, "Emirati Arabic speech synthesis using deep learning models," in *Proceedings of the 12th International Conference on Natural Language Generation*, Association for Computational Linguistics, 2021, pp. 114–118.

[9] A. Altakhaineh, S. Al-Hajj, and A. Al-Khairy, "Improving emotion recognition from Arabic speech using transfer learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6644–6648.

[10] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, June 2021, pp. 5834–5838.

[11] K. Lee, D. Kim, J. Park, and J. Lee, "Speaker diarization using deep clustering with speaker-adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3525–3539, 2021.

[12] M. Kim and K. Lee, "Autoencoder-based speaker diarization for overlapped speech," *Computer Speech & Language*, vol. 67, p. 101219, 2021.

[13] Q. Xie, X. Zhang, W. Hu, and B. Xu, "Learning to Diarize from scratch by exploiting speaker information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2046–2061, 2021.

[14] C. Zhang, X. Li, L. Li, and Y. Li, "A deep embedding learning framework for single-channel speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3157–3171, 2021.

[15] Z. Cai, J. Ma, and Y. Sun, "Self-supervised learning for End-to-End speaker diarization," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 4320–4324.

[16] Q. Xie, X. Zhang, W. Hu, and B. Xu, "Speaker diarization with separated embeddings by Hierarchical clustering," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 3137–3141.

[17] J. Li, X. Li, L. Li, Y. Li, and Y. Liu, "Improving speaker diarization by augmenting DNN with domain specific knowledge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1114–1127, 2020.

[18] Z. Wu, C. Zhang, and Z. Huang, "A learning-based framework for joint speaker diarization and identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3357–3371, 2020.

[19] R. Sadeghi and H. Sameti, "Towards robust and scalable speaker diarization," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 4295–4299.

[20] L. Cai, Y. Zhang, and Y. Ma, "A multitask learning framework for speaker diarization and verification," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 4504–4508.

[21] X. He, Y. Sun, X. Wang, and Z. Li, "Multi-head self-attention for improved speaker diarization in adverse environments," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH),* 2020, pp. 4188–4192.

[22] S. Srivastava and S. K. Mandal, "An unsupervised deep metric learning approach for speaker diarization," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2968–2972.

[23] Z. Li, Y. Li, Z. Li, and K. Yu, "Enhancing speaker diarization performance by data augmentation for short duration speech segments," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2938–2942.

[24] X. Liao, Y. Wang, L. Xie, and B. Zhang, "Self-supervised learning for speaker diarization," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 1723–1727.

[25] X. Gong and W. S. Zheng, "Joint speaker diarization and speech separation with Permutation-free training," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 1738–1742.

[26] M. Jia, H. Liu, L. Xie, and B. Zhang, "Combining textual and acoustic information for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 214–226, 2020.

[27] Y. Wang and B. Zhang, "Unsupervised speaker diarization via deep autoencoder clustering," I*EEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2084–2094, 2020.

[28] P. Zhou, W. Zhang, H. Hu, and L. Xie, "End-to-End speaker diarization with cluster-Tailored warmup learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1006–1018, 2021.

[29] U. Rémi, D. David, R. Albert, L. Laëtitia, A. Anissa-Claire, *et al.*, "A semi-automatic approach to create large gender-and age-balanced speaker Corpora: Usefulness of speaker diarization and identification," in *13th Language Resources and Evaluation Conference*, Marseille, France. 2022, pp. 3271–3280.

[30] S. Chamishka, I. Madhavi, R. Nawaratne, *et al.*, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *Multimed Tools Appl.*, vol. 81, pp. 35173–35194, 2022. https://doi.org/10.1007/s11042-022-13363-4

[31] R. A. Hameed, W. J. Abed, and A. T. Sadiq, "Evaluation of hotel performance with sentiment analysis by deep learning techniques," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 9, pp. 70–87, 2023. https://doi.org/10.3991/ijim.v17i09.38755

[32] J. Q. Kadhim and A. H. Sallomi, "Enabling deep learning and Swarm optimization algorithm for channel estimation for low power RIS assisted wireless communications," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 12, pp. 171–194, 2023. https://doi.org/10.3991/ijim.v17i12.39411

[33] H. A. Abu-Alsaad, "CNN-based smart parking system," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 11, pp. 155–170, 2023. https://doi.org/10.3991/ijim.v17i11.37033

[34] S. T. Ahmed and S. M. Kadhem, "Using machine learning via deep learning algorithms to Diagnose the lung disease based on chest imaging: A survey," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 16, pp. 95–112, 2021. https://doi.org/10.3991/ijim.v15i16.24191

[35] M. L. Prasetyo, A. T. Wibowo, M. Ridwan, M. K. Milad, S. Arifin, M. A. Izzuddin, R. D. N. Setyowati, and F. Ernawan, "Face recognition using the convolutional neural network for barrier gate system," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 10, pp. 138–153, 2021. https://doi.org/10.3991/ijim.v15i10.20175

[36] O. Kenai, S. Djeghiour, N. Asbai, and M. Guerti, "Forensic gender speaker recognition under clean and noisy environments," *Procedia Computer Science*, vol. 151, pp. 897–902, 2019. https://doi.org/10.1016/j.procs.2019.04.124

[37] A. Vitiello and R. S. Alkhawaldeh, "DGR: Gender recognition of human speech using one-dimensional conventional neural network," *Scientific Programming*, vol. 2019, p. 7213717, 2019. https://doi.org/10.1155/2019/7213717

[38] Mohammad Farukh Hashmi, Abeer Ali Alnuaim, Mohammed Zakariah, Chitra Shashidhar, Wesam Atef Hatamleh, Hussam Tarazi, Prashant Kumar Shukla, and Rajnish Ratna, "Speaker gender recognition based on deep neural networks and ResNet50," *Wireless Communications and Mobile Computing*, vol. 2022, p. 4444388, 2022. https://doi.org/10.1155/2022/4444388

[39] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 5214–5218, 2018. https://doi.org/10.1109/ICASSP.2018.8461471

[40] B. Merritt, "Comparing segmental and suprasegmental features in speaker gender perception: An acoustic distance approach," *The Journal of the Acoustical Society of America*, vol. 153, no. 3, 2023. https://doi.org/10.1121/10.0018899

[41] B. Taylor, K. Wheeler-Hegland, and K. J. Logan, "Impact of vocal fry and speaker gender on listener perceptions of speaker personal attributes," *Journal of Voice*, 2022. https://doi.org/10.1016/j.jvoice.2022.09.018

7 AUTHORS

Thaer Mufeed Taha is a PhD Student, received his bachelor's degree in computer science from the Department of Computer Science and Mathematics, University of Mamoun, Iraq in 2013. Then, he graduated with a master's degree in computer science and information technology from the Department of Computer Science and Information Technology, University of Mansoura, Egypt in 2016. Currently, he is a PhD student at Sfax University. His research focuses on Artificial intelligence, Machine learning, and Deep learning (E-mail: thaer.alamer85@gmail.com).

Dr. Zaineb Ben Messaoud is currently an assistant professor at the Higher Institute of Computer and Multimedia of Gabès, Tunisia, and an active member of the ATISP Research Lab. Her research interests include speech and medical image processing. She received her master's degree in 2007 and then her PhD in 2012 from the National School of Engineering of Sfax. Her academic mailing address is: zeineb.benmessaoud@isimg.tn

Prof. Dr. Mondher Frikha is currently a full professor at the National School of Electronics and Telecommunications, University of Sfax, Tunisia. He is also a director of the 'Advanced Technologies of Image and Signal Processing research lab. His research interests include digital signal and image processing, Speech and audio processing, pattern recognition, and IA applications. He received a Master of Applied Sciences in electrical engineering from the University of Ottawa Canada in 1991. He then worked as a head project at the Industriel Land Agency in Tunisia. In 2003, he started pursuing his graduate research and obtained in 2007 his PhD degree from the National School of Engineering of Sfax, Tunisia. His academic mailing address is: mondher.frikha@enetcom.usf.tn