

Conception of speech emotion recognition methods: a review

Abdelkader Benzirar¹, Mohamed Hamidi², Mouncef Filali Bouami¹

¹Laboratory of Applied Mathematics and Information Systems, Multidisciplinary Faculty of Nador, Mohammed Premier University, Oujda, Morocco

²Team of Modeling and Scientific Computing, Multidisciplinary Faculty of Nador, Mohammed Premier University, Oujda, Morocco

Article Info

Article history:

Received Apr 8, 2024

Revised Oct 2, 2024

Accepted Oct 7, 2024

Keywords:

Arabic speech emotion recognition

Classification algorithms

Feature extraction techniques

Human-computer interaction

Speech emotion recognition

ABSTRACT

In recent years, speech emotion recognition (SER) has emerged as a pivotal tool for understanding and enhancing human-computer interaction (HCI), thus garnering significant attention from researchers due to its diverse range of applications. However, SER systems encounter numerous challenges, particularly concerning the selection of appropriate features and classifiers for emotion recognition. This paper provides a concise survey of the field of speech emotion recognition, elucidating its classification algorithms and various feature extraction techniques across multiple languages. Additionally, it explores the limitations and weaknesses inherent in speech emotion recognition systems. Furthermore, the paper endeavors to categorize recent research endeavors in Arabic speech emotion recognition, employing diverse modeling approaches and extraction methods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mohamed Hamidi

Team of Modeling and Scientific Computing, Multidisciplinary Faculty of Nador

Mohammed Premier University

Oujda, Morocco

Email: m.hamidi@ump.ac.ma

1. INTRODUCTION

Recognizing emotions can be accomplished through various modalities, encompassing text and speech. Text emotion recognition (TER) is a specialized research domain centered on identifying and categorizing emotions conveyed through written communication, including social media posts [1]. Speech emotion recognition (SER) is an interdisciplinary field aimed at automatically detecting and classifying emotions from speech, with applications in areas like human-computer interaction (HCI), healthcare, education, and entertainment. SER systems generally consist of three main steps: preprocessing, feature extraction, and classification. Preprocessing enhances speech quality and segments it into manageable units. Feature extraction identifies key acoustic, prosodic, and linguistic elements that convey emotion. Classification uses algorithms to label emotions based on these features. SER is challenging due to the complexity of human emotions and variability in speech across speakers, languages, and contexts, requiring careful design and optimization for accurate results [2]. In 1997, researchers created the danish emotional speech (DES) database, containing recordings of two men and two women expressing five emotions, which were evaluated by twenty listeners to identify the emotions [3]. A 2001 study on emotion recognition used the Spanish IESSDB corpus and the RAMSES engine, leveraging semi-continuous hidden markov models (HMMs) and low-level features for analysis [4]. In 2005, a study applied two classification methods, HMM and support vector machines (SVM), to classify five emotions from the DES database. Features included fundamental frequency, energy, formant frequencies, mel-frequency cepstral coefficients (MFCCs), and mel frequency sub-band energies, achieving notable recognition rates [5]. In 2007, researchers delved into the domain of human-robot interaction (HRI) by developing an intelligent robot capable of comprehending and

responding to human emotions [6]. A significant breakthrough in emotion recognition came in 2013 with the use of deep belief networks (DBN) to extract unsupervised audio-visual features for emotion classification. The field advanced further in 2017, when deep neural networks (DNN) were combined with voice activity detection (VAD) to efficiently remove silent segments from speech signals, enhancing recognition accuracy [7], [8]. In 2019, researchers combined bidirectional long-short term memory (BiLSTM) with VAD and an attention model to filter out silence and non-emotional parts of speech, improving focus on emotional content [9]. In 2022, Atmaja and Sasou [10] applied four data augmentation techniques glottal source extraction, silence removal, impulse response convolution, and noise addition on JTES and IMOCAP databases, showing that combining these methods improved speech emotion recognition performance. They also explored self-supervised learning (SSL) for training models without external labels [11]. In 2023, the wav2vec 2.0 model was implemented on the Italian "Emozionalmente" database, outperforming human accuracy in vocal emotion recognition and demonstrating potential for integration into conversational agents [12]. In addition to wav2vec 2.0, models like YAMnet and VGGish have been used for speech emotion recognition, but speech SSL PTM embeddings showed superior performance. Notably, x-vector embeddings combined with extreme gradient boosting (XGBoost) outperformed other models, including wav2vec 2.0, unispeech-SAT, wavLM, and ECAPA [13], [14]. This study investigates the field of SER across multiple languages, focusing on the use of various classification algorithms. While earlier studies have explored the impact of traditional machine learning models like SVM and HMM on SER, they have not explicitly addressed the challenges of applying these models across diverse linguistic contexts, particularly in less-studied languages such as Amazigh and Arabic. Furthermore, many existing studies focus predominantly on well-represented languages, often neglecting the performance and adaptability of SER systems in these underrepresented languages. By reviewing the evolution of feature extraction techniques and the shift from basic tools to advanced deep learning classifiers, this paper seeks to fill these gaps, offering new insights into the performance and cross-linguistic adaptability of SER systems.

The remainder of this paper is structured as follows: section 2 delves into related works, providing an overview of existing research in the field. Section 3 offers a comprehensive overview of speech emotion recognition. In section 4, various approaches to speech emotion recognition are discussed in detail. Section 5 highlights the limitations and weaknesses inherent in speech emotion recognition. Section 6 focuses specifically on studies conducted in Arabic speech emotion recognition. Section 7 presents the results and discussion of the different studies that we mentioned before. Finally, the paper ends with a conclusion.

2. RELATED WORKS

A study conducted by Nogueiras *et al.* [4] utilized the HMM in conjunction with pitch and energy features to classify seven emotional states: anger, disgust, fear, joy, sadness, surprise, and neutrality. The findings revealed that using instantaneous pitch led to over 80% accuracy in speech emotion recognition. In another study, Lin *et al.* [5] applied HMM and SVM classifiers with features like fundamental frequency, formant frequencies, MFCCs, and mel sub-band energies, alongside sequential forward selection (SFS) for feature optimization. The HMM classifier achieved impressive recognition rates, reaching 98.9% for female subjects, 100% for males, and 99.5% for gender-independent cases. Harár *et al.* [8] used a DNN with VAD to recognize three emotional states angry, sad, and neutral in the Emo-DB dataset, achieving a 96.97% recognition rate. Catania [12] applied the wav2vec 2.0 model to the Italian Emozionalmente database, achieving 83% accuracy in speaker-dependent cases and 81% in speaker-independent cases, outperforming results from the Emovo dataset. Phukan *et al.* [14] compared eight pre-trained models, including wav2vec 2.0, x-vector, and ECAPA, using XGBoost, random forest (RF), and fully convolutional network (FCN) across databases like CREMA-D, TESS, SAVEE, and Emo-DB, with speaker recognition models showing the best performance. Huang *et al.* [15] explored semi-convolutional neural networks (CNNs) in unsupervised and semi-supervised settings, using an objective function to learn affect-relevant features, which significantly outperformed non-discriminative ones across four databases. Asghar *et al.* [16] developed an Urdu emotion database, applying K-nearest neighbor (KNN), SVM, and RF classifiers with features like MFCC, linear prediction coefficients (LPC), and energy, improving accuracy from 66.5% to 76.5% after excluding disgust. Xia and Zhao [17] used a CNN-BiLSTM with an attention model and 3D MFCC features, achieving a 6% accuracy boost. Atmaja and Sasou [18] employed late fusion of nine pre-trained models to recognize shared emotions from multilingual speech (English and Spanish), achieving a top Spearman score of 0.537 with SVM classification. Pham *et al.* [19] proposed a hybrid data augmentation approach combined with generative adversarial networks (GANs) for emotion recognition on the Emo-DB dataset. They utilized 3D log mel-spectrogram features with an ADCRNN model, achieving 87.12% accuracy with traditional methods and 88.47% with GANs. Zhang *et al.* [20] introduced a capsule network (CapsNet) approach for SER, enhanced by data augmentation, achieving 91.67% accuracy, which increased to 93.33% when data augmentation was applied, effectively addressing deep learning challenges and data scarcity.

3. SPEECH EMOTION RECOGNITION

3.1. Speech emotion recognition system architecture

The SER task involves speech processing and computational paralinguistic analysis with the objective of identifying and categorizing emotions conveyed in spoken language. Figure 1 illustrates the various phases involved in speech emotion recognition. Where, pre-processing: involves standardizing the volume and intensity of vocal signals. Feature extraction: consists of the extraction of features such as MFCC, LPC and linear prediction cepstral coefficients (LPCC). Training: the model is trained on a large dataset of labeled emotional states like happy, angry, and surprised. In this step, the model learns to associate the extracted features with the corresponding emotions in the training data. Emotion recognition: the model uses the extracted features from the test audio to predict the most probable emotion. To do this, it compares the features to the learned models during training and assigns the emotion label with the highest probability.

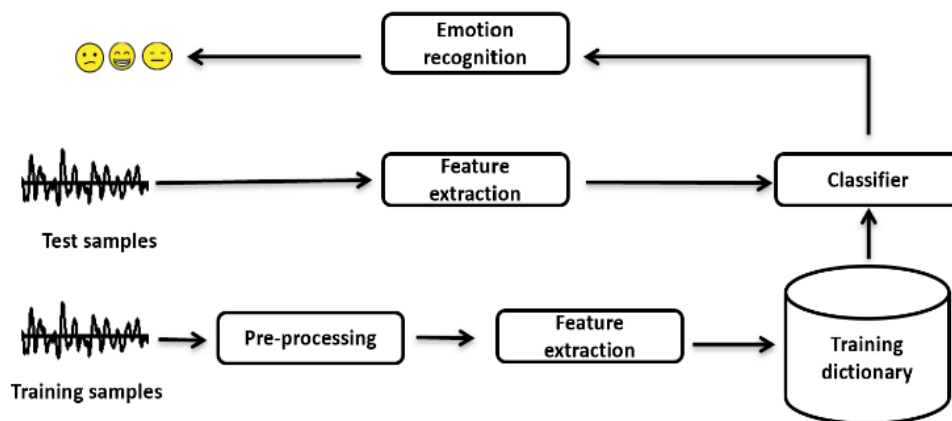


Figure 1. Speech emotion recognition system architecture

3.2. Feature extraction techniques

Feature extraction is the pivotal process of converting the speech signal into a set of parameters that facilitate the identification and classification of various speech sounds and emotions. Below, we will evoke some feature extraction methods.

3.2.1. Mel frequency cepstral coefficients

MFCC are a widely used method for feature extraction in speech processing, effectively capturing key spectral properties through 10 to 12 coefficients. While MFCC is popular, it is highly sensitive to noise, which can impair the accuracy of speech recognition systems. This sensitivity stems from its reliance on spectral features, making it vulnerable to distortion by background noise. Therefore, improving the robustness of MFCC in noisy environments remains a crucial focus in speech processing research [21]-[26].

3.2.2. Linear prediction coefficients

LPC models a speech signal by predicting each sample as a weighted sum of previous samples. This method effectively captures the vocal tract's shape and is vital for identifying formant frequencies, which contribute to a voice's distinct timbre. LPC coefficients are therefore critical in speech analysis for precise signal modeling and synthesis. Additionally, LPCC serve as a fundamental feature in various speech processing applications [22], [27]-[29].

3.2.3. Linear prediction cepstral coefficients

Perceptual linear prediction (PLP) enhances the short-term speech spectrum analysis by incorporating psychophysical adjustments that align more closely with human auditory perception. It utilizes parameters from a filter bank of 18 filters, distributed according to the Bark scale, which reflects the nonlinear frequency perception of the human ear. Covering a range from 0 to 5,000 Hz, these filters effectively capture the critical aspects of speech. This approach ensures that PLP provides a more accurate representation of how humans perceive sound [22].

4. SPEECH EMOTION RECOGNITION APPROACHES

4.1. Naive Bayes classifier

The Naive Bayes classifier (NB) is a probabilistic classifier rooted in Bayes' theorem. Researchers have employed this classifier in numerous studies focused on traditional sentiment analysis tasks. It relies on calculating conditional probabilities using (1) [30]:

$$P(A|B) = P(A) \cdot P(B)/P(B) \quad (1)$$

where:

A is a class, and B is an independent variable or event.

P(A|B): represents the posterior probability of B depending to class A.

P(B|A): represents the likelihood of B when class is B.

P(A): is the prior information of the class A.

P(B): is the evidence of the independent variable B.

4.2. Hidden markov model

The HMM is a widely-used classifier in speech emotion recognition, effectively modeling the dynamic nature of speech. Studies show that HMM achieves improved performance when using logarithmic frequency power coefficients as features. This method has been found to outperform traditional techniques like LPCC and MFCC in emotion recognition accuracy [31]-[33].

4.3. Support vector machine

SVM is known for its simplicity and computational efficiency, making it a popular choice in machine learning. Despite its straightforward structure, it excels in classification tasks with high precision. Research indicates that SVM often surpasses other models in classification accuracy, making it a reliable tool in various applications [34].

4.4. K-nearest neighbor

KNN is a popular supervised algorithm used for both classification and regression tasks. It groups data points based on feature similarity, assuming nearby points in the feature space share the same label or value. KNN typically uses Euclidean distance to measure closeness, making predictions based on the nearest neighbors [2], [35].

$$d(a, b) = \sqrt{\sum_{k=1}^n (a_k - b_k)^2} \quad (2)$$

where a and b are two points in Euclidean space, while a_k and b_k are Euclidean vectors and n is the n-th space.

4.5. Artificial neural network

The artificial neural network (ANN) is modeled after biological neural systems, with feed-forward networks being widely used in classification tasks. These networks consist of interconnected neurons across layers, where each neuron connects to those in the previous layer. This structure enables the network to process input data and learn patterns for decision-making [36].

4.6. Recurrent neural network

Recurrent neural networks (RNNs) are powerful deep learning classifiers, particularly effective for tasks involving sequential data. They excel in applications like speech emotion recognition, speech recognition, and language translation by utilizing information from previous inputs. RNNs are known for their impressive results, handling complex data patterns with ease. Their versatility and strength have established RNNs as essential tools in advanced machine learning and artificial intelligence (AI) applications [30], [37].

4.7. Long short-term memory networks

LSTM networks improve upon RNNs by addressing gradient explosion and vanishing gradient issues, resulting in higher accuracy. LSTMs use specialized gates input, output and forget to effectively manage information flow over long sequences. This makes them especially powerful for tasks requiring long-term dependency understanding. Consequently, LSTMs are favored for applications needing robust sequence modeling [17], [30], [37].

4.8. Convolutional neural network

CNNs are highly effective in deep learning, particularly in image and speech recognition, due to their ability to learn complex data patterns. They utilize convolutional layers for feature detection and pooling

layers for dimensionality reduction while maintaining crucial information. Batch normalization and dropout layers are often employed to stabilize training and prevent overfitting, respectively. This combination of layers ensures CNNs can efficiently process and learn from large datasets. Additionally, DBNs are valuable in speech emotion recognition, offering strong capabilities in this domain [17], [29], [38]-[40].

5. LIMITS AND WEAKNESSES OF SPEECH EMOTION RECOGNITION

The speech emotion recognition has several limits and weaknesses that will be presented at the following:

5.1. Emotion variability

Emotions are complex and can vary significantly between individuals, reflecting differences in personal experiences and psychological states. Cultural norms and values also play a crucial role in shaping how emotions are expressed and perceived, leading to variations across different societies. Additionally, the context in which an emotion is experienced can influence its expression, making it difficult to establish a universal framework for emotion identification. As a result, the interpretation of emotions is highly subjective and can differ not only across languages but also within diverse cultural and social contexts [41].

5.2. Data scarcity and quality

SER systems necessitate extensive and varied datasets of speech signals annotated with reliable emotion labels. Nonetheless, assembling such datasets is challenging due to their rarity, expense, and the time required for collection and labeling. Furthermore, the quality of speech signals may be compromised by factors like noise, distortion, speaker variability, and channel variations, all of which can diminish the performance of SER systems [42].

5.3. Feature extraction and selection

The optimal feature set for SER is still debated, as no single set has been universally accepted as best. The effectiveness of features varies based on the emotion model, dataset, and classification algorithm used. Some features may introduce redundancy or noise, complicating the analysis and potentially reducing SER performance. Therefore, careful selection and evaluation of features are essential for developing efficient and accurate SER systems [43].

5.4. Classification algorithms

Algorithm performance in SER varies based on the emotion model, dataset, and features utilized. While some algorithms perform well under specific conditions, they may struggle in others, making a universal solution challenging. Overfitting and underfitting are common issues that can hinder SER effectiveness. Therefore, careful selection and fine-tuning of algorithms are crucial for optimizing SER systems according to the task and data characteristics [44].

6. ARABIC LANGUAGE-BASED SER APPROACH

In this section, we delve into several investigations within the realm of Arabic speech emotion recognition. Khalil *et al.* [45] focused on detecting anger in human-human dialogues, particularly in call centers, using classifiers like SVM, NB, KNN, and decision tree (DT), and features such as fundamental frequency, formants, energy, and MFCC, achieving 77% accuracy with SVM. Meftah *et al.* [46] studied Arabic speech emotion recognition, analyzing emotions like sadness, happiness, and anger with rhythm metrics and the KSUEmotions corpus, finding that sadness had the highest classification accuracy using multilayer perceptron (MLP) and SVM. Hifny and Ali [47] enhanced Arabic speech emotion recognition with an attention-based CNN-LSTM-DNN model, showing a 2.2% improvement over the deep CNN baseline. Cherif *et al.* [48] investigated emotion detection in the Algerian dialect, focusing on emotions like happy, angry, neutral, and sad. They employed an LSTM-CNN classifier with MFCC as the feature extraction technique, using a manually annotated corpus of Algerian television broadcasts. The model achieved an accuracy of 93.34%. Aljuhani *et al.* [49] developed a speech emotion recognition system for the Saudi dialect, using SVM, MLP, and KNN models with various feature extraction methods, achieving 77.14% accuracy with SVM. Mohamed and Aly [50] showed improved recognition accuracy with MLP and Bi-LSTM models on the BAVED dataset. Tajalsir *et al.* [51] used LSTM and CNN to enhance emotion recognition in human-computer interaction. Alamri and Alshambari [52] reported 95% accuracy using CNN with MFCC on an Arabic YouTube dataset.

7. RESULTS AND DISCUSSION

Table 1 illustrates the evolution of feature extraction techniques and methods used in SER systems. The earliest study in 2001 utilized basic features such as pitch and energy combined with HMM, achieving an accuracy of 80%. As research progressed, more sophisticated features and combinations were explored. For example, a 2005 study introduced a range of features including energy, fundamental frequency, MFCCs, and formant frequencies, coupled with HMM and SVM, leading to remarkable accuracy rates, with up to 100% for male speakers. From 2017 onwards, deep learning methods became increasingly prominent. A 2017 study employing DNN achieved 96.97% accuracy, indicating the potential of deep learning in capturing complex emotional cues in speech. The shift towards deep learning continued, with the introduction of CNN in 2020, which demonstrated an impressive range of accuracies between 92% and 98%. The adoption of advanced techniques such as 3D log mel-spectrograms in 2021, combined with sophisticated models like deep attention-based dilated convolutional-recurrent neural networks (ADCRNN), further enhanced performance, achieving an accuracy of up to 88.47% with GAN-based methods. The trend towards leveraging deep learning models is evident in 2022 and 2023, where methods like wav2vec 2.0, CapsNet, and hybrid approaches achieved accuracies exceeding 93%. The table highlights a clear trend towards the increasing adoption of deep learning models and more complex feature sets in recent years.

Table 1. Summary of speech emotion recognition researches

Ref	Year	Feature extraction techniques	Methods	Results
[4]	2001	Pitch and energy	HMM	80%
[5]	2005	Energy, fundamental frequency (F0), MFCC1 and MFCC2, the first four formant frequencies (F1 to F4), and five mel frequency sub-band energies (MBE1 to MBE5)	HMM and SVM	98.9% (female) 100% (male) 99.5% (gender-independent)
[8]	2017	-	DNN	96.97%
[15]	2014	Affect-salient features	CNN	92%-98%
[19]	2021	3D log mel-spectrogram (MelSpec)	ADCRNN and HAD (hybrid data augmentation)	87.12% (traditional method) 88.47% (GAN-based method)
[16]	2022	MFCC, LPC, energy, spectral flow, spectral centroid, spectral attenuation and zero crossing	KNN, SVM and RF	76.5%
[12]	2023	-	Wav2vec 2.0	81%-83%
[14]	2023	-	XGBoost, RF, FCN, wav2vec 2.0, UniSpeech-SAT, data2vec, wavLM, wav2clip, YAMNet, ECAPA and x-vector	-
[17]	2023	MFCC	CNN and BLSTM	67.44%
[18]	2023	-	XLS-R 53, XLS-R 53 SP (Spanish), XLS-R 53 EN (English), XLSR-300M, XLSR-1B, XLSR-1B EN (English) XLSR-1B SP (Spanish), XLSR-2B and wav2vec 2.0	Spearman rank correlation coefficient: 0.537 (test set) 0.524 (validation set)
[20]	2023	MFCC and ZCR	CapsNet and data augmentation	93.33%

Table 2 presents a clear progression in the techniques and methods used for speech emotion recognition in Arabic language from 2018 to 2023. Early studies, such as the one in 2018, relied on traditional machine learning models like SVM, DT, KNN, and NB, with MFCC and basic acoustic features, achieving a maximum accuracy of 77.2%. As time progressed, more sophisticated features and deep learning models were introduced, significantly improving performance. By 2021, studies using LSTM-CNN architectures with MFCC achieved much higher accuracy (93.34%), indicating the superior ability of deep learning models to capture temporal dynamics in speech. The trend continued into 2022 and 2023, where more complex feature sets like chromogram, mel-scaled spectrogram, and advanced deep learning models like CNN, LSTM, and DNNs were employed. These approaches achieved even higher accuracies, such as 96.81% with LSTM and 95% with CNN, highlighting the increasing effectiveness of deep learning approaches combined with comprehensive feature extraction techniques. This progression underscores the importance of both advanced feature extraction and deep learning methods in achieving high accuracy in speech emotion recognition tasks.

Table 2. Summary of Arabic speech emotion recognition researches

Ref	Year	Feature extraction techniques	Methods	Results
[45]	2018	Fundamental frequency, formants, energy and MFCC	SVM, DT, KNN and NB	77.2% (the best accuracy with SVM)
[46]	2020	Pitch, intensity, formants, jitter, shimmer, harmonics-to-noise ratio and rhythm metrics	MLP neural networks and SVM	Phase1: 49% (with MLP) 52% (with SVM) Phase2: 83.67% (with MLP) 83.95% (with SVM)
[47]	2020	MFCC and Log Mel-filter bank energies (LFBE)	Attention-based CNN-LSTM-DNN model and deep CNN model	Attention-based CNN-LSTM-DNN model: 87.2% Deep CNN: 85% 93.34%
[48]	2021	MFCC	LSTM-CNN	93.34%
[49]	2021	MFCC, mel spectrogram and Spectral contrast	SVM, MLP, and KNN	77.14% (best accuracy obtained with SVM)
[50]	2021	wav2vec2.0 and HuBERT	MLP and Bi-LSTM	Wav2vec2.0: 89% HuBERT: 87% (HuBERT base) 83% (HuBERT large)
[51]	2022	MFCC, chromagram, mel-scaled spectrogram, spectral contrast and tonal centroid features (tonnetz)	LSTM and DNN	LSTM: 96.81% DNN: 93.34%
[52]	2023	MFCC and zero-crossing rate (ZCR)	Machine learning algorithms (SVM and KNN). Deep learning algorithms (CNN and LSTM).	95% (CNN with MFCC)

8. CONCLUSION

Recent advancements in SER underscore the importance of deep learning techniques, especially when paired with advanced feature extraction methods like MFCC, CNN-LSTM architectures, and self-supervised models such as wav2vec 2.0 and HuBERT. These approaches significantly outperform traditional methods, delivering higher accuracy and capturing the nuances of emotional speech. This shift from traditional machine learning to deep learning has led to major improvements in SER's accuracy and adaptability across various languages and contexts. Our future research will focus on developing a speech emotion recognizer for the Amazigh language, addressing its unique phonetic and prosodic features. We aim to create a comprehensive Amazigh dataset and experiment with advanced deep learning and hybrid models. Cross-linguistic analysis will explore model transferability between languages like Arabic and Amazigh, potentially contributing to more universal SER systems and expanding their application across diverse linguistic environments.

REFERENCES




- [1] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.
- [2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.
- [3] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, Sep. 1997, pp. 1695–1698, doi: 10.21437/Eurospeech.1997-482.
- [4] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Sep. 2001, pp. 2679–2682, doi: 10.21437/Eurospeech.2001-627.
- [5] Y. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *2005 International Conference on Machine Learning and Cybernetics*, 2005, pp. 4898–4901, doi: 10.1109/ICMLC.2005.1527805.
- [6] K. H. Hyun, E. H. Kim, and Y. K. Kwak, "Emotional feature extraction based on phoneme information for speech emotion recognition," in *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 2007, pp. 802–806, doi: 10.1109/ROMAN.2007.4415195.
- [7] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 3687–3691, doi: 10.1109/ICASSP.2013.6638346.
- [8] P. Harar, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, Feb. 2017, pp. 137–140, doi: 10.1109/SPIN.2017.8049931.
- [9] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *2019 IEEE International Conference on Signals and Systems (ICSSys)*, Jul. 2019, pp. 40–44, doi: 10.1109/ICSSYS.2019.8811080.
- [10] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, Aug. 2022, doi: 10.3390/s22165941.
- [11] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124396–124407, 2022, doi: 10.1109/ACCESS.2022.3225198.

- [12] F. Catania, "Speech emotion recognition in Italian using wav2vec 2.0 and the novel crowdsourced emotional speech corpus emozionalmente," *Authorea Preprints*, May 2023, doi: 10.36227/techrxiv.22821992.v1.
- [13] O. C. Phukan, A. B. Buduru, and R. Sharma, "Transforming the embeddings: a lightweight technique for speech emotion recognition tasks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023-August, pp. 1903–1907, 2023, doi: 10.21437/Interspeech.2023-2561.
- [14] O. C. Phukan, A. B. Buduru, and R. Sharma, "A comparative study of pre-trained speech and audio embeddings for speech emotion recognition," *arXiv preprint arXiv:2304.11472*, 2023, [Online]. Available: <http://arxiv.org/abs/2304.11472>.
- [15] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the 22nd ACM international conference on Multimedia*, Nov. 2014, pp. 801–804, doi: 10.1145/2647868.2654984.
- [16] A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, and K. Fatima, "An Urdu speech corpus for emotion recognition," *PeerJ Computer Science*, vol. 8, p. e954, May 2022, doi: 10.7717/peerj-cs.954.
- [17] Y. Xia and L. Zhao, "CNN-BLSTM with attention model for speech emotion recognition." Oct. 04, 2023, doi: 10.21203/rs.3.rs-3392008/v1.
- [18] B. T. Atmaja and A. Sasou, "Ensembling multilingual pre-trained models for predicting multi-label regression emotion share from speech," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Oct. 2023, pp. 1026–1029, doi: 10.1109/APSIPAASC58517.2023.10317109.
- [19] N. T. Pham *et al.*, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *Expert Systems with Applications*, vol. 230, p. 120608, Nov. 2023, doi: 10.1016/j.eswa.2023.120608.
- [20] H. Zhang, H. Huang, and H. Han, "MA-CapsNet-DA: Speech emotion recognition based on MA-CapsNet using data augmentation," *Expert Systems with Applications*, vol. 244, p. 122939, Jun. 2024, doi: 10.1016/j.eswa.2023.122939.
- [21] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Amplitude modulation features for emotion recognition from speech," in *Interspeech 2013*, Aug. 2013, pp. 2420–2424, doi: 10.21437/Interspeech.2013-563.
- [22] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system : a comparative study," *arXiv preprint arXiv:1305.1145*, 2013, [Online]. Available: <http://arxiv.org/abs/1305.1145>.
- [23] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 498–502, doi: 10.1109/CONFLUENCE.2016.7508171.
- [24] F. Reggiswarashari and S. W. Sihwi, "Speech emotion recognition using 2D-convolutional neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, pp. 6594–6601, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6594-6601.
- [25] H. Aouani and Y. Ben Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020, doi: 10.1016/j.procs.2020.08.027.
- [26] R. Y. Rumagit, G. Alexander, and I. F. Saputra, "Model comparison in speech emotion recognition for Indonesian language," *Procedia Computer Science*, vol. 179, pp. 789–797, 2021, doi: 10.1016/j.procs.2021.01.098.
- [27] M. Hamidi, H. Satori, O. Zealouk, and K. Satori, "Amazigh digits through interactive speech recognition system in noisy environment," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 101–109, Mar. 2020, doi: 10.1007/s10772-019-09661-2.
- [28] S. Ananthi and P. Dhanalakshmi, "SVM and HMM modeling techniques for speech recognition using LPCC and MFCC features," in *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014: Volume 1*, 2015, pp. 519–526, doi: 10.1007/978-3-319-11933-5_58.
- [29] B. P. Das and R. Parekh, "Recognition of isolated words using features based on LPC , MFCC , ZCR and STE , with neural network classifiers," *International Journal of Modern Engineering Research (IJMER)*, vol. 2, no. 3, pp. 854–858, 2012.
- [30] S. Bodapati, H. Bandarupally, R. N. Shaw, and A. Ghosh, "Comparison and analysis of RNN-LSTMs and CNNs for social reviews classification," in *Advances in Applications of Data-Driven Computing*, 2021, pp. 49–59.
- [31] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6715–6719, doi: 10.1109/ICASSP.2019.8683172.
- [32] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Detection of stress and emotion in speech using traditional and FFT based log energy features," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, 2003, vol. 3, pp. 1619–1623, doi: 10.1109/ICICS.2003.1292741.
- [33] M. Hamidi, O. Zealouk, H. Satori, N. Laaidi, and A. Salek, "COVID-19 assessment using HMM cough recognition system," *International Journal of Information Technology*, vol. 15, no. 1, pp. 193–201, Jan. 2023, doi: 10.1007/s41870-022-01120-7.
- [34] Y. Pan, P. Shen, and L. Shen, "Feature extraction and selection in speech emotion recognition," 2005.
- [35] F. M. J. Mehedi Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines using ng NLP and supervised KNN classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 1, pp. 463–470, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.
- [36] M. Iqbal, S. Ali, M. Abid, F. Majeed, and A. Ali, "Artificial neural network based emotion classification and recognition from speech," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020, doi: 10.14569/IJACSA.2020.0111253.
- [37] I. Zyout and M. Zyout, "Sentiment analysis of student feedback using attention-based RNN and transformer embedding," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, pp. 2173–2184, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp2173-2184.
- [38] M. O. Adebisi, T. T. Adeliyi, D. Olaniyan, and J. Olaniyan, "Advancements in accurate speech emotion recognition through the integration of CNN-AM model," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 3, pp. 606–618, Jun. 2024, doi: 10.12928/telkomnika.v22i3.25708.
- [39] H. Boulal, M. Hamidi, M. Abarkan, and J. Barkani, "Amazigh CNN speech recognition system based on mel spectrogram feature extraction method," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 287–296, Mar. 2024, doi: 10.1007/s10772-024-10100-0.
- [40] H. Asil and J. Bagherzadeh, "Proposing a new method of image classification based on the AdaBoost deep belief network hybrid method," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 5, pp. 2650–2658, Oct. 2019, doi: 10.12928/telkomnika.v17i5.11797.
- [41] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, Feb. 2014, doi: 10.1016/j.specom.2013.07.011.
- [42] A. Shilandari, H. Marvi, H. Khosravi, and W. Wang, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," *Signal, Image and Video Processing*, vol. 16, no. 7, pp. 1955–1962, Oct. 2022, doi: 10.1007/s11760-022-02156-9.




- [43] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23745–23812, Jul. 2021, doi: 10.1007/s11042-020-09874-7.
- [44] M. B. Akçay and K. Oğuz, "Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.
- [45] A. Khalil, W. Al-Khatib, E.-S. El-Alfy, and L. Cheded, "Anger detection in Arabic speech dialogs," in *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, Mar. 2018, pp. 1–6, doi: 10.1109/ICCSE1.2018.8374203.
- [46] A. H. Meftah, M. Qamhan, Y. Alotaibi, and S.-A. Selouani, "Emotional speech recognition using rhythm metrics and a new Arabic Corpus," in *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, Feb. 2020, pp. 57–62, doi: 10.1109/CSPA48992.2020.9068710.
- [47] Y. Hifny and A. Ali, "Efficient Arabic emotion recognition using deep neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6710–6714, doi: 10.1109/ICASSP.2019.8683632.
- [48] R. Y. Cherif, A. Moussaoui, N. Frahta, and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, Mar. 2021, pp. 1–6, doi: 10.1109/WiDSTaif52235.2021.9430224.
- [49] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic speech emotion recognition from Saudi Dialect Corpus," *IEEE Access*, vol. 9, pp. 127081–127085, 2021, doi: 10.1109/ACCESS.2021.3110992.
- [50] O. Mohamed and S. A. Aly, "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset," *arXiv preprint arXiv:2110.04425*, 2021, doi: 10.48550/arXiv.2110.04425.
- [51] M. Tajalsir, S. M. Hernandez, and F. A. Mohammed, "ASERS-LSTM: Arabic speech emotion recognition system based on LSTM model," *Signal & Image Processing: An International Journal*, vol. 13, no. 1, pp. 19–27, Feb. 2022, doi: 10.5121/sipij.2022.13102.
- [52] H. Alamri and H. Alshanbari, "Emotion recognition in Arabic speech from Saudi Dialect Corpus using machine learning and deep learning algorithms." Jun. 13, 2023, doi: 10.21203/rs.3.rs-3019159/v1.

BIOGRAPHIES OF AUTHORS






Abdelkader Benzirar (Ph.D. student)    received his master's degree in intelligent systems and networks at the Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University of Fez (Morocco) in 2015. He is currently pursuing his Ph.D. in speech emotion recognition. His research interest includes machine learning algorithms and speech emotion recognition methods. He can be contacted at email: abdelkader.benzirar.d23@ump.ac.ma.



Prof. Dr. Mohamed Hamidi    is a Professor in the Department of Computer Science at the Multidisciplinary Faculty of Nador, Mohammed I University, Oujda, Morocco. He received his Ph.D. in Computer Science in 2020 from the Faculty of Sciences, Dhar El Mahraz, Sidi Mohammed Ben Abdellah University, Fez. He obtained his Bachelor's degree from the Multidisciplinary Faculty of Nador (FPN), Mohammed I University, in 2011, and his Master's degree from the Faculty of Sciences, Tetouan, Abdelmalek Essaadi University, Morocco, in 2013. His current research interests include machine learning, formant frequencies, pattern recognition, speech and language processing, speech recognition, speech security, interactive voice systems, and voice over IP. He can be contacted at email: m.hamidi@ump.ac.ma.



Prof. Dr. Mouncef Filali Bouami    received an M.Sc. in Electronics from the University of Fez, Morocco in 1998 and a Ph.D. degree from the University of Granada, Spain in 2005 after having defended a doctoral thesis on the modeling of RBF neural networks using T-Norm and T-Conorm operators and weights parameterization. Since 2010 he has been a Senior Lecturer at the Poly-Disciplinary Faculty of Nador, Mohammed premier University, Morocco. His research interest includes machine learning algorithms, text classification and speech recognition methods. He can be contacted at email: m.filalibouami@ump.ac.ma.