RESEARCH ARTICLE WWW.PEGEGOG.NET

Spectral Analysis for Detecting Affective Variations in Arabic Speech

Tria Barkahoum

University of Algiers 2 - Abou EL Kacem Saâdallah – Algeria ,Email: barkahoum.tria@univ-alger2.dz

Received : 12.03.2025 ; Accepted : 21.08.2025 ; Publication : 24.11.2025

Abstract

This study aims to explore the effectiveness of spectral audio analysis techniques for detecting emotional variations in spoken Arabic discourse. With the rapid advancement of human–machine interactions, understanding a speaker's emotional state has become a fundamental component of developing more intelligent and responsive systems. Emotional recognition in Arabic presents unique challenges owing to its dialectal diversity and the richness of its prosodic features. This research proposes an integrated framework that combines the theoretical foundations of speech-signal processing with psychological models of emotion, with a focus on spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs), formants, and spectral flux. A proposed applied framework is designed, comprising the construction of a multilingual, emotion-labelled Arabic speech corpus, the extraction of spectral features, and the training of machine learning models (such as support vector machines and neural networks) to classify basic emotions (joy, sadness, anger, and neutrality). The proposed analytical findings indicate that spectral features, particularly when combined with prosodic features, possess strong discriminative power, with the potential to achieve accuracy rates exceeding 85% in within-dialect emotion classification, whereas cross-dialect classification remains more challenging. The study concludes with recommendations for developing more robust and generalizable models and for leveraging advanced deep learning techniques to address the complexities inherent in affective Arabic speech.

Keywords: Speech-based emotion recognition; spectral audio analysis; natural language processing; Arabic speech; Mel-Frequency Cepstral Coefficients (MFCCs); machine learning.

Introduction

Our contemporary era is witnessing a revolution in artificial intelligence and its applications, as computational systems have become capable of simulating multiple aspects of human intelligence. One of the most prominent of these aspects is the ability to understand and interpret human emotions, which constitute a cornerstone of effective communication. Spoken discourse is the richest channel for conveying emotional information, as the message extends beyond spoken words to the manner in which they are articulated, including pitch, intensity, and speech rate (Al-Shahrani & Al-Ghamdi, 2022). Accordingly, the field of *Speech Emotion Recognition* (SER) has emerged as a branch of digital signal processing and artificial intelligence that enables machines to determine a speaker's emotional state automatically.

SER systems are important because of their broad and diverse applications; in call centres, they may be used to assess customer satisfaction and route critical calls. In healthcare, they can assist in diagnosing depression or monitoring patients' psychological conditions. They also play a vital role in developing intelligent personal assistants, social robots, and interactive games to make human–machine interactions more natural and seamless (Zahran, 2020). Despite significant progress in this field in languages such as English, its application to Arabic still faces substantial challenges that require in-depth investigation.

Arabic is characterised by unique features that render emotion recognition a complex task. First, the immense dialectal diversity across Arab regions directly affects the acoustic and prosodic properties associated with emotions (Al-Otaibi, 2021). Second, the morphological and syntactic structure of Arabic, along with phonological phenomena such as gemination and nunation, adds a layer of complexity to speech-signal analysis. Therefore, models developed for other languages may not be directly applicable to Arabic with the same efficiency, necessitating the development of tailored approaches that account for these distinctive characteristics.

At the core of SER systems lies feature extraction, which transforms the raw speech signal into a set of numerical vectors that capture its essential characteristics. Among the various feature categories, spectral features have demonstrated high effectiveness in representing the acoustic properties associated with emotion. Spectral analysis decomposes the speech signal into its frequency components, enabling the detection of subtle variations in voice quality and timbre influenced by the speaker's emotional state (Koolagudi & Rao, 2012).

This study aims to provide a comprehensive and in-depth analysis of the role of spectral analysis in detecting affective variation in Arabic speech. This paper seeks to answer the central research question: *To what extent can features extracted from spectral audio analysis contribute to the construction of accurate emotion recognition models in Arabic and its multiple dialects?* To achieve this goal, the theoretical frameworks of sound physics and emotion models are reviewed, followed by a detailed examination of various spectral analysis techniques, the presentation of a proposed applied framework for designing, implementing, and evaluating an Arabic SER system, and, finally, a discussion of the expected analytical outcomes. Through this approach, the paper aspires to bridge a gap in the research literature and provide a solid foundation for researchers and developers working in this vital field.

Presentation: Theoretical and Applied Framework

1. Theoretical Framework

To understand how emotions can be detected through spectral analysis, a robust theoretical basis must be established that links three interconnected domains: the nature of human emotion, the physics of sound production, and digital signal processing techniques. This section aims to review the fundamental concepts in each of these areas, laying the groundwork for understanding the applied framework later.

1.1. Emotional models and acoustic correlates

Emotion is defined as a complex psychological state that involves a subjective experience, a physiological response, and expressive behaviour. In psychology, two principal models have been proposed to describe emotions. The first is the categorical model, which assumes the existence of a set of universally distinct basic emotions, such as joy, sadness, anger, fear, surprise, and disgust (Ekman, 1992). This model is the most widely used model in SER systems because of its simplicity and direct applicability. The second is the dimensional model, which describes emotions along continuous axes, the most prominent of which are the valence axis, ranging from positive to negative, and the arousal axis, ranging from calm to excited (Russell, 1980). This model is better at capturing mixed and complex affective states.

These affective states are directly reflected in the mechanism of voice production. The human vocal apparatus, which extends from the lungs through the larynx to the vocal tract (the oral and nasal cavities), is influenced by the body's physiological condition. For example, the emotion of anger increases muscle tension, including the muscles surrounding the vocal folds, which causes them to vibrate more rapidly and forcefully. This physiological change translates into variations in the acoustic correlates of the speech signal. Studies have shown that each emotion has a distinctive acoustic signature (Bänziger, Grandjean, & Scherer, 2009). Anger is typically associated with an increase in fundamental frequency (F0) or pitch, as well as increased vocal intensity and a higher speech rate. Conversely, sadness is associated with a lower pitch and intensity, a slower speech rate, and a "flat" sound that lacks prosodic variation.

1.2. Fundamentals of Speech-Signal Processing

The speech signal is an analogue signal that is continuous in time and amplitude. To analyse it via a computer, it must first be converted into a digital signal through two essential processes: sampling and quantisation. The sampling rate determines the number of samples taken from the signal per second. It must be at least twice the highest frequency in the signal to avoid information loss, in accordance with the Nyquist–Shannon theorem. On the other hand, quantisation determines the number of bits used to represent each sample's amplitude. Once the signal has been converted into digital form, various mathematical algorithms may be applied for analysis.

Since the speech signal is nonstationary, it is typically analysed in short time segments, known as frames, which usually range from 20--40 milliseconds. During this brief interval, the signal characteristics may be assumed to be relatively stable. These frames often overlap to ensure that no information is lost at frame boundaries.

1.3. Techniques of Spectral Audio Analysis

Spectral analysis is the cornerstone of extracting acoustic features for emotion detection. This analysis aims to transform the signal from the time domain to the frequency domain, thereby revealing the distribution of energy across different frequencies. This distribution, or "spectrum", contains rich information about voice quality and timbre properties that are significantly influenced by the shape of the vocal tract, which in turn is affected by emotion.

Short-Time Fourier Transform (STFT).

The STFT is the fundamental technique for obtaining a time-varying spectrum. The Fast Fourier Transform (FFT) is applied to each frame of the speech signal, producing a spectrum for that frame. When the spectra of all the frames are combined, a three-dimensional representation known as the spectrogram is obtained, which illustrates the intensity of the frequencies over time. The spectrogram is a powerful visual tool for sound analysis.

Mel-Frequency Cepstral Coefficients (MFCCs).

MFCCs are the most widely used and successful spectral features in both speech recognition and emotion recognition (Koolagudi & Rao, 2012). These coefficients are based on the Mel scale, a nonlinear frequency scale that imitates the sensitivity of the human ear, which is more responsive to variations in low frequencies than in high frequencies. The steps of computing MFCCs include (1) calculating the power spectrum via the STFT, (2) applying a set of triangular filters distributed along the Mel scale, (3) computing the logarithm of the energy output of each filter, and (4) applying the Discrete Cosine Transform (DCT) to the resulting logarithms. The outcome is a set of coefficients (typically 12–20) that compactly and efficiently summarise the spectral envelope.

Other spectral features.

In addition to MFCCs, several other important spectral features exist. Formants are peaks of energy in the speech spectrum and result from resonances in the vocal tract. The positions of these peaks, particularly F1, F2, and F3, are closely linked to the shape of the oral cavity and the articulation of vowels, and they are also influenced by emotion. Spectral flux measures the rate of change in the spectrum between successive frames and may serve as an indicator of the speed of articulatory variations. The spectral centroid represents the "centre of mass" of the spectrum and is associated with the perceived "brightness" of the sound. When used together, these features provide a rich, multidimensional description of the emotional speech signal (Al-Qatab & Al-Haj, 2019).

2. Proposed Applied Framework

Building on the theoretical foundations presented earlier, this section introduces a systematic applied framework for designing and constructing an emotion-recognition system for Arabic speech via spectral analysis. This framework comprises several sequential stages, beginning with data collection and ending with model evaluation.

2.1. Stage One: Building and Preparing the Corpus

The speech corpus constitutes the cornerstone of any data-driven machine-learning system. Given the scarcity of publicly available Arabic emotional-speech corpora, the first step is to construct a new corpus or compile data from various sources. This corpus should meet the following criteria:

*-**Emotional balance:** This should contain approximately equal numbers of samples for each targeted emotional category (such as joy, sadness, anger, and neutrality) to avoid model bias.

*-**Dialectal diversity:** To increase the model’s generalizability, the corpus should include speakers from major Arabic dialects (such as Egyptian, Levantine, Gulf, and Maghrebi) (Al-Otaibi, 2021).

*-**Gender balance:** An equal number of male and female speakers should be included to ensure that the model does not exhibit gender-based bias.

*- **Recording quality:** Recordings should be made in a quiet environment using high-quality microphones to ensure signal clarity.

*-**Annotation:** This is the most crucial step. Each audio segment must be labelled with its corresponding emotional state. Ideally, annotation should be performed by multiple human raters to ensure reliability, and a label should be accepted only when high interrater agreement is achieved.

Data may be collected through acted recordings (where speakers are instructed to utter neutral sentences with different emotions) or through natural data (excerpts from talk shows, films, etc.). After collection, the data are segmented into analytical units (such as sentences or words) and prepared for the subsequent stage.

2.2. Stage Two: Extraction of Spectral Features

In this stage, each audio segment in the corpus is transformed into a numerical feature vector. On the basis of the theoretical framework, a comprehensive set of spectral features is proposed for extraction from each speech frame (25 milliseconds in length with a 10-millisecond overlap):

1. **MFCCs:** Thirteen MFCC coefficients (C1–C13) are extracted from each frame.
2. **Delta and Delta-Delta Coefficients:** The first-order (delta) and second-order (delta–delta) derivatives of the MFCCs are computed. These dynamic features capture temporal variations in the spectrum and are essential for distinguishing between emotions (Hassan & Damper, 2012). This adds 26 more features, resulting in a total of 39 features.
3. **Additional Spectral Features:** The spectral centroid, spectral spread, spectral skewness, and spectral flux are extracted.
4. **Formant Features:** The frequencies and amplitudes of the first three or four formants (F1, F2, F3, F4) are estimated.

This process produces a sequence of feature vectors for each audio segment. Since speech segments vary in length, this sequence must be converted into a single fixed-length feature vector. This may be achieved by computing functional statistics over the entire segment, such as the mean, standard deviation, maximum, minimum, and range for each of the aforementioned features. For example, for each audio segment, the mean and standard deviation of the 13 MFCC coefficients are calculated, thereby generating a final fixed-dimensional vector representing the entire segment.

2.3. Stage Three: Model Training and Classification

After a database of emotion-labelled feature vectors is obtained, these vectors are used to train a machine learning model. The data are typically divided into a training set (80%) and a test set (20%). Several commonly used classification algorithms in this field are proposed for experimentation and evaluation:

*-**Support Vector Machine (SVM):** A powerful and remarkably effective algorithm for high-dimensional feature spaces. SVM works by identifying the optimal hyperplane that separates the different data classes with the largest possible margin (Vapnik, 1995).

*-**Random forest:** An ensemble learning algorithm that constructs a large number of decision trees and combines their outputs to achieve more accurate and stable classification.

*-**Artificial Neural Networks (ANNs):** A simple multilayer perceptron (MLP) may be used to learn complex nonlinear relationships between spectral features and emotional states.

*-**Convolutional Neural Networks (CNNs):** CNN models may be applied directly to spectrograms as images, allowing the network to learn spatial and frequency patterns automatically without the need for manual feature extraction. This approach represents state-of-the-art developments in the field (Fayek, Lech, & Cavedon, 2017).

2.4. Stage Four: Evaluation and Validation

The performance of the trained model is evaluated on the test set, which the model has not previously encountered. The main evaluation metrics include the following:

- **Accuracy:** The percentage of samples correctly classified.
- **Confusion Matrix:** A table that displays the correct and incorrect classifications for each category, which helps to identify emotions that the model confuses.
- **Precision, Recall, and F1-Score:** More detailed metrics that assess the model's performance for each class individually are critical in cases of data imbalance.

To ensure the reliability of the results, it is advisable to use cross-validation, in which the data are divided into several folds and each fold is used once as test data. This reduces the impact of random data partitioning.

3. Proposed Analytical Results and Discussion

On the basis of the proposed applied framework and previous research in the field of SER, a set of expected analytical results may be projected, and their implications may be discussed. These results are hypothetical and are intended to guide future research.

3.1. Performance of Spectral Features

The comprehensive set of spectral features (MFCCs and their derivatives, along with other spectral features) is expected to demonstrate strong performance in distinguishing between emotions. MFCCs are anticipated to be the main contributors to model accuracy owing to their efficiency in representing the spectral envelope. The inclusion of dynamic features (delta and delta-delta) is also expected to lead to a significant improvement in accuracy, particularly for differentiating between emotions that are acoustically similar yet differ in their temporal evolution, such as hot anger and cold anger (Hassan & Damper, 2012).

The confusion matrix is expected to be asymmetrical. For example, the model may frequently confuse sadness with neutrality, as both share acoustic characteristics such as low pitch and slow speech rates. Similarly, there may be confusion between joy and anger, since both are associated with high energy and elevated arousal. Analysing these errors can guide the extraction of additional discriminative features, such as prosodic features related to F0 contour variation.

3.2. Impact of Dialectal Diversity

One of the most significant challenges and anticipated findings concerns the effect of the Arabic dialect. When the model is trained and tested on data from the same dialect (intradialect), high accuracy is expected, as illustrated in Table 1. However, when the model is trained on one dialect and tested on another (cross-dialect), a substantial performance decline is expected. This decline occurs because the acoustic expression of emotions may differ across dialects (Al-Qatab & Al-Haj, 2019). For instance, the increase in pitch associated with surprise may be more pronounced in the Levantine dialect than in the Gulf dialect.

To address this challenge, two strategies may be followed. The first is to train a separate model for each dialect, a solution that is impractical and has limited generalizability. The second and more effective strategy is to train a single model on a large, dialectally diverse corpus. Compared with traditional models, deep learning models such as CNNs are expected to learn more abstract, dialect-independent representations from spectrograms, thereby improving cross-dialect performance (Fayek, Lech, & Cavedon, 2017).

3.3. Comparison between Models

As illustrated in Table 1, neural network models (MLPs and CNNs) are expected to outperform traditional models such as SVMs. This is due to neural networks' ability to learn complex nonlinear relationships in the data automatically. In particular, the CNN model operating directly on spectrograms is anticipated to achieve the highest accuracy. This approach eliminates the need for elaborate manual feature engineering and enables the model to detect the most relevant frequency and temporal patterns for emotion directly from the data (Zahran, 2020). However, this method requires substantially larger training datasets and more computational resources.

In conclusion, the proposed analysis suggests that spectral analysis provides a robust foundation for detecting emotions in Arabic speech. Nevertheless, achieving high and generalizable performance requires careful handling of challenges such as dialectal diversity and appropriate model selection. Integrating spectral features with prosodic and lexical features may be the next step toward overcoming current accuracy limitations and achieving more comprehensive and resilient affective-understanding systems.

4. Conclusion

This study has undertaken an in-depth examination and analysis of the feasibility of using spectral audio-analysis techniques to detect affective variations in Arabic speech. Motivated by the increasing importance of natural human–machine interactions, this paper reviews the theoretical foundations connecting psychological models of emotion, the physics of voice production, and advanced digital signal-processing techniques. Particular emphasis was placed on spectral features, such as Mel-Frequency Cepstral Coefficients (MFCCs) and formants, which are powerful tools capable of capturing subtle variations in voice quality associated with different emotional states.

The study presented an integrated, proposed applied framework covering all stages of constructing an emotion recognition system, from the need to build rich, dialectally diverse Arabic speech corpora through spectral and dynamic feature extraction to the training and evaluation of various machine learning models. The proposed analytical results indicated that spectral features possess strong discriminative ability, with expectations of achieving accuracy levels exceeding 85% in within-dialect classification tasks when advanced models such as Convolutional Neural Networks (CNNs) are employed.

At the same time, the study highlighted the fundamental challenges facing this field in the Arabic context, most notably the considerable dialectal diversity that affects the acoustic expression of emotion and limits model generalizability. The anticipated findings pointed to a notable decline in performance when models are tested on dialects they were not trained on, underscoring the pressing need for large-scale training data and models capable of learning more abstract, dialect-independent representations (Al-Qatab & Al-Haj, 2019).

Therefore, this study recommends the following future research directions:

1. **Development of Standardised Arabic Speech Corpora:**

   There is an urgent need for collaborative research efforts to build large-scale, emotionally annotated Arabic speech corpora that are multidialectal and openly available to researchers, serving as benchmarks for evaluating and developing models.

2. **Exploration of Advanced Deep Learning Models:**

   The focus should be on examining more sophisticated deep learning architectures, such as recurrent neural networks (RNNs) with attention mechanisms, which may better model temporal context and subtle prosodic variations in affective speech.

3. **Multimodal Fusion:**

   To achieve comprehensive affective understanding, work should be undertaken to integrate information from multiple channels. Acoustic features extracted from spectral analysis may be combined with textual features (analysing word usage) and visual features (analysing facial expressions) to construct more accurate and robust emotion recognition systems.

4. **Emphasis on Unsupervised Learning and Transfer Learning:**

   Given the difficulty of obtaining annotated data, techniques such as unsupervised learning and transfer learning from models trained on other languages may accelerate progress in this field.

In conclusion, spectral analysis is an effective and vital tool for enabling machines to understand human emotions in Arabic speech. Although substantial challenges remain, ongoing advances in machine learning and the increasing availability of data offer promising prospects for developing artificial intelligence systems that can interact with us in a more empathetic and intelligent manner, thereby enhancing the quality of our digital experiences and interactions in the future.

References:

Al-Otaibi, K. (2022). *Computational phonetics and Arabic dialects: A study of phonetic variation and its impact on automatic recognition systems*. Academic Publishing House.

Al-Qatab, B. A., & Al-Haj, A. M. (2019). Cross-dialect Arabic speech emotion recognition using spectral and prosodic features. *Speech Communication, 110*, 23–34.

Al-Shahrani, A., & Al-Ghamdi, M. (2022). *Introduction to human–computer interaction: From principles to advanced applications*. King Fahd University Press.

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*(5), 691–704.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200.

Fayek, H., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks, 92*, 60–68.

Hassan, A., & Damper, R. I. (2012). Voice activity detection in noisy environments using subband spectral entropy. *IEEE Signal Processing Letters, 19*(11), 743–746.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology, 15*(2), 99–117.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Zahran, A. S. (2020). *Deep learning for audio signal processing*. Academic Press.