Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Computational Modeling of Intonation Patterns in Arabic Emotional Speech

Aitor Arronte Alvarez[1], *Elsayed Issa*[2], *Mohammed Alshakhori*[2]

[1]University of Hawaii, Center for Language and Technology, USA
[2]University of Arizona, USA

arronte@hawaii.edu, elsayedissa@email.arizona.edu, mkalshakhori@email.arizona.edu

Abstract

The expression of emotion in speech communication has been frequently studied from the analysis of F_0 contours. Global features such as the mean level and range of F_0, as well as the slope of the contour, have been related to the degree of activation of an emotion. However, the existence of specific patterns associated with basic human emotions has not been empirically demonstrated, and studies generally find no conclusive answer to this question. In this paper we present a computational study of emotional speech in Arabic. A computational method for obtaining tonal contours based on F_0 peak-to-peak and valley-to-valley distances is presented that is able to capture tonal rhythm (macro-rhythm). We introduce a model for extracting frequent tonal patterns to characterize emotions based on the typology of the patterns. Comparative analysis between sets of neutral and emotional utterances shows that distinctive patterns exist for anger, sadness, happiness, and surprise, that are completely absent in neutral speech. Findings highlight the effectiveness of the comparative computational methodology presented to discover macro-rhythmic emotion-specific patterns.

Index Terms: speech emotion recognition, macro-rhythm, pattern discovery, intonation, Arabic.

1. Introduction

The study of emotion in speech communication has been addressed from various scientific disciplines and with different research foci [1]. In the acoustic analysis of speech, emotion has been frequently studied from the examination of global features extracted from the fundamental frequency (F_0) contour, such as the mean, range, and global trend [2, 3, 4]. These features have been connected to the degree of activation of an emotion [4, 5], and studies suggest that emotions with higher arousal (anger, happiness) may be difficult to capture with global intonational features due to the high variability of the F_0 contour [4]. Other authors have found that while intonation patterns signal the perception of emotions, no one-to-one relationship between conveyed emotion and intonation pattern was found [6]. Overall, individual characteristics of speech emotions have been addressed empirically from global descriptors. In this work, we present a computational method for finding relevant intonation patterns that are especific to human emotions using macro-rhythm [7].

In the Autosegmental Metrical (AM) model of intonational phonology, pitch accents and boundary tones define prominence and phrasing [8, 9]. In the prosodic typology model described by Jun [7] macro-rhythm was defined as phrase-medial tonal rhythm whose unit is equal to or slightly larger than a prosodic word (PWord) [7]. In contrast to traditional speech rhythm where temporal patterns are formed by sequences of syllables, morae, or the alternation of weak and strong syllables, tonal rhythm is solely marked by changes in pitch, and therefore presented as a sequence of high and low (H, L) points in the F_0 contour. Macro-rhythm as a tonal temporal pattern has been used in the cross-linguistic study of prosodic typologies [7, 10].

Taking into consideration how the slope of the contour is affected by emotions and their arousal level, we hypothesize that macro-rhythm should be able to capture emotion-specific patterns that could be used for characterizing basic human emotions, and for making direct comparisons between neutral speech utterances and emotional ones. This approach could facilitate the creation of tonal inventories of speech patterns associated with particular emotions, allowing for more analytical and descriptive comparisons, and for the development of speech emotion recognition (SER) features to be used in machine learning systems.

In this paper we consider tonal rhythm as the significant changes in the F_0 slope that occur within an Intonational Phrase (IP). The goal of this work is first to determine whether tonal rhythm is a good descriptor of the emotion expressed in a speech utterance, and secondly, if unique patterns exist for a given emotion set.

In order to compare tonal rhythms between the different emotions and a neutral state, we develop a method for obtaining F_0 peak-to-peak and valley-to-valley distances. Frequent tonal patterns of the contours are then obtained using a variant of the Bide algorithm [11], that discovers and extracts closed maximal patterns [12, 13] directly from the contour vector. Comparisons of maximal patterns are then made to obtain inventories of unique patterns by emotion.

Prior work in the computational modeling of prosody has concentrated in the automatic classification and discovery of intonational phrase boundaries [14], the automatic classification of prosodic events [15], automatic pitch detection [16], prominence detection [17] and more recently with the advent of vector representation models in the last decade, and more generally, deep learning algorithms, phrasing detection [18], and melodic contour extraction [19]. Even though deep learning methods have achieved a high degree of accuracy in classification tasks, and distributed representations are able to synthesize complex relationships and dependencies between prosodic events in a compact vectorial form, their results often present difficulties in their explanation and interpretation. Also, vector representations make assumptions about context that in some cases, are harder to apply and may limit the quality of the results.

In the research presented in this paper we use a sequential pattern mining algorithm to extract frequent macro-rhythmic patterns from Arabic speech utterances. Sequential pattern mining algorithms are a class of data mining methods for discovering hidden frequent patterns in sequential data [20]. Since we are dealing with sequences of F_0 intervals from IPs, sequential data mining algorithms will allow us to extract meaningful patterns that can be quantified and individually analyzed. As shown in a previous work by two of the authors of this paper

10.21437/SpeechProsody.2022-125

[21], intonation patterns extracted with this method could be used as the input for end-to-end speech recognition systems.

2. Computational Method

The computational approach presented in this paper is composed of the following steps, as shown in Figure 1.

- From the raw audio signal a vector of F_0 points is extracted.
- The initial F_0 vector is transformed into a vector of intervals expressed in cents.
- A contour simplification function is applied to the vector of intervals to obtain peak-to-peak and valley-to-valley distances and slopes.
- A variant of the Bide algorithm is used to extract maximal patterns of macro-rhythm from the closed set of patterns.
- Following a similar method as the anti-corpus approach utilized in Music Information Retrieval [22], unique macro-rhythmic patterns are obtained for each emotion.

2.1. F_0 extraction

From the collection of raw audio files, we extract F_0 using Praat's [23] Python interface Parselmouth [24] using ranges of F_0 between 75-650 Hz (floor, and ceiling), resulting in a vector of frequencies per utterance. Unvoiced audio segments were linearly interpolated and octave jumps eliminated.

2.2. Contour approximation and representation

In order to extract macro-rhythm from the F_0 contour, intervallic distances between 50 ms consecutive frames were obtained in cents. A contour approximation function was developed for finding relevant changes in F_0 movement. The contour approximation method detects changes in F_0 that are greater than a semitone (100 cents). Once semitone distances were obtained, distances between local minima and maxima points were computed in cents, and coded in intervallic form to avoid differences in speaker ranges.

Initially, we coded distances between local minima and maxima using integers, where 1 represented 100 cents (a semitone). After an initial evaluation of the pattern extraction algorithm, and observing the large amount of patterns obtained, we represented changes in intervallic distances using 1 for 200 cents (1 whole tone, 2 semitones), using the negative sign - for descending distances. For instance, a sequence of [3,-3, 2, -2] represents and ascending 6 semitones, followed by 6 descending semitones, and 4 ascending and 4 descending semitones.

Since the goal is to code and represent distances between L and H points in the contour, this coding scheme resulted in clearer results.

2.3. Pattern discovery

The main objective of descriptive pattern mining is to extract hidden patterns from sequential data. The problem of sequential pattern mining was defined as the problem of mining subsequences in a set of sequences [25]. More formally, given a set of distinct items $I = \{i_1, i_2, ..., i_k\}$, a sequence S is defined as an ordered list of events $\langle e_1, e_2, ...e_m \rangle$ where e_i is an item such that $e_i \in I$ for $1 \leq i \leq m$. If sequence S_a is contained in sequence S_b, S_a is then a subsequence of S_b, and we call S_b a supersequence. The absolute support of a sequence S_α is the number of times that S_α appears in a sequence database (SDB), and the relative support the percentage of S_α contained in SDB. We say then that S_α is frequent on SDB if S_α appears with a frequency above a certain support threshold or minimum support (min_sup). And we say that S_α is closed if there is no other proper supersequence of S_α with the same support. In this context a sequence and a pattern refer to the same object.

2.3.1. Maximal patterns

We use the BIDE algorithm to obtain the set C of closed patterns from the entire database of utterances. A set of maximal patterns is defined as the set $M \subseteq C$ such that every pattern in M is not strictly a subpattern of any other closed pattern in C.

The set C tends to be of large cardinality and therefore contain many patterns that may be subpatterns of other superpatterns with different support. In cases where the search space does not need to be that large, maximal patterns not only reduce the search space, but also provide more informative patterns. Since we are dealing with IPs, finding the largest possible macro-rhythmic pattern that it is not contained in another subpattern will allow us to make inferences about the typology of specific emotions with more information and descriptive power.

2.3.2. Unique distinctive patterns

From the set M we create subsets by emotion and compare them with neutral patterns following the concept of anti-corpus described in [22] which defines a distinctive pattern as the one that is overrepresented in a corpus compared to an anti-corpus. In our case, we limit this definition even further to define unique patterns as those that are present in a corpus given a minimum support and completely absent in an anti-corpus. This restriction will highlight further the uniqueness of certain patterns when comparing emotion sets with a neutral one acting as the anti-corpus.

We make the software implementation of the method presented available to the public [1].

3. Data: KSUEmotions Corpus

To perform our experiments The King Saud University Emotions (KSUEmotions) corpus was used [26], which contains emotional speech recorded by speakers from Saudi Arabia, Syria, and Yemen. Speakers recorded sixteen sentences extracted from newspaper articles, where each speaker is asked to record each sentence in five different emotions: neutral, happiness, sadness, surprise and anger. The total number of audio files in the KSUEmotions corpus is 3276 as table 1 shows.

Table (1) *Statistics of the KSUEmotions Corpus*

	Corpus
Num. of audio files by male speakers	1639
Num. of audio files by female speakers	1637
Total number of audio files	3276
Total number of hours	5 hr. and 10 m.

The verification of the emotions followed two phases. First, a human perceptual test is done. Nine human raters listened to

[1] https://github.com/aitor-arronte/robust-speech-emotion-recognition

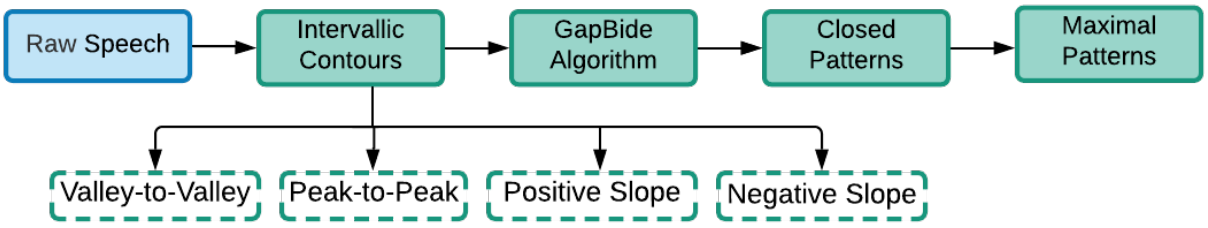

Figure (1) *Schematic diagram of pattern recognition and extraction*

the audio files and assigned the appropriate emotions and gave a score on a scale from 1 to 5 where (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Only audio files that obtained the highest scores have been selected. Additionally, during the human perception test, five emotions were considered for the baseline corpus: neutral, sadness, happiness, surprise and questioning. Since we are interested in understanding the basic human emotions following the Ekman model [27, 28] questioning was discarded.

Table (2) *distribution of audio files with respect to speakers' gender and emotion type.*

Emotion	Male	female
Neutral	328	328
Happiness	328	328
Sadness	327	326
Surprise	328	328
Anger	168	168

3.1. Arabic intonation

The corpus uses Educated Standard Arabic (ESA) (see: [29]) and pronounced by Saudi, Yemeni, and Syrian speakers. Arabic is a stressed-timed language in which the stress depends on the weight of the syllable [30]. Every content word has a prominent stressed syllable with distinctive phonetic properties such as a higher-pitch, loudeness, and longer vowel duration. Thus, F_0 associated with the stressed syllable will interact with the intonation encoded at the post-lexical level [31, 32].This gives prominence to the syllable and the whole word associated with. Some Arabic varieties exhibit greater F_0 excursion and longer duration in focused constituents, and pitch range compression in post-focus materials; however, pre-focus components have no systematic behavior [33, 34, 35].

4. Analysis of Results

In order to evaluate macro-rhythm as a prosodic feature for modeling emotion types, we quantify it using the Macro-rhythm Variation Index (MRVI) measure proposed in [7], which is defined as the sum of the standard deviations of the rising slope, falling slope, peak-to-peak, and valley-to-valley distances. A high MRVI value indicates weak macro-rhythm and therefore more variability in their peak-to-peak and valley-to-valley distances. MRVI values closer to 0 will indicate more regularity and stronger macro-rhythm. We conduct statistical analysis on this measure, as well as on F_0 mean and range to confirm previous findings on global features [2, 3, 5].

Table (3) *Results of ANOVA tests*

One-way ANOVA	F-value	p-value
F0 range $\sim$ emotion	20.07	0.001 ***
F0 mean $\sim$ emotion	162.5	0.001 ***
Macro Rhythm Index $\sim$ emotion	90.97	0.001 ***

4.1. Statistical analysis

One-way ANOVAs were performed on F_0 mean, F_0 range, and on MRVI using emotion (5 levels) as a factor. Min-max normalization was applied to MRVI. The tests were run using the statistical programing language R [36]. As we can see in Table 3, results were significant for all 3 measures ($p \leq 0.001$). Post-hoc Tukey tests (α = 0.05) were conducted to determine which levels of the factor show significant differences. For MRVI, all pairwise comparisons were significant with the exception of neutral-happiness ($p = 0.076$). In Figure 3, we can see how the MRVI median values for neutral and happiness are close to each other, with means of 0.283 and 0.263 respectively. This indicates similar macro-rhythm, with no significant differences as the test shows. Since happiness is considered a higher arousal emotion, this result is somehow surprising. Also, we should note that anger is the emotion with the strongest macro-rhythm with a MRVI mean of 0.195. The weakest conveyed emotion in terms of macro-rhythm is sadness with an MRVI mean of 0.353. Pairwaise comparisons on F_0 mean showed non-significant differences for sadness-neutral, and surprise-happiness pairs. These results are more expected and coincide with previous findings in the literature since both pairs share similar arousal categories, low for sadness-neutral, and mid/high for surprise-happiness. F_0 range results showed findings compared with previous literature, and all comparisons were significant except for the pair neutral-sadness ($p = 0.99$). High-level arousal emotions had higher F_0 range means (anger=279.5 Hz, happiness=382.7 Hz, and surprise=485.2 Hz) , which was predicted based on the previous research.

The statistical analysis showed some unexpected results on macro-rhythm. Low-arousal emotions (sadness, and neutral) showed higher degree of variability in macro-rhythm as expressed in MRVI values, and therefore weaker macro-rhythm when compared with high-arousal emotions such as surprise, and anger. Happiness seems to be closer to neutral speech. This may indicate a cultural difference, or a different degree of activation in the emotion that the annotation does not capture. Even though F_0 range is larger for high-arousal emotions, results on MRVI highlight more predictability from these emotions in terms of the distances from peaks and valleys, and the size of the slopes, resulting in stronger macro-rhythm.

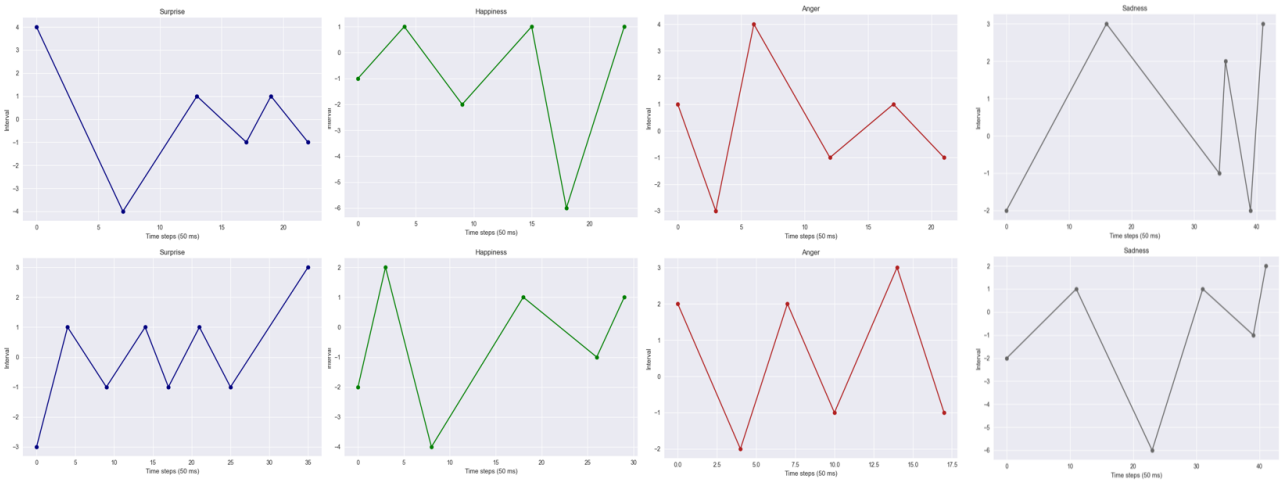

Figure (2) *Unique maximal patterns by emotion. Interval distances on the y axis, and time units (50 ms) in the x axis.*

Figure (3) *Boxplots for the Macro-Rhythm Variation Index. Horizontal lines indicate median values.*

4.2. Pattern extraction

Taking as input the audio files from KSUEmotion corpus, we apply the computational model described in section 2 to obtain maximal patterns by emotion type. Once maximal patterns are extracted, we compare all patterns associated with a given emotion against the neutral ones acting as the anti-corpus, resulting in subsets of unique patterns by emotion. The parameters of the algorithm were set to 6 for minimum support, and 5 for the minimum length of a pattern. In other words, we only accept as frequent maximal patterns those that appear within a given emotion set with a frequency greater or equal than 6, and that contain at least 5 slopes (peaks and valleys). Of particular importance here is the concept of anti-corpus, since F_0 patterns that are completely absent in neutral speech highlight their emotional content. In this context then, the frequency of appearance of a pattern is crucial, since it determines the relevance within an emotional subset. This approach is the dominant view in the frequent pattern mining literature [37].

Image 2 presents the top-2 most frequent unique patterns by emotion. Following the statistical analysis, where anger and surprise had the strongest macro-rhythm, we can see that both emotions had more regular temporal patterns in their peak-to-peak, and valley-to-valley distances, especially when we compare them with sadness patterns, which show less temporal regularity. Anger patterns seem to be characterized by regular, and steeper F_0 slopes, while surprise contain sudden excursions followed or preceded by shorter, regular intervals. Sadness on the other hand, contain patterns with more irregular slopes and intervallic distances, which is supported by the weaker MRVI as seen in Figure 3. Patterns that belong to happiness seem harder to place or characterize, since they do have a weaker macro-rhythm. Anger had the least percentage of unique maximal patterns with 11.3% of the total maximal patterns, but with higher frequency of occurrence compared to sadness, happiness, and surprise. This may be the cause of the stronger macro-rhythm.

Overall, results on macro-rhythm suggest that high-arousal emotions such as anger, and surprise, tend to have stronger macro-rhythm than low-arousal emotions such as sadness. Happiness seems to vary between regular and irregular patterns, and it is harder to characterize. The pattern extraction method provides a more descriptive analysis by emotion of the top-k most frequent patterns that are completely absent in neutral speech, and by comparing sets of patterns by emotion we can find patterns that are specific to an emotion and completely absent to another. Since the goal of this paper was to determine the suitability of macro-rhythm for the characterization of emotions in speech, the proposed approach was able to uncover particular patterns that can be further analyzed, providing more descriptive power to the statistical analysis. We speculate that this approach could be used in the automatic classification of emotions, since it produces statistical patterns that can uniquely identify an emotion. This however, goes beyond the scope of this work and should be tested in a machine learning study.

5. Conclusions

In this paper, we presented a computational method for extracting macro-rhythmic patterns from speech utterances. It was shown how macro-rhythm could be used as a significant feature in the analysis of emotions in speech, and how the comparative computational approach can highlight particular emotion-specific patterns. Results show that high-arousal emotions tend to have greater F_0 ranges, but more regular tonal patterns, and therefore stronger macro-rhythm.

6. References

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[2] A. Paeschke, M. Kienast, W. F. Sendlmeier *et al.*, "F0-contours in emotional speech," in *Proc. 14th Int. Congress of Phonetic Sciences*, vol. 2, 1999, pp. 929–932.

[3] A. Paeschke and W. F. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[4] A. Paeschke, "Global trend of fundamental frequency in emotional speech," in *Speech Prosody 2004, International Conference*, 2004.

[5] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech communication*, vol. 46, no. 3-4, pp. 252–267, 2005.

[6] S. J. Mozziconacci and D. J. Hermes, "Role of intonation patterns in conveying emotion in speech," in *Proceedings of the 14th International Conference of Phonetic Sciences*, 1999, pp. 2001–2004.

[7] S.-A. Jun, "Prosodic typology: By prominence type, word prosody, and macro-rhythm," *Prosodic typology II: The phonology of intonation and phrasing*, pp. 520–539, 2014.

[8] M. E. Beckman, "The parsing of prosody," *Language and cognitive processes*, vol. 11, no. 1-2, pp. 17–68, 1996.

[9] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[10] C. Prechtel, "Macro-rhythm in english and spanish: Evidence from radio newscaster speech," in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 675–679.

[11] J. Wang and J. Han, "Bide: Efficient mining of frequent closed sequences," in *Proceedings. 20th international conference on data engineering*. IEEE, 2004, pp. 79–90.

[12] C. Luo and S. M. Chung, "Efficient mining of maximal sequential patterns using multiple samples," in *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 2005, pp. 415–426.

[13] P. Fournier-Viger, C.-W. Wu, and V. S. Tseng, "Mining maximal sequential patterns without candidate maintenance," in *International Conference on Advanced Data Mining and Applications*. Springer, 2013, pp. 169–180.

[14] M. Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech & Language*, vol. 6, no. 2, pp. 175–196, 1992.

[15] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 216–228, 2007.

[16] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level." in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, 01 2009, pp. 81–84.

[17] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, "Cross-language prominence detection," in *Speech Prosody 2012*, 2012.

[18] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] D. Kocharov and A. Menshikova, "Distributed representation of melodic contours," in *Proceedings of Speech Prosody*, 2018.

[20] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A survey of sequential pattern mining," *Data Science and Pattern Recognition*, vol. 1, no. 1, pp. 54–77, 2017.

[21] A. A. Alvarez and E. S. A. Issa, "Learning intonation pattern embeddings for arabic dialect identification," in *Proc. Interspeech*, 2020, pp. 472–476.

[22] D. Conklin, "Discovery of distinctive patterns in music," *Intelligent Data Analysis*, vol. 14, no. 5, pp. 547–554, 2010.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer program," Version 6.1.38, retrieved 2 January 2021, 2021.

[24] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[25] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the eleventh international conference on data engineering*. IEEE, 1995, pp. 3–14.

[26] A. H. Meftah, M. A. Qamhan, Y. Seddiq, Y. A. Alotaibi, and S. A. Selouani, "King saud university emotions corpus: Construction, analysis, evaluation, and comparison," *IEEE Access*, vol. 9, pp. 54 201–54 219, 2021.

[27] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[28] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[29] A.-S. M. Badawi, *Mustawayat al-arabiyya al-muasira fi Misr*. Dar al-maarif, 1973.

[30] J. C. Watson, *Word stress in Arabic*. Wiley-Blackwell, 2011.

[31] D. Chahal and S. Hellmuth, "The intonation of lebanese and egyptian arabic," 2014.

[32] K. Norlin, "A preliminary description of cairo arabic intonation of statements and questions," *Speech Transmission Quarterly Progress and Status Report*, vol. 1, pp. 47–49, 1989.

[33] S. Hellmuth, "Acoustic cues to focus and givenness in egyptian arabic," *Instrumental studies in Arabic phonetics*, vol. 319, p. 301, 2011.

[34] S. J. Hellmuth, *Intonational pitch accent distribution in Egyptian Arabic*. University of London, School of Oriental and African Studies (United Kingdom), 2006.

[35] M. S. Alzaidi, Y. Xu, and A. Xu, "Prosodic encoding of focus in hijazi arabic," *Speech Communication*, vol. 106, pp. 127–149, 2019.

[36] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[37] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data mining and knowledge discovery*, vol. 15, no. 1, pp. 55–86, 2007.