INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Acoustic and Prosodic Correlates of Emotions in Urdu Speech

Saba Urooj[1], *Benazir Mumtaz*[2], *Sarmad Hussain*[1], *Ehsan ul Haq*[1]

[1]Center for Language Engineering, University of Engineering and Technology, Pakistan
[2]University of Konstanz, Germany

{saba.urooj|sarmad.hussain|ehsan.ulhaq}@kics.edu.pk,
benazir.mumtaz@uni-konstanz.de

Abstract

Emotional speech corpora exhibit differences in duration, intensity and fundamental frequency. We investigated acoustic as well as prosodic correlates of emotional speech in Urdu. We recorded a corpus of 23 sentences from four speakers of Urdu covering four emotional states. Main results show that: a) sadness exhibits lowest utterance rate, lowest intensity and narrow pitch range, b) anger exhibits highest utterance rate, highest intensity and wider pitch range, and c) happiness exhibits higher utterance rate and wider pitch range as compared to neutral and sadness; but no significant differences are found between the intensity and pitch range of anger and happiness. The analysis also shows differences in terms of pitch or phrase accents and boundary tones.

Index Terms: pitch accents, boundary tones, acoustic correlates, fundamental frequency

1. Introduction

Testing of synthesized speech carried out in terms of intelligibility and naturalness [1] shows that synthesized speech is intelligible but less natural. Natural sounding text-to-speech (TTS) systems also require information about emotion using the prosody of a language [2]. Understanding the prosodic features for emotions is a difficult process as similar changes can occur for different emotions [3], e.g., the fundamental frequency (F0) rises for both anger and joy. The current work aims to analyze the pitch contours of emotional utterances to derive a set of intonation rules for integration in our Urdu TTS.

The study in [4] claims that pitch patterns are not statistically significant across emotions. The only difference that can be found is the frequency distribution of different accents across all emotions. The same is true for Urdu i.e. repetition of basic phrasal accent L*Ha across all emotions where Ha functions as a boundary tone for the accentual phrase (AP). Further, [5] could not establish any significance correlation between pitch or phrase accents and emotion. If this is true, the question thus arises how pitch accent L* spoken with happiness differs from the pitch accent L* spoken with anger or sadness. Similarly, how the boundary tone Ha on unaccented syllables spoken with happiness differs from the boundary tone Ha on the unaccented syllables spoken with anger or sadness. The answer lies in our results that these tones are different in terms of their pitch register i.e., the neutral emotion has the lowest F0 followed by sadness, happiness and anger emotions for both L* and L*Ha pitch accents. Also, Ha of happiness has highest F0 value as compared to Ha of neutral and sadness.

Another question is whether the position of words in a sentence matters in carrying emotional information or not. The study in [6] claims that words at the final position of a sentence exhibit more emotion related information as compared to the words in a sentence suggesting that boundary tones also show differences across emotions. The work in [4] found that L% boundary tone outnumbers all other types of boundary tones in all emotions. Our results stand in contrast to this as same is not true for sadness emotion where H% boundary tone outnumbers L% (see Table 1). The study in [5] tried to establish a correlation between ToBI labels and emotion taking pitch accent, phrase accent and boundary tones as features and found boundary tones as the most striking feature. The study claims that positive emotions are positively correlated with the simple declarative contour while the negative emotions are negatively correlated with this contour. Our study stands in accordance with the first claim as happiness is ending on L% for 82% of the time. Our study stands in contrast to the second claim as we have a different result for one negative emotion i.e., anger. We present and discuss these findings in more detail below.

2. Literature review

The analysis of emotional speech is done on multiple levels including acoustic, lexical or prosodic level [7]. Past research has mainly focused on the contribution of acoustic features such as variation in duration, pitch and loudness as measures for determining emotions. Among these features, it has been established that pitch plays the most important part while loudness is the least important [8].

Some of the past studies have analyzed speech databases for speech synthesis and speech production purposes and others have done so for the speech perception and speech recognition purposes. The production-based study in [9] found that segmental duration is longer in sad speech of the female speaker as compared with anger and joyous speech, whereas male speaker's data showed shortest duration for the sad speech. Furthermore, mean F0 was lower in sad speech of both male and female speakers as compared to their anger and joyous speech. Another production study [10] states that anger affects increase of pitch ranges and mean values of intensity or energy. Sadness typically decreases the mean value of the pitch contour and the pitch ranges. Speech spoken in sadness has lower pitch contour.

The perception-based study of Hindi emotional speech in [11] found that sad emotion has the maximum duration i.e. 3.2

http://dx.doi.org/10.21437/Interspeech.2021-910

seconds whereas anger emotion has the minimum duration i.e. 1.6 seconds. For intensity, anger shows highest intensity followed by happiness, surprise, sadness, natural and fear emotions. The study in [11] also found that F0 curve at the end of the utterances falls for anger and sadness and rises for happiness.

Although it has been established that pitch plays the most important role, the studies have mostly looked at global statistical measures of pitch, like range, mean, or variability. The speech synthesizers/recognizers might give better results, if it could be analyzed how emotional states can be differentiated on the basis of intonation patterns.

The basic phrase of Urdu prosody is the accentual phrase AP. It has been found that L*Ha is the most common tonal pattern. Moreover, L% is the most commonly occurring boundary tone in a simple declarative sentence of Urdu [12].

3. Method

3.1. Materials

In the current study, 23 semantically neutral sentences were selected from the CLE Urdu Digest Corpus [13]. In order to elicit an emotion, the semantically neutral target sentences were embedded in the context of a dialogue designed to trigger a specific emotion. Example dialogues for eliciting sadness, happiness and anger emotions for sentence (1) is given in (1a), (1b), and (1c) respectively.

(1) Urdu: ʊs vəqt̪ vo ʃɑed̪ ʧʰe sɑl kɑ t̪ʰɑ.
English: That time he probably six years was
He was probably six years old at that time.

(1a) Asma: Did you find out about the death of Farhan's father?

Saima: Yes dear, I am very sad to hear about his father's death. Poor Farhan, he has suffered a lot! His mother passed away when he was very young. ***He was probably six years old at that time.*** And now his father also left him alone in this entire world.

(1b) Salma: I am very happy for your son's success. What is the secret behind his success?

Asma: It is his passion that has shown this day. Today he has won the first position in the All Pakistan Debating Competition. He has been a very good debater from a very young age. He also won the first prize in the debating competition at a very young age. ***He was probably six years old at that time***. Thank God I am blessed with such a good son.

(1c) Farah: I don't know what is missing in my training that Nadeem is so annoying.

Nadia: It's all your fault. Your unconditional pampering has spoiled him. He has been spoiled since childhood. Do you remember the time when he first ran away from the home? ***He was probably six years old at that time.*** He started blackmailing you at such a young age.

The average length of the 23 target sentences was seven words and all the target sentences selected in this study end with an auxiliary verb. In order to avoid boundary effects, the target sentences were never placed dialog initially or finally. Moreover, no dialog was designed for the neutral emotion. The neutral sentences were presented to participants with fillers to avoid boundary effects.

3.2. Participants

Four professional radio speakers (two males and two females) were selected for the recording. These participants, ranging in age between 25-45 years with 14 -16 years of education, are from Lahore, Pakistan and speak Urdu for communicating at home and outside. They can usually understand Punjabi as it is spoken in their environment.

3.3. Procedure

Participants were recorded individually in four sessions (one for each emotion) on four different days where each session lasted for approximately three hours. The participants were asked to record one emotion in one session to best elicit the target emotion. Productions were recorded via microphone using PRAAT software with a sampling rate of 48 kHz. To maintain the speech quality, sufficient recording time was given to each speaker, with a number of breaks during each recording session. Speakers are provided with the text to read ahead of the recording sessions, so that they are comfortable in acting out the different scenarios.

Each session began with the presentation of dialogues written on screen which participants were instructed to read. A linguist also sat inside the booth and participated in the dialogues as a prompter and also did the initial assessment of the emotion spoken. To verify whether the resulting speech corpus expresses the intended emotion, a forced choice test among anger, happiness, and sadness was performed as a follow-up. The listeners for the perception tests were four university students (two male, two female). Happy speech was identified most successfully among the three kinds of emotional speech for female speakers whereas sad speech was identified most successfully for male speakers. The results of identification of emotion by listeners for the female speech were anger: 90%, happiness: 92%, and sadness: 87%; results for the male speech were anger: 94%, happiness: 87%, and sadness: 97%. The rejected sentences were recorded again in the re-recording sessions and assessed by the linguist participating in the dialogue. As the emotions are elicited by actors and not naturally, these are anticipated to be somewhat exaggerated compared to spontaneous speech. However, perceptual testing still confirmed the emotional content independently.

The resulting spoken corpus of emotions was annotated at syllable, word and intonation tiers. A sample sentence from the annotated speech corpus is provided in Figure 1 where Tier 1 is for words, Tier 2 is for syllables, and Tier 3 is for intonation. The emotion is annotated in the fourth tier.

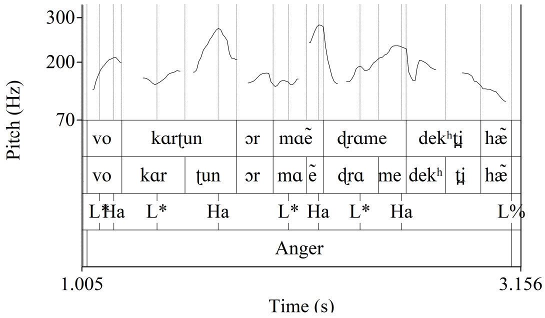

Figure 1: *Annotation of emotional speech corpus at multiple tiers*

The acoustic features analyzed in the data are duration, F0 and intensity. Several utilities are designed for the automatic calculation of the acoustic parameters. It was found during manual verification that fundamental frequency was not picked up correctly by the utility for some of the F0 minima values and hence they were manually reviewed.

3.4. Data treatment

We calculated a series of linear-mixed effects regression (LMER) models for multiple outcome variables i.e. utterance rate, syllable duration of pitch and phrase accents, the intensity of pitch and phrase accents, F0 of pitch accents and boundary tones, and F0 range of overall utterances. For syllable duration, intensity, and F0 models, emotion and pitch/phrase accents were selected as fixed effects whereas for utterance rate and F0 range models only emotion was selected as fixed effects (using the packages lme4 and lmerTest in [14], [15]). The pitch and phrase accents focused on in the above-mentioned models are L*, Ha, and L*Ha. However, for the F0 model, two boundary tones (L%, H%) along with the pitch and phrase accents were also analyzed. For each model participants and items were added as crossed random effects. Moreover, Posthoc tests were conducted using the emmeans function in R with Tukey correction [16] for all models. Note that F0 was calculated in semitones (st) to account for differences in gender.

4. Results

We have done the analysis on acoustic and prosodic level. On acoustic level, we have calculated: a) utterance rate (number of syllables divided by total duration of the sentence), and b) overall F0 range of the utterances. On prosodic level, we have calculated the syllable duration and syllable intensity and F0 according to the pitch and phrase accent types (L*, Ha, L*Ha). The syllables containing L*Ha are the ones which are monosyllabic words. The syllables that contain L* are from bi-syllabic and tri-syllabic words which contain L* on their first syllable and Ha on their last syllable. The comparison of final boundary tones (L%, H%) for all emotions has also been conducted.

4.1. Utterance rate

The results of the utterance rate model showed a significant main effect of emotion on utterance rate (x2=348, df = 3, p < 0.0001). Post-hoc comparisons with Tukey correction of p-values (emmeans-function in R) showed that all emotions differ significantly from others at utterance rate level including anger vs. happiness (ß = 0.42, p < 0.0001), anger vs. neutral (ß = 1.22, p < 0.0001), anger vs. sadness (ß = 0.89, p < 0.0001), happiness vs. neutral (ß = 0.80, p < 0.0001), happiness vs. sadness (ß = 0.46, p < 0.0001), and neutral vs. sadness (ß = -0.33, p < 0.0001). The anger emotion had the highest utterance rate followed by happiness, sadness, and neutral emotions respectively as illustrated in Figure 2.

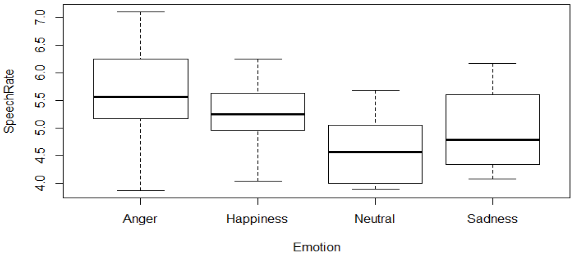

Figure 2: *Speech rate of four emotions averaged across all speakers*

4.2. Syllable duration of accent types (L*, Ha, L*Ha)

LMER model with syllable duration as outcome variable showed a significant interaction between tone and emotion (x2 = 45.4, df = 15, p< 0.0001). To investigate the nature of the interaction, data was split according to tones and the effect of emotions was investigated for these subsets. Results showed that anger emotion has the shortest syllable duration across all pitch accents. In the syllables with L*, the significant differences in duration were noticed between anger vs. neutral (β=-0.01, p < 0.01) and anger vs. sadness emotions (β=-0.01, p < 0.001). In the syllables with Ha, significant differences were found between anger vs. happiness (β=-0.01, p < 0.05), anger vs. neutral (β=-0.04, p < 0.001), and anger vs. sadness emotions (β=-0.02, p < 0.001). In the syllables with L*Ha, significant differences were reported between anger vs. neutral (β=-0.04, p < 0.001) and anger vs. sadness (β=-0.05, p < 0.001).

Moreover, the syllables with Ha were found longer in neutral emotion as compared to happiness (β=-0.02, p < 0.001) and sadness emotions (β=-0.01, p < 0.05). Similarly, the syllables with L*Ha were longer in neutral (β=-0.04, p <0.0001) and sadness (β=-0.04, p < 0.0001) emotions as compared to happiness emotion.

4.3. Intensity of accent types (L*, Ha, L*Ha)

The results of the intensity model showed that there is a significant interaction between emotion and tone (x2 = 35.9, df = 15, p< 0.001). Post-hoc comparisons reported that all pitch accents (L*, Ha, L*Ha) have highest intensity in anger emotion as compared to happiness, neutral and sadness (p<.0001 for all comparisons). Similarly, all pitch accents showed higher intensity in happiness emotion as compared to neutral and sadness emotions (p<.0001 for all comparisons). However, no difference in intensity was found for neutral and sad emotions across all pitch accents (p>0.1 for all comparisons).

4.4. F0 range of utterance

The F0 range in semitones was measured for all emotions. The results showed the significant main effect of emotions on F0 range (x2 = 445, df = 3, p < 0.0001). Post-hoc comparisons showed that anger has wider pitch range than neutral (β=14.9, p < 0.0001) and sadness (β=12.8, p < 0.0001). Similarly, happiness emotion has wider range than neutral (β=15, p < 0.0001) and sadness (β=13, p < 0.0001). However, no significant differences were found between anger and happiness (p =0.9) as well as neutral and sadness (p =0.1).

4.5. F0 of pitch and phrase accents (L*, Ha, L*Ha) and boundary tones (L%, H%)

The results of F0 analysis showed that there is a significant interaction between emotion and tone (x2 = 211, df = 15, p< 0.001). Post-hoc comparisons reported that the neutral emotion has the lowest F0 followed by sadness, happiness and anger emotions for both L* and L*Ha pitch accents. All emotions differ significantly from all others for L* and L*Ha tones (p <.0001 for all comparisons) except the anger and happiness emotion for L*Ha tone (ß = 1.17, p=0.09). Moreover, the happiness emotion has highest F0 for Ha tone as compared to neutral (ß = 4.15, p<.0001) and sadness (ß = 6.58, p<.0001). Neutral emotion has higher F0 for Ha tone as compared to sadness emotion (ß = 2.43, p<.0001). No significant differences were found between anger and happiness emotions (p > 0.05).

H% boundary tone was observed at the end of happiness and sadness sentences. The counts were extracted to see the percentage of L% and H% boundary tones across all emotions as shown in Table 1 below.

Table 1: *Boundary tones across emotions in overall data*

Emotion	H%	L%
Anger	0	100%
Happiness	18%	82%
Neutral	0	100%
Sadness	64%	36%

Analysis was also conducted to see how these L% and H% boundary tones differed across all emotions. It was found that neutral emotion has lowest F0 for L% boundary tone as compared to anger (ß = 6.27, p <.0001), happiness (ß = 7.15, p <.0001) and sadness (ß = -3.76, p <.0001)(see Figure 3). Sadness has lower L% boundary tone as compared to happiness (ß = 3.38, p <.001) and anger (ß = 2.5, p=.005) and there is no significant difference between anger and happiness emotion for L% boundary tone (ß = -0.88, p=.4). Moreover, we also found H% boundary at the end of sadness and happiness emotions. Results showed that happiness emotion has higher H% boundary tone (ß = 2.94, p=0.02) as compared to sadness emotion. The F0 contour for single sentence is shown in Figure 3.

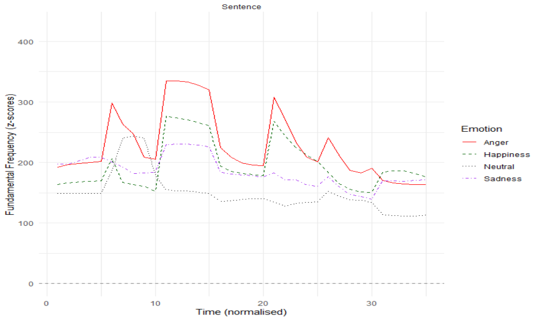

Figure 3: *F0 curve of one sentence "vo karʈun ɔr maẽ ɖrame dekhʈ̪I hæ̃/They (children) watch cartoons and mothers watch dramas" for all four emotions*

5. Discussion

In this study, we statistically analyzed speech data to find out the important acoustic correlates of emotion. As far as utterance rate is concerned, the results indicate that utterance rate follows the following order:

Anger>happiness>sadness>neutral.

In addition, syllable duration was shortest for anger emotion across all pitch accents. Our results are consistent with the findings of previous studies of Japanese [9] and Hindi [11] languages. At prosodic level, it was found that syllables containing L* and L*Ha accents (accented syllables) showed significant differences in duration between anger vs. neutral, and anger vs. sadness emotions. However, the accented syllables of anger and happiness did not show any significant differences. Moreover, syllable with Ha was found longer in neutral emotion compared to happiness and sadness emotion. This shows that duration mostly decreases for anger and happiness emotion as compared to neutral. This tendency of decrease is greater in unaccented syllables as compared to accented syllables in both female and male data.

For intensity, it was found that all pitch/phrase accents have highest intensity in anger emotion as compared to happiness, neutral and sadness. Further, all pitch accents showed higher intensity in happiness emotion as compared to neutral and sadness. Similar results were reported in [10] and [11]. However, no difference in intensity was found for neutral and sadness emotions across all pitch accents suggesting that intensity.

The F0 range at utterance level shows that anger and happiness have wider pitch ranges whereas neutral and sadness have narrow pitch ranges. This is also supported by [10] and [11]. However, no significant differences were found between anger and happiness suggesting that F0 is not a strong clue to differentiate anger from happiness.

The analysis of boundary tones suggests that neutral emotion has lowest L% boundary tone followed by sadness. However, there is no significant difference between the L% boundary tones of anger and happiness. Moreover, H% boundary found at the end of happiness and sadness emotions showed that H% boundary tone is higher in happiness emotion as compared to sadness emotion. This rise of F0 curve was also observed at the end of happy sentences of Hindi language [11]. The percentages of boundary tones suggest that anger and neutral emotions contain L% boundary tone whereas happiness emotion has L% boundary tone in 82% of the data while H% boundary tone was found in 18% of the data. H% boundary tone was found mostly at the end of sadness sentences which is 64% of the data and H% boundary tone was found in 36% of the data (see Table 1).

6. Conclusions

In this paper, we have presented analysis of an emotional speech corpus recorded in Urdu. The emotions considered are anger, happiness, sadness and neutral. The analysis is performed using acoustic and prosodic parameters. The quality of the emotional corpus recorded for this purpose is evaluated using forced choice perception tests. We are currently integrating the results of our research in Urdu TTS system to improve its naturalness.

7. Acknowledgements

This work is funded by the DAAD German Academic Exchange Service under the project "Urdu Text-to-Speech: Integrating Prosody of Emotions".

8. References

[1] K. S. Shahid, T. Habib, B. Mumtaz, F. Adeeba, and E. U. "Subjective Testing of Urdu Text-to-Speech (TTS) System, LANGUAGE & TECHNOLOGY, 2016, p. 65.

[2] N. Campbell, "Prosody and the selection of units for concatenation synthesis," in Proceedings of the Second ESCA/IEEE Workshop in Speech Synthesis, 1994, pp. 61-64.

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

[4] R. Stibbard, "Automated extraction of ToBI annotation data from the Reading/Leeds emotional speech corpus," in ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000.

[5] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in Eighth European Conference on Speech Communication and Technology, 2003.

[6] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," International journal of speech technology, vol. 2, no. 16, pp. 143-160, 2013.

[7] J. E. Cahn, "Generating expression in synthesized speech," Massachusetts Institute of Technology, Doctoral dissertation 1990.

[8] R. W. Frick, "Communicating emotion: The role of prosodic features," Psychological bulletin, vol. 97, no. 3, 1985.

[9] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," Speech communication, vol. 40, no. 1-2, pp. 161-187, 2003.

[10] J. E. Noad, S. E. Whiteside, and P. Green, "A macroscopic analysis of an emotional speech corpus," in Fifth European Conference on Speech Communication and Technology, 1997.

[11] S. S. Agrawal, "Emotions in Hindi speech-analysis, perception and recognition," in International Conference on Speech Database and Assessments (Oriental COCOSDA). IEEE, 2011.

[12] S. Urooj, B. Mumtaz, and S. Hussain, "Urdu intonation," Journal of South Asian Linguistics, vol. 10, pp. 3-22, 2020.

[13] S. Urooj, S. Hussain, F. Adeeba, F. Jabeen, and R. Parveen, "CLE Urdu digest corpus," in LANGUAGE & TECHNOLOGY, 2012, p. 47.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," in arXiv preprint arXiv:1406.5823, 2014.

[15] A. Kuznetsova, P. B. Brockhoff, and R. H. Christensen, "lmerTest package: tests in linear mixed effects models," Journal of statistical software, vol. 82, no. 13, pp. 1-26, 2017.

[16] R. Lenth, H. Singmann, J. Love, B. Buerkner, and M. Herve, "Emmeans: Estimated marginal means, aka least-squares means," R package version , vol. 1, no. 1, p. 3, 2018.