forum acusticum 2023

ACOUSTIC MARKERS IMPACTING DISCRIMINATION ACCURACY OF EMOTIONS IN VOICE

Baiba Trinite **Anita Zdanovica** **Ilva Magazeina**
Anita Jansone **Daiga Kurme** **Evija Lavrane**
Voice and Speech Research Laboratory, Liepaja University, Latvia

ABSTRACT

The quality of communication depends on how accurately the listener perceives the intended message. In addition to understanding the words, listeners are expected to interpret the speaker's accompanying emotional tone. However, it is not always clear why a neutral voice can be perceived as affective or vice versa. The present study aimed to investigate the differences between the acoustic profiles of angry, happy, and neutral emotions and to identify the acoustic markers that can lead to misperception of emotions conveyed through the voice.

The study employed an encoding-decoding approach. Ten professional actors recorded the Latvian word /laba:/ in neutral, happy, and angry intonations, and thirty-two age-matched respondents were asked to identify the emotion conveyed in the heard voice sample. A complete acoustic analysis was conducted for each voice sample using PRAAT, which included fundamental frequency (F_0), intensity level (IL), spectral (HNR) and cepstral parameters (CPPs), and duration of a produced word (DPW). The vocal expressions of emotions were analyzed from both encoding and decoding perspectives.

The results showed statistically significant differences in the acoustic parameters that distinguish vocally expressed happy and angry emotions from neutral voices and acoustic parameters that were different between happy and angry emotions.

*Corresponding author: baiba.trinite@liepu.lv

Keywords: *emotions, voice, acoustic analysis.*

1. INTRODUCTION

Voice is a critical aspect of communication as it provides nonverbal information about the speaker's emotional state, health, social status, age, and gender. Communication quality depends on how the speaker's message, including linguistic and paralinguistic components, is perceived. Decoding, or the perception of the voice, is the final output of the encoding process of voice production. The correct recognition of vocal emotion relies on sharing the same knowledge about what a vocal emotion sound like [1]

The modulation of acoustic parameters in vocal expression can convey specific emotions to listeners [2, 3], and the expression of emotions in the voice is differentially patterned [4]. Numerous acoustic variables are involved in expressing emotions in the voice signal, including fundamental frequency, voice intensity or sound pressure level, energy distribution in the frequency spectrum, location of formants, and temporal parameters [3-9].

The previous research investigating acoustic characteristics of emotions has been mainly focused on English, German, Italian, Swedish, Spanish, and other languages [3-13] while there are no such studies in the Latvian language. Moreover, previous research has not sufficiently addressed the misperception of emotions based on vocal signal acoustic features. Therefore, the study aimed to fill those gaps by investigating differences between acoustic profiles of angry, happy, and neutral emotions and finding acoustic markers that lead to misperception of emotions in voice.

The current study is a pilot investigation, where one word, extracted from a larger corpus of Latvian speech samples representing happy, angry, and neutral tones of voice (N = 270), was analyzed.

2. MATERIAL AND METHODS

Ten professional actors (5 females and 5 males) participated in the study by recording the words, phrases, and text samples pronounced in neutral, happy, and angry intonation according to written descriptions of various emotional states provided to the actors before the recording session.

The recordings were conducted in a soundproof room using a calibrated head-worn AKG C520 microphone placed 5 cm away from the mouth. The audio signals were captured using an audio interface Scarlett Solo (Focusrite Plc) connected to a MacBook Air. The voice signals were recorded at a sampling rate of 44.1 kHz, a resolution of 16 bits, and saved in wav format using the software PRAAT (v.6.1.31).

The listening task was built with the open-source Python Kyvi framework and presented as a mobile application on the PC tablet Lenovo TB-X606X. The experimental design was based on an oddball paradigm, i.e., voice signals representing one emotional state were grouped in one block, including deviant voice samples. For example, 90 happy voice samples, 5 angry and 5 neutral voice samples served as oddballs. The experimental blocks (happy, neutral, and angry voices) were randomly shuffled and voice stimuli inside the block were presented in pseudorandom order. The developed application had a user interface with the task to chose one of three emotional valences after listening to the voice stimulus.

Thirty-two age-matched participants without hearing disorders were invited to participate in the study. The mean age of female participants was 28.8 years (SD = 12.3) and 29.3 years (SD = 12.1) for males. After a short practice session, they listened to recorded voice samples using AKG K240 headphones and were asked to determine the voice emotion. All responses were automatically saved as a text file .csv.

One word /laba:/ (meaning "right," with the direction connotation) was selected from all linguistic units used in the experiment for further analysis. This word was chosen because it had two open syllables, two voiced consonants, and no consonant clusters. The word was pronounced in neutral (n = 10), happy (n = 10), and angry (n = 10) intonation.

Acoustic analysis of the voice samples was conducted using PRAAT software (v.6.1.31) and Phonanium scripts: Vocal fundamental frequency (v.02.04), Vocal intensity level (v.02.04), and Cepstrography and quefrency-domain analyses (v.01.02). Minimum (min), 10th percentile (P10), 25th percentile (P25), 50th percentile (P50), mean (M), 75th percentile (P75), 90th percentile (P90), maximum (max), standard deviation (SD), interquartile range (IQR), range between 10th and 90th percentiles (P90-P10) were estimated for fundamental frequency (F_0), vocal intensity level (IL), and smoothed cepstral peak prominence (CPPs). Phonanium script Sound spectrography and acoustic voice markers (v.02.03) was used for measurements of harmonic-to-noise ratio (HNR) and the PRAAT Voice report determined the duration of a produced word (DPW).

Statistical analysis was done using SPSS software (v.28; SPSS Inc., New York, NY). The Shapiro-Wilk test was performed for each parameter to assess the data distribution. All parameters were found to be normally distributed; however, many of them had significant outliers. Therefore, both parametric and non-parametric methods of data analysis were used. The differences in acoustic parameters between neutral, happy, and angry emotions were investigated using paired samples t-tests for normally distributed data and Related-Samples Friedman's Two-way analysis with Post Hoc Kruskal-Wallis tests, with Bonferroni corrections applied for non-parametric data. Associations between acoustic parameters and the number of detected emotions were analyzed using both Pearson product-moment correlation and Spearman's rank-order correlation methods.

3. RESULTS

The word /laba:/ was spoken by ten professional actors to convey neutral, happy, and angry emotions. A total of 30 voice samples, consisting of 10 neutral (N), 10 happy (H), and 10 angry (A) voices, were recorded and analyzed. The results, shown in Table 1, demonstrate statistically significant differences ($p < .05$) in acoustic parameters between neutral and angry, neutral and happy, and angry and happy voices.

Table 1. Statistically significant differences found in acoustic parameters of vocally expressed different emotions.

Emotion comparison	Acoustic parameters ($p < .05$)
Neutral vs. Angry	F_0 M, F_0 P90, F_0 max, F_0 SD, F_0 IQR, F_0 P90-P10 IL min, IL P10, IL P25, IL P50, IL M, IL P75, IL P90, IL max, IL SD, IL IQR, IL P90-P10 HNR CPPs SD, CPPs IQR
Neutral	F_0 P50, F_0 P75, F_0 max

vs. Happy	IL P25, IL P50, IL M, IL P75, IL P90 CPPs overall
Happy vs. Angry	DPW IL min, IL P50, IL M, IL P75, IL P90, IL max, IL SD

The mean discrimination accuracy for neutral emotions was 76.6% (N-N), for happy emotions was 55.3% (H-H), and for angry emotions was 76.6% (A-A). Neutral voices were perceived as happy in 7.5% of cases (N-H) and as angry in 15.9% (N-A). Happy voices were identified as neutral in 35.6% (H-N) and as angry in 9.1% of cases (H-A). Angry voices were perceived as neutral in 20.3% of cases (A-N) and as happy in 3.1% (A-H). The actors' gender did not have a statistically significant impact on the discrimination accuracy of emotional valence.

The correlation method was used to determine the relationships between the acoustic parameters of vocal signals and the number of correctly detected emotions by 32 listeners. Moreover, relationships between acoustic parameters and the number of misperceived vocal emotions were analyzed. Only statistically significant associations ($p < .05$, $p < .01$) were reported in Tables 2, 3, and 4. Strong positive and negative correlations between $r = .634$ and $r = .888$ were found between the acoustic parameters of recorded voice samples and decoded neutral, happy, and angry emotions.

Table 2. Statistically significant associations between acoustic parameters and neutral voice stimuli.

Parameter	N-N	N-H	N-A
DPW	-.792**		
F_0 IQR		.786**[a]	
F_0 SD		.653*[a]	
F_0 P90-P10		.653*[a]	
IL P10		-.830**[a]	
CPPs P25	-.696*		.634*[a]
CPPs overall	-.679*		
CPPs P50	-.713*		
CPPs M	-.693*		.684*[a]
CPPs P75	-.731*		.665*[a]
CPPs P90			.665*[a]
CPPs max			.772**[a]

**. Correlation was significant at the 0.01 level (2-tailed).
*. Correlation was significant at the 0.05 level (2-tailed).
[a]. Spearman's rank-order correlation.

Table 3. Statistically significant associations between acoustic parameters and happy voice stimuli.

Parameter	H-H	H-N	H-A
F_0 P75	.685*		
F_0 P90	.759*	-.656*	
F_0 max	.859**	-.759*	
F_0 IQR	.729*[a]		-.683*[a]
F_0 SD	.798**	-.684*	
F_0 P90-P10	.888**[a]	-.732*[a]	
IL P25			.695*[a]
IL P50			.695*[a]
IL M			.634*[a]
IL P75			.665*[a]
IL P90			.708*[a]
IL max			.652*[a]
CPPs P10	-.730*	.767**	
CPPs P50		.644*	
CPPs M		.699*	
CPPs P75		.666*	

**. Correlation was significant at the 0.01 level (2-tailed).
*. Correlation was significant at the 0.05 level (2-tailed).
[a]. Spearman's rank-order correlation.

Table 4. Statistically significant associations between acoustic parameters and angry voice stimuli.

Parameter	A-A	A-N	A-H
DPW		-.716*[a]	
IL P25	.677*[a]	-.695*[a]	
IL P50	.707*[a]	-.689*[a]	
IL P90-P10			-.755*[a]

*. Correlation was significant at the 0.05 level (2-tailed).
[a]. Spearman's rank-order correlation.

In order to investigate the relationships between the median values of fundamental frequency, intensity level, harmonic-to-noise ratio, and CPPs in neutral, happy, and angry vocal expressions, a correlation analysis was carried out. Statistically significant correlations were found between CPPs P50 and IL P50 ($r = .782$, $p = .008$) and between F_0 P50 and HNR ($r_S = .855$, $p = .002$) in neutral voice samples. In happy and angry voice samples, a significant correlation was found between F_0 P50 and HNR ($r = .642$, $p = .045$; $r_S = .685$, $p = .029$).

4. DISCUSSION

Fundamental frequency, energy, and speech rate were the most common acoustic cues investigated in the studies of the differentiation of emotions [5]. Therefore, it was interesting to investigate these parameters in voice samples of the Latvian language because the perception of emotions can be culturally and linguistically dependent [4]. In addition, the acoustic analysis of emotional voice stimuli was expanded by HNR because HNR was related to the clarity of emotional expression [10], and it showed valence-dependent sensitivity during the neural decoding of aggressive and joyful vocalizations [14]. Voice quality is a central aspect of emotional vocalization [9], and CPPs is one of the most promising measures for the acoustic measurement of overall voice quality [15]. Therefore, also CPPS was included in the analysis of perceived emotional vocalizations.

4.1. Differences between acoustic profiles of neutral, happy, and angry emotions

The first objective of this study was to examine differences between acoustic profiles of neutral, happy, and angry vocal expressions (Table 1). The data analysis revealed that neutral and angry voices exhibited statistically significant differences in mean and ceiling values, range and variation of fundamental frequency, and intensity level. Angry voices also had higher intensity floor values than neutral voices. Furthermore, variations in harmonic-to-noise ratio and CPPs were also found to differ between neutral and angry voices. Discriminators between neutral and happy voices included higher fundamental frequency ceiling values and values above the 50th percentile, higher intensity values above and below P50, and overall CPPs.

The role of fundamental frequency in characterizing affective emotions has been well-established in previous research. For example, mean F_0 is considered a classic indicator of arousal [9]; higher mean F_0, F_0 SD, and higher F_0 range characterize happiness [1, 7, 10], while anger is characterized by increased mean, variance, and range of F_0 [1]. Both happiness and anger are also characterized by a higher pitch compared to neutral voice. [10].

The review of empirical data on acoustic patterning of basic emotions shows that angry speech has been described by an increase in mean intensity [2]. Also, Ekberg et al., investigating acoustic features distinguishing emotions in Swedish speech, found that loudness was significantly higher for anger and happiness than other emotions [10]. The present study supports these findings, as F_0 and intensity were found to be discriminators between neutral and both affective emotions.

The duration of produced words and intensity differences were statistically significant acoustic markers for differentiating angry and happy emotions. Words with angry emotions were pronounced slower than happy, which aligns with a study investigating emotion decoding in Italian pseudowords [11]. The stronger intended intensity yielded a slower speech rate [7]. Interestingly, fewer acoustic differences were found between angry and happy voices compared to neutral-affective voice models. This is consistent with the findings of another study where anger and happiness did not differ significantly from each other in any parameter [10].

4.2. Acoustic markers contributing misperception of emotions in voice

The second objective of this study was to identify acoustic markers that contribute to the misperception of emotions in voice. The results showed that the discrimination accuracy of neutral, angry, and happy emotions in vocal expression ranged between 57% and 77%. These findings are consistent with the accuracy of vocal emotion recognition reported by Scherer [5]. Vocally expressed affective emotions had a higher discrimination accuracy than neutral voice, which agrees with previous studies indicating that portrayals with strong emotion intensity yielded higher decoding accuracy than portrayals with weak intensity [7]. Consistent with Grichkovtsova et al. [6], the study found that the voice quality of anger was more accurately perceived than happiness.

The study's results suggest that lower CPPs values and slower speech were more strongly associated with neutral voices, based on ratings from 32 participants (Table 2). In addition, increased values of F_0 above P50, high F_0 variability, wide frequency range, and decreased CPPs floor values were found to be acoustic markers of vocally expressed happy emotions (Table 3). The listeners' decision about angry voices was more determined by voice loudness, specifically the median and first quartile of vocal intensity level (Table 4). These findings suggest that listeners use emotion-specific patterns of cues to decode emotions, with the fundamental frequency being an essential marker of happy emotions and voice intensity being important for the expression of anger, in line with previous research [7, 11].

Confusion patterns can be valuable in identifying the degree of similarity or proximity between different emotion categories [5]. The confusion matrices obtained from misperceived neutral voices demonstrated that higher pitch (F_0) variability, broader F_0 range, and decreased loudness

(IL) are associated more with happy than neutral voices. Higher CPPs values make listeners hear angry emotions in words produced in a neutral mood and neutral emotions in words produced in happy intonation. At the same time, neutral voice samples with lower CPPs scores were more precisely identified as neutral. CPPs is an acoustic measure of overall voice quality, where higher values represent a more periodic voice signal and a more harmonic spectrum [15]. The results demonstrate a strong positive correlation between the median values of CPPs and intensity in neutral voice samples. Therefore, the variations in intensity can explain the perception of angry emotion instead of neutral. Neutral words produced at a higher intensity had higher CPPs and were perceived as angry, while quieter words had lower CPPs and were more convincingly rated as normal.

However, it was more challenging to explain the role of CPPs in the neutralization of happy emotions. Hillenbrand et al.'s works describe the strong correlation between perceptual measures of breathiness and CPP magnitude [16]. In non-pathological voices, breathy phonation can be related to various socio-pragmatic functions, such as interpersonal relationships and emotions [17]. According to Murray and Arnott [18], a breathy voice can be associated with anger and happiness. Therefore, we can speculate that happy voices that were portrayed with less breathiness, i.e., with higher CPPs, were perceived by listeners as more neutral than happy.

The perceived decrease in pitch ceiling values and pitch range leads to neutralizing happy emotions, but a decrease in speech rate and loudness leads to neutralizing angry emotions.

Anger and happiness are emotions that have similar arousal but opposite valence. Our study found that discrimination errors between these two emotions were low, with happy emotions identified as angry in 9.1% of cases and angry emotions identified as happy in 3.1% of cases. When the pitch was more stable and the loudness was higher, the listener tended to perceive happy emotions as angry. On the other hand, more narrow loudness range is perceived as belonging to the happy emotions rather than angry.

4.3. Limitations of the study

This small pilot study investigated differences in acoustic profiles and markers of misperceived angry, happy, and neutral emotions in one word in the Latvian language /laba:/. However, the small number of analyzed voice samples determines the current study's limitations. Furthermore, the analysis was limited to a small set of acoustic parameters, including F_0, IL, CPPs, HNR, and speech rate. In future research, it would be beneficial to extend the acoustic analysis to include more spectral parameters and examine a larger corpus of Latvian speech samples to explore the acoustic markers of neutral, happy, and angry voices across linguistic units of different lengths, such as words, phrases, and text.

4.4. The potential relevance of the study

The studies investigating differences between acoustic profiles of emotions of different valences and identifying acoustic markers that lead to misperception of emotions in voice can be helpful in the improvement of communication training programs, where individuals can be trained to improve their ability to interpret and convey emotions through voice accurately. The study findings can contribute to speech therapy techniques working with individuals who struggle to express or recognize emotions through voice. By targeting specific acoustic markers and using biofeedback, therapists can develop tailored interventions for patients. Finally, the findings can be incorporated into artificial intelligence systems, helping developers to improve the systems' ability to detect and respond to user emotions accurately, thereby enhancing human-machine communication.

5. CONCLUSIONS

Overall, the study found that certain acoustic parameters can be used to distinguish between vocally expressed happy and angry emotions from neutral voices in the Latvian language. Specifically, fundamental frequency, intensity, CPPs, and HNR were important markers for distinguishing between emotional and neutral voices.

The study also found that the discrimination accuracy of emotions in vocal expressions was high, with affective emotions being better decoded than neutral emotions.

The study further revealed that specific acoustic parameters were associated with specific emotions. For example, F_0 was found to be a main acoustic marker for happy emotions, while intensity was associated with angry emotions. In addition, higher F_0 variance and range were found to lead to the perception of a neutral voice as happy, while more pronounced CPPs values led to the perception of a neutral voice as angry. In contrast, lower F_0 values and more pronounced CPPs values led to the neutralization of happy voices, while faster speech rate and lower intensity led to the neutralization of angry emotions.

6. ACKNOWLEDGMENTS

The Latvian Science Council funds the study, project "Affective and disordered vocal stimuli neural processing during mobile task: an EEG study", No. lzp-2021/1-0159.

7. REFERENCES

[1] S. Schaerlaeken, and D. Grandjean: "Unfolding and dynamics of affect bursts decoding in humans," *PLos ONE*, vol. 13, no. 10, pp. 1-21, 2018.

[2] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," In: Lewis, M. Haviland, J. (Eds.), *Handbook of emotions*, second ed. New York: Guilford, 2000, pp. 220-235.

[3] M. Belyk and S. Brown: "The acoustic correlates of valence depend on emotion family," *Journal of Voice*, vol. 28, no. 4. Pp. 523.e9-523.e18, 2013.

[4] R. Banse and K. Sherer: "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614-636, 1996.

[5] K. R. Scherer: "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227-256, 2003.

[6] I. Grichkovtsova, A. Lacheret, and M. Morel: "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, no. 3, pp. 414-429, 2012.

[7] P. N. Juslin and P. Laukka: "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, no. 4, pp. 381-412, 2001.

[8] M. Van Mersbergen and A. E. Payne: "Cognitive, emotional, and social influences on voice production elicted by three different stroop tasks," *Folia Phoniatrica et Logopaedica*, pp. 1-9, 2020.

[9] S. Patel, K. R. Sherer, E. Bjorkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological Psychology*, vol. 87, pp. 93-98, 2011.

[10] M. Ekberg, G. Stavrinos, J. Andin, S. Stenfelt, and O. Dahlstrom: "Acoustic features distinguishing emotions in Swedish speech," *Journal of Voice*, Article in Press, 2023.

[11] E. Preti, C. Suttora, and J. Richetin: "Can you hear what I feel? A validated prosodic set of angry, happy, and neutral Italian pseudowords," *Behavior Research Methods*, vol. 48, pp. 259-271, 2016.

[12] M. Guzman, S. Correa, D. Munoz, and R. Mayerhoff: "Influence on spectral energy distribution of emotional expression," *Journal of Voice*, vol. 27, no. 1, pp. 129.e1-129.e10.

[13] K. Hammerschmidt and U. Jurgens: "Acoustic correlates of affective prosody," *Journal of Voice*, vol. 21, no. 5, pp. 531-540.

[14] S. Fruhholz, W. van der Zwaag, M. Saenz, P. Belin, A. K. Schobert, P. Vuilleumier, and D. Grandjean: "Neural decoding of discriminative auditory object features depends on their socio-affective valence," *Social Cognitive and Affective Neuroscience*, vol. 11, no. 10, pp. 1638-1649, 2016.

[15] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, P. Corthals: "Acoustic measurement of overall voice quality: a meta-analysis," *Journal of Acoustic Society of America*, vol. 126, pp. 2619-2634, 2009.

[16] J. Hillenbrand, R. A. Cleveland, and R. Erickson: "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hering Research*, vol. 37, pp. 769-778, 1994.

[17] M. Hejna, P. Šturm, L. Tylečkova, and T. Boril: "Normophonic breathiness in Czech and Danish: are females breathier than males?" *Journal of Voice*, vol. 35, no. 3, pp. 498.e1-498.e22, 2020.

[18] I. R. Murray and J. L. Arnott: "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2. pp. 1097-1108, 1993.