The face, sound and expressiveness of voice quality settings

Juliana Andreassa, Alice Crochiquia, Rafael Scarpelli, Mateus Pires
LAEL, PUC-SP, Brazil

https://doi.org/10.36505/ExLing-2024/15/0005/000630

Abstract

In this paper, we investigate associations between vocal and facial gestures and meaning effects in utterances produced by an actor with labial, mandibular and tongue-tip phonetic settings of voice quality. The corpus comprises 12 videos. Each video refers to an utterance produced with a different voice quality setting. Four kinds of analysis were performed: a perceptual analysis of voice quality settings and prosodic features by four experts on the use of the VPA system; an automatic analysis of facial AUs and related basic emotions, valence, and arousal states; an analysis of acoustic parameters extracted with the Prosody Descriptor Extractor Script; and a perceptual test to investigate 30 listeners' judgements of semantic criteria based on their listening to the audio stimuli. The Exploratory multivariate statistical analysis was applied to correlate the quantitative and qualitative variables concerned.

Keywords: voice quality, facial expression, multimodality.

Introduction

Voice qualities are expressive gestures. For centuries, metaphors have described them based on the sensations (Fónagy 2000) accompanying sound production (proprioceptive, tactile, motor sensations) and the auditory impressions they give rise to.

Facial gestures are especially relevant to consider alongside voice in the attribution of paralinguistic and extralinguistic meaning in speech as the changes in the facial plan can directly affect the acoustic properties in speech (Banse and Scherer 1994), and the use of particular voice quality settings, vocal prosody and their acoustics effects can lead to the correlation to facial expressions, such as Lip Spreading and the aural perception of a smile and emotions with neutral or positive valence (Madureira and Fontes 2023). Unsurprisingly, the perception of emotions is far more efficient in bimodal situations, where people have both auditory and visual information. However, some expressions of emotions may be perceived somewhat more accurately in one channel in isolation than the other, like anger in the visual modality and surprise with audio (Abelin 2008).

The experiment presented in this paper has the following aim: to investigate associations between vocal gestures and their acoustic output, facial gestures,

and meaning effects in utterances produced by an actor with labial, mandibular and tongue tip phonetic settings of voice quality.

Material and methods

Corpus and subject

The corpus contains 12 utterances. They were recorded with a Canon 60D camera connected to a lapel camera with a 24mm lens in full HD files with lightning quality. The speaker was always in a frontal position.

The research subject is a 33-year-old actor from the State of São Paulo, Brazil. The speaker was told to produce the sentence: "*O ignorante rejeita, o sábio duvida, o sensato reflete*" (The ignorant one rejects, the wise one doubts, the sensible one reflects), using a different setting for each repetition. Each utterance was produced with a distinct voice quality setting. The chosen settings were (1) Minimised Labial Range; (2) Extensive Labial Range; (3) Lip Rounding; (4) Lip Spreading; (5) Labiodentalization; (6) Closed Jaw; (7) Open Jaw; (8) Protruded Jaw; (9) Extensive Mandibular Range; (10) Minimised Mandibular Range; (11) Advanced Tongue Tip/Blade; (12) Retracted Tongue Tip/Blade.

Perceptual-semantic analysis test: corpus, descriptors and judges

Only the audios were selected to be included in the perceptual-semantic test. The test was presented entirely online to 30 Brazilian Portuguese native speakers, male and female, aged between 20 and 60. The judges were asked to listen to 12 audios to evaluate the speaker using the following adjectives: pleasant/unpleasant, fragile/strong, relaxed/tense. The semantic descriptors chosen by the judges for each utterance were rated in positive, negative, and neutral levels from -3 to 3 on a scale, and the measures related to these levels were weighted and yielded a value for each of the polar semantic descriptors.

Acoustic analysis

The Prosody Descriptor Extractor script (Barbosa 2020) for Praat (Boersma and Weenink 2021) was used for the acoustic analysis. It extracts 22 acoustic measures related to frequency and intensity acoustic parameters.

Perceptual voice quality analysis

Four phoneticians carried out the perceptual voice quality analysis using the VPA protocol (Laver and Mackenzie Beck 2007). In this study, we have included just the first 12 settings described in the protocol, which cover labial, mandibular, and tongue-tip phonetic settings.

Facial expression analysis

For the facial analysis, the software FaceReader 8.1 was used. This software characterizes the facial Action Units (AUs) based on the FACS system (Ekman

and Friesen, 1971) and associated basic emotions, Valence, and Arousal. For this study, 7 emotions (Happiness, Sadness, Anger, Surprise, Fear, Disgust, and Contempt), 20 AUs, Arousal (activated and passive), and Valence (positive and negative) were considered.

Principal component analysis

The variables were analyzed using the Multi-Factor Analysis (MFA) (Josse, Pagès and Husson 2008). All the variables were normalized by z-score.

Results

Table 1 shows the significant variables and the degree of correlation of the variables in Dimensions 1, 2, 3, and 4. Positive values are highlighted in light orange, and negative values in light blue. Based on the percentages of correlation between variables with positive values and variables with negative values, variables were associated.

Dim. 1		
quanti	co.	p. value
AU26	0,8222	0,001
AU10	0,8027	0,0017
AU27	0,7799	0,0028
AU25	0,7737	0,0032
Disgust	0,7528	0,0047
f0base	0,7419	0,0057
Arousal	0,7131	0,0092
AU07	0,7009	0,0111
Laxness	0,6726	0,0165
f0min	0,6592	0,0197
AU09	0,6164	0,0328
Surprise	0,5869	0,0449
Happiness	0,5816	0,0473
sdf0peak	-0,5889	0,0439
Tenseness	-0,6148	0,0334
AU24	-0,6157	0,0331
quali	R2	p.value
MEJ	0,3687	0,0363

MEJ = Mandibular Extensive Range

Dim. 2		
quanti	co.	p. value
AU02	0,7556	0,0045
f0SAQ	0,6983	0,0115
df0posmean	0,6698	0,0172
f0sd	0,664	0,0185
AU05	0,6601	0,0195
Pleasantness	0,5945	0,0415
AU01	0,5929	0,0422
f0peak_ratc	-0,61	0,0353
Unpleasantness	-0,6157	0,0331
quali	R2	p.value
MOJ	0,4963	0,0105

MOJ = Open Jaw

Dim. 3		
quanti	co.	p. value
AU12	0,791	0,0022
Valence	0,7664	0,0036
AU06	0,682	0,0146
Happiness	0,6238	0,0302
Sadness	-0,619	0,0317
AU43	-0,65	0,022
AU15	-0,684	0,0142
AU17	-0,747	0,0053
quali	R2	p.value
MPJ	0,3908	0,0297

MPJ = Protruded Jaw

Dim. 4		
quanti	co.	p. value
jitter	0,7093	0,0098
sll.TAShigh	0,6972	0,0117
Strength	0,6896	0,0131
emph	-0,63	0,028
Fragility	-0,816	0,0012
quali	R2	p.value
TIA	0,3424	0,0457

TIA = Advanced Tongue Tip/Blade

Table 1. Significant variables, percentages of correlation (co.) and p.values.

Conclusions

Based on the correlation percentages and the voice quality settings of the stimuli, six frames containing the variables related to the 3 semantic descriptors included in the Perceptual Test (Pleasantness, Strength, and Tenseness) were derived.

Frame 1 - Pleasantness: Labial Extensive Range, Lip Spreading, Open Jaw, Happiness, Surprise, AU06 (Cheek Raiser), AU12 (Lip Corner Puller), Positive Valence, Interquartile Semi-Amplitude of f0 (f0SAQ), Mean of f0 Positive First Derivative (df0posmean), and Standard Deviation of f0 (f0sd). Features related to utterances 2, 4, and 7.

Frame 2 - Unpleasantness: Labial Minimized Range, Close Jaw, Mandibular Extensive Range, Advanced Tip Blade and Retracted Tip Blade, AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU24 (Lip Pressor), AU43 (Eyes Closed), AU 07 (Lid Tightner), Sadness, Disgust, Negative Valence, Rate of f0 peaks (f0peak_rate). Features related to utterances 1, 6, 9, 11, and 12.

Frame 3 - Fragility: Labial Minimized Range, Labiodentalization, Mandibular Minimized Range, Advanced Tip Blade, AU06 (Cheek Raiser), AU12 (Lip Corner Puller), and Spectral Emphasis (emph). Features related to utterances 1, 5, 10, and 11.

Frame 4 - Strength: Labial Extensive Range, Lip Rounding, Mandibular Protruded Jaw, Retracted Tip Tongue, AU10 (Upper Lip Raiser).Features related to utterances 2, 3, 6, 8, and 12.

Frame 5 -Laxness: Lip Rounding, Open Jaw, Lingual Advanced Tip, AU25 (Lip Part), AU26 (Jaw Drop), and f0 Minimum (f0min). Features related to utterance 3, 7 and 11.

Frame 6 - Tenseness: Labial Minimized Range, Labial Extensive Range, Mandibular Minimized Range, Close Jaw, AU24 (Lip Pressor) and Standard Deviation of F0 peaks (sdF0peaks), Jitter and (sILTAShigh) and. Features related to utterances 1, 2, 6, and 10.

References

Abelin, Å. 2008. Seeing glee but hearing fear? Emotional McGurk effect in Swedish. Proceedings of Speech Prosody. May 6, 9.

Banse, R., Scherer, K. R. 1994. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, vol. 70, no. 3, pp. 614–636.

Barbosa, P. A. 2020. Prosody Descriptor Extractor (for Praat). [Online].

Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer. [Online].

Ekman, P., Friesen, W. V. 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124–129.

Ekman, P., Friesen, W. V., Hager, J. C. 2002. Facial Action Coding System.

Fónagy, I. 2000. Languages within languages: an evolutive approach. John Benjamins.

Josse, J., Pagès, J., Husson, F. 2008. Testing the significance of the rv coefficient. Computational Statistics amp; Data Analysis, vol. 53, no. 1, pp. 82–91.

Laver, J., Mackenzie Beck, J. 2007. Vocal profile analysis scheme-VPAS. Queen Margareth University College-QMUC, Speech Science Research Centre, Edinburgh, Handout.

Madureira, S., Fontes, M.A.S. 2023. Multimodal impressions of voice quality settings: the role of vocal and visual symbolism. Frontiers in Communication, vol.8. [Online].