Voice Quality of European Portuguese Emotional Speech

Ana Nunes[1], Rosa Lídia Coimbra[2], and António Teixeira[3]

[1] Universidade de Aveiro, Portugal
[2] Dep. Línguas e Culturas, Universidade de Aveiro, Portugal
[3] Dep. Electrónica, Telec. & Informática/IEETA, Univ. Aveiro, Portugal

Abstract. In this paper we investigate parameters related to voice quality in European Portuguese (EP) emotional speech. Our main objectives were to obtain, to our knowledge for the first time, values for the parameters commonly contemplated in acoustic analyses of emotional speech and investigate if there is any difference for EP relative to the results obtained for other languages. A small corpus contemplating five emotions (joy, sadness, despair, fear, cold anger) and neutral speech produced by a professional actor was used. Parameters investigated include fundamental frequency, jitter, shimmer and Harmonic Noise Ratio. In general, results were in accordance with the consulted literature regarding F0 and HNR. For jitter and shimmer our results were, in certain aspects, similar to the ones reported in a study of emotional speech for Spanish, another Latin language. From our analyses, and taking into consideration the reduced size of our corpus and the use of an actor as informant, no clear EP characteristic emerged, except for a possible, needing confirmation, difference regarding joy, with values similar to neutral speech.

1 Introduction

Emotions influence physiological state, with important effects on speech production and especially on the phonation process. These effects are reflected in varied and complex voice quality related parameters, such as fundamental frequency (F_0) and jitter. Some parameters are identical in several (or all) languages; others are part of the specificities of a language or speaker.

"Increased emotional arousal is accompanied by greater laryngeal tension and increased subglottal pressure which increases a speaker's vocal intensity". For example, Darwin observed that angry utterances sound harsh and unpleasant because they are meant to strike terror into an enemy [1].

Anger is usually associated with an increase in mean F_0 and energy. Anger also includes "increases in high frequency energy and downward-directed F_0 contours. The rate of articulation usually increases" [2].

The increase of F_0 mean and range is also a characteristic of fear, also with high frequency energy; sadness shows a decrease in mean F_0, F_0 range and mean energy; joy, a positive emotion (one of the few that are usually studied), has an increase in mean F_0, F_0 range, F_0 variability, mean energy, and an increase in high frequency energy [2].

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 142–151, 2010.

"Understanding a vocal emotional message requires the analysis and integration of a variety of acoustic cues" [3].

1.1 Voice Quality and Emotion

Johnstone & Scherer (1999) [4] present studies in which emotional vocal recordings were made using a computer emotion induction task and an imagination technique. Voice quality acoustic parameters included F_0 minimum, F_0 range, jitter and spectral energy distribution. The emotions studied were: tense, neutral, irritated, happy, depressed, bored and anxious.

The authors report that: "values for jitter are correlated with F_0 floor, thus indicating that period to period F_0 variation tends to be larger with higher F_0. This tendency is absent for anxious and tense speech though, which is in agreement with previous findings of a reduction of jitter for speakers under stress". Happy speech presents significantly higher values of jitter than all other emotions. Also as expected, F_0 floor was found to be lowest for the emotions bored and depressed, and highest for happy and anxious speech.

Zovato et al. (2004) [5], used three basic simulated emotional styles (besides neutral, they had anger; happiness and sadness). An Italian female professional speaker recorded 25 sentences; the main point was to investigate the correlation between emotions and acoustic parameters (F_0 minimum, maximum, mean and range, plus RMS energy). They also applied a perceptual test to 10 volunteers to evaluate the corpus. It was shown that there was some confusion in discerning the pairs neutral/sad and happy/angry.

Recently, Toivanen et al. [6] investigated how well voice quality conveys emotional information to be perceived by humans and computers. They used nine professional actors to produce their data, simulating: neutral; sadness; joy; anger and tenderness states, in which they extract a vowel from the entire running speech (approximately one minute). They analyzed vowel [a]. The samples were presented to 50 listeners to recognize the emotion and classified using automatic methods. Human listeners were better than the machine at recognizing anger.

Drioli et al. [7] analyzed F_0, duration, intensity, jitter, shimmer, HNR and other voice quality indexes such as Hammarberg Index. The authors utilized Praat voice report. Regarding irregularities, and for stressed vowels, they report a high shimmer value for anger; higher jitter values for joy and surprise (with anger in third place). The HNR is lower for anger and joy.

Chung, in 2000, investigated acoustical properties of Korean emotional speech. The author measured: F_0 parameters (mean, maximum, minimum, mean of the 20% lowest values, range), jitter, shimmer, speaking rate and spectral distribution. The analysis showed that joy increases F_0 mean, whereas sadness enhances the decrease of F_0 minimum. The increase of F_0 maximum and of F_0 range was found to be "a good indicator of the general emotional arousal". "The jitter and the shimmer values seem to increase under the emotional tension (...). However, these variations (...) were not statistically significant in the case of Korean data" [8].

Voice quality aspects are very often described qualitatively. In quantitative investigations, the most studied parameters relate to F_0. More recently, the list of investigated parameters expanded to include jitter, shimmer, HNR, glottal source parameters, etc.

1.2 Crosslinguistic Information on Emotional Speech

Probably beginning with Darwin [1], it is known that facial expressions are more universal than prosody, even though studies that only contemplate prosody or non-verbal aspects revealed that anger is reasonably well perceived, but the same does not occur with joy. It is also well known that speakers are commonly better on perceiving emotions in their own language.

Many emotion theorists defend that emotions are mostly learned and affected by social environment. As a result, emotions are conjectured to vary considerably across cultures.

Scientific studies have been made crossing speakers and listeners of several origins. However, according to [9], the languages and cultures studied so far are not actually very diverse. Moreover, only a few specific emotions have been studied systematically, usually basic emotions [10, p. 244].

1.3 Related Work for European Portuguese

Not much research was conducted on the subject of emotional speech for European Portuguese. There is no corpus, big or small, of EP emotional speech available. Even the recent work on emotional speech synthesis of Portuguese [11,12] was based on information published for other languages, complemented by extraction of glottal parameters (such as open quotient) from a German database.

The corpus used in the present work was previously partially used in a recent Master thesis work [13]. The parameters investigated were related to F_0 and articulation. Parameters such as jitter were not contemplated.

1.4 Objectives and Paper Structure

Our main objectives were: to obtain values for the parameters commonly contemplated in acoustic analyses of emotional speech; to compare the obtained parameter values with the ones reported for other languages, in order to determine which follow general tendencies and which are characteristic of the EP language.

In the next section, corpus and information on the extraction of voice quality related parameters are presented. Section 3 presents results for the different parameters contemplated in our study, ending with the joint analysis of four of them. The paper ends with discussion, main conclusions and suggestions for future work.

2 Material and Method

The corpus that was recorded and analyzed is composed of two sentences - one simple, the other more complex - both extracted from the Portuguese version of the naturalistic dialogue 'The human voice' by the French writer Jean Cocteau. The simple was "O melhor será tomares conta deles" (You've better take care of them) and the complex "Não tenho com certeza a voz de uma pessoa que esconde qualquer coisa" (I don't really have the voice of a person who hides something). The chosen sentences do not present by themselves any emotional charge, so that the actor may interpret them according to the intended values: joy, despair, anger, fear, sadness and the neutral form.

The informant is a professional actor, male, 42 years old, with a wide experience in national television and theatre (actor of Portuguese National Theater Company). He has also performed several cartoon voices and has been the voice of well-known advertisements.

Recording sessions took place in a soundproof booth at the Municipal Theatre of Guarda (North Interior of Portugal), all in the same day, with a gap of one hour between the two recording sessions. A microphone AKG C 451 B and a DAT recorder Tascam DA-P1 were used, and an experienced sound technician was present. Between each utterance there was no more time than the needed for respiratory pauses. We consider that these circumstances allowed the production to get closer to spontaneous emotion, since the actor did not have much time to concentrate, and thus the register became less acted.

Annotation and Feature Extraction - The utterances were first annotated at word and phone levels, using SAMPA transcription in SFS (Speech Filing System). The limits of each segment were marked and a broad phonetic transcription was made, considering phenomena such as elision, sandhi and epenthesis. All data were processed in Praat software [14] which allowed the extraction of the needed elements using Praat Voice Report function. Statistical analyses were made in SPSS (v. 16) and R. As part of our parameters departs from a Normal distribution, non-parametric tests were employed.

3 Results

The following subsections present the results of our analyses regarding the most commonly studied parameters relevant to voice quality: F_0, jitter, shimmer, and harmonic noise ratio, complemented with autocorrelation.

3.1 F_0 Parameters

Four different F_0 related parameters were investigated. The results for F_0 minimum, F_0 max, F_0 mean and F_0 standard deviations are presented in Fig. 1 as function of emotion.

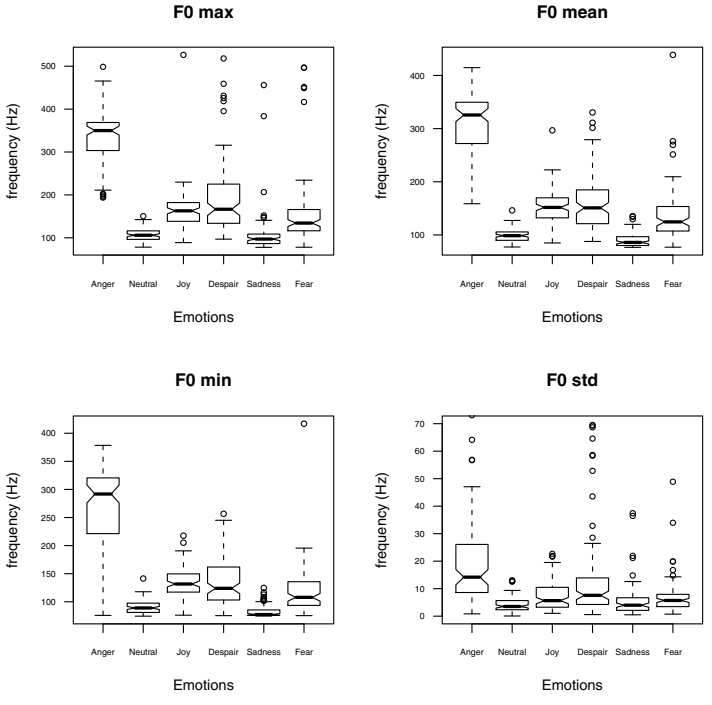

Fig. 1. Effect of emotions on four fundamental frequency (F_0) related parameters. From top left: maximum, minimum, mean and standard deviation.

The analysis of different F_0 parameters shows that anger is clearly differentiated, presenting an average value near 300 Hz and the highest standard deviation and range.

Joy and despair present similar values on the four F_0 parameters, with mean around 150 for F_0 mean and F_0 max. One difference between the two is the higher range of values for despair.

Fear has values of F_0 a little lower than the previous pair. Standard deviation is also lower. Sadness presents the lower values for the parameters, some similar to the neutral.

The Kruskal-Wallis (KW) test (p=0.01 corrected for multiple comparisons) confirms as significant the factor emotion for all four parameters: $[\chi^2(5) = 338.65, p < 0.001]$ for minimum; $[\chi^2(5) = 370.72, p < 0.001]$ for maximum, $[\chi^2(5) = 408.46, p < 0.001]$ for mean; and $[\chi^2(5) = 140.11, p < 0.001]$ for standard deviation.

For F_0 maximum, minimum and mean, post-hoc tests showed as significantly different all pairs except despair-fear, despair-joy, fear-joy, and neutral-sadness. For F_0 standard deviation, also the following pairs were not significantly different: fear-neutral, fear-sadness and joy-sadness. At least some pairs are difficult to differentiate based on F_0 parameters. The standard deviation presents the lowest discrimination power, the other three show similar power.

3.2 Irregularities: Jitter and Shimmer

To compensate for the intonation related variations of F_0, as we used speech from sentences, we only contemplated PPQ5 parameter for jitter. Influence of emotion in PPQ5 is presented in Fig. 2.

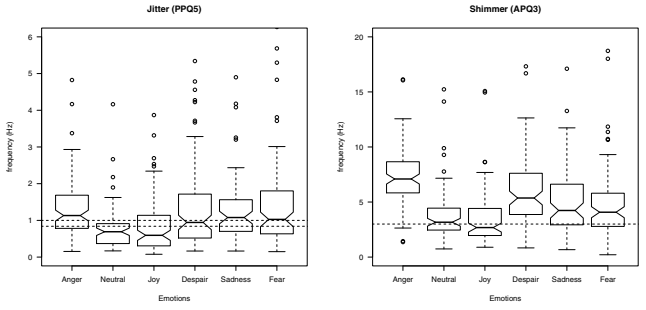

Fig. 2. Effect of emotions on jitter PPQ5 and shimmer APQ3. Horizontal lines represent two thresholds usually associated with normality.

Comparing with F_0 related parameters, differences in jitter are not so evident. Our results showed that higher jitter values are associated with despair, fear, anger and sadness, more negative emotions. The neutral speech and joy present lower or similar values.

Only joy presents smaller PPQ5 than the normality threshold indicated in MDVP . The other emotions don't present values significantly lower or higher than both thresholds.

The KW test confirms as significant the factor emotion $[\chi^2(5) = 53.9012, p < 0.001]$. Post-hoc multiple comparisons test after KW showed as different: joy-anger, joy-fear, joy-sadness and neutral-sadness. Concerning jitter values, joy appears clearly lower than three of the other emotions. Jitter seems a relevant factor to detect joy.

Analysing shimmer parameters, in Fig. 2 at the right, it was clear that they are particularly high for anger, followed by the group integrating despair, sadness and fear.

Looking at the threshold of normality, only neutral and joy are not significantly above. The shimmer values for anger and despair are clearly in the region usually considered as pathologic. The KW test confirms as significant the factor emotion $[\chi^2(5) = 123.99, p < 0.001]$. Multiple comparisons test after KW shows as significant the following differences: anger-fear, anger-joy, anger-neutral, despair-joy and despair-neutral. That is, anger only does not present significantly higher shimmer values than sadness and despair, related emotions; despair also has significantly higher shimmer values than joy and neutral; other emotions present no significant differences. Shimmer only differentiated anger and despair from all the remaining.

3.3 Harmonic Noise Ratio (HNR)

HNR values are presented in Fig. 3. A horizontal line represents a common threshold, 12 dB. While joy presents a value higher than 12 dB, most of the emotions present values around that value and some, like anger and despair, values significantly smaller.

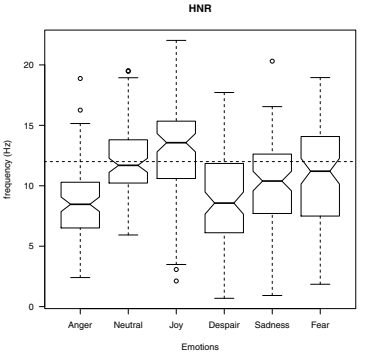

Fig. 3. Effect of emotions on Harmonic Noise Ration (HNR). Horizontal line represents a common normality threshold.

The KW non-parametric ANOVA confirms as significant the factor emotion $[\chi^2(5) = 84.23, p < 0.001]$. Multiple comparisons test after KW shows as significant the following differences: anger-fear, anger-joy, anger-neutral, despair-joy, despair-neutral and joy-sadness. The situation is similar to the one reported for shimmer, with the addition of a new significant difference between joy and sadness. In HNR besides the salient differences of anger and despair (now the lowest values) we also have differences for the positive emotion joy, with significantly higher values of HNR than all others. The values of HNR for anger and despair are in a region potentially classifiable as pathologic.

3.4 Parameters Combination

Inspired by multidimensional representations of voice parameters (ex: Kay Elemetrics MDVP and Hoarsness Diagram) in Fig. 4 we combine F_0 (only mean, as no significant differences were obtained by applying the other three parameters), jitter, shimmer and HNR in a plot. For normalization, values of each parameter were dividided by the mean.

Based on the figure and taking in consideration the results from statistical tests presented before, we have: (1) anger differing from neutral by all four parameters. Jitter, shimmer and F_0 are increased, HNR decreases; (2) despair also differing by the four parameters, but with smaller differences than for anger; (3) sadness is essentially only different by irregularities (Jit & Shim); (4) fear similar to despair, with HNR closest to neutral values and smaller differences; (5) joy, a positive emotion, with only distinct F_0 values.

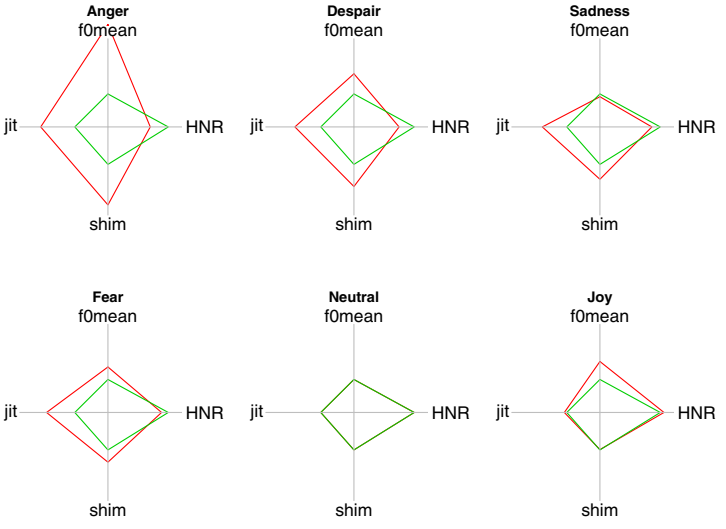

Fig. 4. Comparing the five emotions and neutral speech based on the four types of parameters used. For F_0, mean was chosen as representative of the four F_0 parameters included in the study.

4 Discussion

Parameters analyzed included several F_0 related measures, jitter, shimmer and HNR. Five emotions plus neutral were contemplated. As the languages and cultures studied so far are restricted, this research presents a small contribution to increasing the diversity in quantitative data regarding the complex relation between voice quality and emotions. In general, results were in accordance with the consulted literature. This is particularly true for F_0 related parameters and emotions such as anger and despair. Our results for joy disagree, at least in some parameters, from some previously reported results.

F_0 maximum and average differentiates anger, sadness and joy as reported in [2] and [15]. Anger presents the highest F_0, joy an equally high value, sadness the lowest value. Sadness, as reported in [15], has values close to neutral. Our measures don't confirm the increase of F_0 for fear. As Scherer (cited by [16]) suggested, our F_0 measures correlate with activation dimension; high activation relates with higher values of F_0.

The comparison of jitter and shimmer values with the literature is more difficult. Firstly, there are fewer studies reporting such parameters; secondly, there is some uncertainty on the exact parameter report (ex: local, PPQ5); thirdly, the process of parameter extraction is not necessarily equivalent. Our option was to compare essentially with [7], also using Praat in their analyzes.

Regarding shimmer, our results, showing that only anger does not present significantly higher shimmer values than sadness and despair, are in agreement with published measures such as [7] by Drioli et al. (2003) showing high shimmer values for anger.

Higher jitter values for joy and surprise (with anger in the third place) were reported [7]. [4] reported higher jitter for happy. In our results, joy appears clearly lower in jitter than three of the other emotions. Jitter seems a relevant factor to detect joy, but values are to the lower side, contrary to the reported higher values of the mentioned study. The similarity of jitter for joy and neutral observed for our EP data is in agreement with the results obtained by Monzo et al. (2007) [17] for Spanish happy and neutral.

In agreement with [7], HNR is lower for anger and for despair, also negative. Contrary to same work, we verified significantly higher values of HNR for joy, placing this emotion very far from fear relative to HNR.

The differences observed for joy can be related to the difficulty in identifying the emotion in perceptual tests conducted with the same material [18], in agreement with the observations of Darwin, presenting joy as more difficult to transmit by voice alone. Joy was often confused with neutral, by native and non-native EP speakers. The corpus is too reduced to generalize, but this question of possible differences in joy expression or relative unsuccessful of the actor in expressing this emotion, recommend follow-up studies.

For some of the emotions, parameters such as HNR and jitter present values in the "pathological" ranges usually considered in voice evaluation. This points to the necessity of controlling the emotional state of the subject to whom a voice evaluation procedure is applied. Being sad or happy has, as demonstrated by our results, very noticeable effects on the "normal" range of several parameters.

The main shortcoming of the study is the corpus size and the use of an actor. Extensions of the work regarding the number of subjects, the extent of speech material and more natural emotional speech are needed.

The parameters analyzed are the most common, providing a good starting picture of the effects of emotions on the voice quality, but do not cover all possibilities. The glottal source parameters, such as Open Quotient, and the spectral parameters should be added.

5 Conclusion

For the first time was investigated several voice quality related parameters for five different emotions in European Portuguese: sadness; happiness; fear; anger and despair. Analyses focused on F_0 related parameters; shimmer; jitter and Harmonic Noise Ratio (HNR).

From our analyses, and taking into consideration the reduced size of our corpus and the use of an actor as informant, no clear EP characteristic emerged. Nevertheless, differences on several parameters were observed for joy that, if confirmed, would constitute a cultural difference.

References

1. Darwin, C.: The Expression of Emotions in Man and Animals. Portuguese translation by Relógio D' Água (2000) (1872)
2. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70(3), 614–636 (1996)
3. Schirmer, A., Kotz, S.A.: Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. TRENDS in Cognitive Sciences 10(1) (2006)
4. Johnstone, T., Scherer, K.R.: The effects of emotion on voice quality. In: International Congress on Phonetic Sciences (ICPhS), San Francisco (1999)
5. Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S.: Towards emotional speech synthesis: A rule based approach. In: ISCA SSW (2004)
6. Toivanen, J., Waaramaa, T., Alku, P., Laukkanen, A.M., Seppänen, T., Väyrynen, E., Airas, M.: Emotions in [a]: a perceptual and acoustic study. Logoped Phoniatr Vocol 31(1), 43–48 (2006)
7. Drioli, C., Tisato, G., Cosi, P., Tesser, F.: Emotions and voice quality: Experiments with sinusoidal modeling. In: VOQUAL (2003)
8. Chung, S.J.: Expression and Perception of emotion extracted from the Spontaneous Speech in Korean and in English. PhD thesis, Sorbonne Nouvelle University (2000)
9. Zinken, J., Knoll, M.A., Panksepp, J.: Universality and diversity in the vocalisation of emotion. In: Isdebski, K. (ed.) Emotions of the human voice. San Diego Plural Publishing (in press)
10. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256 (2003)
11. Cabral, J.: Transforming Prosody and Voice Quality to Generate Emotions in Speech. Dissertação de mestrado, IST/UTL (2006)
12. Cabral, J., Oliveira, L.C.: EmoVoice: a system to generate emotions in speech. In: InterSpeech, pp. 1798–1801 (2006)
13. Rodrigues, A.: As Emoções na Fala (Emotions in Speech). Masters dissertation, Universidade de Aveiro, PT (2007)
14. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
15. Cowie, E.D., Cowie, R., Schroeder, M.: The description of naturally occurring emotional speech. In: ICPhS, pp. 2877–2880 (2003)
16. Airas, M., Alku, P.: Emotions in short vowel segments: Effects of the glottal flow as reflected by the normalized amplitude quotient. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 13–24. Springer, Heidelberg (2004)
17. Monzo, C., Alías, F., Ignasi, I., Gonzalvo, X., Planet, S.: Discriminating expressive speech styles by voice quality parameterization. In: XVIth International Conference on Phonetic Sciences (ICPhS), pp. 2081–2084 (2007)
18. Nunes, A.M.B., Roussel, N., Rodrigues, A., Coimbra, R.L., Teixeira, A.: Cross-linguistic effects on the perception of emotions. In: ICPLA (2008)