An Acoustic Study of Telephone Speaking Style in Brazilian Portuguese

Renata Regina Passetti[1], *Plinio Almeida Barbosa*[1]

[1] Speech Prosody Studies Group, Department of Linguistics, State University of Campinas, Brazil

re.passetti@gmail.com, pabarbosa.unicampbr@gmail.com

Abstract

This paper presents an acoustic analysis of the telephone speaking style in a Brazilian Portuguese corpus. The corpus contains 80 recordings from 20 speakers (10 male and 10 female) in two non-simultaneous conditions: face-to-face (non-mediated speaking style) and via mobile phone (telephone speaking style) and were recorded in silent and noisy environments. The acoustic analysis is divided considering global and local parameters, which are tested with respect to their correlation with condition (face-to-face, mobile phone), environment (silent, noisy) and speakers' gender (male, female). For global parameters, spectral emphasis and F_0-peak production rate in noisy environment tend to be higher in comparison with their values in silent environment. For local parameters, the use of mobile phone in noisy environments causes an increase of the F_0-median. The F_0 standard-deviation is higher in this condition when compared to face-to-face condition. Concerning the differences between genders, F_0-median of male speakers tends to be higher in mobile condition when in noisy environments. Even though females' values for F_0 standard-deviation are higher in each recording condition than males', males' values for this parameter differ more across face-to-face vs. mobile phone conditions if compared to women's.

Index Terms: telephone speaking style, forensic phonetics, speaker comparison, Brazilian Portuguese.

1. Introduction

This work aims at studying the telephone speaking style by analyzing a set of prosodic parameters that are modified during the act of speaking on a mobile and their correlation with three factors we believe are related to a telephone speaking style.

The speech production context may influence the acoustic signal, as requires certain acoustic or gesture-visual adaptations from the speaker, with the purpose of enhancing the message intelligibility for the listener. These adaptations can occur as a result of factors such as environment noises [1], composition of audience [2], degradation of signal quality [3], emotional expression [4] and can modify both segmental and suprasegmental parameters of the speech signal.

In Forensic Phonetics, most of speaker comparison tasks deal with inter- and intraspeaker variation due to channel quality differences between the compared recordings. In the majority of the cases, the investigation involves the comparison between a telephone recording and a face-to-face recording [5]. In Brazil, the use of telephone recordings as criminal evidence has increased in the last few years. A survey presented by ISTOÉ Magazine [6] showed that, in 2007, the Brazilian Police intercepted more than 20,000 telephone lines, resulting in recorded speech samples of over 100,000 individuals. Data collection presented in 2016 by Brazilian National Telecommunications Agency [7] has registered 255.23 million of active mobile phone lines in Brazil, which also indicates an increasing growth in the use of this communication medium.

The telephone transmission effect has recurrently been theme of studies [3, 8, 9], but less attention was paid to the influence the act of speaking on a mobile phone would have in intraspeaker variation. Some studies mention some behavioral modifications that occur during the use of a mobile phone. Byrne & Foulkes [8] called this as "speaker effects" and explain that, for some speakers, they can result in the adoption of a "telephone voice". By analyzing the telephone effect on the variability of fundamental frequency (F_0), Hirson et al. [5] used speech samples recorded at the end of the telephone line and compared them to face-to-face recordings. The results showed an increase of 5 Hz in speakers' F_0 in the telephone situation. De Jong et al. [10] found similar results in a study about the telephone effect on fundamental frequency. By comparing the F_0-mean for spontaneous speech samples in three recording conditions (face-to-face, telephone studio quality and telephone-intercepted quality), they found out that telephone communication significantly leads to an increase in speakers' F_0 when compared to the values in face-to-face condition.

Even though these studies have pointed out some modifications caused by telephone context, no studies have analyzed acoustical and perceptual modifications on the speech signal caused by a telephone speaking style. For this reason, we believe the present study will shed some light on questions regarding this theme and will contribute to the improvement of forensic tasks involving the comparison between telephone and face-to-face speech samples.

2. Method

To obtain speech material for the telephone speaking style and without the mobile phone filtering distortions, which could impair the correct extraction of acoustic parameters, the corpus was obtained by directly recording both mobile phone and face-to-face interactions, as summarized below.

2.1. Speakers

Twenty subjects were recorded (10 male and 10 female) at a sample rate of 44.1 kHz/16 bits. The speakers were students of the State University of Campinas and with ages from 19 to 28 years (mean of 23.7 yrs). They were all speaking a variety of São Paulo countryside dialect.

2.2. Recordings

The recordings consist of informal interviews with each speaker individually, carried out by the first author, and were conducted in two non-simultaneous conditions: face-to-face (non-mediated speaking style) and via mobile phone (telephone speaking style). They were also obtained in silent and noisy environments.

The recordings in silent environments were made in a sound treated studio with a noise average intensity of 42 dB_{SPL}. The recordings in noisy environments were made outdoor (in the entry of Student's Restaurant of UNICAMP) with a noise average intensity of 87 dB_{SPL}.

All the recordings were made using a Yoga HM20 headset microphone connected to a Zoom H4N digital recorder. The microphone was placed 50 mm from the speaker's mouth. The mobile calls were made on an Apple IPhone 5s mobile phone and the receiver was a Nokia 110 mobile phone. Both mobile phones were operating on the Vivo network.

Each interview lasted 5 minutes, totalizing 200 minutes of speech material (20 speakers × 2 recording conditions × 5 minutes of speech). The recording environments are illustrated in Figure 1.

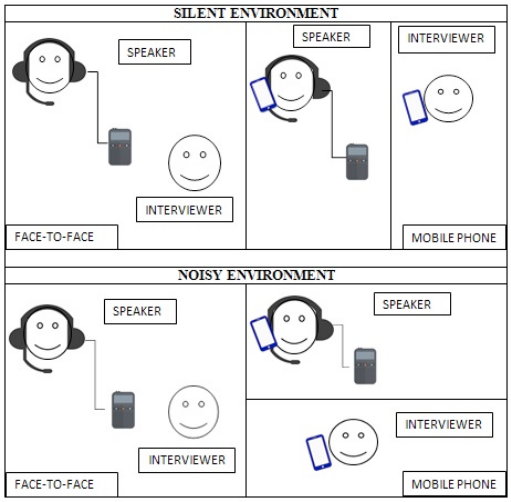

Figure 1: *Schematic view of recording conditions face-to-face (left) and mobile phone (right) in silent (top) and noisy (bottom) environments.*

2.3. Acoustic analysis

The recordings were annotated in Praat [11] with the creation of a 2-tier TextGrid. The first tier contains the orthographic transcription of the recordings. The second tier delimits phonetic syllables from the onset of a vowel to the offset of the following vowel [12, 13]. The segmentation of vowel-to-vowel units (VV units) was performed by a Praat script, BeatExtractor [12] and each VV segment was manually labeled with an ASCII-based notation.

The acoustic analysis includes the computation of global and local parameters. The global parameters are computed considering the whole audio file, and the local parameters are computed considering the VV segment domain.

A Praat script, ProsodyDescriptorNew [14], was implemented to automatically extract global and local parameters. The statistical descriptors for each group of parameters are outlined below.

2.3.1. Global parameters

Speech rate is measured by VV units per second, including pauses.

Articulation rate is measured by VV units per second excluding pauses.

Normalized duration of salient peaks is computed by the statistical descriptors mean, standard-deviation, and skewness. The statistical descriptors of salient peaks duration were normalized using the z-score technique according to equations 1a and 1b.

$$z = \frac{dur - \sum \mu_i}{\sqrt{\sum \mathrm{var}_i}} \quad (1a)$$

Where dur is the VV unit duration and μ_i and var_i are, respectively, the reference mean and variance of each phone within the corresponding VV unit.

$$z^i_{smoothed} = \frac{5 \cdot z^i + 3 \cdot z^{i-1} + 3 \cdot z^{i+1} + 1 \cdot z^{i-2} + 1.z^{i+2}}{13} \quad (1b)$$

where *i* represents each VV unit.

Spectral emphasis is calculated according to Eriksson et al. [15] by expression 2, where $L_{spectrum}$ is the intensity of the whole spectrum (up to Nyquist frequency) and L_0 is the intensity in the band 0 – 400 Hz.

$$Spectral\ emphasis\ (dB) = L_{spectrum} - L_0 \quad (2)$$

F_0-peaks rate is computed after using a 2 Hz lower-band filter and is calculated by F_0-peaks per second.

Global intensity is measured by the ratio of global intensity median to maximum intensity of the whole spectrum.

Baseline is computed in semitones according to Lindh and Eriksson [4] by expression 3.

$$Baseline\ (st) = F_0 median - 1.43 \cdot F_0 SD \quad (3)$$

2.3.2. Local parameters

Duration of each VV unit is calculated in milliseconds.

Fundamental frequency is computed by the statistical descriptors median, standard-deviation and skewness in Hertz and semitones. The use of semitones is justified by trying to approximate the F_0 variation from the auditory perception of intonation and the reference value adopted for the conversion between frequencies in Hertz and semitones is 1 Hz.

3. Results

The software R (version 3.3.2) [16] was used to perform the statistical analysis.

3.1. Multivariate analysis of variance (MANOVA) for global parameters

To investigate which global parameters signal telephone speaking style as a function of condition, environment and gender, the statistical test MANOVA was selected. Pillai's statistical method (V) was adopted [17, pp. 842].

The results of a 3-Way MANOVA for factors "condition", "environment" and "gender" revealed that speakers modify this set of parameters only as a function of modifications in the levels of the "environment" factor (silent, noisy) (V (10) = 0.5262; p = **$4.2 \cdot 10^{-8}$**) and its interaction with "gender" factor (V (10) = 2.551; p = **0.01**), irrespective of speaking style.

Table 1 shows the global parameters that are significant for the distinction of the levels.

Table 1: *Results of 3-Way MANOVA for comparison between variance values of global parameters and significant factors* (ns = not significant; asterisk means interaction between factors)

Global parameters		**3-Way MANOVA**	
		Environment	Environment * Gender
Speech rate		ns	ns
Articulation rate		ns	ns
z-scores of salient peaks duration	Mean	ns	ns
	Standard-deviation	ns	ns
	Skewness	ns	ns
Spectral emphasis		F (1) = 34.697; p = **$3.23 \cdot 10^{-8}$**	F (1) = 3.775; p = 0.05
F_0-peaks rate		F (1) = 4.577; p = **0.03**	ns
Baseline (st)		ns	F (1) = 3.139; p = 0.08
Global intensity		F (1) = 4. 057; p = 0.05	ns

The results for interaction between environment and gender are marginally significant for two parameters. On the other hand, spectral emphasis (mean 6.63 dB for silent and 9.49 dB for noise) and F_0-peaks rate (mean 1.07 peaks/s for silent and 1.13 peaks/s for noise) are distinct across environment levels. According to Figure 2, speakers' spectral emphasis increases and is more variable in noisy environments (an increase of 2.86 dB compared to silent environments). The F_0-peaks production rate is slightly higher in noisy environments – an increase of 0.06 peak/s compared to silent environments. As a correlate of intonation, this modification of F_0 can signal a rhythmic variation in noisy environments as a way of speakers to compensate the environmental noise and make themselves understood.

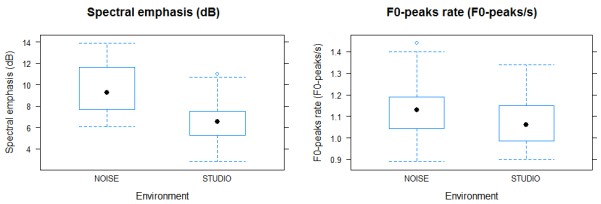

Figure 2: *Spectral emphasis (dB) (left) and F_0-peaks production rate (F0-peaks/s) (right) as a function of environment.*

3.2. Scheirer-Ray-Hare (SRH) test for local parameters

To investigate if there is an interaction between local parameters and the factors "condition", "environment" and "gender" which can characterize the telephone speaking style the Scheirer-Ray-Hare (SRH) was used – a non-parametric alternative to 2-Way ANOVA due to not obeying the conditions of non-normality and heteroscedasticity of the residuals. Table 2 shows the local parameters that are significant for the distinction of levels.

Table 2: *Results of SRH test for comparison between variance values of local parameters and significant factors* (ns = not significant; asterisk means interaction between factors)

Local parameters		**SRH test**	
		Cond.*Environ.	Cond. * Gender
Duration (ms)		ns	ns
F_0	Median (st)	H (1) = 44; **$p < 2.2 \cdot 10^{-16}$**	H (1) = 20.8; **$p < 1.0 \cdot 10^{-5}$**
	Standard-deviation (st)	ns	H (1) = 31; **$p < 2.2 \cdot 10^{-16}$**
	Skewness (st)	ns	ns
	Median (Hz)	H (1) = 44; **$p < 2.2 \cdot 10^{-16}$**	H (1) = 21.3; **$p < 2.2 \cdot 10^{-16}$**
	Standard-deviation (Hz)	H (1) = 12; **$p = 4.7 \cdot 10^{-4}$**	H (1) = 44; **$p < 2.2 \cdot 10^{-16}$**
	Skewness (Hz)	ns	ns

Fundamental frequency median in semitones and Hertz and its standard-deviation in Hertz signal as distinct the interaction between factors "condition" and "environment". As can be seen in Figure 3, the F_0 median is higher in noisy environments.

Comparing the average of this parameter in telephone condition in silent (mean 179 Hz) and noisy environments (mean 191 Hz), there is an increase of 12 Hz.

Considering the difference in semitones, although there is no difference between the levels of "condition" factor in noisy environments (mean 90 st both in face-to-face and mobile phone), F0 median values in mobile phone condition are slightly higher (mean 89 st) when compared to face-to-face condition (mean 88 st) in silent environments.

For F_0 standard-deviation in Hertz, data presented in Figure 3 show slightly higher values for mobile phone condition. In noisy environment, the mean of F_0 standard-deviation is 4.5 Hz for mobile phone (against 3.9 Hz for this condition in silent environment) and 4.1 Hz for face-to-face interactions (against 3.4 Hz for this condition in silent environment).

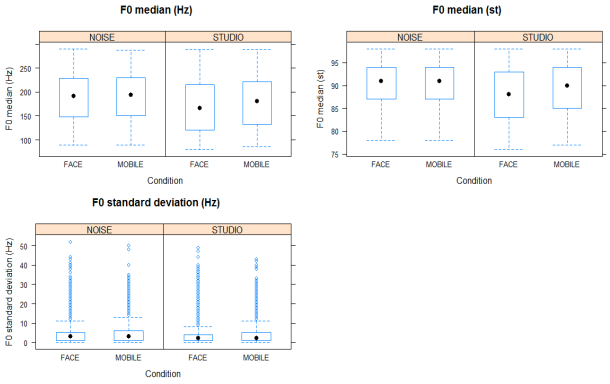

Figure 3: *F_0 median in Hertz (top left), semitones (top right) and F_0 standard deviation in Hertz (bottom left) as a function of condition and environment.*

As for factors "condition" and "gender", the following parameters signal interaction between them: F_0 median in Hertz and semitones, and F_0 standard-deviation in Hertz and semitones. By analyzing Figure 4, we can see that F_0 median in Hertz and semitones is higher for both males (mean 147 Hz/86 st in mobile phone against 139 Hz/85 st in face-to-face) and females (mean 224 Hz/94 st in mobile phone against 220 Hz/93 st in face-to-face) in mobile phone condition. The difference between the levels of "condition" is higher for males' F_0 median in Hertz – a difference of 8 Hz.

Considering the F_0 standard-deviation in Hertz and semitones, data in Figure 5 show higher values for females (mean 5 Hz/0.4 st in mobile phone against 4.7 Hz/0.38 st in face-to-face) comparing to males (mean 3.4 Hz/0.38 st in mobile phone against 2.8 Hz/0.33 st in face-to-face). However, among male speakers, there is a greater variation for this parameter in mobile condition.

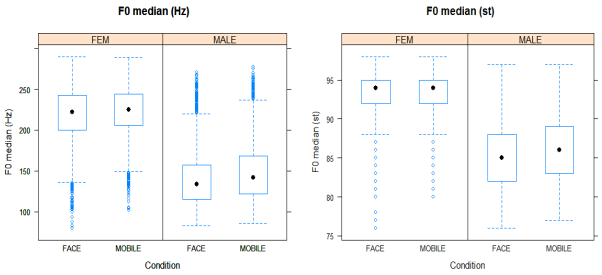

Figure 4: *F_0 median in Hertz (left) and semitones (right) as a function of condition and gender.*

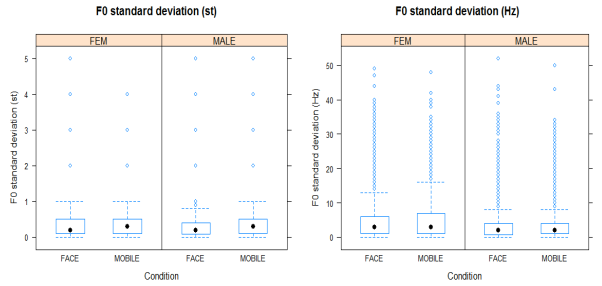

Figure 5: *F_0 standard deviation in semitones (left) and Hertz (bottom right) as a function of condition and gender.*

4. Discussion and conclusions

First, considering the influence of global parameters on signaling a telephone speech style, the results show that none of the parameters seems to be modified over the telephone speaking style. On the other hand, speaking style is affected by changes in environment. The results show that the speech produced in noisy environments has both higher effort levels (higher spectral emphasis) and a F_0-peak production rate, which suggest changes in vocal effort and intonation as a way for speakers to acoustically compensate the environment noises and to be understood by their listeners. These results have forensic purposes, since most of forensic tasks deal with a disputed recording containing speech samples produced in noisy environments or with a background noise. Therefore, identify parameters which are modified by environmental noise and quantify these modifications can lead to robust methods in forensic analysis.

As regards the effect that local parameters have on signaling a mobile phone speaking style, the statistical analysis shows that fundamental frequency is the parameter that significantly changes while speaking on mobile phones. These changes are influenced by environment and differ across genders. Speakers' fundamental frequency median increases and is more variable in noisy environments, which means that when speaking on mobile phone in this environment, their voices have higher pitch levels. Concerning to differences between gender in telephone speaking style, the results show that speech produced by male speakers in a mobile phone interaction has a higher fundamental frequency if compared to the speech produced by this gender in face-to-face interactions. On the other hand, in mobile phone interactions, women's F_0 standard-deviation values are higher than men's. Applied to a forensic context, in which speaker comparison tasks are carried out most of the time by comparing a questioned recording obtained by telephone interceptions with a reference recording obtained by directly recording a suspect, these results shed some light on what should be considered in a comparative analysis. Regardless the technical effects in telephone recordings [8], a phonetic analysis of a mobile phone speech sample should also take in account the changes in speaking style due to the telephone context. This concern is especially important because, as our findings suggest, the telephone speaking style modifications are more pronounced for male speakers and the majority of the prison population in Brazil (and in most of the countries) is formed by men. In Brazil, according to this statistics, men represent more than 94% of the prison population [18].

5. Future work

The present work is part of the first author's PhD research entitled "An acoustic-perceptual study of the telephone speaking style with implications for speaker verification in Brazilian Portuguese" and is intended to describe the telephone speaking style both acoustically and perceptually. For the later purpose, we are currently applying a perceptual task that will investigate the listeners' discrimination capacity between the telephone speaking style (mobile phone) and the non-mediated speaking style (face-to-face). The analysis of the perceptual dimension will correlate the listeners' choices in the perceptual test to the analyzed acoustical parameters and will investigate the possible "acoustic cues" that listeners could be using to support their choices in distinguishing between these two speaking styles.

6. Acknowledgements

The authors acknowledge FAPESP grant #2015/12174-9. The opinions, hypothesis and conclusions expressed in this article are the authors' own and do not necessarily reflect the view of FAPESP.

7. References

[1] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow and M. A. Stokes. "Effects of noise on speech production: acoustic and perceptual analyses". *Journal of the Acoustical Society of America*, vol. 84, pp. 917–28, 1988.

[2] R. M. Uchansky. “Clear Speech”. In: D. B. Pisoni and R. E. Remez (Orgs.) *The Handbook of Speech Perception*. Oxford: Blackwell Publishing, pp. 207 – 235, 2005.
[3] H. Künzel. “Beware of the ‘telephone effect’: the influence of the telephone transmission on the measurement of formant frequencies”. *Forensic Linguistics*, vol. 1, no. 8, pp. 80-99, 2001.
[4] J. Lindh and A. Eriksson. “Robustness of Long Time Measures of Fundamental Frequency”. *Proceedings of the Interspeech 2007*. Anwerp, Belgium, 27-31 August, pp. 2025-2028, 2007.
[5] A. Hirson, J. P. French and D. Howard. “Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics”. J. Windsor Lewis (Org.), *Studies in General and English Phonetics in Honor of Professor J. D. O’Connor*. London: Routledge, 1995
[6] M. Simas Filho and H. Marques. “Por dentro do grampo”. *Revista ISTOÉ*. São Paulo, ed. 1964, jun. 2007. Online: http://istoe.com.br/8109_POR+DENTRO+DO+GRAMPO/. Accessed in 23 Aug. 2016.
[7] Brasil. Agencia Nacional de Telecomunicações. “Brasil registra 255,23 milhões de acessos em maio”. *Telefonia móvel*. Online: http://www.anatel.gov.br/institucional/index.php/component/content/article?id=1231. Accessed in 23 Aug. 2016.
[8] C. Byrne and P. Foulkes. “The ‘Mobile Phone Effect’ on Vowel Formants”. *The International Journal of Speech, Language and the Law* – Equinox Publishing, vol. 17, no.1, pp. 83-102, 2004.
[9] R.R Passetti. “O efeito do telefone celular no sinal da fala: uma análise fonético-acústica com implicações para a verificação de locutor em português brasileiro”. *MSc Dissertation in Linguistics*. State University of Campinas, 2015.
[10] G. de Jong, T. Hudson, F. Nolan and K. McDougall. “The telephone effect on F0”. Paper presented at The IAFPA 2011 conference, Vienna, Austria, 2011.
[11] P. Boersma and D. Weenink. “Praat: doing phonetics by computer” (Version 5.1.37) [*Computer program*]. Online: http://www.praat.org.
[12] P. A. Barbosa. *Incursões em torno do ritmo da fala*. Pontes, 2006.
[13] P.A. Barbosa, P. Arantes, A. R. Meireles and J. M. Vieira. “Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors”. *Interspeech 2005*, pp. 1441-1444, 2005.
[14] P. A. Barbosa. *ProsodyDescriptorNew.* Praat Script, 2016.
[15] A. Eriksson, G. C Thunberg and H. Traunmüller. “Syllable prominence: A matter of vocal effort, phonetic distinctness and topdown processing”. P. Dalsgaard, B. Lindberg, H. Benner, and T. Zheng-Hua [Eds.], *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech ’01),* pp. 399-402. Aalborg, Denmark, 2001.
[16] The R foundation for statistical computing, “The R project for statistical computing”, Online: http://www.r-project.org/, 2008.
[17] A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. Sage Publications (CA), 2012.
[18] Brasil. Ministério da Justiça, 2014. Departamento Penitenciário Nacional. *Levantamento Nacional de Informações Penitenciárias. INFOPEN – Dezembro 2014*.