Characterizing Variation in Fundamental Frequency Contours of Professional Speaking Styles

Luciana Castro, Ben Serridge, João Moraes, Myrian Freitas

Universidade Federal do Rio de Janeiro, Brazil

lucianacastro@ufrj.br, serridge@alum.mit.edu,
jamoraes2@br.inter.net, myfreitas@uol.com.br

Abstract

Recent perception experiments confirm that listeners are clearly able to distinguish between professional speaking styles, even when semantic content is removed by low-pass filtering the original speech signal. The objective of this study is to evaluate whether acoustic metrics that characterize the overall fundamental frequency contour also vary significantly according to speaking style. The analysis is based on a corpus of Brazilian Portuguese speech representing TV news broadcasters, politicians, religious leaders, and interview subjects on a TV talk show. Of the metrics proposed in the literature and evaluated here, statistical analysis shows that only the mean fundamental frequency and the percentage of dynamic tones exhibit statistically significant differences across speaking styles.

Index Terms: fundamental frequency, prosody, speaking style, professional voice

1. Introduction

It has been noted that speakers are able to adjust their speaking style to the communicative situation, and furthermore that in the case of certain professions, the speaking style can eventually become associated with that profession to such an extent that it acquires a distinct functional role [1]. It has also been shown that listeners can distinguish between professional speaking styles even when semantic information has been removed by low-pass filtering the speech signal [2].

Although much of the literature that attempts to characterize the melodic variations among speaking styles, professional or otherwise, is subjective in nature and difficult to validate or reproduce, some authors have attempted to isolate specific metrics that can be used to characterize speaking styles. The standard deviation of the fundamental frequency, for example, has been suggested as an indicator of dynamic speech: the higher the standard deviation, the more lively the voice, and the more pleasant and interesting to listen to [3]. Other metrics, including interquantile range and mean melodic movement, as well as metrics related to the classification of tones as static, rising, or falling, have been defined in an attempt to characterize speaking styles [4], while other characteristics of the fundamental frequency contour, such as the presence of pitch reset [5], have been identified and framed in terms of quantitative metrics in the context of other research [6].

In this study, several metrics defined in the literature are calculated for four different speaking styles and then analyzed statistically in order to evaluate to what extent the metrics reflect differences between the speaking styles in a statistically meaningful way.

2. Methodology

The data used in these experiments consist of the speech of twenty Brazilian professionals: five TV news anchors, five Catholic priests, five politicians, and five interview subjects. The speakers, all native speakers of Brazilian Portuguese with no apparent communication disorders, are all male between the ages of 35 and 78. All recordings were captured directly from regularly televised programs on Brazilian television, and each recording consisted of one minute of speech sampled from normal use situations: the TV news anchors and the interview subjects in the television studio, the politicians on the senate floor during debate, and the Catholic priests in the church, from which mass is broadcast live on television.

2.1. Corpus

The speech data was collected by a portable computer coupled to an external sound card, in turn linked to the digital decoder supplied by the cable TV company.

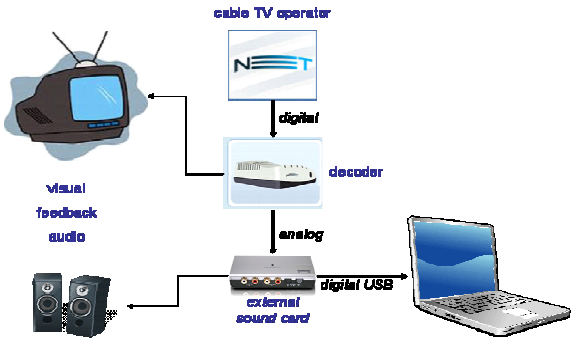

Figure 1: *Diagram of the setup used to capture speech from television.*

The audio signal was sampled at 22 kHz / 16 bits and stored in PCM (.wav) format for later processing by the acoustic analysis software Praat [7]. Transcription of the speech was performed in stages, and by using Praat it was possible to partially automate the transcription process [8]. A total of 20 minutes of speech were analyzed, containing 2780 words and 5861 syllables. All alignments were verified manually by listening to the original speech and inspecting the spectrogram, the final result being a multi-tiered representation of the transcribed speech: the first tier containing only the phonetic sequences and pauses; a second tier containing the time-aligned orthographic transcription; and, finally, a tier containing the time-aligned phonetic

transcription. In the context of the metrics described here, the phonetic transcription is used only to assist in the extraction of pitch information from segments labeled as vowels, and to identify which of these correspond to stressed vowels.

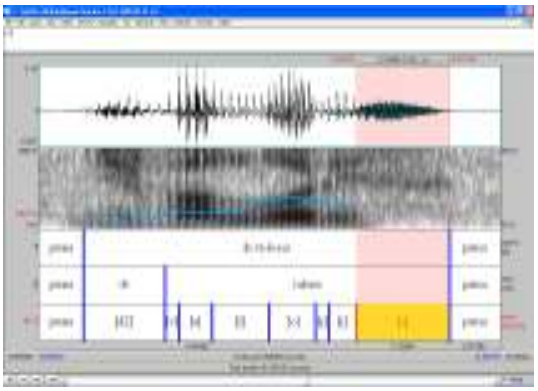

Figure 2: *Transcription of the corpus in three tiers.*

2.2. Metrics

The underlying assumption of the present work is that this ability to distinguish speaking styles derives from perceived differences in the fundamental frequency contour of the different speaking styles. If so, it should be possible to define metrics that quantify these differences, and furthermore, show that the measured differences are statistically significant; that is, that even when taking into consideration the variation between individual speakers, there is still sufficient variation between classes of speakers (speaking styles) to explain the ability of listeners to distinguish among the classes.

As a starting point for the metrics described below, Praat was used to calculate the fundamental frequency for individual frames of speech, spaced at intervals of 10 msec [9]. All segments not labeled as a vowel in the phonetic transcription, even those that might have yielded valid F_0 measurements, were removed from the analysis, as were vowels for which more than half of the frames did not yield a value for F_0. The remaining frames were then used to calculate the metrics described in this study, including standard statistics such as the average and standard deviation of the fundamental frequency for each speaker.

Several other metrics have also been defined in the literature for use in evaluating differences between speaking styles. In [4], Goldman et al. note that although the fundamental frequency range is typically measured as the difference between the highest and lowest fundamental frequency values in a given sample, this statistic is not robust when measured this way, as it tends to fluctuate significantly even within a single utterance. A better metric is the interquantile range, which is the range measured after excluding the highest and lowest 5% of F_0 values. The authors also propose a metric called mean melodic movement, expressed in semitones per second, which measures the sum of the absolute value of changes in fundamental frequency from one frame to the next. This parameter can be measured considering only the values within vowels (intra-vocalic), considering only the differences between vowels (inter-vocalic), or considering all frames for which an F_0 value exists. Table 1 summarizes the set of metrics described thus far.

Table 1: *Metrics that characterize the overall fundamental frequency contour.*

Measure	Definition	Description
F0avg	mean of the fundamental frequency	The mean value of F_0 for a given speech sample, measured either in Hz or semitones.
F0sdev	standard deviation of fundamental frequency	The standard deviation of the values of F_0
Range	vocal range	The difference between the maximum and minimum value for F_0
IR	interquantile range	Measured by the same procedure used to calculate the vocal range, but disregarding the lowest and highest 5% of F_0 values [4]
MM	mean melodic movement	The sum of the absolute value of all the displacements of the fundamental frequency, upward and downward, during a given period, divided by the time in seconds; expressed in semitones per second [4]

One shortcoming of metrics like interquantile range, standard deviation, and mean melodic movement is that they do not discriminate in any way between changes in pitch that are perceptually relevant and those that are not. Interquantile range, for example, does not tell us anything about how fast the speaker changed from the lowest to the highest F_0 value; likewise, standard deviation tells us only that the values vary more or less widely around the mean, but nothing about how that variability is spaced over time; while mean melodic movement is based on the sum of individual pitch movements, each of which may or may not be perceptually relevant. Whatever the perceptually relevant prosodic cues are that listeners use when they distinguish one speaking style from another, if those cues are relatively sparse in the speech signal, the impact of those cues on the overall metric will likely be diluted by the myriad other perceptually irrelevant fluctuations in F_0.

Table 2: *Perceptually motivated metrics.*

Measure	Definition	Description
PR	pitch reset	The difference between the average F0 of the last stressed vowel in a phonetic sequence and the first stressed vowel of the subsequent phonetic sequence
%PR	percentage of pauses with PR > 2 ST	Percentage of pauses for which there is a pitch reset of more than two semitones
PRavg	average pitch reset	The average value of PR across all pauses
%DT	percentage of dynamic tones	The percentage of vowels in which the F_0 range is greater than two semitones
%RT	percentage of rising tones	The percentage of vowels in which the F_0 range is greater than two semitones and in which the maximum F_0 value occurs after the minimum F_0 value
%FT	percentage of falling tones	The percentage of vowels in which the F_0 range is greater than two semitones and in which the maximum F_0 value occurs before the minimum F_0 value

An attempt was made, therefore, to define metrics that are perceptually motivated, while at the same time objective, reproducible, and subject to automatic calculation by scripts (as opposed to manual annotation). Table 2 lists these metrics, which are described in more detail below.

Pitch reset refers to a phenomenon reported in the literature whereby the fundamental frequency tends to decline over the course of a phonetic sequence, only to be "reset" to the original level at the start of a new breath group. In this study, pitch reset is calculated (for each pause) as the difference between the average F_0 of the first stressed vowel after the pause and the average F_0 of the last stressed vowel before the pause [10]. Based on this definition, two metrics were calculated: the percentage of pauses that exhibit a pitch reset of more than two semitones, and the pitch reset averaged over all pauses. Note that the latter can have a negative value, if the last stressed vowel before a pause has an average F_0 higher than that of the first stressed vowel after the pause. Note also that, in order to avoid introducing subjective (not reproducible) criteria into the definition of the metric, reset was calculated for all pauses, not only for pauses considered to be boundaries between breath groups.

In an attempt to measure the dynamism of a speaking style in a perceptually motivated way, the tones in the corpus corresponding to stressed vowels were classified as either static, rising, or falling [4]. A static tone is defined as one in which the difference between the maximum and minimum pitch values is less than two semitones. Non-static tones are considered dynamic tones and are further classified as either rising or falling, according to whether the maximum value appears after or before the minimum value, respectively.

Finally, each of the metrics described above and listed in Tables 1 and 2 was calculated for each of the 20 one-minute recordings in the corpus. One way analysis of variance, in which the speaking styles correspond to "treatments", was then performed on each metric to estimate to what extent the null hypothesis (that all speaking styles have the same mean) holds. In this study, results are considered statistically significant when $p < 0.05$.

3. Results

The results of the ANOVA analysis on the data from this corpus show that average fundamental frequency is a statistically significant indicator of speaking style ($p = 0.015$), with the political and religious speaking styles making the most use of higher frequency ranges, while TV news anchors and interview subjects tend to use lower registers. It has been observed that the use of high registers creates the perception of emphasis and the sensation of a more marked oral production [11], which may explain the more frequent use of relatively high pitch values by the political and religious speaking styles. Table 3 shows the absolute values of average F_0 for the different speaking styles, along with other metrics aimed at capturing the dynamism of the speaking style.

Table 3: *Observed values and ANOVA p-value for the global fundamental frequency metrics defined in Table 1.*

Speaking Style	**F0avg (Hz)**	**IR (ST)**	**F0sd (ST)**	**inter-vocalic MM (ST / sec)**
ANOVA p-value	**0.015**	0.088	0.079	0.113
Interview	155	12.2	4.0	21.5
TV news	139	14.1	4.3	25.6
Political	213	13.8	4.3	21.6
Religious	186	15.3	5.0	19.7

The interquantile range does not distinguish, in any statistically significant way, among the four speaking styles. Likewise, the standard deviation does not vary in a statistically significant way across speaking styles, although both metrics rank the speaking styles in the same order, with the religious speaking style showing the highest interquantile range and standard deviation, and the speech of interview subjects showing the lowest, with political speech and TV news in the middle.

Mean melodic movement was calculated in three ways – intra-vocalic, inter-vocalic, and considering all frames for which an F_0 value exists – but none of the three variants demonstrate statistically significant variation across speaking styles. Of the three variants, the one that exhibited the greatest difference among speaking styles is the intervocalic movement, perhaps because it involves fewer measurements (one per phone boundary rather than one per frame), and as such is less susceptible to noise from perceptually irrelevant movement within vowels.

Table 4: *Observed values and ANOVA p-value for the perceptually motivated metrics, for each speaking style.*

Speaking Style	% F0 reset	Avg F0 reset (ST)	% dynamic tones	% rising tones	% falling tones
ANOVA p-value	0.082	0.065	0.016	0.099	0.669
Interview	42%	3.1	46%	25%	21%
TV news	55%	2.6	54%	26%	28%
Political	55%	4.1	67%	40%	27%
Religious	36%	1.2	62%	34%	28%

As in the case of other metrics, statistical analysis shows that neither of the two measures of pitch reset is statistically significant, although both come close ($p = 0.082$ and 0.065, respectively). However, it is interesting to note that the rank order of the speaking styles is distinct from that of interquantile range and standard deviation. In fact, the metrics related to pitch reset place the political and religious speaking styles at opposite ends of the spectrum, with the political speaking style having an average pitch reset of 4.1 semitones,

and the religious speaking style having an average pitch reset of only 1.2 semitones. Given the methodological difficulties in calculating pitch reset in a consistent, objective, and reproducible way (e.g. the choice of boundaries for which pitch reset should be calculated, the choice of what F0 value to use to represent the start and end of a declination, etc.), the fact that the two metrics related to pitch reset both show near significant variation across speaking styles suggests that the way in which speakers reset pitch levels between phonetic sequences may very well be perceptually relevant to the characterization of speaking styles.

Finally, statistical analysis of the percentage of dynamic tones across speaking styles shows that the differences in values for this metric are statistically significant (ANOVA $p = 0.016$). A further look at the types of dynamic tones shows that the three speaking styles employ falling tones in roughly the same proportion, the difference in the percentage of dynamic tones being influenced primarily by the relative frequency of rising tones, used much more in the political and religious speaking styles (40% and 36%, respectively) than in the TV news and interview styles (26% and 25%, respectively). These results are consistent with those of [4], who observe a higher incidence of dynamic tones in radio speech than in read speech.

4. Conclusions

In this study, several acoustic metrics proposed in the literature were measured over a corpus of 20 one-minute samples representing four professional speaking styles. Two metrics exhibited significant differences across speaking styles: the average of the fundamental frequency and the percentage of tones with F_0 movement greater than two semitones. The presence and degree of pitch reset, while not statistically significant as per the ANOVA analysis, does appear to be correlated to speaking style, such that refinements in the way this metric is measured may qualify it as an indicator of speaking style. Other proposed metrics for measuring how dynamic a given speaking style is – the interquantile range, the standard deviation, and the mean melodic movement – proved not to be statistically significant indicators of speaking style.

Previous research has shown that temporal aspects of speaking styles, specifically the length of pauses and phonetic sequences, are highly correlated with speaking style [12]. Participants in the perception study described in [2] may have leveraged temporal cues more than variations in fundamental frequency to discriminate between speaking styles. Furthermore, temporal aspects of the changes in fundamental frequency are likely to be critical to their association with speaking style, though most of the metrics evaluated here do not take the time domain into consideration explicitly. Certainly, further research is required to quantify the relative importance of both temporal and fundamental frequency components of the cues that listeners use to distinguish speaking styles. One approach would be to evaluate the relative impact of different prosodic cues by modifying the speech signal used in the perception experiment, as in [13]. Another would be to perform a regression analysis to determine the relative importance of each metric in characterizing speaking styles. In either case, the methodology described in this study, whereby metrics are defined and then evaluated statistically against a corpus of speech collected in real-use situations, should provide a reference for future research aimed at quantifying the prosodic differences between speaking styles.

5. References

[1] Léon, P. *Précis de phonostylistique*. Paris, Nathan, 1993.

[2] Castro, L. O comportamento dos parâmetros duração e freqüência fundamental nos fonoestilos político, sermonário e telejornalístico. *PhD Thesis*, Federal University of Rio de Janeiro, 2008.

[3] Hincks, R. Standard deviation of F0 in student monologue. *Proceedings of FONETIK 2004*, Stockholm, p. 132-135, 2004.

[4] Goldman, J.-Ph., Auchlin, A., Avanzi, M., Simon, A.C. ProsoReport: an automatic tool for prosodic description. Application to a radio style. In: *Proceedings of Speech Prosody*, Campinas, p. 701-704, 2008.

[5] Ladd, D. R. Declination "reset" and the hierarchical organization of utterances. *J. Acoust. Soc. Am*, v. 84, n.2, p. 530-544, ago. 1988.

[6] Werner, S.; Keller, E. Prosodic aspects of speech. In: Keller, E. (org.). *Fundamentals of speech synthesis and speech recognition: basic concepts. State of the art and future challenges.* Chichester: John Wiley, 1994. p. 23-40.

[7] Boersma, P.; Weenink, D. *Praat: doing phonetics by computer*, 2008. Computer program available at http://www.praat.org/.

[8] Serridge, B.; Castro, L. Faster time-aligned phonetic transcriptions through partial automation. *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics*, Atenas, p. 189-192, ago. 2008.

[9] Boersma, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings* 17: 97-110.

[10] Heldner, M.; Edlund, J.; Björkenstam, T. Automatically extracted F0 features as acoustic correlates of prosodic boundaries. *Proceedings of FONETIK 2004*, Stockholm, p. 52-55, 2004.

[11] Behlau, M. *Voz: o livro do especialista*. Volume I. Rio de Janeiro: Revinter, 2001.

[12] Castro, L.; Moraes, J. A. The temporal structure of professional speaking styles in Brazilian Portuguese. In: *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics*, Athens, p. 53-56, 2008.

[13] Scherer, K. R.; Ladd, D. R.; Silverman, K. E. A. Vocal cues to speaker affect: testing two models. *Journal of the Acoustic Society of Ametica*, v. 76, n. 5, nov. 1984.