Listeners' Ability to Identify Professional Speaking Styles Based on Prosodic Cues

Luciana Castro, Myrian Freitas, João Moraes, Ben Serridge

Universidade Federal do Rio de Janeiro, Brazil

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

Abstract

The goal of this study is to evaluate to what extent listeners are capable of distinguishing between professional speaking styles based only on prosodic cues. The four speaking styles contemplated in the study – TV news, religious, political, and interview speech – were recorded in the context of normal use from Brazilian television and low-pass filtered to remove semantic information. The participants of a perception study were presented with one minute samples of the filtered speech and asked to choose which of two speaking styles each sample represents. The results of the perception experiment demonstrate that listeners are able to identify the speaking style with 90% accuracy, proving that even when semantic and lexical information is removed from the signal, there remains sufficient information in the prosodic cues to allow listeners to identify these speaking styles.

Index Terms: prosody, speech perception, speaking style, professional voice

1. Introduction

The relation between prosody and the characterization of professional speaking styles is noted by Léon, who observes that certain speaking styles over time become permanently associated with particular professions [1]. Fónagy asserts that "it is perfectly possible to recognize, from another room, through the walls that absorb the words and let pass only the melodic and rhythmic structure, the scientific conversation, TV news, a sports report, not to mention political discourse or a sermon" [2]. Several previous studies investigate the relation between prosody and professional speaking styles [3], [4], [5], [6]; however, not all professional speech is associated with a characteristic speaking style or to an established discourse tradition, which can be considered a limitation on the identification of a profession based only on the interpretation of the voice. Tielen evaluated listeners' ability to identify the profession of nurses, managers, and information desk clerks from the original, unaltered speech signal, semantic information having been eliminated by recording speech read from neutral texts, with inconclusive results [7]. Léon writes: "To what extent can the voice characterize a profession? When semantic information is removed, what are our chances of distinguishing the voice of a nurse from that of a business manager and from that of an employee responsible for giving out information? … It turns out that certain professions, like certain individuals, are easier to characterize than others" [1].

In the present study, to validate the hypothesis that prosodic cues alone convey sufficient information for listeners to correctly identify the speaking style, a perception study was performed using a corpus of Brazilian Portuguese speech, as part of a broader research effort to characterize these professional speaking styles [8]. The samples presented to listeners represent professions with strong discourse traditions and were selected from speech recorded in the context of real use. This paper describes the methodology and the results obtained from the perception study.

2. Methodology

The methodology of this study draws on previous work that investigates the ability of listeners to distinguish between different speaking situations based on prosodic information. Several studies, in particular, influenced the present work: Mixdorff and Pfitzinger evaluated the ability of listeners to distinguish between spontaneously produced speech and the read version of a transcription of that same passage [9]. In the experimental tradition of evaluating the communicative power of supra-segmental aspects of speech, Scherer, Ladd, and Silverman altered recorded speech in an effort to isolate prosodic cues and evaluate speech based only on the supra-segmental structure that remains after segmental elements have been eliminated [10]. In similar studies that evaluate listeners' ability to distinguish different languages based on prosodic cues alone, Ohala and Gilbert performed a study comparing the ability of listeners to correctly identify the language being spoken when presented with a triangular pulse train rendering of the fundamental frequency [11]. More recently, Dufter and Reich performed a similar study on romance languages, presenting listeners with low-pass filtered speech with a cutoff of 400Hz [12]. In both studies, listeners were able to recognize the language at levels significantly above chance, but low enough to "justify characterizing the task as difficult" [11].

In this study, in order to remove semantic information from the speech signal, as in [12], the recordings were low-pass filtered using the pass Hann band option in Praat [13].

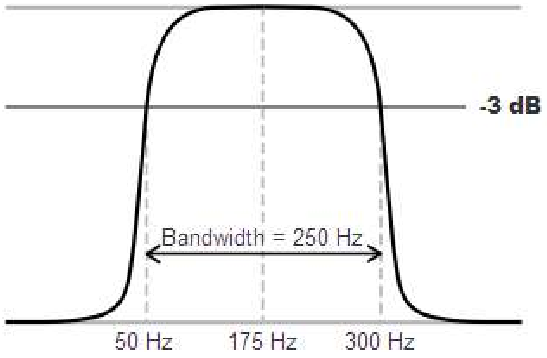

Figure 1: *Amplitude vs. frequency for the Hann band-pass filter used to filter the speech.*

The upper limit of the band was set to 300 Hz, representing approximately the highest F0 value observed in the corpus, ensuring that all relevant pitch information was captured while at the same time resulting in a signal in which no individual words can be made out.

2.1. Corpus

The corpus consists of recordings of continuous speech from television news anchors, Catholic priests, and politicians. In addition to these representatives of professional speaking styles, a fourth set of samples were collected from the speech of interview subjects responding to questions on a television talk show. These recordings represent a control group in the sense that the speech, more conversational in tone, does not appear to be clearly associated with any particular profession.

The speakers, with ages ranging from 30 to 78, are native speakers of Brazilian Portuguese with no apparent speaking disorders. Although the speakers are natives of several different geographic regions in Brazil, none of the speakers presents an overtly strong regional dialect.

Each recording consists of one minute of speech captured during the exercise of each speaker's profession: the TV news anchors and the interview subjects in the studio, the politicians in the senate, and the religious leaders in the church. The speech samples for the religious speaking style were selected from the homily, or sermon, a portion of the Catholic ceremony during which the priest adopts a more conversational style, in contrast to other parts of the ceremony, such as readings from scripture or prayer, which are delivered using a speaking style particular to the Catholic church that Léon characterizes as "melodic cliché" [1]. In the case of the political speaking style, care was taken to select recordings from political discourse delivered by senators during debate, as opposed to the rote reading of clauses in a bill. In the case of television news, all samples were taken from news anchors in the studio rather than from reporters on the street. Finally, the interview speech can be characterized as small monologues, as each speaker answered questions poised by the interviewer, in contrast to the more frequent turn taking characteristic of everyday interpersonal conversation.

All of the recordings were captured directly from Brazilian broadcast television. The recordings were made on a laptop computer with an external sound card, connected to the analog audio output of a digital cable set-top box, and digitized at a sampling rate of 22 kHz. From observing the video, it is apparent that none of the speakers was reading directly from a script, though it is assumed that the TV news anchors were working with the support of a teleprompting machine, and presumably all of the speakers had to a greater or lesser degree rehearsed their speech, and any attempt to arbitrarily classify the speech as either "read" or "spontaneous" would be to some extent futile.

2.2. Perception experiment

Twenty native speakers of Brazilian Portuguese, all university graduates ranging in age from 24 to 39, were recruited to participate as volunteers in the study. The tests were administered individually in a quiet room, with participants using headphones to listen to the utterances and a web browser to step through the questions, hosted locally as HTML pages on a laptop PC. Participants were allowed to listen to each utterance only once and instructed to attempt to identify the speaking style and mark their answers on a sheet of paper.

Each of the participants in the study was presented with 12 one minute filtered speech samples, three examples drawn from each of the four speaking styles. In addition, a practice question was prepared using a recording from a cattle auction, serving to demonstrate the test procedure and to accustom the participants to the modified speech. The answers to the practice question were discarded and do not figure in the results presented here.

To eliminate ordering effects, the utterances were presented in random order, such that no two participants received them in the same order. Furthermore, the order in which the two possible responses were presented was also randomized, with 50% of the questions presenting the correct answer as the first option, and the remaining questions presenting the correct answer as the second option.

On each question, subjects were asked to choose between two speaking styles, one corresponding to the speaking style being presented, and the other chosen randomly from the other three speaking styles. Note that, although the recognition rate is naturally higher when given a choice of two speaking styles rather than four, the null hypothesis is also proportionally higher (50% vs. 25%), so the conclusions drawn from statistical analysis of the test results are equally valid for the binary choice task as they would be if each listener were asked to choose from among all four speaking styles.

Table 1 illustrates the randomization applied during the creation of the test suite, by showing the placement of recording number 6 (which happens to be an example of the political speaking style) in each of the 20 tests. For example, the second participant in the study was presented with recording #6 as the fourth question in the test and asked to choose between "interview subject" and "politician", in that order. In contrast, the same recording was presented to the next participant in the study as the last question in the test, and in this case the two alternatives were "politician" and "priest".

Table 1. *The placement of recording #6, corresponding to the political speaking style, in the 20 test scenarios, illustrating its position in each test and the two options from which the listener was asked to choose.*

Test #	Question #	Utterance #	Speaking Style	Option 1	Option 2
1	12	6	Politician	Politician	TV news
2	4	6	Politician	Interview	Politician
3	12	6	Politician	Politician	Priest
4	11	6	Politician	TV news	Politician
5	11	6	Politician	Priest	Politician
6	9	6	Politician	Politician	Priest
7	6	6	Politician	Politician	Interview
8	6	6	Politician	Priest	Politician
9	12	6	Politician	Politician	Interview
10	10	6	Politician	Interview	Politician
11	7	6	Politician	TV news	Politician
12	3	6	Politician	TV news	Politician
13	1	6	Politician	Politician	TV news
14	3	6	Politician	Interview	Politician
15	9	6	Politician	Interview	Politician
16	3	6	Politician	Politician	Priest
17	9	6	Politician	Politician	TV news
18	9	6	Politician	Priest	Politician
19	7	6	Politician	Politician	TV news
20	9	6	Politician	Priest	Politician

After participating in the test, the subjects were asked to fill out a questionnaire in order to evaluate the degree of familiarity with the three professional speaking styles. Of the participants, 85% reported watching television news regularly, 85% reported having watched political debates or speeches, and 90% reported having attended Catholic mass. In short, the participants in this study were fairly familiar with the speaking styles under investigation, a prerequisite for the accurate recognition of them. As noted by Cagliari: "It's not sufficient that prosodic manifestations be perceptible for them to have an impact in interactive communication situations; it's necessary, furthermore, that there be consensus between speakers and listeners as to their meaning" [14].

3. Results

The results of the perception test confirm the hypothesis that these professional speaking styles can be recognized based on prosodic cues alone. Participants correctly identified the professional speaking style with 90% accuracy, a highly significant result ($p < 0.001$) considering the null hypothesis of 50% recognition. Furthermore, the results were quite consistent across speaking styles, as shown in Table 2. Even interview speech, not characterized as a professional speaking style, was recognized with a high degree of accuracy, perhaps because participants were able to identify this style as not representing any one of the other three speaking styles. While there are differences in the recognition accuracy across the speaking styles, ranging from the religious speaking style at 96.7% to the political speaking style at 86.7%, these differences are not statistically significant, as the overall recognition accuracy of 90% is within the 95% confidence range for each of the four speaking styles.

Table 2. *Recognition accuracy for each speaking style and the associated p-value. In tables 2-4, the p-value is derived from a two-tailed chi-square test with one degree of freedom.*

Speaking Style	Correct	Incorrect	% Accuracy	p-value
Interview	52	8	86.7	1.3 x 10^{-8}
Political	52	8	86.7	1.3 x 10^{-8}
Religious	58	2	96.7	4.9 x 10^{-13}
TV news	54	6	90.0	5.8 x 10^{-10}
Total	**216**	**24**	**90.0**	**2.8 x 10^{-35}**

In addition to recognition accuracy being consistently high across speaking styles, an analysis of the confusion matrix shows that no one pair of speaking styles is responsible for the majority of errors. The only confusion for which a significantly high number of errors occurs is the case of the political speaking style misidentified as the religious speaking style, highlighted in Table 3. Even in this case, the recognition accuracy is 70%. A chi-square analysis yields a p-value of 0.074 ($p > 0.05$), insufficient to reject the null hypothesis.

Table 3. *Confusion matrix showing the number of errors and (in parentheses) the p-value for each confusion. Note that the expected number of errors in the null hypothesis is 10, as each pair was presented once to each of the 20 participants.*

Recognized Presented	Interview	Political	Religious	TV news
Interview	-	5 (2.5 x 10^{-2})	2 (3.5 x 10^{-4})	1 (5.7 x 10^{-5})
Political	1 (5.7 x 10^{-5})	-	6 (7.4 x 10^{-2})	1 (5.7 x 10^{-5})
Religious	1 (5.7 x 10^{-5})	1 (5.7 x 10^{-5})	-	0 (7.7 x 10^{-6})
TV news	3 (1.8 x 10^{-3})	2 (3.5 x 10^{-4})	1 (5.7 x 10^{-5})	-

Note that the confusions are not symmetric: when presented with (filtered) political speech, listeners misidentified it as the religious speaking style in 30% of the cases; but when presented with the religious speaking style, only 1 in 20 listeners misidentified it as political. In fact, summing the values in the confusion matrix across the diagonal and running a chi-square analysis for each of the six possible confusions shows that none of the confusion pairs can be explained by the null hypothesis. Table 4 shows the error rates and p-values obtained from this analysis.

Table 4. *Confusion matrix showing the number of errors and (in parentheses) the p-value for each confusion. Note that the expected number of errors in the null hypothesis is 20.*

Speaking Style	Interview	Political	Religious	TV news
Interview	-			
Political	6 (9.6 x 10^{-6})	-		
Religious	3 (7.6 x 10^{-8})	7 (3.9 x 10^{-5})	-	
TV news	4 (4.2 x 10^{-7})	3 (7.6 x 10^{-8})	1 (1.9 x 10^{-9})	-

As described in section 2.3, several precautions were taken (e.g., the randomization of the tests) to ensure that unwanted factors would not interfere systematically in the results of the experiment. Analysis of the results on a per-utterance basis show that the lowest recognition accuracy for a single utterance was 75%, while only two utterances were identified correctly by 100% of the listeners. Similarly, none of the individual participants in the study correctly identified fewer than 75% of the 12 utterances, and five participants correctly identified 100% of the utterances. In both the per-utterance and per-participant analysis, the individual error rates are all within the 95% confidence interval for the results as a whole. In other words, there were no outliers in the study that might otherwise have distorted the results.

Another factor with the potential to distort the results is the ability of participants to learn as they proceed through the study. A lack of familiarity with the test procedure might lead to lower recognition accuracy on the first few samples, while learning during the course of the test may favor samples in the second half of the test. As described above, to neutralize the impact of lack of familiarity with the test procedure, a single recording from the cattle auction speaking style was used as a practice question before beginning the study itself. To quantify to what extent learning played a factor in the results of the study, a linear regression was applied to points of the curve formed by the recognition rate as a function of the position of the sample in the test. As shown in Figure 1, there was a slight increase in recognition accuracy (0.73% per utterance) over the course of the test. Note that because of the randomization of both the order of the utterances in each test and of the choice of incorrect response for each utterance, the presentation order is not correlated to any particular utterance or to a given combination of speaking style and incorrect response.

To validate the significance of the apparent learning effect, a t-test was applied comparing the recognition accuracy of the first six questions to that of the last six questions, the null hypothesis being that there is no effect. The t-test resulted in a p-value of 0.086 ($p > 0.05$), insufficient to reject the null hypothesis but nonetheless suggesting that there may be a minor learning effect. However, because of the random ordering of the questions in each test, it is highly unlikely that a learning effect of this magnitude would interfere with the overall results.

Finally, analysis of the recognition accuracy as a function of the order in which the correct answer was displayed on the answer sheet reveals that the position of the correct answer does not have a significant effect on the results, with questions for which the correct answer was the second option resulting in a slightly higher accuracy (91.2%) than questions in which the correct answer was the first choice (88.7%).

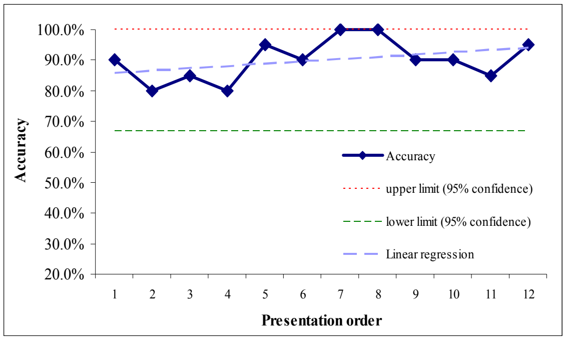

Figure 2: *Recognition accuracy as a function of the question number.*

4. Conclusions

The results of this study demonstrate that listeners are in fact able to discriminate quite well between these professional speaking styles, even when lexical, grammatical, and semantic information has been stripped from the speech signal. There must be, therefore, sufficient information in prosodic cues alone to allow listeners to distinguish between speaking styles. However, it is far from clear *which* prosodic cues contribute most to the correct identification of the speaking style, and even less clear which acoustic features in particular are relevant in characterizing each speaking style. These questions are investigated in [15] and in [8]. The empirical analysis of the prosody of professional speaking styles thus provides a context for future studies aimed at resolving these and other open research questions.

5. Acknowledgements

We would like to thank the participants in the perception study, who volunteered their time and energy to this project. Also we would like to express our gratitude to Uli Reich and to our lab colleague Manuela Colamarco for their guidance and practical help in filtering the utterances used in this study.

6. References

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

[2] Fonagy, J. Prosodie professionnelle et changements prosodiques, Le Français Moderne, XLIV(3), p. 193-228, 1976.

[3] Bhatt, P.; Léon, P. Melodic patterns in three types of radio discourse. In: Proceedings of The ESCA Workshop of Speaking Styles, Barcelona, v. 11 p./1-5, 1991.

[4] Herment-Dujardin, S.; Hirst, D. Emphasis in English: a perceptual study based on modified synthetic speech. Proceedings of Speech Prosody, Aix-en-Provence, p. 379-382, 2002.

[5] Botinis, A. et al. Produced pauses, perceived pauses and thematic units. In: Proceedings of FONETIK 2004, Stockholm, p. 112-115, 2004.

[6] Braga, D. e Marques, M. A. The pragmatics of prosodic features in the political debate. In: Proceeding of Speech Prosody 2004, Nara, p. 321-324, 2004.

[7] Tielen, M. Perception of the voices of men and women in relation to their profession. Proceedings of the Tutorial and Research Workshop on Speaker Characterization in Speech Technology, Edinburgo, p. 192-197, 1990.

[8] 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.

[9] Mixdorff, H.; Pfitzinger, H. Analysing fundamental frequency contours and local speech rate in map task dialogs. Speech Communication, v. 46, p. 310-325, 2005.

[10] 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.

[11] Ohala, J. J.; Gilbert, J. B. Listeners' ability to identify languages by their prosody. In: Léon, P. and Rossi, M. [Ed], Problémes de prosodie, Vol. II: Experimentations, modeles et fonctions. Ottawa, Didier, p. 123-131, 1981.

[12] Dufter, A.; Reich, U. "Rhythmic differences within Romance: identifying French, Spanish, European and Brazilian Portuguese." Proceedings of 15th ICPhS, Barcelona, p. 2781-2784, 2003.

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

[14] Cagliari, L. C. Da importância da prosódia na descrição de fatos gramaticais. In: Ilari, R. (org.). Gramática do português falado. Campinas, Editora da Unicamp. p. 37-60, 2002.

[15] 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.