INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Relationships between speech timing and perceived hostility in a French corpus of political debates

Charlotte Kouklia, Nicolas Audibert

Laboratoire de Phonétique et Phonologie, UMR7018 CNRS/Sorbonne-Nouvelle, Paris, France
charlotte.kouklia@sorbonne-nouvelle.fr, nicolas.audibert@sorbonne-nouvelle.fr

Abstract

This study investigates the relationship between perceived hostility and speech timing features within extracts from Montreuil's City Council sessions in 2013, marked by a tense political context at this time. A dataset of 118 speech extracts from the mayor (Dominique Voynet) and four of her political opponents during the City Council has been analyzed through the combination of perception tests and speech timing phenomena, estimated from classical timing-related measurements and custom metrics. We also develop a methodological framework for the phonetic analysis of non-scripted speech: a double perceptive evaluation of the original dataset (22 participants) allowed us to measure the difference of hostility perceived (dHost) between the original audio extracts and their read transcriptions, and the five speakers produced the same utterances in a controlled reading task to make the direct comparison with original extracts possible. Correlations between dHost and speech timing features differences between each original utterance and its control counterpart show that perceived hostility is mainly influenced by local deviations to the expected accentuation pattern in French combined with the insertion of silent pauses. Moreover, a finer-grained analysis of rhythmic features reveals different strategies amongst speakers, especially regarding the realization of interpausal speech rate variation and final syllables lengthening.

Index Terms: expressive speech, politics, hostility, speech timing.

1. Introduction

Since Aristotle's Rethoric, it is known that politicians have to give the image of moral, benevolent and competent individuals to reach or maintain a high status [1]. For contemporary research in psycho-sociology [2], politicians should also appear to be dominant, especially in political debates, where it becomes an essential quality. Dominance signals may include a strong display of self-control and more or less tacit hostility, implying suppressed aggressive expressions such as contempt, sarcastic irony or irritation.

The phonetic analysis of expressive speech produced in ecological contexts faces two major difficulties, due to the absence of a systematic variation on the same utterances that do not enable comparisons "all other things equal". Regarding the phonetic analysis itself, the comparison of utterances with different segmental contents and syntactic structures to get insights on a factor of variation is likely to lead to erroneous conclusions. Moreover, with variable lexical contents, the perceptual evaluation necessary to draw a link between speech characteristics and the paralinguistic values conveyed can hardly be interpreted. Indeed, in such a case perceptual ratings depend both on the semantic contents of the utterance and on the way it is uttered. In order to overcome those limitations and make a controlled phonetic analysis of naturally occurring expressive speech, we also develop a methodological framework presented here.

As for the expression of hostility, the most salient phonetic characteristics of hot anger are largely described in the literature. However, divergent descriptions can be found among studies for suppressed sub-categories such as cold anger, associated with less activated emotional reaction and therefore expressed with less specific features [3][4]. Nonetheless, suppressed forms of anger in French have been described by an increased speech rate, temporal perturbations such as unstable rhythmic patterns, strengthening of secondary accents and a smaller syllabic Vowel/Consonant duration ratio [5][6]. Such characteristics tend to go against the canonic rhythmic structure of French, known for its isochrony and gradual syllable and vowel lengthening towards word and intono-syntactic boundaries [5][7].

The aim of this study is to analyze the specificities of political expressive speech, and more precisely how certain rhythmic features can be related to the perception of hostility. As classical timing-related measurements do not enable fine-grained analysis of rhythmic irregularities, we develop a set of custom metrics at the interpausal run level, including differences between consecutive units and runs to better account for the dynamics of rhythmic phenomena.

2. Methodology

2.1. The original dataset

Debates were extracted from the TV archive of Montreuil's City Council in 2013 broadcasted by the local channel TV Montreuil, during Dominique Voynet's last months as mayor before the new elections. During that period, and since a couple of years already, the political context was very tense and sessions took place in a particularly stormy atmosphere. In order to avoid overlapping speech, a turn taking regulation system only allowed one microphone to be active at a time.

The present study focuses on 5 speakers: the mayor, D. Voynet (Ecologist Party) et 4 of her male opponents: G. Le Chequer, F. Molossi, J.-J. Serey and A. Tuaillon, all members of left wing parties. The opponents have been selected on the basis of the high frequency of their interventions and their recurring and lively exchanges with the mayor. They also have been selected for their dynamic speech style, characterized by a constant improvisation based on prepared speech.

Six sessions of the year 2013 have been analyzed (24 hours in total). The first steps performed on the video recordings were manual speaker diarization and labeling of events related to turn-taking and expressiveness.

http://dx.doi.org/10.21437/Interspeech.2017-293

The extracts were cut in order to avoid the presence of strong verbal cues to the intentions of the speaker, such as insults or mentions of the names of the interlocutors. Truncated extracts are cut at the end of propositions, equivalent to major continuation boundaries [8][9][10]. Extracts, comparable to short or truncated sentences, do not exceed 12 seconds (mean duration: 8.7 seconds), to avoid inducing in further perceptual evaluations biases due to a too large amount to contextual cues [11]. A subset of 125 extracts (25*5 speakers) was selected for the present study.

2.2. The reread dataset

As phonetic analysis based on comparison of utterances with different segmental contents and syntactic structures is likely to lead to erroneous conclusions, a control condition was recorded from simplified orthographic transcriptions. All 5 speakers were asked to produce a read counterpart of their own stimuli as neutrally and intelligibly as possible. Recordings were done in a quiet environment, either in their home or office, with a hypercardioid headset microphone (AKG C520) connected to a M-AUDIO (M-TRACK II) sound card.

A simplified orthographic transcription was prepared to facilitate the reading and ensure the understandability of the extract, applying recommendations by Blanche-Benveniste [12]: addition of punctuation marks, suppression of filled pauses such as hesitations, repetitions and revisions. To facilitate a natural reading for the speakers without inserting artificial final boundaries, the transcription was extended to the next interpausal run in this production task, recorded stimuli being cut afterwards to match the verbal contents of the original ones. An example is presented below:

Original stimulus: *donc euh aujourd'hui euh la justice euh a a rendu euh euh son plaidoyer euh*

Extended transcription, cut part in brackets: *donc, aujourd'hui, la justice a rendu son plaidoyer [et le directeur en question a été réhabilité].*

English translation: so, (um) today (um) justice (um gave) gave its (um um) defense (um) [and the director in question has been cleared]

In order to avoid interference with the semantic contents and heterogeneous reading strategies from one speaker to another, speakers were asked to produce this reading in the way of a dictation, after having listened to the original version in a self-rating task (see [13] for results). Although this previous knowledge of their original productions could have leaded to a partial imitation, it appeared to help speakers satisfy their curiosity and focus on the controlled production task as requested. Given the specificities of syntactic structures in the original dataset, some extracts were reread with misplaced or forgotten determinants and negation marks, making direct comparison with original stimuli impossible. As a result, a total of 7 extracts had to be removed from the reread dataset.

2.3. Perceptual evaluations

Two perception tests were conducted [13]: on the original data (audio condition), and on the simplified orthographic transcriptions of the stimuli (written condition). Perceptual scores in the written condition can be considered as ratings of the semantic contents of utterances. After carefully weighting pros and cons, this condition was preferred to an audio evaluation of stimuli recorded in both conditions (original and reread), that would have implied biases whatever the considered evaluation protocol.

In both conditions, participants were requested to rate on Likert-scales the degree of hostility expressed by the stimuli presented in random order, as well as the target audience (speaking to a particular interlocutor or to the whole audience) associated with a degree of certainty. The test on audio data was run with Praat [14] and taken by 11 native French listeners (8 F, 3 M, mean age 24). The written test (8 F, 3 M, mean age 26) was conducted online through a *Google Form.* In the audio condition, participants were given the possibility to replay each stimulus once. In this study, we focus only on ratings of perceived hostility.

Mean hostility ratings in audio condition are noted hereafter $host_{Audio}$, while mean hostility ratings in written condition are noted $host_{Written}$.

2.4. Segmentation and annotation

Stimuli from both datasets were segmented in phones, syllables and words with the *Easy Align* plugin [15] for Praat [14], boundaries being manually corrected when necessary. Interpausal runs were defined as sequences of phones separated by a silent break superior to a threshold. Most recent studies on silent pauses and spontaneous speech [18][20][19][20] consider a 200 ms threshold, while Duez [9] proposed a variable threshold value based on segmental durations to account for speech rate specificities. However, given the large amount of local speech rate variation observed, this approach turns out to be hardly applicable on our data. Given the considerable amount of speech uttered with a very high speech rate, the silent pauses duration threshold was set to 140 ms, i.e. the lowest value used in [9].

2.5. Extraction of speech timing features

All speech timing-related features were extracted on each interpausal run for all stimuli (in both original and reread version). As an interpausal run generally include several words, retained features give information on the number of units and the distribution of duration-related parameters.

Extracted features can be subdivided into the four categories listed below. (1) Basic counts of occurrences of segmental categories: vowels, consonants, silent pauses, and filled pauses, with a variable for pauses observed at the beginning or at the end of a run, as well as syllables count; (2) Mean and standard deviation of the durations of units listed in cat. 1, of the preceding and following silent pause and of syllable constituents, as well as the V/C ratio (as used by Fónagy [6]) inside each syllable; (3) Speech rate related measurements: speech rate including pauses and articulatory rate in both segments and syllables per second, PVI and nPVI [21]; (4) Metrics specifically designed to capture deviations to the canonical accentuation pattern in French, in which the lengthening of words final syllables is expected to increase gradually in the course of an interpausal run [5].

Rhythmic features in category 4 are computed both relatively to vowel durations, and to rime durations. They are based on duration measurements taken on the first and final syllables of words included in the run. At the word level, the ratio between durations on the first and final syllable of each word, that captures increases or decreases in the relative lengthening of the final syllable, is denoted as *wordLgRatio*. Values of *wordLgRatio* are expected to be lower in the case of a didactic accent increasing the duration on the first syllable. Durations on final syllables of the first and last word in the run are used to compute the lengthening slope (referred to hereafter as *lgSlope*), which reflects the increase rate of lengthening throughout each interpausal run.

2.6. Comparison between pairs of stimuli

At the perceptual level, the degree of hostility added by the audio condition in addition to the semantic load (dHost) is computed for each pair as: $dHost = host_{Audio} - host_{Written}$.

Timing-related characteristics of stimuli were analyzed considering the difference between each original extract and its reread version. The first step of the pairing procedure was to combine measurements taken on each interpausal run into features relative to a whole utterance, either in original or reread condition. The mean of each feature was computed to reflect the tendency across interpausal runs. To account for possible rhythmic breaks, the retained measurement of inter-runs variability was the mean difference between consecutive runs (i.e. a measurement similar to the PVI applied at the interpausal run level).

The final step was to compute pair-differences between values extracted from the original stimulus vs. its reread version. For the sake of readability, pair-differences computed on a particular variable are referred to using the raw variable name in the following, indicating only the kind of statistic used to group values relative to different runs.

3. Results

3.1. Time-related metrics in original vs. reread extracts

Comparisons between the original extracts and reread ones were performed using a paired t-test for each timing-related variable. Those comparisons were conducted both: (1) on mean measurements across all interpausal runs of each utterance (columns labeled "mean") and (2) on mean differences between two consecutive interpausal runs (columns labeled "runDiff"). Table 1 presents a set of 8 variables retained on the basis of significant differences found either in utterances comparisons or timing-related behavior differences between each consecutive interpausal run. In addition to the significance level, effects sizes measured by Cohen's d are presented to enable comparisons of effect magnitude between different dependent variables.

Measurements related to a same time-related phenomenon are highly intercorrelated (for instance the duration of interpausal runs is strongly linked with the number of segments or syllables they're composed of). As a consequence, among sets of similar variables we picked the variable for which the comparison between original and reread condition yielded the strongest effect.

3.2. Production-perception correlations

As a first step, correlations between the 8 time-related variables retained and $host_{Audio}$ were computed for both mean measurements and mean differences between consecutive runs. Corresponding correlation coefficients are presented in the column labeled "timing vs. $host_{Audio}$" of Table 1 (this kind of correlation is noted $corr_{Audio}$ in the presentation of results). Correlations with dHost are presented in column "timing vs. dHost" (noted $corr_{Diff}$).

3.3. Main results

For mean values across interpausal runs of utterances, the main differences between original and reread productions relate to rhythmic aspects such as vocalic and syllabic duration variation. Indeed, leaving pauses duration apart, the largest differences are found for differences of (1) V/C ratio variability, due to differences in onset and rime relative duration means ($p<.001$, $d=.41$); (2) mean nPVI on vowels; (3) mean wordLgRatio; (4) runDuration; (5) lastWordFinalDuration. Correlations with hostility ratings (for $corr_{Audio}$ and to a lesser extent in the first four cases $corr_{Diff}$), indicate that utterances rated as more hostile are realized with a more stable V/C ratio, a more instable intra-run speech rate, a weaker relative accentuation of final syllables at the word level, shorter interpausal runs (in relationship with the number of inserted silent pauses), and a shorter accentuation at the end of interpausal runs. Looking at differences between consecutive interpausal runs on those variables, large differences are observed as well as correlations (mainly in $corr_{Audio}$) for VCratioVar and lastWordFinalDuration, with a rather large negative correlation for VCratioVar indicating higher hostility ratings when the V/C ratio is more constant.

Table 1: *Comparison between original and reread condition (***= p<.001, **=p<.01, *=p<.05, n.s. = non-significant; Cohen's d value reported in parentheses) and correlations between the 8 retained timing variables and perception.*

	original vs. reread		**timing vs. $host_{Audio}$**		**timing vs. dHost**	
variable	**mean**	**runDiff**	**mean**	**runDiff**	**mean**	**runDiff**
runDuration	** (.25)	n.s.	.0148	-.1656	-.0864	.0068
filledPausesDurationVar	*** (.35)	*** (.37)	-.2141	-.2258	-.0537	-.0192
lastWordFinalDuration	* (.19)	* (.23)	-.0331	-.0161	-.0508	.0672
nPVIvowels	*** (.35)	* (.24)	.0258	.0852	.0340	-.0740
silentPausesDurationVar	* (.19)	n.s.	.0428	.0425	.0319	-.0410
speechRate (phonemes/s)	n.s.	*** (.44)	.2180	.0270	.0004	.1075
VCratioVar	*** (.43)	** (.33)	-.1851	-.1717	.0043	-.0093
wordLgRatio	** (.27)	n.s.	.2106	.1322	.0452	.1202

Comparisons between original vs. reread extracts yield significant difference in terms of duration variability for both kinds of pauses, though the difference is larger for filled pauses. Both perceptual variables show negative correlations with pauses duration variability, with much larger correlations for $corr_{Diff}$ than for $corr_{Audio}$ regarding filled pauses: utterances in which such pauses are more constant in duration are the least perceived as hostile. However, it should be outlined that reread extracts include no filled pauses and fewer short silent pauses.

Mean speech rate is the only retained variable that does not distinguish productions from original to re-read across utterances, though mean speech rate is one of the variables that is the most correlated with $host_{Audio}$, indicating that utterances realized with a faster speech rate are perceived as more hostile. Such tendencies could be explained by the highly formal nature of the communication situation. Indeed, even when expressing hostility, politicians may have to exercise control over their production in order to maintain their image. On the other hand, differences between speech rates in consecutive runs yield the largest effect when comparing original vs. reread extracts, with a much higher correlation in $corr_{Diff}$ than in $corr_{Audio}$. This last result indicates that speech rate variation between an interpausal run and the following one play a key role in perceived hostility carried by the audio level.

In addition to the case of speech rate, the higher correlations in $corr_{Diff}$ vs. $corr_{Audio}$ for variables lastWordFinalDuration and speechRate suggest that information carried by the acoustics and relevant for the perception of hostility might be partly masked by the semantic information. Changes in correlations sign between both conditions also indicate that this information on hostility is not always congruent with the semantics.

Looking at inter-speakers differences, the most remarkable result is observed for the V/C ratio variability: contrarily to all her four opponents, the mayor D. Voynet is perceived as more hostile when this ratio is more unstable across syllables, which might be related to her specific role in the council. Similar inter-speaker differences are also found for PVI measurements, D. Voynet and J.J. Serey being perceived more hostile on their more unstable productions. Finally, large inter-speakers differences are found for the variability of filled pauses durations, with correlations in $corr_{Audio}$ ranging from -.075 for D. Voynet to -.522 for G. Le Chequer.

4. Discussion

Our results underline the impact of timing irregularities and deviation from the canonical accentuation pattern in French on the perception of hostility. Comparison of correlations between $corr_{Audio}$ and $corr_{Diff}$ suggest that the hostility specifically carried by the acoustic realization is largely congruent with that carried by the utterance as a whole (as indicated by correlations sign), and in most cases has a lesser magnitude. However, focusing on differences between consecutive runs, results also reveal patterns of variations specific to the acoustic level in terms of speech rate changes.

In addition to the links we could draw between timing-related speech parameters and perceived hostility in naturally occurring expressive utterances, the methodology we developed enables an analysis of expressivity independently of the semantic level, potentially revealing more subtle phenomena. This method can easily be extended to include acoustic parameters such as F0-based features. However, the paralinguistic information carried by acoustics and semantics are obviously correlated (in our data, r=.764 between $host_{Audio}$ and $host_{Written}$). If one assumes that the large majority of this information is carried by the semantic level, which in our results would be consistent with differences observed between the two kinds of correlations for most variables, this analysis might reveal only perceptually irrelevant phenomena. In order to obtain a direct evaluation of the paralinguistic information specifically carried by the acoustics and get more insights on the way it relates with the semantic information carried by the lexicon, we plan to carry out a perceptual evaluation of delexicalized versions of the original extracts.

Although their absence in the reread productions might artificially boost the statistical effects on their duration, filled pauses seem to play a major role in the rhythmic organization of the original extracts. A more extensive analysis of filled pauses-related rhythmic events should enable us to analysis more thoroughly inter-speakers differences [17][18][20][20].

Differences in speech rate from one interpausal run to the following one could be explained using a fine-grained manual categorization of silent pauses types (hesitation, demarcation or focalization). As shown by Ferré [18] and Béchet et al. [20], different rhythmic strategies from one interpausal block to another can be expected depending on the type of silent break. Similarly, the strong correlations between perceived hostility and speech rate differences between adjacent runs could be related to the use of focalization pauses, reported in the literature as frequent in French political speech [9][20]. Ferré [18] also describes in spontaneous speech a strong pragmatic use of *displaced demarcation pauses*, i.e. intra-constituent pauses, for instance between a conjunction and the following syntagm in a subordinate. Following annotation guidelines by Duez [9], a total of 49 occurrences of intra-constituent pauses is found in the original dataset against only 5 in the reread dataset, in line with this description.

Focusing only on duration-based measurements, this study puts light into the multiple and heterogeneous accentual reinforcements in hostile political speech. A multiparametric analysis including f0 contours would be necessary to distinguish simple focalizations from expressive emphasis [22] and expressions of hostility. Indeed, aggressive speech is often characterized by abrupt and repetitive descending contours [6][4].

In order to get further in our understanding of expressivity in political speech, multidisciplinary approaches can be precious. For instance, Béchet's study [20] on silent pauses in French political debates was the results of a collaboration between phoneticians, interactionists and discourse analysts. In psychology, Poggi et al. [23] developed a typology of discriminating moves from the analysis of hostile Italian debates. In their view, speech and expressive strategies depend on the part of ethos that is aimed at the opponent (Benevolence, Competence, Dominance). Such a typology could be of great help in explaining and categorizing the various paralinguistic cues found in the speaker's productions.

5. Acknowledgements

The authors would like to thank the 5 speakers of this dataset for their outstanding involvement in this study, and the listeners for their participation. This work was supported by the Labex EFL program (ANR-10-LABX-0083).

6. References

[1] P. Charaudeau. “Le charisme comme condition du leadership politique”, *Revue française des sciences de l’information et de la communication*, vol. 7, 2015.

[2] I. Poggi and F. D’Errico, “Dominance Signals in Debates”, *Human Behavior Understanding*, Springer, pp. 163–174, 2010.

[3] R. Banse and K.R. Scherer, “Acoustic profiles in vocal emotion expression”, *Journal of Personality and Social Psychology*, vol. 70, no. 3, 1996, pp. 614–636.

[4] K.R. Scherer, “Vocal communication of emotion: A review of research paradigms”, *Speech communication*, vol. 40, no. 1, 2003, pp. 227–256.

[5] P. Léon, *Précis de phonostylistique: parole et expressivité*. A. Colin, 2005.

[6] I. Fónagy, *La vive voix: essais de psycho-phonétique*. Payot, 1983.

[7] V. Lucci, “L’accent didactique,” *Studia Phonetica Montréal*, vol. 15, pp. 107–121, 1980.

[8] P. Delattre, “Les Dix Intonations de base du français,” *The French Review*, vol. 40, no. 1, pp. 1–14, 1966.

[9] D. Duez, *La pause dans la parole de l’homme politique*. Paris: CNRS, 1991.

[10] D. Duez, “Silent and Non-Silent Pauses in Three Speech Styles,” *Language and Speech*, vol. 25, no. 1, pp. 11–28, 1982.

[11] S. Morange and M. Candea, “Aux frontières de l’écoute. Durée des échantillons et choix des auditeurs : deux variables déterminantes dans la construction des tests de perception,” *Frontières. Du linguistique au sémiotique,* Paris: Lambert-Lucas, pp. 79–96, 2010.

[12] C. Blanche-Benveniste, *Approches de la langue parlée*, Paris: Editions OPHRYS, 2010.

[13] C. Kouklia, “Approches méthodologiques pour l’étude de la parole expressive en politique ”. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, no. 32, 2016.

[14] P. Boersma, and D. Weenink, Praat: doing phonetics by computer [Computer program], version 6.0.18, retrieved June, 5, 2016.

[15] J.P. Goldman, Easy align [computer program]. Retrieved May, 7, 2010.

[16] P. B. de Mareüil, G. Adda, M. Adda-Decker, C. Barras, B. Habert, and P. Paroubek, “Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques,” *TIPA. Travaux interdisciplinaires sur la parole et le langage*, 2013.

[17] M. Candea, *Contribution à l’étude des pauses silencieuses et des phénomènes dits “d'hésitation” en français oral spontané. Etude sur un corpus de récits en classe de français*. PhD dissertation, University Paris 3, 2000.

[18] G. Ferré, “Les pauses démarcatives déplacées en anglais spontané. Marquage kinésique et prosodique,” *LIDIL - Revue de linguistique et de didactique des langues*, vol. 26, pp. 155–169, 2003.

[19] G. Ferré, “Gesture, Intonation and the Pragmatic Structure of Narratives in British English Conversation,” *York Papers in Linguistics*, vol. 2, pp. 55–90, 2005.

[20] M. Béchet, M. Sandré, F. Hirsch, A. Richard, F. Marsac, and R. Sock, “De l’utilisation de la pause silencieuse dans le débat politique télévisé. Le cas de François Hollande,” *Mots. Les langages du politique*, no. 103, pp. 23–38, 2013.

[21] E. Grabe and E.L. Low, “Durational variability in speech and the rhythm class hypothesis,” *Papers in laboratory phonology*, vol. 7, pp. 515–546, 2002.

[22] I. Guaïtella. “Proéminences et éminences : savoir-faire discursif, faillances et défaillances des hommes politiques” *TIPA. Travaux interdisciplinaires sur la parole et le langage*, no. 30, 2014.

[23] I. Poggi, F. D’Errico and L. Vincze, “Discrediting moves in political debates,” In *UMMS 2011 – Second International Workshop on User Models for Motivational Systems: the affective and the rational routes to persuasion, Girona, Spain, Proceedings*. Springer LNCS, 2011, pp. 84–99.