



Audio Quotation Marks For Natural Language Understanding

Simon Boutin^{1,2}, Réal Tremblay², Patrick Cardinal¹, Doug Peters², Pierre Dumouchel¹

¹École de Technologie Supérieure, Montréal, Canada

²Nuance Communications, Montréal, Canada

simon.boutin.1@ens.etsmtl.ca, real.tremblay@nuance.com, patrick.cardinal@etsmtl.ca

doug.peters@nuance.com, pierre.dumouchel@etsmtl.ca

Abstract

Detecting the presence of quotations in speech is a difficult task for automatic natural language understanding. This paper presents a study on the correlation between three prosodic features present in a voice command and the presence or absence of quotations. These features consist of intra-word pause durations, F0 reset and F0 continuity. A combination of lexical and prosodic extraction tools was used to extract these features. The two-sample Kolmogorov-Smirnov test was then used to compare the distributions of the collected measures. The results show a correlation between these features and the presence or absence of quotations. Moreover, the results show that it is possible to use these features to differentiate direct from indirect quotations.

Index Terms: prosody, quotation, Kolmogorov-Smirnov, natural language understanding

1. Introduction

The goal of human-machine dialog systems is to interact with computer systems in a more natural and effective way than with conventional interfaces. Current systems consist of three loosely coupled components: the automatic speech recognition system (ASR), the natural language understanding system (NLU) and the dialog system or conversational agent (CA). In addition to the lexical information extracted by the ASR system, the original signal may contain prosodic cues that could help the NLU system. However, prosodic cues are not often used by NLU systems, which are usually driven entirely by lexical information. Prosody concerns sound intonation, intensity and duration, which of course is absent at the lexical level.

The goal of this paper is to determine the possibility of improving a NLU system's performance for quotation detection. To achieve that, a correlation between some intra-word prosodic features and the presence or absence of quotations is investigated. If there is such a correlation, the prosodic information could be eventually combined with the existing lexical information for enhanced quotation detection.

In the literature, quotations can often be classified as a type of dialog acts (DA). Some authors already combine lexical and prosodic models to classify and detect DA [1, 2, 3, 4]. However, the literature does not address the case where the author of the quotation is the narrator himself, which is the present focus. A quotation is used for the assignment of specific words and thoughts to others. It represents a discourse in a discourse. Quotations are divided into three categories [5]: direct speech (quotation is expressed word for word), indirect speech (what was said is expressed in our own words) and free indirect speech (characters speak to us without narrator as intermediary). En-

glish speakers can use prosodic cues - "audio quotation marks" by analogy - to identify quotation boundaries [6]. Changes are frequently observed on the global pitch, intensity, and speech rate. However, not all quotations are prosodically marked.

Prosody can help to detect the end of direct speech [7]. The start of a quotation is often linguistically marked with a quotation verb, but no such linguistic mark indicates its end. The pitch reset seems to be the most appropriate feature to detect the end. Prosody can also differentiate direct and indirect speech [8]. The global pitch range is much greater for direct quotations. A significant difference also exists in the quantity of pitch reset relative to a preceding sentence.

Finally, a correlation has been studied with human annotators between prosody and statements in quotation marks [9]. Sentences read aloud differ at the prosodic level relative to their counterparts when they contain passages enclosed with quotation marks. Four different strategies are observed: a main break, a drawing break, a change in voice quality and the pitch accents movement. In fact, speakers do not rely on a single strategy. The current study represents an attempt to quantify these kinds of observations.

The rest of this paper is organized as follows. Section 2 introduces the definition of quotation and makes hypotheses. Section 3 describes the corpus used for the present experiment. Section 4 presents the two-sample K-S test. Section 5 describes the methodology. Section 6 presents the experiment and results, followed by the interpretation and discussion in Section 7. Finally, a conclusion summarizes the experiment.

2. Definition of Quotation

As explained in the literature review, a quotation is used for the assignment of specific words and thoughts to others. It represents a discourse in a discourse. In this experiment, authors of quotations and their narrator correspond to the same entity. Two quotation categories are defined inspired by the literature: direct and indirect, i.e. with or without the immediate presence of words on the left boundary announcing its presence. Table 1 shows an artificial quotation example of each category.

This paper tests two hypotheses. First, that there is usually a more pronounced pause period, F0 reset and F0 continuity at the start of a quotation relative to other concepts. Here, the word "concept" is used to refer to a word or phrase representing a semantic component of an utterance. Second, that these same features are generally more pronounced at the start of a direct relative to an indirect quotation.

Sentence Quotation Quoted Type	Text Mary have a nice day have a nice day Narrator Direct
Sentence Quotation Quoted Type	Text Mary and tell her that we're leaving we're leaving Narrator Indirect

Table 1: Examples of direct and indirect quotations

3. Experimental Corpus

To investigate the above hypotheses, a corpus of 39 584 utterances was used. This corpus contained 10 621 words starting a quotation (10 621 records) and 97 682 words in records without quotations (28 963 records). Extraction of F0 reset and F0 continuity features from this corpus was attempted. However, utterances with invalid results due to pitch halving and doubling during automatic extraction (a known problem in the literature [10]) were discarded. Audio records came from potentially different English speakers for each record. Their textual transcriptions and concepts assigned for each word were available from an automated NLU process. Concepts were manually corrected for commands containing quotations. Unfortunately, this sample set was slightly biased, because the data were previously filtered by an automated NLU system. This filtering maximized the probability of automatically extracting the appropriate concepts. To have an unbiased sample set, a random sampling would have been necessary. However, the subsequent manual annotation of a random sample set was prohibitive. The use of this biased sample set represents a practical means to overcome this tradeoff.

4. Two-Sample K-S Test

The two-sample Kolmogorov-Smirnov test (K-S test) [11] is a nonparametric test to determine if two sample sets derive from the same one-dimensional continuous probability distribution. It quantifies a distance between their empirical distribution functions (K-S statistic). To achieve this, the empirical distribution function is calculated for each sample set as observations X_i of two random variables taken to be independent and identically distributed:

$$F_{n_1}(x) = \frac{1}{n_1} \sum_{i=1}^m I_{X_i \leq x} \quad F_{n_2}(x) = \frac{1}{n_2} \sum_{i=1}^m I_{X_i \leq x} \quad (1)$$

where $F_{n_1}(x)$ and $F_{n_2}(x)$ are empirical distribution functions for the first and second sample sets of size n_1 and n_2 respectively, $I_{X_i \leq x}$ is the characteristic function and equals 1 if $X_i \leq x$ and equals 0 otherwise. The K-S statistic is then calculated as follows:

$$D_{n_1, n_2} = \sup_x |F_{n_1}(x) - F_{n_2}(x)| \quad (2)$$

where \sup is the supremum function. The p-value [12] is a function of the observed sample results that is used for testing a statistical hypothesis. The null hypothesis, specifying that samples are taken from the same distribution, is rejected or not by comparing the p-value to a level of confidence α usually assigned to 0.05 or 0.01. It is rejected if the p-value is

less than this threshold, and the test result is declared "statistically significant". Otherwise, the null hypothesis is accepted. The p-value is difficult to calculate and statistical software is required, often using numerical methods rather than exact formulas. The p-value and the K-S statistic were calculated in the Python programming language using the *ks_2samp* function of the *scipy.stats* library for this experiment.

5. Method Description

It was necessary to use appropriate tools for extracting features of voice commands to verify hypotheses in this experiment. A quality assessment was performed on these tools to determine their effectiveness. A lexical extraction tool has been combined with a prosodic extraction tool to collect intra-word prosodic features. The following sections describe these two tools and the studied prosodic features.

5.1. Lexical Extraction Tool

Lexical characteristics were automatically extracted with a voice recognition system developed by Nuance named Nuance Recognizer version 10.2.4 (NR10). Time resolution was 10 msec for the extracted measures. NR10 needs an audio and a grammar as input to work. For the needs of this experiment, the given grammar only contained the same previous recorded automatic transcription of the relevant record. This approach optimizes the alignment of words and pauses relative to the signal.

5.2. Prosodic Extraction Tool

Prosodic features were automatically extracted with an application designed by Paul Boersma and David Weenink from the University of Amsterdam named Praat version 3.5.49 (Praat). Praat is an open source software specialized for speech analysis. It allows the creation and execution of scripts required for the automated extraction of prosodic features. Time resolution was 1 msec for the extracted measures.

5.3. Pause Duration

Pause duration was the only prosodic feature requiring the combination of Praat and NR10 measures. The lexical extraction from NR10 did not always detect short duration pauses. On the other hand, Praat was very sensitive to noise and filled pauses, often not considering them as pauses. An algorithm was designed to minimize the error on the automatic extraction of pauses, combining the strengths of the two tools.

This algorithm starts by obtaining the pauses extracted by NR10 before each word, even if many go undetected (duration of zero). The algorithm then tries to expand each of these pauses with those extracted by Praat. It takes the minimum value between the NR10 and Praat pause start. These pauses must overlap, but a margin of up to 50 msec of no pause was allowed between the end of the Praat and the beginning of the NR10 pause. The new pause start became this minimum value. The new pause end became the maximum value between the NR10 and Praat pause end. These pauses must overlap, and no margin was allowed.

Praat sometimes generated intermittent pauses when the signal was noisy. The algorithm then tried to cover these intermittent pauses to form only one large pause. To succeed, the noise interval between two intermittences must be no more than 30 msec. This 30 msec noise interval and the 50 msec

pause margin above are ad hoc values chosen based on the observation of several voice commands used in algorithm development. The algorithm proceeded to the right and left of the original pause. Finally, the new values were considered as the official automatic measures once the execution of the algorithm was completed. Figure 1 shows an example of final pauses combined from NR10 and Praat.

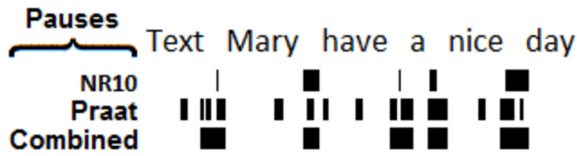


Figure 1: Example of final pauses combined from NR10 and Praat

5.4. Features derived from F0

A line was generated for each side of the intra-word boundary using a simple linear regression model applied to F0 estimates in windows to the left and right side of a pause. These lines roughly represented the F0 values measured for this window.

F0 reset correspond to the difference in pitch between the two sides of the intra-word boundary. This feature is used to capture the well-known tendency of speakers to reset pitch at the start of a new major unit. The reset is typically preceded by a final descent in pitch associated with the end of such a unit. Thus, a larger reset is observed at the boundaries of these units relative to the other parts of speech. Figure 2 shows an example of F0 reset.

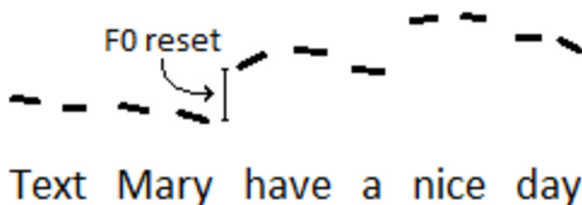


Figure 2: Example of F0 reset

6. Experiments

The data was partitioned into two classes: quotations (Q) and others (O). To collect data for Q, an algorithm extracts only the transcripts belonging to domains “email”, “sms”, “reminder” or “note”, containing at least once the concept “title” or “text” but also containing lexical material prior to those concepts. The class O consist of data in others domains with at least two words.

For quotations, features were extracted before each quotation-initial word. For other concepts, all words of the voice command were considered, except the first since there was no previous word for comparative purposes. Table 2 shows the Praat settings used for the pauses extraction, followed by the Table 3 with the Praat settings used for the F0 extraction.

Minimum pitch (Hz)	100.0
Time step (s)	0.001
Subtract mean	Yes
Silence threshold (dB)	-30.0
Minimum silent interval duration (s)	0.001
Minimum sounding interval duration (s)	0.001

Table 2: Praat settings for automatic pauses extraction

Time step (s)	0.01
Minimum pitch (Hz)	75.0
Maximum pitch (Hz)	600.0

Table 3: Praat settings for automatic F0 extraction

7. Results

This section presents the experimental results. Tables 4, 5 and 6 show the pauses duration, F0 reset and F0 continuity distributions statistics respectively.

	Quantity	Mean	Std dev
Quotations	10 621	285.214	213.510
Direct quotations	8 644	311.212	323.713
Indirect quotations	1 977	171.544	224.993
Other concepts	97 682	103.513	307.809

Table 4: Pauses duration distributions statistics

	Quantity	Mean	Std dev
Quotations	8 806	10.383	32.587
Direct quotations	7 216	10.092	32.518
Indirect quotations	1 590	11.706	32.869
Other concepts	80 084	3.651	26.439

Table 5: F0 reset distributions statistics

	Quantity	Mean	Std dev
Quotations	9 715	104.593	718.351
Direct quotations	7 938	96.646	718.465
Indirect quotations	1 777	140.092	716.765
Other concepts	85 663	64.961	673.825

Table 6: F0 continuity distributions statistics

8. Interpretation

The results showed that the mean of three prosody features were higher before quotations relative to other concepts. However, the standard deviation for each distribution was very high. This is in large part due to some extreme values in the data set. To verify whether these results are nevertheless significant, the K-S test was applied to these three features with a threshold of $\alpha = 0.05$. The three null hypotheses - that the sample sets from quotations and other concepts derive from the same distribution - were rejected since p-values were less than α for each of these cases. The results therefore support the first hypothesis formulated in Section 2, namely, that there is usually a more

pronounced pause period, F0 reset and F0 continuity prior to a quotation. Table 8 shows p-values and K-S statistics calculated with the Python *ks_2samp* function.

	P-value	K-S statistics
Pauses duration	0.0	0.455
F0 reset	4.252e-259	0.194
F0 continuity	8.188e-32	0.064

Table 7: K-S test for quotations and other concepts

The results showed that the mean pauses duration was higher before the start of direct relative to indirect quotations. However, the opposite was observed for the two other features. Applying the K-S test again, Table 8 shows the relevant statistics.

	P-value	K-S Statistics
Pauses duration	2.560e-126	0.301
F0 reset	2.472e-7	0.078
F0 continuity	0.027	0.038

Table 8: K-S test for quotations and other concepts

The three null hypotheses, that the sample sets from direct and indirect quotations derive from the same distribution, were rejected since p-values were less than α for each of these cases. The results support the second hypothesis formulated in Section 2 for the pauses duration, while it is not supported for the F0 reset and F0 continuity.

9. Discussion

F0 reset observations suggest that speakers generally increase the pitch when they start to dictate a quotation, whether direct or indirect. However, this finding was also observed to a lesser degree for other concepts. This suggests that speakers tend to increase the pitch when they start a new word, starting a quotation or not.

Surprisingly, the F0 reset seemed less pronounced at the start of direct relative to indirect quotations. This can correspond to the often low pitch of the word end preceding the start of the indirect quotation. For example, in the command “Remind me at nine to [watch the tv]”, the word “to” ends with a low pitch. The range of words was very limited preceding indirect quotations, and this word was often present. The low pitch of the previous word end could explain this counter-intuitive observation, regardless the pitch of the next word start.

Positive F0 continuity was observed to be more pronounced for quotations relative to other concepts. This suggests that speakers tend to gradually decrease the pitch at the word end preceding a quotation, and then gradually increase at its start. However, this finding was also observed in a less degree when the speaker ended a word to start another, starting a quotation or not.

F0 continuity seemed less pronounced for direct relative to indirect quotations. This could be explained again by the word end pitch preceding the indirect quotation. The words “to” and “that” were the two most frequently observed words preceding indirect quotations. The word “that” ends with a fricative having no associated F0 value. Thus, the algorithm ignored that part and tried to recover the last F0 values. The end of these

two words was potentially a line of slope close to zero due to ignoring the final fricative, resulting in a bias for the F0 continuity measure.

10. Conclusion

In this paper, the correlation between three prosodic features and the presence or absence of a quotation was studied. This experiment differs from the literature since the quotation’s author and its narrator correspond to the same entity. The features have been studied through a combination of lexical and prosodic extraction tools. The results support the first hypothesis specifying that there is usually a more pronounced pause duration, F0 reset and F0 continuity at the quotation start relative to other concepts. The results support the second hypothesis for the pauses duration on direct and indirect quotations, but the opposite is observed for the F0 reset and F0 continuity. Finally, the lack of available data did not permit a study of the quotation end. The next step is to train a NLU system with these prosodic features to verify whether its performance can be improved for quotation detection.

11. Acknowledgments

This research was jointly funded by FRQNT, CRSNG and Nuance Communications.

12. References

- [1] M. Mast, R. Kompe, S. Harbeck, A. Kiessling, H. Niemann, E. Noth, E. Schukat-Talamazzini, and V. Warnke, “Dialog act classification with the help of prosody,” in *Proceedings of Fourth International Conference on Spoken Language, 1996. ICSLP 96.*, vol. 3, Oct 1996, pp. 1732–1735.
- [2] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, and et al., “Can prosody aid the automatic classification of dialog acts in conversational speech?” 1998.
- [3] A. Stolcke, E. Shriberg, R. Bates, N. Coccaro, D. Jurafsky, R. Martin, M. Meteer, K. Ries, P. Taylor, and C. Ess-Dykema, “Dialog act modeling for conversational speech,” in *Papers from the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, 1998.
- [4] V. Warnke, R. Kompe, H. Niemann, and E. Nth, “Integrated dialog act segmentation and classification using prosodic features and language models,” 1997.
- [5] G. Leech and M. Short, “Style in fiction: A linguistic introduction to english fictional prose,” in *Pearson Longman*, 2007.
- [6] G. Klewitz and E. Couper-Kuhlen, “Quote - unquote? the role of prosody in the contextualization of reported speech sequences,” in *Pragmatics*, 2008, vol. 9, pp. 459–485.
- [7] M. Oliveira and D. Cunha, “Prosody as marker of direct reported speech boundary,” in *Proceedings of Second International Conference on Speech Prosody*, 2004.
- [8] W. Jansen, M. Gregory, and J. Brenier, “Prosodic correlates of directly reported speech: Evidence from conversational speech,” in *Prosody in Speech Recognition and Understanding. Molly Pitcher Inn*, 2001.
- [9] E. Kasimir, “Prosodic correlates of subclausal quotation marks,” in *Papers in phonetics and phonology / Marzena Zygis & Susanne Fuchs (ed.)*, 2008, vol. 49. [Online]. Available: <http://www.zas.gwz-berlin.de/175.html>
- [10] E. Shriberg, A. Stolcke, D. Hakkani-Tr, and G. Tr, “Prosody-based automatic segmentation of speech into sentences and topics,” 2000.
- [11] G. Corder and D. Foreman, *Nonparametric Statistics: A Step-by-Step Approach*. Wiley, 2014.
- [12] R. Thisted, “What is p-value?” 2000.