Insights Into the Importance of Linguistic Textual Features on the Persuasiveness of Public Speaking

Alisa BARKAR
LTCI, Institut Polytechnique de Paris, Telecom Paris
Palaiseau, France
alisa.barkar@telecom-paris.fr

Mathieu Chollet
mathieu.chollet@glasgow.ac.uk
School of Computing Science, University of Glasgow
Glasgow, United Kingdom

Beatrice Biancardi
CESI LINEACT
Nanterre, France
bbiancardi@cesi.fr

Chloé CLAVEL
LTCI, Institut Polytechnique de Paris, Telecom Paris
Palaiseau, France
chloe.clavel@telecom-paris.fr

ABSTRACT

In both professional and private life, there is a growing need for public speaking skills. With this background, our research project's long-term aims are to develop tools that can analyse public speeches and provide useful feedback. The impact of audio and visual characteristics on the automatic analysis of speech quality has been widely explored in the existing literature. However, only a few studies have focused on textual features. In response to this shortcoming, this paper investigates the importance of textual content for the automatic analysis of public speaking. We created an open-source Python library of textual features and integrated them as inputs of simple machine learning models for automatic public-speaking analysis, and persuasiveness prediction, in particular. The best result (accuracy of 61%) is obtained using a logistic regression. We then evaluated the impact of these features on persuasiveness prediction using both correlation analysis and Explainable AI methods. This evaluation was conducted on the French data set 3MT_French, including student performances in the "Ma Thèse en 180 Secondes" competition.

CCS CONCEPTS

• **Computing methodologies** → *Classification and regression trees*; *Modeling and simulation.*

KEYWORDS

Public speaking, multimodal system, deep learning, explainable artificial intelligence, behavioural models

1 INTRODUCTION

Various public speaking training systems have been recently developed which leverage modern artificial intelligence techniques to provide training feedback to users, such as Poised (AI-Powered Communication Coach)[1] and Speaker Coach [Microsoft 2022]. Some of these systems are now even integrated into software such as Zoom, Teams, or PowerPoint. These systems are based on deep learning models (as described in [Microsoft 2022]) and can provide automatic evaluations of speaking performance using performance-related *dimensions* such as the speaker's level of confidence, the clarity of speech, or the positive or negative perception of the performance. A wide variety of public speaking performance dimensions (to name but a few: level of insecurity, hesitancy, monotony, persuasiveness or self-confidence) have also been studied in the academic literature [Strangert and Gustafson 2008], [Scherer et al. 2012] and [Chen et al. 2015]. Each dimension is typically annotated by humans on a 5- or 7-point Likert scale and correlations between multimodal features and some of these dimensions are then studied in various corpora of public speaking. [Nguyen et al. 2012] used audio features (*e.g.*, F0 statistics, average pause time) and visual features (*e.g.*, motion energy, facial expression, Kinect skeletal data [Zhang 2012]) and investigated the correlation between them and audience final ratings for a dataset of student presentations. In [Dinkar et al. 2020b], it is the paralinguistic content that is studied. They examined the impact of filler words ("umm" or "uhh") on listeners' perceptions of speaker confidence in film reviews recorded in English.

Literature related to public speaking performance offers standard feature sets for audio (*e.g.*, GeMAPS set of parameters related to voice frequency/energy/amplitude/spectrum [Eyben et al. 2016]) or visual cues (skeletal data, action units (AU), facial expressions). Comparatively, some studies incorporate verbal features in the analysis of speaking performance, although those are still relatively scarce. For example, [Park et al. 2014] incorporated verbal features (*e.g.*, uni-grams, bi-grams) in addition to visual, audio, and paraverbal cues (*e.g.*, articulation rate, fillers) and investigates the ability of features to predict, – using a Support Vector Machine (SVM) – the level of persuasiveness in recorded film reviews in English (rated by human raters on a 7-point Likert scale). Persuasiveness was also found to be important in the related domain of evaluation of argumentative essays [Carlile et al. 2018]. [Chen et al. 2015] took into account textual features such as the presence and number of subjective/neutral words from the MPQA subjectivity lexicon [Wilson et al. 2005] and the ASSESS sentiment lexicon [Klebanov et al. 2013], as well as the pointwise mutual information (PMI) in bi/tri-grams. They found a significant correlation between lexical features and overall performance ratings by human raters on English presentations. [Yang et al. 2020] also found that emotional features (using LIWC lexicon [Pennebaker 2022]) are important for the perception of the speakers' charisma in clips from prepared talks, educational lectures, and interviews.

[1] https://www.poised.com/about-us

The current state of research on the analysis of oratory performance reveals certain limitations: *i)* existing work does not offer a standardised set of textual features, as has been done in the case of audio or visual modalities; *ii)* existing automatic systems (i.e. deep-learning-based systems) do not focus on the explanation of the predicted scores of performance; *iii)* textual features analysis is dominated by English studies , consequently French, have comparatively been rarely studied in the context of public speaking analysis. The work presented in this paper takes a first step towards addressing these three limitations by providing a python library of textual features for automatic analysis of public speaking[2], and by conducting a comprehensive analysis of the textual features that matter in prediction in a French public dataset. Our main focus is on the explanation of the model performance and we addressed the following research questions: **Q1.** Can the quality of a public speaking performance be assessed reliably on the sole basis of textual features (more specifically, *form* features)? **Q2.** Which textual elements could be used to explain a listener's perception of the speech?

Public speaking is multi-faceted and can be evaluated from a number of perspectives; we have chosen to start with the *dimension* of *persuasiveness* (rather than, e.g., *self-confidence*, *engagement*), as we assume that this dimension is one for which textual content of the presentation plays an important role (this assumption is supported by [Carlile et al. 2018]). We focused on textual features related to *form* of the presentation (i.e. language level, vocabulary, negative/positive words, etc.) to make our study independent of the topic addressed in the speech. We thus did not consider textual features that capture information about the content of the performance (*e.g.*, word count features such as uni-/bi- grams in [Park et al. 2014]). To construct the feature set characterising a speech's *form*, we drew on features that have been shown to be related to language level in the automatic evaluation of essays [Vajjala 2016]. We also proposed new features in order to characterise the vocabulary richness, the level of fluidity of the discourse as well as the use of words in relation to affective, cognitive, and perception processes. Since we were interested in understanding which features had the most impact on the prediction of the persuasiveness level, we tested several standard classifiers and performed feature importance analysis. For that we calculated Spearman's correlation and, additionally, to universally compare an impact of features in several models, we used Shapley values [Lundberg and Lee 2017] instead of coefficients of the classification models. Our research distinguishes itself from conventional text classification tasks as we examined domains evaluated based on all modalities, not solely textual content. Consequently, our results also reveal the proportion of information contributed by text in performance assessment. Lastly, our work represents the initial steps in studying French public speech, acknowledging the significant influence of cultural differences on feature prediction.

2 DATA

We leveraged the 3MT_French dataset [Biancardi et al. 2022], which consists of annotated 3-minute video recordings of presentations in the French scientific public speaking competition "Ma Thèse en 180 seconds". This dataset includes presentations from both female (135 speakers) and male (113 speakers) participants, covering diverse topics and showcasing the thesis works of French PhD students. Among the evaluated dimensions related to performance , our focus was on assessing the persuasiveness level based on the full video. To evaluate persuasiveness level raters were asked to rate the speakers' ability to construct a convincing message with solid reasoning using a 5-point Likert scale. Each video was evaluated by three different viewers through the Amazon Mechanical Turk [Mason and Suri 2011] crowd-sourcing platform. To mitigate the impact of low inter-rater agreement (measured by intraclass correlation coefficient (ICC) [Bartko 1966]), we applied the root mean square (RMS) to the ratings provided by the three annotators as the final scoring method, similar to the approach followed in [Dinkar et al. 2020a] (for more details, refer to the paper [Biancardi et al. 2022]). In order to analyse the speech transcripts, we processed the data set using a speech transcription library [3]. However, it is important to note that this automatic transcription lacks punctuation, stuttering, or pause fillers. Additionally, such systems have a tendency to improve the text by correcting incorrect grammar constructions, which may affect our analysis. Finally, we categorised performances into two classes w.r.t. the calculated median of persuasiveness score. Data points with human-evaluated scores equal to or higher than the median were classified as "high-quality," while those with scores lower than the median were classified as "low-quality." This approach allowed us to create balanced data set, however, it causes a borderline effect when some of the performances with scores close to the median are barely distinguishable. This can potentially cause lower accuracy scores in the prediction.

3 METHODOLOGY

We present the schema of our experimental pipeline in Figure 1 with three main steps: *i.* extract set of features $\{\bar{x_i}\}$; *ii.* build the model to predict scores $\{y_i\}$ using $\{\bar{x_i}\}$; *iii.* analyse the impact of various features from $\{\bar{x_i}\}$ on the result of prediction $\{\hat{y_i}\}$ by calculating Spearman's correlation and SHAP values ($\overline{sv_i}$). Each step is then detailed in the following subsections.

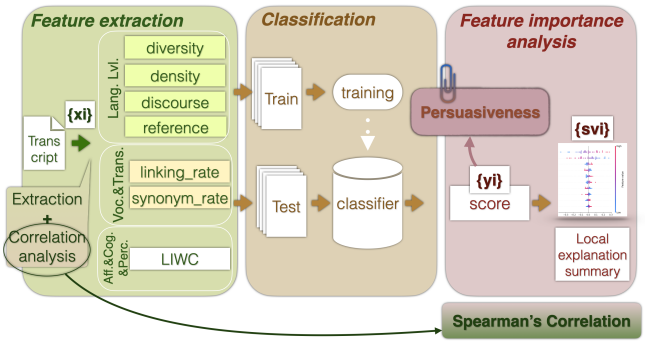

Figure 1: Experimental pipeline in three stages

[2]https://github.com/anonympapers/textual_features_importance.git

[3]https://pypi.org/project/youtube-transcript-api/

3.1 Feature Extraction.

We extracted a set of 78 features that are presented below. Exact formulas for all extracted features and all code implemented in Python can be found on anonymous GitHub[4]. We used spacy [Honnibal and Montani 2017] to provide Part-of-speech (POS) for words and French tagger [Labrak and Dufour 2022] for finer tags in French.

3.1.1 Characterizing Language level. The proposed library provides *diversity, density, discourse, referential* features initially designed to evaluate the language level of essays and taken from [Vajjala 2016]. The **5 *diversity* features** are word-level features measuring lexical diversity based on the type-token ratio (*TTR*) (the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language) and its different variants measuring both global diversity (*CorrectedTTR*, *RootTTR*, *BilogTTR*) and local diversity (*MTLD* measures the average length of continuous text sequence that maintains the *TTR* above a threshold of 0.72 (we used the threshold number from [Vajjala 2016])). The **21 *density* features** consisted of three main categories : i) **9** lexical variation features (mainly ratios of the number of different POS to the total number of adjectives, nouns, verbs, and adverbs). For example, *POS_squaredVerbVar1* which equals to $(nb\_Verbs)^2$ divided by *nb_types_Verbs*; ii) **11** general POS/tag features; and iii) **1** verb tag feature (ratio of the number of different POS to the total number of words). Those features represent the distribution of different POS in the transcript. To measure the overlapping in-between sentences we used **8 *discourse* features** (the ratio of the number of appearances of the same words in (subsequent) sentences to the total number of sentences). For example, *globalContentWordOverlap* is the number of the same words within any two sentences. As sentences are difficult to segment in public speaking transcripts, we chose to consider sequences of 10 words instead of a sentence unit. To measure reference level we calculated **8 *referential* features** calculated as the ratio between the number of pronouns/personal pronouns/determiners to the total number of sentences/nouns/words, for example, *DISC_RefExprPronounsPerNoun* which is the ratio of *nb_Pronouns* to the *nb_Nouns*.

3.1.2 Characterizing vocabulary richness and transitions within speech. To evaluate vocabulary richness and transitions within speech we proposed new features. First, we focused on conjunctions and added **6 *linking_rate* features** which represent the diversity of transitions between different parts of public speech. Those features are calculated as the ratio between the number of linking words/types of linking words to the total number of words/sentences, for example, *conjunctToSent* equal to the ratio of *conjunctNum* to the *nb_Sent*, where *conjunctNum* – the total number of conjunctions within the transcript. Also we measured the ratio of cases when two subsequent sentences started with the same conjunction. We used **4 *synonym_rate* features**. We divided nouns and verbs into groups of synonyms (for that we used [Bird et al. 2009] toolkit) then we calculated the ratio of the number of those groups of synonyms to the total number of nouns/verbs. For example, *synonymToNouns* is a ratio of *nb_GroupsOfSynomyns* to the *nb_Nouns*. Additionally, we measured the average size of those synonym classes for nouns and verbs.

3.1.3 Characterizing language in relation to the affective, cognitive and perception processes. In the proposed feature set, we integrated features characterizing the use of words in relation to affective, cognitive and perception processes using [Pennebaker 2022] software tool (with the French dictionary [Piolat et al. 2011]). We used **5 *LIWC* features** where each feature represents the percentage of words corresponding to a given LIWC category: positive, negative, anxious, angry, sad (for affective processes); cognition, insight, cause, divergence, tentative, certainty, inhibition, inclusion, exclusion (for cognitive processes) and perceptive processes.

3.2 Classification

We tested several classification models such as Support Vector Machine (SVM), Random Forest classification (RFC), and Logistic Regression (LR)[5]. For each pair of (dimension; model), we found the best set of parameters[6] by using the Grid Search method with the split onto 80%/20% train/test data. To assess the quality of prediction, we used a leave-one-out (LVO) method and calculated the average accuracy score (AAcc). AAcc is calculated as the ratio of the total number of truly predicted negative or positive values to the total number of predictions. This accuracy metric fitted the evaluation because we obtained balanced data set with a separation of classes.

3.3 Feature importance analysis

First, we applied the correlation analysis used in the literature (e.g. [Strangert and Gustafson 2008], [Scherer et al. 2012]). We calculated Spearman's correlation between each feature from $\{\bar{x_i}\}$ and the class of the performance $\{y_i\}$. To analyse the importance of different categories, we calculated the average absolute value of Spearman's correlation of features within each category. To evaluate the role of each feature, we calculated Shapley values which by definition are a difference between performance with and without considered features. In our case, we used an approximation called SHAP (SHapley Additive exPlanations) described by [Lundberg and Lee 2017]. This mathematical tool allows us to numerically evaluate players' contributions (features in our case) to achieve their common goal (model prediction). We used the method called Kernel SHAP from the SHAP library. It calculates the Shapley value approximation by replacing feature values with values from the background data set. For each feature for each example from the test data set: $\forall i : \bar{x_i} \longmapsto \overline{sv_i}$, where $\overline{sv_i}$ – is a vector of SHAP values $\overline{sv_i}_j \forall j \in \bar{x_i}$, we plot obtained SHAP values and call this a *local explanation summary*. Each row of the local explanation summary corresponds to the analysed feature. The position of each point on the row represents a SHAP value of corresponding features in the sample from the test data set (positive values indicate the positive impact of the feature on the model prediction). The point's colour indicates the feature's value (big values are in red and small in blue).

[4] https://github.com/anonympapers/textual_features_importance.git

[5] We additionally tested Naive Bayes (NB) and K Nearest Neighbours (KNN). The Table with results is in the Appendix

[6] List of optimal parameters can be also found in the Appendix

4 RESULTS

Classification. In Table 1 one can find the AAcc of the models with the best parameters[7]. We obtained significant out-performance of LR and SVM on the Persuasiveness[8]. Even though the best prediction score (AAcc score of 61%) is not extremely high, it is comparable with state-of-art results (in [Park et al. 2014], an accuracy of 66.29% was obtained with SVM when using verbal and para-linguistic features). It is important to note that in 3MT_French corpus, speakers have been trained, resulting in presentations that are generally of good quality. It is therefore difficult to distinguish between low and high persuasiveness, as the boundary between the two classes is relative.

Feature importance analysis. With Spearman's correlation analysis we observed that features of *linking_rate* (e.g. *conjunctToSent*, *conjunctToWords*), *LIWC* (e.g. *anx*, *anger*), *synonym_rate* (e.g. *synonymToNouns* , *synonymToVerbs*), *density* (e.g. *POS_numNouns*, *POS_nounVar*) categories had the biggest correlation. We obtained the biggest average correlation (around 0.13) for *synonym_rate* and *linking_rate* categories. As a comparison, [Yang et al. 2020] obtained the absolute correlation values of lexical features (i.e. LIWC features) with charisma ratings ranged from 0.27 to 0.40 when considering all speakers. [Larrimore et al. 2011] obtained the biggest correlation 0.116 for Word Count features from LIWC categories. In [Chen et al. 2015] lexical features had Pearson's correlation with holistic ratings around 0.3.

Alternatively, we present obtained local explanation summary of SHAP analysis for the LR model prediction in the Figure 2. Interestingly, we obtained results have slight difference with correlation analysis. For example, while *synonym_rate* and *linking_rate* categories had the biggest Spearman's correlation scores, only *conjunctTypesToTotal* feature had relatively significant SHAP value within all features presenting those categories. On the other hand, similarly to the Spearman's correlation analysis, with SHAP analysis we observed *density* (e.g. *POS_correctedVV1*, *POS_squaredVerbVar1*), *discourse* (e.g. *globalStemOverlapCount*, *localContentWordOverlap*) and *LIWC* (e.g. *negemo*) categories to be the most important for the persuasiveness of speech. We observed that high SHAP values (red) of *POS_correctedVV1* or *conjunctTypesToTotal* impact positively (positive SHAP values on the x-axis) the prediction of the model. This means that by taking into account high values of these features, the model tends to classify performance as more successful. Similarly, we observe high SHAP values for low (blue) values of the *negemo* or *globalStemOverlapCount* feature. This means that a low percentage of negative words or global overlapping within the speech leads to a higher persuasiveness prediction.

5 CONCLUSIONS AND PERSPECTIVE

Q1 - Predictability of the perception of persuasiveness on the basis of textual content. The proposed feature set that widely describes the form of the discourse leads to encouraging prediction scores. Best results were obtained by LR and SVM models.

[7]We obtained AAcc for SVM, RFC and LR with and without feature selection and with different accuracy functions in the Grid Search for the parameters, those results could be found in the Appendix

[8]We also obtained AAcc for other dimensions available in the data set. Results are presented in the Appendix

Table 1: Average accuracy scores (AAcc) and their confidence intervals (in brackets) for LVO for different classification models.

	SVM	RFC	LR
Pers.	0.57(0.50; 0.64)	0.54(0.48; 0.62)	0.61(0.54; 0.67)

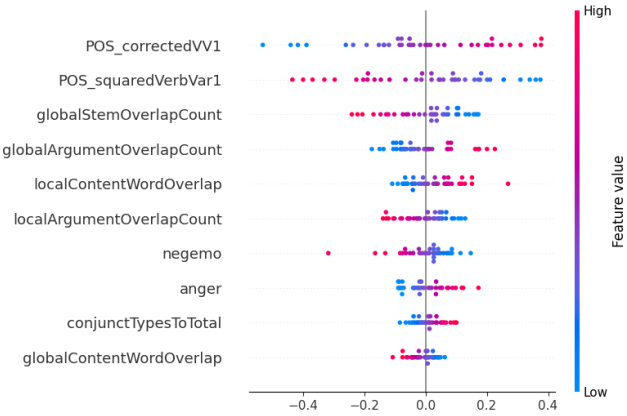

Figure 2: Local explanation summary for LR (without FS) predicting classes in persuasiveness dimension.

Q2 - Most important textual features. We observed that features of discourse, density and LIWC categories have the biggest impact on the model decisions. We also observed that negative emotions impacts negatively on the persuasiveness of speech which is consistent with findings of [Yang et al. 2020]. Additionally, we obtained new findings on the negative impact of global stem overlapping and positive impact of the higher variation of verbs.

Future work will be dedicated to studying the interplay between textual content and other modalities (*e.g.*, prosody, gestures, facial expressions) and its impact on the perception of persuasiveness. We also plan to integrate paralinguistic features such as fillers. This study is the first step in our project to develop a pedagogical system for public speaking training. Our understanding of the importance of features will help us in the future to provide reliable feedback and to develop a series of exercises for public speaking training.

ACKNOWLEDGMENTS

This research was partly funded under the ANR REVITALISE grant ANR-21-CE33-0016-02.

REFERENCES

John J. Bartko. 1966. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports* 19 (1966), 11 – 3.

Beatrice Biancardi, Mathieu Chollet, and Chloé Clavel. 04 October 2022. *Introducing the 3MT_French Dataset to Investigate the Timing of Public Speaking Judgements.*

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In *Annual Meeting of the Association for Computational Linguistics.*

Lei Chen, Chee Wee Leong, Gary Feng, Chong Min Lee, and Swapna Somasundaran. 2015. Utilizing multimodal cues to automatically evaluate public speaking performance. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (2015), 394–400.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020a. The importance of fillers for text representations of speech transcripts. arXiv:2009.11340 [cs.CL]

Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, and Chloé Clavel. 2020b. How confident are you? Exploring the role of fillers in the automatic prediction of a speaker's confidence. (05 2020), 8104–8108. https://doi.org/10.1109/ICASSP40776.2020.9054374

Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Phuong Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE transactions on affective computing* 7, 2 (April 2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417 Open access.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

Beata Beigman Klebanov, Jill Burstein, and Nitin Madnani. 2013. Sentiment Profiles of Multiword Expressions in Test-Taker Essays: The Case of Noun-Noun Compounds. *ACM Trans. Speech Lang. Process.* 10, 3, Article 12 (jul 2013), 15 pages. https://doi.org/10.1145/2483969.2483974

Yanis Labrak and Richard Dufour. 2022. ANTILLES: An Open French Linguistically Enriched Part-of-Speech Corpus. (Sept. 2022). https://hal.archives-ouvertes.fr/hal-03696042

Laura Larrimore, Li Jiang, Jeff Larrimore, David M. Markowitz, and Scott Gorski. 2011. Peer to Peer Lending: The Relationship Between Language Features, Trustworthiness, and Persuasion Success. *Journal of Applied Communication Research* 39 (2011), 19 – 37.

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. (2017). https://doi.org/10.48550/ARXIV.1705.07874

Winter Mason and Siddharth Suri. 2011. A Guide to Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior research methods* 44 (06 2011), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Microsoft. 2022. Rehearse your slide show with speaker coach. (2022). https://support.microsoft.com/en-gb/office/rehearse-your-slide-show-with-speaker-coach-cd7fc941-5c3b-498c-a225-83ef3f64f07b [Online; accessed 01-March-2023].

Anh-Tuan Nguyen, Wei Chen, and G.W.M. Rauterberg. 2012. Online feedback system for public speakers. *2012 IEEE Symposium on E-Learning, E-Management and E-Services* (2012), 1–5.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. (2014), 50–57. https://doi.org/10.1145/2663204.2663260

Boyd Ryan L. Booth Roger J. Ashokkumar Ashwini Francis Martha E. Pennebaker, James W. 2022. Linguistic Inquiry and Word Count: LIWC-22. Pennebaker Conglomerates. (2022). https://www.liwc.app

Annie Piolat, R. Booth, Cindy Chung, Morgana Davids, and James Pennebaker. 2011. La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation. *Psychologie Francaise - PSYCHOL FR* 56 (09 2011), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002

Stefan Scherer, Georg Layher, John Kane, Heiko Neumann, and Nick Campbell. 2012. An audiovisual political speech analysis incorporating eye-tracking and perception data. (2012).

Eva Strangert and Joakim Gustafson. 2008. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. (2008).

Sowmya Vajjala. 2016. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *CoRR* abs/1612.00729 (2016). arXiv:1612.00729 http://arxiv.org/abs/1612.00729

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. (Oct. 2005), 347–354. https://aclanthology.org/H05-1044

Zixiaofan Yang, Jessica Huynh, Riku Tabata, Nishmar Cestero, Tomer Aharoni, and Julia Hirschberg. 2020. What Makes a Speaker Charismatic? Producing and Perceiving Charismatic Speech. *Speech Prosody 2020* (2020).

Zhengyou Zhang. 2012. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia* 19, 2 (2012), 4–10. https://doi.org/10.1109/MMUL.2012.24

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009