ANALYZING THE INTERACTION BETWEEN THE READER'S VOICE AND THE LINGUISTIC STRUCTURE OF THE TEXT: A PRELIMINARY STUDY

Benedetta Iavarone[1,2], Maria Sole Morelli[3], Dominique Brunato[2], Shadi Ghiasi[4], Enzo Pasquale Scilingo[4], Nicola Vanello[4], Felice Dell'Orletta[2], Alberto Greco[4]

[1] Scuola Normale Superiore, Pisa, Italy; [2] Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy;
[3] Fondazione Toscana Gabriele Monasterio, Pisa, Italy; [4] Dipartimento di Ingegneria dell'Informazione,
Research Center "E. Piaggio", University of Pisa, Pisa, Italy

benedetta.iavarone@sns.it, msmorelli@monasterio.it, dominique.brunato@ilc.cnr.it, shadi.ghiasi@centropiaggio.unipi.it, enzo.scilingo@unipi.it, nicola.vanello@unipi.it, felice.dellorletta@ilc.cnr.it, alberto.greco@unipi.it

Abstract: **In this study, we present a preliminary analysis of the relationship between the linguistic profile of a text and the voice properties of the reader aiming to improve the speech-based emotion recognition systems. To this aim, we recorded the speech signals from a group of 32 healthy volunteers reading aloud neutral and affective texts and used the BioVoice toolbox to compute some of the main speech features. The selected texts were analyzed to quantify their lexical, morpho-syntactic, and syntactic content. Correlation and Support Vector Regressor analyses between linguistic and speech features have shown a significant modulation of some voice acoustic properties performed by the linguistic structure of the text. Particularly, a significant effect was shown on some specific speech features often used for the assessment of human emotional state (e.g., F0). This suggests that the lexical, morpho-syntactic, and syntactic properties could play an important role in the emotional dynamics of a person.**

Keywords: **Speech analysis, linguistic profile, emotions, Support Vector Regressor**

I. INTRODUCTION

Human speech is the result of fine control of up to eighty muscles from respiratory, laryngeal, pharyngeal, palatal, and orofacial groups [1]. Such control is a complex process that involves both somatic and autonomic nervous systems (ANS) activity. This latter is the main responsible for the regulation of bodily functions and is the primary mechanism of emotional regulation [2]. Alterations in the respiratory activity induced by the ANS manifest changes in the emotional state of the speaker by influencing the voice spectrum characteristics such as the fundamental frequency (F0 - the frequency of vibration of the vocal folds), and its formants (F1, F2, F3 - resonance frequencies of the vocal tract) [3]. Hence, speech processing represents one of the most promising tools in the affective computing field for a non-invasive assessment of the speaker's emotional state [4]. Indeed, voice signal analysis has been successfully used to explore several psychological dimensions of the speaker: emotion [5], mood [6], stress [7, 8], and personality [9] have been widely studied. To effectively characterize the affective prosody, several previous studies have developed and applied analytic methods to measure changes in pitch, loudness, speech rate, and pause [10]. However, the use of these features to infer the emotional state of a speaker remains an extremely complex task. One important and still little studied source of complexity could be the interaction between the speaker's hidden emotional state and the linguistic and semantic properties of what the speaker is saying. The combination of such linguistic and speech information in computational models could improve the accuracy of inferring the speaker's emotional state. Indeed, a text is characterized by many levels of information (linguistic, lexical, stylistic). By annotating these levels, it is possible to extract many features modeling the lexical, grammatical, and semantic phenomena to construct a linguistic profile that characterizes language variations within and across texts [11]. The linguistic profile has been used for different applications, such as registry and genre variation [12], or the study of psycholinguistic phenomena. In [13], the authors have shown that linguistic features can be effectively used to predict the human perception of sentence complexity, intended as processing difficulty of the language. Linguistic aspects and their effect on human processing effort and perception of complexity were studied also in [14], where the authors demonstrate that linguistic aspects from context play an important role in the perception of complexity and cognitive processing effort. Recently, Singh et al. [15] have proposed a deep learning hierarchical model for emotion recognition, combining text analysis computed by ELMo v2 with prosody, voice quality, and spectral features. However, formal modeling of the relationship between prosodic and linguistic features has not been investigated yet. In this preliminary study, we aim at studying whether the acoustic features, commonly used to characterize

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

speech production prosody, are significantly influenced by the linguistic structure of the pronounced text. To this aim, we analyzed speech signals and linguistic profiles of texts with different levels of arousal and valence. We apply correlation and regression methods to understand how the linguistic profile and structure of the texts interact with the speech production of the same texts.

II. METHODS

A group of 33 healthy volunteers was enrolled in the study (17 females), aged between 26.6 and 30.0. None of them suffered from heart diseases, mental disorders, or phobias. Each participant gave their written informed consent, and the study was approved by the Ethical Committee of the University of Pisa. We selected four texts, two describing different medieval tortures and two describing text types and writing styles. Based on the topics covered, two texts were classified as high arousal and negative valence, whereas the other two were neutral. Moreover, before starting the experiment, a group of 10 subjects, other than those enrolled in this study, evaluated the texts in terms of arousal and value, confirming the arousal and valence levels supposed apriori based on the reading topic. Each participant was asked to read aloud one neutral and one affective text, randomly chosen [16]. All texts have similar lengths to make the duration of the reading similar among subjects. The speech signal and other physiological signals such as electrocardiogram and electrodermal activity (not considered in this study) were recorded during the reading task.

A. Linguistic analysis

The texts were divided into sentences, using the full stop as a splitting criterion, i.e., identifying a sentence as the part of text between two full stops. After the splitting, neutral texts contained a total of 25 sentences, with an average sentence length of 28 tokens; affective texts contained a total of 40 sentences, with an average sentence length of 21 tokens. Each sentence was analyzed from a linguistic point of view and represented as a vector of ~140 features, a subset of the ones described in [11] that model a wide range of properties extracted from different levels of linguistic annotation. The features capture on one hand complex information like the syntactic phenomena (subordination, structure, and length of dependency relations, structure of the verbal predicates) or morpho-syntactic phenomena (distribution of grammatical categories across the text, aspects about the verb conjugation), on the other hand, they capture raw properties, like the length of the text and its components (sentences and words). The features can be grouped based on the linguistic aspects they describe and are further discussed below.

(1) Raw Text Properties. Features on the length of the text and of the sentences and the words that are in it; **(2) Lexical Variety**. Features on how varied the vocabulary of a text is, determined as the percentage of diverse and nonrepeated words over the total number of words; **(3) Morpho-syntactic information**. Features on: *(i)* the distribution in the text of grammatical categories (e.g., adjectives, nouns, determiners, pronouns); *(ii)* the ratio of content words (nouns, verbs, adjectives, and adverbs) over the total number of words in a text; *(iii)* the inflectional morphology, i.e., the distribution, for verbs and auxiliaries, of a set of inflectional features (e.g., mood, tense); **(4) Verbal Predicate Structure**. Features on: *(i)* the distribution of verbal heads, i.e., the average number of propositions (main or subordinate) co-occurring in a sentence; *(ii)* the distribution of verbal roots, i.e., the percentage of verbal roots out of the total of sentence roots; *(iii)* verb arity, i.e., the average number of instantiated dependency links sharing the same verbal head; **(5) Global and local parsed tree structures**. Features on: *(i)* the average depth of the syntactic tree, i.e., the average of the longest dependency link in a sentence. *(ii)* the average number of tokens per clause, where the number of clauses is the ratio between the number of tokens in a sentence and the number of verbal or copular heads; *(iii)* length of dependency links, i.e., the number of words occurring between the syntactic head and its dependent; *(iv)* the average depth of complement chains (a list of consecutive complements); *(v)* the order of the subject and the object in a sentence; **(6) Syntactic relations**. Features on the percentage distribution of 37 universal dependency relations; **(7) Subordination phenomena**. Features on: *(i)* the distribution of main clauses vs. subordinate clauses; *(ii)* the distribution of subordinates in post-verbal and preverbal position; *(iii)* the average number of subordinates recursively embedded in the top subordinate clause.

B. Speech signal processing

To analyze the speech time series and extract from each sentence acoustic parameters, we used the BioVoice toolbox [17]. The toolbox detected first only voiced parts of each segment. Then, F0, F1, F2, and F3 were calculated. In each voiced frame, F0 is estimated with a two-step procedure: first, Simple Inverse Filter Tracking (SIFT) was applied to signal time windows of fixed length related to the F0 range; secondly, F0 is adaptively estimated on signal frames of variable length inversely proportional to F0, through the Average Magnitude Difference Function (AMDF) within the range provided by the SIFT [18]. To extract formants values, Autoregressive Power Spectral Density (AR PSD) was considered. Furthermore, in each sentence, the total time duration of reading, the overall voiced duration, and the average voice duration were extracted.

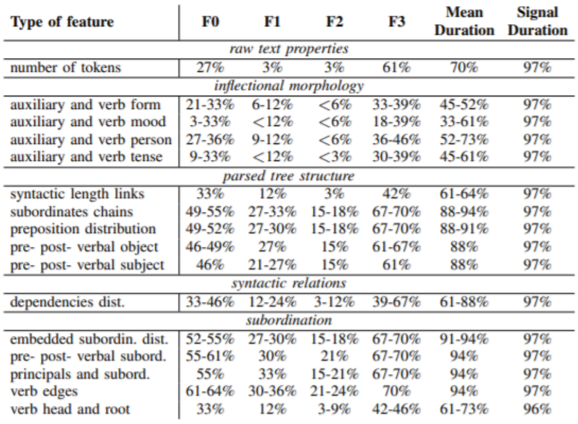
C. Statistical analysis and modeling of the features

Before running the analyses, we scaled the frequency features of the voice in each sentence, as they are subject-dependent. For each subject and each frequency feature (F0, F1, F2, and F3), we computed F_i^{scaled} , as $F_i^{scaled} = F_i / \overline{Fi}_{neu}$ where F_i represents the frequency feature of interest (in neutral or emotional test in each sentence) and $\overline{Fi}_{neu}$ the mean of the frequency of the corresponding neutral texts, computed for all time duration. As a first analysis, we examined the relationship between linguistic features and speech features. In this way we could understand which linguistic aspects of the text are most related to speech production, discovering the underlying interaction between linguistic structure and speech. To do so, we correlated each linguistic feature with every speech one, using Spearman's correlation coefficient. We selected all pairwise correlations that had a correlation coefficient different from zero and a p-value < 0.05. Afterward, we tested the predictive strength of the linguistic profile. We implemented a regression model to predict acoustic parameters, using as input to the model the linguistic features. We employed a Support Vector Regressor (SVR) implemented with a Radial Basis Function (RBF) kernel and standard parameters. To account for within-subject repetitions, we used leave-one-out cross-validation, training the model on all subjects minus one, and testing on the left-out subject. The baseline was calculated by running the model with only the length of sentences as input feature.

III. RESULTS

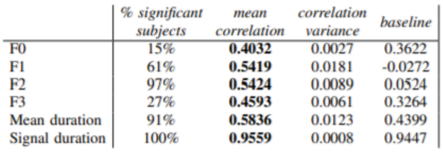
Table I shows a summarized representation of the correlation results between speech frequency features and linguistic features. Linguistic features are grouped according to their function and the linguistic aspect they describe. We report the percentage of subjects for which the features in the group were significantly correlated with acoustic features; when two percentages are presented, they indicate the minimum and the maximum number of subjects for which the different linguistic features of the group were significant. Overall, linguistic features within the same group were significant for a similar or the same number of subjects. As expected, acoustic features that reflect the length of the sentences (Mean and Signal Duration) were always correlated with linguistic features that encode aspects of sentence length, for most subjects. We found significant correlations for a high number of subjects for F0 and F3 and many linguistic aspects, while F1 and F2 were the least correlated with linguistic features. The highest correlations were found with features regarding subordination phenomena and the structure of the parsed tree, especially for F3, with up to 70% of subjects showing a significant correlation. Most linguistic features that show significant correlations are related to different aspects of language complexity, such as the length of sentences, syntactic structures (e.g., longer dependency links), or the verbal morphology (e.g., a past verbal tense may be perceived as more complex than the present tense). In Table II, we report the results for the prediction of the acoustic features using the SVR model.

TABLE I
CORRELATION SUMMARY RESULTS



Type of feature	F0	F1	F2	F3	Mean Duration	Signal Duration
raw text properties						
number of tokens	27%	3%	3%	61%	70%	97%
inflectional morphology						
auxiliary and verb form	21-33%	6-12%	<6%	33-39%	45-52%	97%
auxiliary and verb mood	3-33%	<12%	<6%	18-39%	33-61%	97%
auxiliary and verb person	27-36%	9-12%	<6%	36-46%	52-73%	97%
auxiliary and verb tense	9-33%	<12%	<3%	30-39%	45-61%	97%
parsed tree structure						
syntactic length links	33%	12%	3%	42%	61-64%	97%
subordinates chains	49-55%	27-33%	15-18%	67-70%	88-94%	97%
preposition distribution	49-52%	27-30%	15-18%	67-70%	88-91%	97%
pre- post- verbal object	46-49%	27%	15%	61-67%	88%	97%
pre- post- verbal subject	46%	21-27%	15%	61%	88%	97%
syntactic relations						
dependencies dist.	33-46%	12-24%	3-12%	39-67%	61-88%	97%
subordination						
embedded subordin. dist.	52-55%	27-30%	15-18%	67-70%	91-94%	97%
pre- post- verbal subord.	55-61%	30%	21%	67-70%	94%	97%
principals and subord.	55%	33%	15-21%	67-70%	94%	97%
verb edges	61-64%	30-36%	21-24%	70%	94%	97%
verb head and root	33%	12%	3-9%	42-46%	61-73%	96%

TABLE II
REGRESSION RESULTS FOR THE PREDICTION OF LINGUISTIC FEATURES



	% significant subjects	*mean correlation*	*correlation variance*	*baseline*
F0	15%	**0.4032**	0.0027	0.3622
F1	61%	**0.5419**	0.0181	-0.0272
F2	97%	**0.5424**	0.0089	0.0524
F3	27%	**0.4593**	0.0061	0.3264
Mean duration	91%	**0.5836**	0.0123	0.4399
Signal duration	100%	**0.9559**	0.0008	0.9447

To evaluate the goodness of the model, we correlated the model's predictions with the actual values of the features that we predicted, calculating the mean Spearman's correlation and its variance over all subjects. Percentages show the number of subjects for which the predictions were significantly correlated. Our predicting model always performed better than the baseline. The robustness of the model is confirmed by the low variance, indicating that the acoustic values predicted are consistent among the different subjects. The prediction of mean and signal duration was significant for almost every subject. This was expected, as these features are directly linked to the length of the sentences, a feature that the model could see in input. The predictions of F1 and F2 were significant for many subjects (>60%). Contrary to what was seen previously in the correlation analysis, where F0 and F3 were obtained significant results for a high number of subjects, when predicting them with the SVR their predictions are significant for a low number of subjects.

IV. Discussion and conclusion

In this preliminary study, we combine the analysis of the linguistic profile of neutral and emotional texts with the speech analysis of the reader. We assumed that the speech signal reflected the emotional state induced by the task and assessed by the SAM. Correlation and regression methods were used to understand how the linguistic profile and structure of the texts interact with speech production. We found a statistically significant relationship between some of the linguistic properties of the text, regarding their syntactic structure, subordination phenomena within the texts and the verbal predicate structure, and the speech features that describes some prosodic aspects of speech often related to the human emotional state (e.g., F0, F3). This could suggest a double possible interpretation: on the one hand, it could suggest that the linguistic structure of the pronounced sentence may be a confounding factor that masks the actual contribution of prosodic features in the estimation of the emotional state. On the other hand, the linguistic structure itself could have a direct influence on the emotional state of the subject. This last hypothesis has already been supported by some studies that have combined the features derived from voice processing with some linguistic features to feed classifiers for the recognition of the emotional state [15, 19]. However, in these studies, the encoding of the text considers the lexical and contextual aspects of language but does not consider other important features considered in our study such as morpho-syntactic or syntactic ones. Indeed, these features could have a strong impact on the emotional state of an individual, because they are related to a variety of psycholinguistic phenomena and could affect the cognitive load and processing difficulty of the language user. Future studies will investigate the selected linguistic features to estimate their actual effect on emotional state prediction. Moreover, we will consider other physiological parameters such as electrocardiogram and electrodermal activity recorded during the reading task to evaluate their correlation with voice and linguistic parameters in affective reading.

REFERENCES

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete time processing of speech signals, 2000.

[2] A. L. Callara, L. Sebastiani, N. Vanello, E. P. Scilingo, and A. Greco, “Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heartrate variability,” IEEE Transactions on Biomedical Engineering, 2021.

[3] Z. Zhang, “Mechanics of human voice production and control.” The Journal of the Acoustical Society of America, vol. 140, no. 4, p. 2614, 2016.

[4] C. S. Hopkins, R. J. Ratley, D. S. Benincasa, and J. J. Grieco, “Evaluation of voice stress analysis technology,” in Proceedings of the 38th annual Hawaii international conference on system sciences. IEEE, 2005, pp. 20b– 20b.

[5] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012.

[6] N. Cummins, et al, “A review of depression and suicide risk assessment using speech analysis,” Speech Communication, vol. 71, pp. 10–49, 2015.

[7] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, “Vocal indices of stress: a review,” Journal of voice, vol. 27, no. 3, pp. 390–e21, 2013.

[8] R. Fernandez and R. W. Picard, “Modeling drivers’ speech under stress,” Speech communication, vol. 40, no. 1-2, pp. 145–159, 2003.

[9] A. Guidi, C. Gentili, E. P. Scilingo, and N. Vanello, “Analysis of speech features and personality traits,” Biomedical Signal Processing and Control, vol. 51, pp. 1–7, 2019.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, mar 2011. [11] D. Brunato, A. Cimino, F. Dell’Orletta, G. Venturi, and S. Montemagni, “Profiling-ud: a tool for linguistic profiling of texts,” in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 7145– 7151.

[12] S. Argamon, “Computational register analysis and synthesis,” Register Studies, Forthcoming, 2019.

[13] D. Brunato, et al., “Is this sentence difficult? do you agree?” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2690–2699.

[14] B. Iavarone, D. Brunato, and F. Dell’Orletta, “Sentence complexity in context,” in Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, 2021, pp. 186–199.

[15] P. Singh, R. Srivastava, K. Rana, and V. Kumar, “A multimodal hierarchical approach to speech emotion recognition from audio and text,” Knowledge-Based Systems, vol. 229, p. 107316, 2021.

[16] S. Ghiasi, G. Valenza, M. S. Morelli, M. Bianchi, E. P. Scilingo, and A. Greco, “The role of haptic stimuli on affective reading: a pilot study,” in 2019 41st Annual EMBC. IEEE, 2019, pp. 4938–4941.

[17] M. S. Morelli, S. Orlandi, and C. Manfredi, “Biovoice: A multipurpose tool for voice analysis,” Biomedical Signal Processing and Control, vol. 64, p. 102302, 2021.

[18] C. Manfredi, L. Bocchi, and G. Cantarella, “A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices,” Biomedical Signal Processing and Control, vol. 4, no. 3, pp. 212–220, jul 2009.

[19] B. T. Atmaja and M. Akagi, “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm,” Speech Communication, vol. 126, pp. 9–21, 2021