VOICE METRICS FOR DISCOURSE QUALITY ANALYSIS

Nicolae JINGA[1], Florica MOLDOVEANU[2], Alin MOLDOVEANU[3], Anca MORAR[4], Irina MOCANU[5], Alexandru BUTEAN[6]

In this paper, we introduce new discourse quality metrics and an evaluation method, and furthermore, we provide a pilot implementation and evaluate it in a specific use case – that of public speaking. Voice analysis and discourse quality metrics are important topics as they can solve various problems of modern-day life. In public speaking trainings, it is important to be able to analyze voice and discourse, to offer the speaker valuable feedback and information on how to improve when speaking in public. We identified several papers that study this issue, in which rhythm, tone, fluency and clarity, as well as biological signals, are analyzed, to detect emotions or anxiety. We defined a set of voice quality metrics that can be implemented based on any speech recognition system. Our pilot implementation showcases the difference of using various Speech-to-Text (STT) APIs over the proposed metrics and how these affect the discourse evaluation process.

Keywords*:* Speech-to-text, speech recognition, public speaking, training system

1. Introduction

Speech recognition solutions have been the focus of researchers for a long time now. More meaningful breakthroughs have happened in the last several years, mostly through machine learning solutions [1][2]. There are various implementations from open-source to run-on-device and cloud-based. The implementations use models that are trained on custom data sets. In this paper, we evaluate the cloud based STT solutions provided by Google, Microsoft, IBM, and Amazon.

Voice analysis is the study of speech for purposes such as speech recognition or, more advanced, discourse understanding. While speech

[1] PhD Student, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: nicolae.jinga@stud.acs.upb.ro

[2] Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: florica.moldoveanu@cs.pub.ro

[3] Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: alin.moldoveanu@cs.pub.ro

[4] Associate Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: anca.morar@cs.pub.ro

[5] Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: irina.mocanu@cs.pub.ro

[6] Associate Professor, Faculty Engineering, “Lucian Blaga” University of Sibiu, Romania, e-mail: alex@butean.com

recognition is extremely well defined and equipped with tools and metrics, voice quality and discourse quality are not clearly defined in the current literature. They are multidimensional structures, which cannot be measured based only on features such as pitch or loudness [3].

Voice quality and discourse quality are essential for various disciplines and applications, such as:

- Psycholinguistics, which studies the psychological processes of the language and speech
- Mood and emotions understanding
- Acoustic discourse analysis
- Anxiety detection
- Exposure therapy and cognitive behavior therapy for glossophobia
- Public speaking training. Such training applications imply that the users speak (alone, or sometimes in front of a virtual audience), and the system evaluates their voice and discourse characteristics. After evaluating certain metrics on the user's speech, scoring related to the user's performance is shown back as feedback. Since speech is at the core of public speaking training, it is natural to be one of the most important factors to analyze.

2. Voice-related discourse quality metrics

There are several papers which describe voice and discourse metrics. In this chapter, we analyze some of them.

In the work described by Qi et al. [4], rhythm is measured in increments of 200 frames of speech, to determine volume fluctuations when speaking. Volume fluctuations are calculated based on waveform segments. Waveform segments are computed based on a predefined ratio between two adjacent segments. It was noted in the paper that the speed of the speech has a great effect on the paper's proposed solution, and thus remains for their solution's evaluation to be improved.

In the research of Seliverstov et al. [5], speech analysis is done in greater detail, and five main levels are singled out and investigated. These are discursive analysis of speech, analysis of the content of speech statements, analysis of the tone of the speech content, psycholinguistic emotions analysis and acoustic analysis of speech. For evaluation, speech is converted into text. The authors attempt to define a broad set of characteristics that might be useful to evaluate speech, such as:

- The number of spoken words and the duration of the audio (difference between first and last word). Evaluation of speech quality is done by a set of readability indices, which are computed via Flesch-Kincaid readability tests, Coleman-Liau index, Dale-Chall readability formula, Automatic Readability Index, Gunning fog index [6].

- Intelligibility and Purity of Speech - defined as having a speech where there are no words or phrases alien to the literary language. The criterion of intelligibility and purity of speech is evaluated by subtracting the number of parasite words (words that do not carry meaning).
- Fluency – computed using the number of words over the duration of the audio
- Emotional coloring – computed with a speech to text classifier by emotional state. Classification is based on Levheim three-dimensional model [6] of basic emotions.
- Sentiment analysis. Dostoevsky software library is used to analyze the tone of the text with three possible outputs: negative, neutral (no bright display of emotions) or positive.
- Speech coherence, branch style of speech and uniqueness of conversational style were only briefly mentioned without many details.

Unfortunately, while the proposed set of metrics is extremely generous, most of the provided definitions and even calculation formulas are rather ambiguous, and no implementation is provided. Hence, the paper doesn’t directly support the implementation of applications that would handle practical use-cases.

Baird et al. [7] study the recognition of emotion in public speaking scenarios with the help of a Long-Short Term Memory Recurrent Neural Network (LSTMN-RNN) [8]. The LSTMN-RNN is utilized with an attention mechanism, and various audio-based feature sets as well as fusion with biological signals are evaluated. The main conclusion is that the audio features alone are suitable enough for the task of predicting emotion in the described context. Fusing biological signals with audio has shown only minimal improvement in most cases but was not consistent across all the feature sets and would require a deeper analysis.

In the work of El-Yamri et al. [9][10], audio features are again the primary focus. The solution iteratively records short audio fragments of 5-10 seconds of the speech. These fragments are used to analyze the voice tone and detect emotions. An undisclosed API was used for the analysis of emotions in the voice, to extract a predominant emotion or a group of emotions. Based on the emotions detected, a table with certain weight values attached to different emotions like boredom, stress, neutral, calmness and happiness is defined. With these data, the speech effectiveness score is computed as Emotion Weight times Confidence Score provided by the API, which shows how confident is the API that the detected emotion is the correct one. After adding these two values, the result is divided by 10000, a number that was not explained further in the article, but the expectation is to normalize the values back to single-digit values. Based on this score, a virtual audience reacts accordingly with levels of boredom, neutrality, or happiness.

Feng et al. [11] aim to detect anxiety during public speaking. Since fear is inherently associated with the public speaking stimuli, machine learning models have been built to classify between fear and neutral. The domain-adversarial neural network [12] and Wasserstein generative adversarial network [13] were trained to detect the degree of fear in a speech sample. There have been promising results based on the research, for the artificial intelligence system to estimate public speaking anxiety in real-time and provide immediate feedback based on this.

Smith et al. [14] assess a public speaking simulator called Virtual Orator, which gives feedback on a speech, using metrics like voice modulation and pitch. Examples of feedback that the application gave were that the speaker talked single-toned or that their voice was too low or too high. There was no provided descriptive information on how these metrics were implemented.

Sülter et al. [15] designed and evaluated an education tool for children. The application does not have an automatic way to compute metrics but asks the users directly through the app to rate three metrics (Nervousness, Heart rate, Palmary Sweat) in three different stages of the speaker's presentation (before, during, after). The video of the presentation (together with the audio file) was recorded and evaluated for clarity at a later point in time, by a teacher and peers.

Vanni et al. [16] provided a meta-analysis on various papers discussing about virtual reality exposure characteristics which trigger anxious reactions and evaluations of a protocol to treat subjects diagnosed with social anxiety and fear of public speaking. In general, the virtual environments reviewed in this meta-analysis were effective in triggering appropriate emotions of the speaker, even when they lacked in aspects like graphical quality and realism (e.g., avatars which composed the virtual audience had a poor appearance fidelity). This survey, however, does not discuss to what extent voice analysis is used in any of the reviewed papers, to adapt to the virtual audience or to give any form of real-time feedback.

As a concluding note, several solutions focus on studying emotions (based on voice analysis) in various ways or use cases, but very few try to define speech metrics that can be used to provide automatic feedback, so the user can improve upon each metric.

3. Proposed metrics

Our proposed metrics for speech analysis in a public speaking training application in VR involve measuring certain characteristics of audio speech [17]. These measurements are done on the text generated from the audio input. First, we use a STT algorithm and apply it continuously over the whole presentation session. This results in a transcript of what the user has been saying during the presentation. With the given transcript, we analyze it with the following **metrics**:

- *Voice rhythm:* number of words per time unit
- *Long pauses:* long pauses between words, or sentences. A long pause is counted if no speech or unintelligible speech is detected for a certain amount of time
- *Voice volume:* measured by the normalized microphone input level
- *Voice clarity:* computed based on the confidence level of the STT algorithm in generating the accurate words
- *Filler words:* the number of words that don't provide value to the presentation (e.g., "ah", "umm", "like", "er", "right", "okay", "you know", etc.) but are rather used by speakers when they are not confident, stutter, or are being emotive. There's an extensive list of words defined in the application. Our voice analysis solution compares the words detected by the STT algorithm with each filler word stored in that list and, in case of perfect matches, considers the current word in the transcript as a filler one.

For practical uses, the proposed metrics can be tracked and computed on any number of time frames. In principle, two modalities are typical and useful to evaluate and give feedback to users:

- *On-the-fly score* – computed continuously and covering a relatively short recent time box (e.g., last 20 seconds). This is useful to provide real time feedback to speakers, or to observe the variation during the whole discourse. The reason for a time box like this is to smooth the values over time and not present the user with big discrepancies on the values.
- *Total-average score* – computed for the whole discourse. This would give a single, simple overview of the user's performance over the whole presentation. For calculation purposes, the total-average score is like an on-the-fly score with a time box set to the whole presentation.

Computing metrics for the On-the-fly and Total-average scores

- *Time-box duration (TBD):* a relatively short, fixed time-box (e.g., 20 seconds), or the duration of the whole discourse
- Number of words (NW): number of words identified during the time-box
- *Voice rhythm*: NW divided by TBD.
- *Long pauses:* number, and ratio. To identify long pauses, we measure pauses as the durations between the end of a word and the start of the next one, and filter pauses above a set threshold. As metrics, we could consider either their number or their ratio (i.e., the sum of the duration of the long pauses divided by TBD).
- *Voice volume:* weighted (by words' length) average of all words' voice volume.

- *Voice clarity*: Typically, voice clarity of each word is provided by speech to text services in terms of confidence in the recognition of each identified specific word.
- *Filler words percentage*: number of detected filler words divided by NW.

4. Pilot Implementation

The pilot was implemented in Unity 3D [18] and the implementation consists of a virtual audience that reacts automatically based on the aforementioned factors. The reaction of the audience can be in five states: “Very Interested”, “Interested”, “Little Interested”, “Not Interested”, “Completely Distracted”. To determine these states, each has assigned an interval of possible values (i.e., 80-100 for Very Interested) and based on On-the-fly score the state is determined. Each individual from the audience reacts accordingly with an animation corresponding to the computed state. The application embeds an STT API for the transcript on which the voice metrics are computed on. The implementation of these metrics relies on four commercial STT services: Microsoft Azure [19], Amazon Web Services [20], IBM Watson [21] and Google Cloud [22]. Any of these STT services provides the following information, for each identified word:

- The time when the detection of a spoken word started
- The time when the detection of a spoken word ended
- The STT confidence level in recognizing that the actual spoken word is correct

In order to compute the proposed metrics, this information can be stored as they are received, for example in an array containing the per-word information. Performance-wise this would only require a relatively low memory space (e.g., for a 5-minute discourse there would be only 500-1000 words). Then, to calculate the on-the-fly metrics, we only need to search from the end of the array and select words from the last *n* seconds (*n* being the time-box duration). This is, of course, extremely fast and light in terms of the computational effort required. For the total average metrics, once the discourse is finished, all the required data is available, and the metrics can be immediately calculated.

In Fig. 1 we illustrate the manner in computing the duration of words and pauses in the proposed metrics. The green lines represent the duration of each word. The duration is computed as the difference between *detected end-time of current word* and *detected start-time of current word*. The blue lines represent the pauses between each word. A pause is the difference between *detected start-time of current word* and *detected end-time of previous word*. Annotated with red are the pauses which surpass the given threshold, which are counted as long pauses. The orange segments represent the time-box duration taken into consideration for the On-the-fly score.

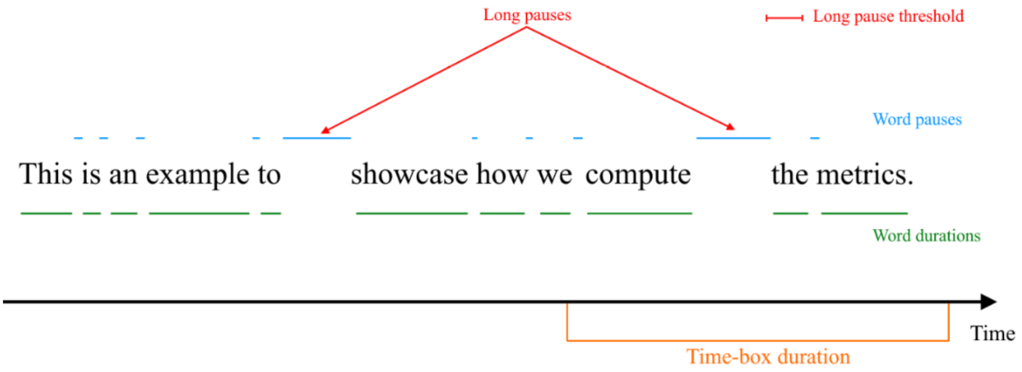

Fig. 1. Example of how words and pauses are detected

The computed values in the time-box duration are used throughout the presentation to compute local average values. Time-box values are used to display a smoothed-out graph line feedback to the user at the end of the presentation, but also during the presentation. This helps to guide the users on certain metrics based on their performance.

5. Consistency evaluation over multiple commercial speech recognition APIs

We tested 4 STT APIs as support for computing the proposed voice metrics, in a public speaking scenario. Three different presentations were recorded, and then used to calculate the metrics with each of the 4 STT solutions. All the presentations were in English.

- The first presentation is a biology-specific presentation where the speaker must present the human spine.
- The second presentation is a math-specific presentation where the speaker must explain basic vector calculus.
- The third presentation is a geography-specific presentation where the user must present notable landforms from all the continents.

In the table below, we aggregated the values specific to each STT API over each presentation. We have monitored the previously defined metrics: rhythm, long pauses, volume, clarity, and filler words, but also the number of detected words.

Table 1.

Aggregated results for each STT API on each presentation

Presentation Type	**Metric**	**Speech-To-Text Engine**			
		Microsoft	**Amazon**	**IBM**	**Google**
BIOLOGY (5 min 23 sec)	Detected words	543	535	534	540

	Rhythm	1.68	1.65	1.65	1.67
	Long pauses	4	4	4	4
	Volume	57	56	59	56
	Clarity	98	97	97	98
	Filler Words	3	3	3	3
MATHS (9 min 47 sec)	Detected words	1127	1095	1108	1123
	Rhythm	1.91	1.86	1.88	1.91
	Long pauses	5	4	4	5
	Volume	59	58	61	57
	Clarity	96	96	94	97
	Filler Words	6	6	6	6
GEOGRAPHY (3 min 11 sec)	Detected words	329	317	318	330
	Rhythm	1.72	1.65	1.66	1.72
	Long pauses	3	3	3	3
	Volume	60	59	62	59
	Clarity	97	95	94	96
	Filler Words	4	4	4	4

Based on the table content, we notice only small differences between the four APIs.

- For **Rhythm**, Amazon and IBM have a bit lower value overall than Microsoft and Google across all three presentations, but nothing too concerning in impacting the overall evaluation.
- **Long Pauses** represents the only metric where a notable difference is found between these four APIs. Microsoft and Google detected 5 long pauses, while Amazon and IBM identified 4 long pauses in the MAT presentation.
- **Volume** has slight differences, and this is due to how each STT marks the timestamp for the beginning and end of each word as we are computing a weighted average based on their recorded length, which is directly tied to these timestamps.
- **Clarity** has a slight difference: Microsoft and Google gave values a little bit above those given by Amazon and IBM, but the values are close, nonetheless.

- **Filler Words** were detected identically across all four APIs.

Overall, the differences in the calculated values for the proposed metrics, with the 4 STT engines, are minimal.

6. Conclusions

We analysed the state of the art for public speaking metrics, defined and detailed our proposed metrics, and went through the implementation details of said metrics in our pilot. The core method which enables us to measure these metrics is the STT algorithm. We compared four different APIs to check if any provide a statistical or significant advantage over the others, obtaining nearly identical results.

Based on these results, we can conclude that any of the analyzed commercial services can be successfully used to implement the metrics for public speaking training purposes.

Future investigations can be done, by evaluating how various accents of the speakers can influence the results of the APIs and of the proposed metrics.

Acknowledgement

This work was supported by the project TRAVEE (Virtual Therapist with Augmented Feedback for Neuromotor Recovery) through a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-III-P2-2.1-PTE-2021-0634, within PNCDI III

REFERENCES

[1] Y. B. Singh and S. Goel, "Survey on Human Emotion Recognition: Speech Database, Features and Classification," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 298-301, doi: 10.1109/ICACCCN.2018.8748379.

[2] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir and B. W. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3114365.

[3] Ben Barsties, Marc De Bodt, Assessment of voice quality: Current state-of-the-art, Auris Nasus Larynx, Volume 42, Issue 3, 2015, Pages 183-188, ISSN 0385-8146, https://doi.org/10.1016/j.anl.2014.11.001.

[4] H. Qi, Z. Zhang, M. He and X. Zhao, "Rhythm analysis of teacher's speech in classroom," 2017 Chinese Automation Congress (CAC), 2017, pp. 1955-1959, doi: 10.1109/CAC.2017.8243090.

[5] Y. A. Seliverstov, A. A. Komissarov, D. A. Tsyrkov, S. S. Torsionov, A. A. Lesovodskaya and A. V. Podtikhov, "Development of an Intelligent Speech Analysis System," 2022 XXV International Conference on Soft Computing and Measurements (SCM), 2022, pp. 192-197, doi: 10.1109/SCM55405.2022.9794875.

[6] H. Lövheim, "A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters", Medical Hypotheses, vol. 78, pp. 341-348, 2012.

[7] A. Baird, S. Amiriparian, M. Milling and B. W. Schuller, "Emotion Recognition in Public Speaking Scenarios Utilising An LSTM-RNN Approach with Attention," 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 397-402, doi: 10.1109/SLT48900.2021.9383542.

[8] Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735

[9] M. El-Yamri, A. Romero-Hernandez, M. Gonzalez-Riojo and B. Manero, "Emotions-Responsive Audiences for VR Public Speaking Simulators Based on the Speakers' Voice," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 2019, pp. 349-353, doi: 10.1109/ICALT.2019.00108.

[10] El-Yamri, M., Romero-Hernandez, A., Gonzalez-Riojo, M. et al. “Designing a VR game for public speaking based on speakers features: a case study.” Smart Learn. Environ. 6, 12 (2019). https://doi.org/10.1186/s40561-019-0094-1

[11] K. Feng, M. Yadav, M. N. Sakib, A. Behzadan and T. Chaspari, "Estimating Public Speaking Anxiety from Speech Signals Using Unsupervised Transfer Learning," 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2019, pp. 1-5, doi: 10.1109/GlobalSIP45357.2019.8969502.

[12] Abdelwahab, M., & Busso, C. (2018). Domain Adversarial for Acoustic Emotion Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1–1. doi:10.1109/taslp.2018.2867099

[13] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in International Conference on Machine Learning, 2017, pp. 214–223

[14] Angela M. Smith and Brendan G. Beal, “Can the use of Virtual Reality Headsets Help Reduce the Fear of Public Speaking”, International Journal of Liberal Arts and Social Science, Vol. 7, No. 5, June 2019, ISSN: 2307-924X, https://ijlass.org/articles/7.5.12.116-121.pdf

[15] Robin E. Sülter, Paul E. Ketelaar, Wolf-Gero Lange, "SpeakApp-Kids! Virtual reality training to reduce fear of public speaking in children – A proof of concept," Computers & Education, Volume 178, 2022, 104384, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2021.104384.

[16] Vanni F, Conversano C, Del Debbio A, Landi P, Carlini M, Fanciullacci C, Bergamasco M, Di Fiorino A, Dell'Osso L. A survey on virtual environment applications to fear of public speaking. Eur Rev Med Pharmacol Sci. 2013 Jun;17(12):1561-8. PMID: 23832719.

[17] N. Jinga, A. Moldoveanu, F. Moldoveanu, A. Morar, O. Mitrut, “VR training systems for public speaking – a qualitative survey”, International Scientific Conference eLearning and Software for Education, Volume 2, 2021, Pages 174-181, doi: 10.12753/2066-026X-21-092

[18] https://unity.com/

[19] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/

[20] https://aws.amazon.com/transcribe/

[21] https://www.ibm.com/ro-en/cloud/watson-speech-to-text

[22] https://cloud.google.com/speech-to-text