



Automatic Assessment of Speech Intelligibility using Consonant Similarity for Head and Neck Cancer

Sebastião Quintas¹, Julie Mauclair¹, Virginie Woisard^{2,3}, Julien Pinquier¹

¹IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France

²IUC Toulouse, CHU Toulouse, Service ORL de l'Hôpital Larrey, Toulouse, France

³Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France

{sebastiao.quintas, julie.mauclair, julien.pinquier}@irit.fr
woisard.v@chu-toulouse.fr

Abstract

The automatic prediction of speech intelligibility is a widely known problem in the context of pathological speech. It has been seen as a growing and viable alternative to perceptual evaluation, which is typically time-consuming, highly subjective and strongly biased. Due to this, the development of automatic systems that are able to output not only unbiased predictions, but also interpretable scores become relevant. In this paper we investigate a method to predict speech intelligibility based on consonant phonetic similarity. The proposed methodology relies on a siamese network to compute similarity scores between healthy and pathological phonemes, and based on the combination of those scores, regresses the intelligibility values. Our experimental evaluation suggests a high baseline correlation value of $p = 0.82$, when applied to our corpus of head and neck cancer. Moreover, further conditioning of the system on specific phonemes in key contexts increased the correlation up to $p = 0.89$. The given methodology also aims to promote interpretability of the predicted intelligibility score, which is highly relevant in a clinical setting.

Index Terms: speech intelligibility, pathological speech, automatic speech processing, head and neck cancer.

1. Introduction

Head and neck cancer (HNC) is a type of cancer with major functional repercussions on breathing, swallowing and speech. Due to this, a communication impairment is likely to appear, impacting the speech-related quality of life. As a result, perceptual evaluation has long been the most used method for therapists to assess disordered speech. On the other hand, perceptual evaluations are very time-consuming, biased and variant, since the evaluation can be conditioned on, for example, patients previously assessed by the same therapist [1]. Due to the biased nature and low reproducibility of these scores, and also due to the increasing rate of oropharyngeal cancer incidence, the development of an automatic assessment that is able to output unbiased and reproducible intelligibility measures becomes of high interest [2, 3].

From the literature, one can distinguish different ways to predict speech intelligibility for pathological speech. These methods can range from approaches such as regressing a score from the word error rate achieved by automatic speech recognition (ASR) systems [4], to extract relevant features from pathological speech, using automatic speech processing technologies [5, 6]. Speaker embeddings, such as *i-vectors* or *x-vectors*, have also proved to be a viable alternative for intelligibility estimation [7, 8]. Similarity estimation systems, such as siamese

networks, have seen a growing use in tasks such as speaker verification [9] and sentence similarity [10]. Recent works, such as [11] and [12], used the aforementioned methodology in a pathological speech context. In both cases, the systems were developed for the detection of children's speech disorder, focusing on the binary task of detecting specific mispronunciations.

In the context of pathological speech, besides being unbiased and reproducible, it is also highly relevant that an automatic approach maintains explainability of the produced estimations, which typically lacks in the automatic systems based on deep learning. An explainable system could provide more relevant cues in a clinical setting and promote more objective measures. The interpretability of the results, that normally lacks in machine learning systems, can also be used to build trust in the implementation of automatic approaches [13].

In the present work, we introduce an intelligibility prediction system based on consonant similarity. There are multiple motivations behind the development of such system, namely that: (i) ASR based intelligibility prediction systems typically underperform in patients with severe speech impairments [4]; (ii) Individual phonemes, especially consonants, are highly relevant for perceptual speech intelligibility [14, 15], either healthy [16] or pathological [17]; (iii) Automatic systems tend to lack explainability [18], which is normally demanded by health practitioners [13]. Given this, we propose an automatic system that predicts speech intelligibility based on consonant similarity, that is able to output not only an objective, but also a fully explainable prediction. Since that in our previous work, we found that there are sentences able to conduct a more accurate intelligibility prediction [7], in the present work we also explore the relevance of specific phonemes in our automatic speech intelligibility score.

The rest of this paper is organized as follows. Section 2 introduces our proposed system and the function used to regress our intelligibility values. Section 3 details our experiments on the French HNC Speech Corpus (C2SI). Section 4 presents an analysis of the obtained results, the post-treatment and the discussion. Finally, section 5 summarizes the results of the work.

2. Methodology

The proposed methodology relies on three steps. The first one corresponds to the feature extraction and data preparation. The second uses a recurrent siamese network in order to compute the phonetic similarity between two phones. Finally, the third step is the computation of the intelligibility score based on the phonetic similarity scores previously achieved.

2.1. Data Preparation

We used individual phones as input to our system, which were obtained via forced alignment using the Montreal Forced Aligner toolkit [19]. For each file, 13 Mel Frequency Cepstral Coefficients (MFCCs) were extracted. Filterbank features, MFCCs + delta coefficients and phonetic posteriors [20] were also experimented, however, in our specific context they conducted to a poorer generalization ability of the system. Interestingly, the work of [21] also found that simple MFCCs conducted to a better intelligibility classification in continuous children’s speech, when used in a recurrent system. The phones were then paired in two categories: same-phoneme and different-phoneme pairs, which were used to train the network. Due to the nature of the corpus used, further explained in subsection 3.1, and the motivation stated in section 1, the phonemes used in this study correspond to the 16 French consonants.

2.2. Siamese Network

The proposed system uses a siamese network to detect phonetic similarities. The system receives as input two phones, one as a reference and another as a test, and computes their similarity. A mispronunciation is detected whenever the test phone is found to be dissimilar from the reference one. The system uses two bilateral Gated Recurrent Units (GRU) as encoders with shared weights, in which the two outputs form embedding representations (see figure 1).

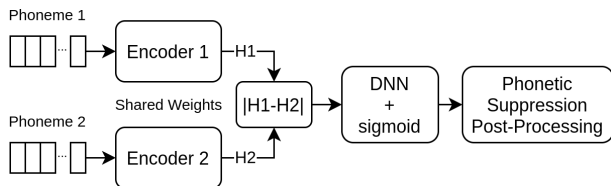


Figure 1: Schematic diagram of the proposed system.

The encoders are comprised of two hidden layers, with a hidden dimension of 100. The embedding representation corresponds to the concatenation of the last two hidden states, of forward and backward context of the bilateral GRU used. This generates an embedding representation of size 200. The absolute difference between the two fixed-length representations is then computed. It is expected that the system models the input phonemes so that same-pair phonemes are closer in the embedding space than different-pair phonemes. Afterwards, a Deep-Neural-Network block is appended. This block is composed of 3 fully-connected layers of size 200. The Rectified Linear Unit function is used as an activation function in all of the layers except for the final one, which used a sigmoid. A dropout rate of 0.25 and batch normalization are applied on every fully-connected layer.

2.3. Intelligibility Estimation

Based on the similarity scores obtained by our system, we compute an intelligibility score for each speaker:

$$I(S_a) = \frac{\sum_{n=1}^{16} \frac{\text{Sim}_a(n)}{\text{Tot}_a(n)}}{16} * 10 \quad (1)$$

This scoring function corresponds to the arithmetic mean of each patient’s individual consonant score. $\text{Sim}_a(n)$ refers to the number of similar phones of one of the 16 consonants

obtained by a speaker (S_a), while $\text{Tot}_a(n)$ is the total number of representations of that consonant issued by a speaker.

3. Experiments and Results

3.1. C2SI Corpus

The present work made use of the French head and neck cancer speech corpus C2SI [22]. The corpus includes a variety of patients that suffer from the oral cavity or oropharyngeal cancer, with different onset tumor locations, and also healthy speakers. All of the patients were asked to record a variety of tasks, and in this study, we used the isolated pseudo-word task. In this task, each speaker was asked to record a set of 52 pseudo-words, nonexistent in the French language [23]. Each pseudo-word was automatically generated so that it respects French phonotactic and orthographic rules, following a $C(C)_1V_1C(C)_2V_2$ structure, where $C(C)_i$ is either a single consonant or a consonant group and V_i a single vowel. Each speaker was attributed a phonetically balanced pseudo-word set that contains at least one instance of each consonant in the applicable contexts of beginning and middle of the word. The correct pronunciation of each pseudo-word was played to each speaker before recording. All audio files were recorded with a sampling rate of 48 kHz. While all the french consonants were present in each set of 52 pseudo-words, either isolated or in consonant groups, there were only 8 vowel representations. Due to this and the other reasons aforementioned in section 1, we focused mainly on consonant similarity in this study.

Table 1: Example of a set of 52 pseudo-words with the aforementioned structure of $C(C)_1V_1C(C)_2V_2$.

banfou	bleja	boucti	brimpli	chessant	choniou
clifant	coгу	crimpin	daillu	dinrant	dredi
fanrsi	flinrpu	fouma	fravi	gabi	glunou
gorvo	guchin	joutu	juro	lanvin	lerda
messo	mouco	nianlo	niejo	noksa	nouillou
pastu	pidant	ploniou	pripin	psila	quiga
rinta	rurnu	sanvrin	scuna	souquin	spaclant
sticho	tangri	tougzu	tradrou	virjant	vumou
yainzi	yaltin	zebou	zouzant		

The intelligibility values, used as targets, were computed based on the independent perceptual evaluation of six different therapists. A score between 0 and 10 was attributed, based on those evaluations, the smaller the score is, the less intelligible the speaker is. A total of 102 speakers, 24 healthy controls (HC) and 78 patients, were used in the present study.

3.2. Experimental Setups

In order to learn phonetic similarity, we used same-phoneme pairs as positive and different-phoneme pairs as negative. 24 HC of the C2SI corpus were used as training data. After force aligning and extracting the features, we obtained a total of 3,323 training phones (consonants only). For each one of the 16 consonants groups, an individual training set was created. For a specific training set, every consonant was paired with every other representation of the same consonant, and also to a random set of 650 different consonants. Out of those formed pairs, a random subset of 50k pairs was extracted. This subset corresponds to the final training set for that same consonant.

3.2.1. System Training

Sixteen different models, one for each consonant, were created. Each individual model was fine-tuned on four hyperparameters: learning rate (between 0.001 and 0.0001), epochs (between 2 and 8), batch size (either 256 or 512) and GRU dropout rate (either 0.0 or 0.25). A binary cross-entropy loss function was used, optimized by the SGD algorithm [24].

3.2.2. System Validation

A set of six patients from the C2SI corpus, unseen during training, was used to validate the proposed model instead of a subset of the HC. The chosen patients had high intelligibility (over 9.8 on a 0 to 10 scale), making them virtually indistinguishable from HC. The validation phones were grouped into the 16 different consonant groups. Each new phone is paired with all the reference phones seen during training. The median similarity is computed between each test and reference consonant groups (see figure 2).

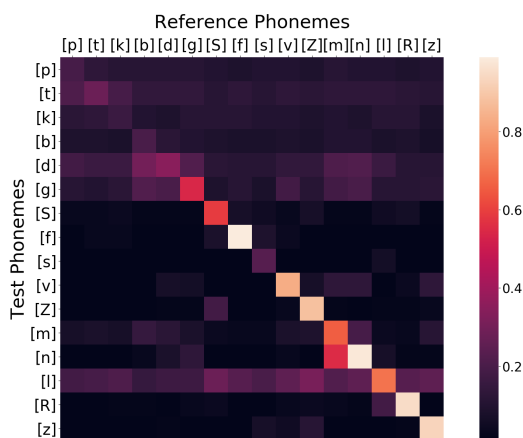


Figure 2: *Siamese network validation heatmap. Consonant phonemes are illustrated by their respective French SAMPA symbols.*

A reliable distinction between all phonetic groups was obtained. This aspect was less evident in the plosive consonants, where the difference between the key consonant to be tested and the other consonants was more subtle when compared to the remaining groups. This was expected due to the nature of forced-alignment, that segments plosives in short time intervals. This in turn provided less contextual information, making the phonetic distinction a slightly more difficult task. This aspect, however, did not prove to be problematic as all speakers were submitted to the same forced-alignment.

3.2.3. Evaluation Scores

To generalize the trained models on the remaining patients of the corpus, we used as a threshold the median values obtained by the validation patients for each consonant group, also known as the diagonal values of figure 2. In order to make the system more robust to the different median values, we added to the threshold the median absolute deviation (MAD) [25]. The classification method is as follows. If a new phone has a similarity score above the threshold ($\tilde{x}_i - MAD_i$), where \tilde{x}_i is the median, it is considered a similar phone, otherwise it is considered dissimilar. The intelligibility score is computed according to equation 1.

By correlating the predicted scores with the given intelligibility reference values, we were able to achieve a Spearman's correlation coefficient of $\rho = 0.82$ (see figure 3).

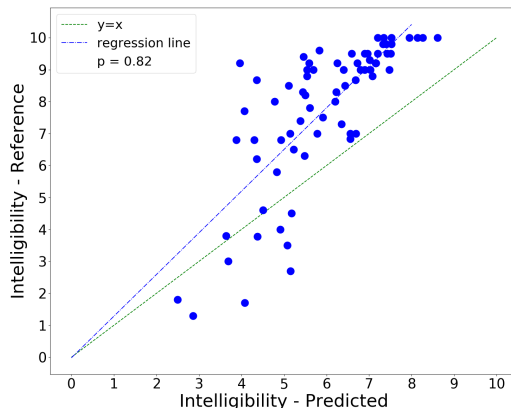


Figure 3: *Intelligibility prediction plot using the proposed consonant similarity approach.*

4. Outlier Analysis

We performed an analysis on the predicted values that had a large deviation from the target intelligibility, all of which had a prediction below target. From the perceptual evaluation of these outliers, we found that the system was underperforming in specific contexts, namely very breathy/hoarse speech and patients with nasalized plosives. All of the cases corresponded to phonetic mispronunciations, and it was expected that the system would classify them as such, since it was neither trained nor validated with speakers that had similar mispronunciations. On the other hand, despite the phonemes being mispronounced, we noticed that in these cases the reference intelligibility values were still high, pointing out that in the aforementioned contexts, those specific mispronunciations had a little affect on the reference intelligibility values.

4.1. Phonetic Suppression Post-Processing

We noticed that by suppressing specific consonants we were able to obtain a better intelligibility estimation. Thus, we made use of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [26]: an ensemble of acoustic parameters tailored for indexing physiological changes in voice production, with an added degree of theoretical significance and explainability, and we went to search for hand-crafted features used in the literature to address the key types of speech impairment aforementioned. We found a set of three acoustic parameters:

- **Slope UV0-500 (mean)** - Mean value of the linear regression slope of the logarithmic power spectrum within 0-500Hz on unvoiced segments. Related to breathy and hoarse voice qualities [27].
- **Loudness (percentile 20)** - Estimate of perceived signal intensity from an auditory spectrum. In our context, we hypothesize that it can help to detect nasalized plosives [28] due to the added intensity found in this type of mispronunciation.
- **LogRel F0-H1-A3 (mean)** - Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3). Relevant feature for breathy/hoarse voice assessment as well [29].

By using this set of features, we can model the speakers with the aforementioned speech pathologies and suppress certain phonetic groups accordingly. The modelization of these speakers according to these features can be found described in table 2, where a threshold indicates the cutoff region for each feature. Results from this phonetic suppression are in table 3.

Table 2: Phonemes used and suppressed in the intelligibility score function according to the GeMAPS features.

Feature name	Suppressed phonemes	Threshold	Used phonemes
Slope UV0-500 (mean)	[b], [d], [g], [z], [Z], [v], [m], [n], [l]	> 2.41	[p], [t], [k], [s], [S], [f], [R]
LogRel F0-H1-A3 (mean)		< 15.00	
Loudness (percentile 20)	[p], [t], [k], [b], [d], [g]	> 0.31	[s], [S], [z], [Z], [f], [v], [m], [n], [l], [R]

Table 3: Correlation results achieved by the proposed methodology and by the phonetic suppression post-processing.

		ρ
Automatic Speech Recognition Approach		0.63
X-vector Speaker Embedding Approach		0.74
Consonant Similarity Approach	Predicted	0.82
	Predicted + Loudness	0.83
	Predicted + LogRelF0-H1-A3	0.86
	Predicted + SlopeUV0-500	0.87
	Predicted + LogRelF0-H1-A3 + SlopeUV0-500 + Loudness	0.89

4.2. Discussion

The results suggest that we can reliably predict speech intelligibility using consonant similarity. Moreover, by conditioning the used consonants on key mispronunciations and the external GeMAPS features, we are able to obtain an even higher level of correlation ($\rho = 0.89$). This aspect points out that, depending on the speech impairment a speaker may have, there are mispronounced phonemes that do a little contribution to the overall intelligibility score: for patients with a high level of hoarseness, all voiced phonemes were classified as non-similar by the system. By suppressing the voiced phonemes, we were able to obtain more accurate predictions. To detect this hoarseness, we used the Slope and LogRel (see table 2). Any patient that had a feature value above a threshold had their voiced phonemes suppressed, and the remaining phonemes were used for the score. The suppression of the full plosive group, displayed on table 2, also lead to more accurate predictions, showing that those mispronunciations did not affect much the perceptual intelligibility estimations (see correlation values on table 3). As expected, the

used features also isolated a few patients that did not have the specific mispronunciations aforementioned, however, the same phonetic suppression poorly affected those intelligibility scores, confirming a certain level of robustness of the chosen features.

The assumption that different phonemes have different degrees of relevance corroborates the fact that for each speaker, there are sentences that are able to convey a better intelligibility estimation than others, concluded in [7]. A deeper feature analysis should be investigated in order to identify other contextual key phonemes that are less important in the intelligibility score. Further robust feature conditioning could help provide more accurate scores and also a more objective and explainable patient-specific information. We hypothesize that an *a priori* knowledge of patient’s specific features (e.g. tumor location, reconstruction type), which are known in a clinical context, could condition the system to obtain even more accurate results. We leave this analysis for future work.

The results were also compared to two previous approaches used on the same corpus (see table 3), one based on the application of a Wagner-Fischer algorithm to the distance between the automatically transcribed (ASR) and ground truth pseudo-word [6], and the other based on *x-vector* speaker embeddings [7]. The same set of speakers and respective pseudo-words (see figure 3) were submitted to both approaches. A significant increase in correlation can be found in both cases. Moreover, an added degree of explainability can be obtained when using our proposed approach, since the intelligibility score can be fully traced back to the amount of similar/dissimilar consonant phones that a speaker has.

The correlation values, illustrated on table 3, also show that despite the subjectivity of the perceptual evaluations, high values can be found when using an automatic objective intelligibility estimation. This aspect mitigates the low reproducibility, variance and subjectivity of the perceptual measures in favor of a reproducible, objective and interpretable automatic prediction. The more objective way to predict speech intelligibility and the resulting added degree of explainability become highly relevant in a clinical context [30].

5. Conclusions

This paper investigated an automatic approach to predict speech intelligibility based on consonant similarity. Our approach made use of a pseudo-word task, in the context of head and neck cancer, that was force-aligned in order to train a siamese network. A base correlation of $\rho = 0.82$ was obtained between the predicted and reference values, which showed a significant correlation gain over two previous approaches. When conditioning the intelligibility prediction on certain consonants, the correlation increased up to $\rho = 0.89$. This showed that depending on the speech impairment experienced by the speaker, there are phonemes that have a greater or smaller importance in the intelligibility score. The proposed system also maintains a high degree of interpretability, since the final intelligibility score is a function of the individual scores of each phoneme, which is relevant in a clinical setting. Future work will investigate the importance of specific phonemes and co-articulations between consonants and vowels on the intelligibility score.

6. Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287.

7. References

- [1] M. Balaguer, T. Pommée, J. Farinas, J. Pinquier, V. Woisard, and R. Speyer, "Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review," *Journal of the Sciences and Specialities of Head and Neck*, 2019.
- [2] C. Middag, *Automatic analysis of pathological speech*. Doctoral Dissertation: Ghent University, Department of Electronics and information systems, Ghent, Belgium, 2012.
- [3] S. Fex, "Perceptual evaluation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 6, no. 2, pp. 155–158, 1992.
- [4] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," *Proceedings of Interspeech*, 2012.
- [5] C. Middag, J.-P. Martens, G. V. Nuffelen, and M. D. Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, 2009.
- [6] C. Fredouille, A. Ghio, I. Laaridh, M. Lalain, and V. Woisard, "Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers," *International Congress of Phonetic Sciences (ICPhS)*, 2019.
- [7] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, "Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer," *Proceedings of Interspeech*, 2020.
- [8] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers," *Proceedings of Interspeech*, 2018.
- [9] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *Proceedings of ICASSP*, 2018.
- [10] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," *Proceedings of Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [11] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features," *Proceedings of Interspeech*, 2019.
- [12] S.-I. Ng and T. Lee, "Automatic detection of phonological errors in child speech using siamese recurrent autoencoder," *Proceedings of Interspeech*, 2020.
- [13] W. K. Diprose, N. Buist, N. Hua, Q. Thurier, G. Shand, and R. Robinson, "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator," *Journal of the American Medical Informatics Association, Volume 27, Issue 4*, 2020.
- [14] G. V. Nuffelen, C. Middag, M. D. Bodt, and J.-P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International Journal of Language & Communication Disorders, Volume 48, Issue 6*, 2008.
- [15] G. Saravanan, V. Ranganathan, A. Gandhi, and V. Jaya, "Speech outcome in oral cancer patients - pre- and post-operative evaluation: A cross-sectional study," *Indian J Palliat Care*, 2016.
- [16] M. Fort, A. Martin, and S. Peperkamp, "Consonants are more important than vowels in the bouba-kiki effect," *Lang Speech*, 2015.
- [17] Crevier-Buchman, V. J. M. S, and B. D., "Intelligibility of french consonants after partial supra-cricoid laryngectomy," *Revue de Laryngologie - Otologie - Rhinologie*, 2002.
- [18] L. G. McCoy, C. T. Brenna, S. S. Chen, K. Vold, and S. Das, "Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based," *Journal of Clinical Epidemiology*, 2021.
- [19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldii," *Proceedings of Interspeech*, 2017.
- [20] J. C. Vasquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Noth, "Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech," *Proceedings of Interspeech*, 2019.
- [21] Y.-S. Lin and S.-C. Tseng, "Classifying speech intelligibility levels of children in two continuous speech styles," *Proceedings of ICASSP*, 2021.
- [22] V. Woisard, C. Astésano, M. Balaguer, J. Farinas, C. Fredouille, P. Gaillard, A. Ghio, L. Giusti, I. Laaridh, M. Lalain, B. Lepage, J. Mauclair, O. Nocaudie, J. Pinquier, G. Pouchoulin, M. Puech, D. Robert, and V. Roger, "C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers," *Language Resources and Evaluation*, 2020.
- [23] M. Lalain, A. Ghio, L. Giusti, D. Robert, C. Fredouille, and V. Woisard, "Design and development of a speech intelligibility test based on pseudowords in french: Why and how?" *Journal of Speech, Language and Hearing Research*, 2020.
- [24] N. Ketkar, "Stochastic gradient descent. in: Deep learning with python," https://doi.org/10.1007/978-1-4842-2766-4_8, 2017.
- [25] T. Pham-Gia and T. Hung, "The mean and median absolute deviations," *Mathematical and Computer Modelling*, vol. 34, pp. 921–936, 2001.
- [26] F. Eyben, K. Scherer, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2015.
- [27] F. Alipour, R. C. Scherer, and E. Finnegan, "Measures of spectral slope using an excised larynx model," *Journal of Voice*, vol. 26, pp. 403–411, 2012.
- [28] K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria," *Journal of Speech, Language and Hearing Research*, vol. 47, pp. 766–783, 2004.
- [29] S. V. Narasimhan and K. Vishal, "Spectral measures of hoarseness in persons with hyperfunctional voice disorder," *Journal of voice: official journal of the Voice Foundation*, 2017.
- [30] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language & Communication Disorders, Volume 48, Issue 6*, 2013.