Can we use speaker embeddings on spontaneous speech obtained from medical conversations to predict intelligibility?

Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, Julien Pinquier

CAN WE USE SPEAKER EMBEDDINGS ON SPONTANEOUS SPEECH OBTAINED FROM MEDICAL CONVERSATIONS TO PREDICT INTELLIGIBILITY?

Sebastião Quintas[1], *Mathieu Balaguer*[1,2], *Julie Mauclair*[1], *Virginie Woisard*[2,3], *Julien Pinquier*[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2]IUC Toulouse, CHU Toulouse, Service ORL de l'Hôpital Larrey, Toulouse, France
[3]Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France

ABSTRACT

The automatic prediction of speech intelligibility is a recurrent problem in the context of pathological speech. Despite recent developments, these systems are normally applied to specific speech tasks recorded in clean conditions that do not necessarily reflect a healthcare environment. In the present paper, we intend to test the reliability of an intelligibility predictor on data obtained in clinical conditions, in the specific case of head and neck cancer. In order to do so, we present a system based on speaker embeddings trained on a multi-task methodology to simultaneous predict speech intelligibility and speech disorder severity. The results obtained on the different evaluation tasks display correlations as high as 0.891 on a hospital patient set, showing robustness to the type of speech material used in these automatic assessments. Moreover, the usage of spontaneous speech during the evaluation shed light on an understudied, but with more ecological validity, type of speech material which displayed promising results. The reliability displayed across the different tasks suggests a direct deployment of the developed systems in a hospital setting.

***Index Terms*—** speech intelligibility, automatic speech processing, speaker embeddings, head and neck cancer, spontaneous speech

1. INTRODUCTION

A functional impairment at communication level is generally expected in the post-treatment of conditions that affect the vocal tract, such as head and neck cancer (HNC) and neurodegenerative diseases with dysarthria symptoms. Since major functional repercussions on the upper aerodigestive tract (breathing, swallowing, and phonation/speech) are likely to appear, a loss of speech intelligibility is commonly found, impacting the patient's quality of life [1]. Since an early tracking and diagnosis are usually correlated to a better prognosis, due to the progressive and timed implementation of post-treatment measures, the perceptual evaluation of speech intelligibility has long been the most common method of disordered speech assessment.

Within the topic of perceptual clinical measures, besides speech intelligibility there is also speech disorder severity, that can be seen as a more global measure that also takes in consideration speech intelligibility. Despite serving two different purposes, both measures tend to share high correlations, one evaluating spoken communication ability (intelligibility) and the other evaluating the degree of the impact of a speech affecting disorder on functional communication (severity). Furthermore, these measures are known to be highly variant, biased and subjective, since the evaluation can be conditioned on several aspects such as prior knowledge of the task being issued (e.g. passage reading tasks), earlier assessments or also on *a priori* knowledge of the patients. Given this, an automatic approach has been seen as a growing alternative that can promote more reliable and less variant predictions [2].

In recent years, the automatic prediction of speech intelligibility applied to pathological speech has seen an increasing number of new approaches and methodologies. These approaches can range from scores based on automatic speech recognition performance [3, 4] to approaches that make use of more traditional signal processing techniques or machine learning methodologies [5, 6]. The speaker embedding paradigm, where speech utterances are represented into fixed-dimensional vectors that have discriminating properties among different speakers, has shown interesting gains on distinct tasks applied not only to speech intelligibility [7, 8], but also to general pathological speech assessment [9, 10]. Despite speech intelligibility being a term not only exclusive to the field of pathological speech, it becomes relevant to distinguish the perceptual evaluation of speech intelligibility from the one applied to the speech-in-noise paradigm [11]. While the definition of intelligibility may be similar to both, the perceptual decoding differs between the two. Traditional intelligibility predictors used in speech perception, such as STOI or E-STOI, require the usage of clean time-aligned signals, even in end-to-end approaches [12], which is unfeasible for pathological speech.

While recent venues in automatic prediction of speech intelligibility display interesting and promising results, the ma-

978-1-6654-7189-3/22/$31.00

jority of these systems are tested on data that hardly mimics real hospital conditions. This aspect comes across as corpora recorded in standardized environments, such as sound-treated rooms with the same microphone, a predefined microphone distance and predefined speech tasks[13, 14]. Since the end goal of these systems is to provide more accurate and unbiased intelligibility estimations, it becomes essential to evaluate their reliability and accuracy across various data sets and clinical scenarios when contemplating their direct implementation. For the sake of simplicity, we will define spontaneous speech as opposed to prepared speech, where utterances contain well-formed sentences close to those found in written documents [15]. Given this, spontaneous speech can be seen as any non-scripted and non-prepared speech material issued by a speaker.

Recorded speech tasks, such as passage reading and pseudo-words, are generally used to perform either the perceptual or automatic assessments [16, 8] of speech intelligibility. Despite this, tasks that involve spontaneous speech still have not seen enough applications on the subject of automatic evaluations [17]. Even if some works touched on this aspect [18, 19], there is still an interesting gap in the literature to be explored, which we will investigate during the course of this work. An automatic assessment using spontaneous speech in real clinical conditions greatly mimics the environment this class of systems would be deployed in, while also using a type of data that more closely represents the real communication ability of a given person: spontaneous speech segments. Hence, in the present work we conduct a series of experiments of an adapted intelligibility prediction system on more ecological hospital data. Given this, we intend to: (i) Analyze the reliability of an embedding-based intelligibility system on data recorded under real clinical conditions (ii) Assess the reliability of the same system when predicting intelligibility based on spontaneous speech segments, obtained from patient-doctor interviews.

The rest of this paper is organized as follows. Section 2 describes our system and global methodology. Section 3 presents our corpus, experiments and results. Section 4 and 5 propose a discussion and perspectives respectively. Finally, section 6 illustrates our main conclusions.

2. METHODOLOGY

Similarly to [8], the automatic intelligibility prediction system made use of the speaker embedding paradigm and an appended shallow neural network. In the present work, however, the system[1] was adapted to predict two perceptual measures in a multi-task setting: speech intelligibility (INT), defined as the degree to which the speaker's message can be understood by a listener, and speech disorder severity (SEV), the degree of intelligibility impairment associated to other speech signal variables such as acoustic-phonetic code emission quality, speech speed and other relevant temporal or prosodic parameters [20]. These two measures, despite sharing high correlations and a certain degree of similarity, serve distinct purposes. While speech intelligibility can directly evaluate the communication ability of a given patient, speech disorder severity serves as a global disease score that encapsulates different aspects of spoken communication.

2.1. Speaker Embeddings

Speaker embeddings are fixed-length representations typically used in speaker verification, speaker diarization and automatic speech recognition. Recently, these embeddings have shown an ability to convey speaker attributes that correlate well with the detection of speech affecting disorders [10]. Hence, a growing usage of these embeddings has been seen for the automatic assessment of pathological speech [7, 8, 9]. Given also the good performance of the speaker embedding paradigm on tasks that typically deal with more spontaneous speech (e.g. speaker diarization in conversational settings) [21], we hypothesize that an embedding based approach could better help predict speech intelligibility in this same spontaneous speech context, as opposed to read speech typically used in clinical evaluations. Since the *x-vector* speaker embeddings outperformed *i-vectors* in intelligibility prediction [8], two classes of speaker embeddings were tried in the present study, both extracted using the Speechbrain toolkit [22].

The first ones are the *x-vector* speaker embeddings, also used in [8] to predict speech intelligibility. This method aims to display discriminative features between speakers [23]. The embedding extractor [24] works by first passing the speech signal through a block of time-delayed neural networks (TDNN) that operates on speech frames with a small temporal context centered at the current frame. Subsequent TDNN layers build on the temporal context of previous layers. A statistic pooling layer aggregates all frame-level outputs into a fixed-length dimension, which is then fed to a fully connected block. The *x-vectors* are extracted from the affine component of the last fully connected layer. The system was pre-trained using voxceleb1 [25] + voxceleb2 [26] data, and was tested on the voxceleb1 test set, achieving an equal error rate (EER) [27] of 3.2%.

For the second class of embeddings tried, the more recent Ecapa TDNN speaker embeddings were experimented [28]. These fixed-length representations build on the concept of the *x-vector* speaker embeddings (TDNN based and a subsequent fully connected block), however, with multiple enhancements that suggest a better performance on speaker verification when compared to other embeddings [29]. The improvements range from the introduction of squeeze and excitation blocks to channel-dependent frame attention. These are hypothesized to enable the network to focus more on speaker

[1] https://gitlab.irit.fr/samova/embedding_intelligibility

characteristics that do not activate on identical or similar time instances, e.g. speaker-specific properties of vowels versus speaker-specific properties of consonants. We hypothesize that these enhancements could provide a more robust speaker embedding for pathological speech assessment, and therefore outperform the previously used *x-vectors*. Similarly to the previously introduced extractor, the Ecapa TDNN system was also pre-trained using voxceleb1 + voxcelebd2 data and tested on the voxceleb1 test set, achieving an EER of only 0.8%.

2.2. Shallow Neural Network

Figure 1 presents a diagram of our shallow neural network. The network receives as input the speaker embeddings that depending on the type can have distinct fixed dimensions (512 for the *x-vectors* and 192 for the Ecapa TDNN). Furthermore, the signal passes through two layers of fixed dimensions, and then finally the two multi-task layers that predict the two different perceptual measures. The system is optimized using a mean squared error (MSE) loss function and an Adam optimizer algorithm. Since the system was developed to predict two distinct measures, the loss function takes both in consideration with equal contributions, meaning a 50% weight for intelligibility and 50% weight for severity. Due to the high correlations typically shared between speech intelligibility and speech disorder severity, we hypothesize that the learning of these two measures together will conduct to a better and more robust estimation of speech intelligibility.

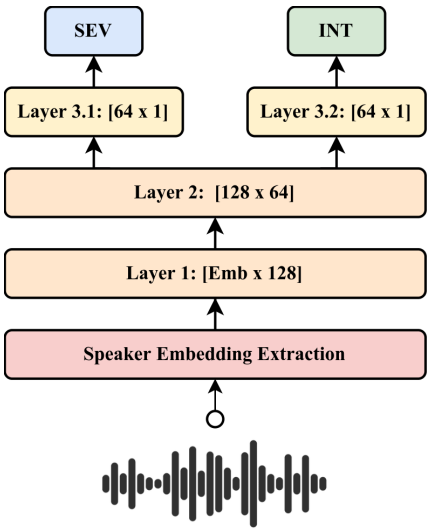

Fig. 1. Schematic diagram of the proposed shallow neural network and corresponding multi-task learning approach.

2.3. Train and validation

Two systems were trained using the C2SI Corpus [14], one for each embedding type. The corpus includes a variety of patients that suffer from oral cavity or oropharyngeal cancer, with different onset tumor locations, and also healthy speakers. Both systems were trained and validated using the segmented passage reading task. A data augmentation scheme, similar to [8], based on temporal distortion [30, 31] that preserves the pitch and spectral envelope, was implemented during training. A total of 98 speakers were used for training and a subset of 10 speakers with varying degrees of intelligibility was randomly sampled to be used as validation. A batch size of 8, a learning rate of 0.001 and a dropout rate of 0.2 were used during the course of 20 epochs.

3. EXPERIMENTS AND RESULTS

3.1. SpeeCOmco Corpus

Our speech and communication in oncology (SpeeCOmco) corpus is a set of 27 patients with varying degrees of intelligibility that recorded different tasks in real clinical conditions [32]. From the corpus population, the mean age corresponds to 66.3 years (min. 38 years, max. 83 years) with a 63% male and 37% female representation. The recording conditions include the usage of non-sound-treated rooms, the use of a headworn microphone commonly used in clinical practice and the presence of some degree of background noise. Given that the recordings took place in a hospital environment, more specifically during clinical appointments in speech therapy, the recording conditions mimic the exact same conditions that the present system would be deployed in. All patients are native French speakers.

For this patient set, the mean intelligibility and speech disorder severity were computed based on the independent perceptual evaluation of six different health professionals. Each speaker was given a score between 0 and 10, the smaller the value, the less intelligible the speech is. The same scale is used for speech disorder severity. The intraclass correlation coefficient (ICC) was computed to evaluate inter-rater reliability. An ICC (Two-way mixed-effects model with absolute agreement) of 0.816 was achieved for the six judges when rating speech intelligibility among all patients and an ICC of 0.852 was achieved for speech disorder severity, showing a good level of agreement between experts.

A variety of recorded tasks were used for this assessment, that are widely accepted and used in pathological speech assessment. These tasks can be found described below:

- **Reading passage (LEC)**. Speakers were asked to read the first paragraph of “La chèvre de M. Seguin”, a tale by Alphonse Daudet that was chosen due to being long enough to include almost all French phonemes. This passage is also well known and widespread in French clinical phonetics [33].
- **Semi-vowel sentences (PHR)**. Speakers had to read two sentences containing the French semi-vowels [w] and [U], absent from the LEC text.

- **Consonant inflection (CSN)**. Speakers were asked to read 17 sentences in the form of *"Le sac eu**C**eu convient"*, where the ***C*** is replaced at each sentence by a different consonant.

- **Pseudo-words (DAP)**. Each speaker had to record a set of 52 pseudo-words, nonexistent in the French language [34]. Each pseudo-word was automatically generated so that it respects French phonotactic and orthographic rules.

- **Spontaneous Speech (SPO)**. In this task, the audio sample came from the recording of an interview between a speech and language pathologist and the speaker. The conversation focused on the daily communication and the limitations perceived by the speaker. In order to obtain the spontaneous speech segments, the entire file was passed through a voice activity detector (VAD). Segments with less than 3 seconds and more than 10 seconds were discarded. This was done in order to minimize the number of artifacts captured. Segments with a larger presence of the therapist's voice were removed as well, however, on occasion, files with some voice overlap were preserved. Given the nature of the speech task, the size of the interview can greatly differ between speakers, and therefore the number of files as well (from 8 to 56).

3.2. Test and results

All 27 patients and their respective 5 recorded speech tasks were assessed using the two developed systems, one using the *x-vectors* and the other using the Ecapa TDNN speaker embeddings (see section 2). With the exception of the spontaneous speech task, whose resulting intelligibility prediction corresponds to the mean of the previously mentioned VAD-segmented files, the remaining tasks were analyzed on a single audio file per speaker. Table 1 illustrates the Spearman's correlation (ρ) and root mean squared error (RMSE) values on speech intelligibility as well as on the complimentary prediction of speech disorder severity. The table also illustrates the different tasks assessed and choice of embeddings. The results suggest high correlations above 0.82 on four out of five tasks when using *x-vectors*. Similarly, for the RMSE values the results also suggest low error values, below 1.5 on three of the tasks going as low as 1.322. The figure 2 displays a plot of the predictions associated to the spontaneous speech task.

4. DISCUSSION

The results generally displayed interesting correlation and error values. Moreover, since the end goal of the present article is to validate the implementation of an intelligibility prediction system clinically, a combination of high correlation and low errors should be envisioned. Despite the ECAPA embeddings showing promising results for speaker verification [29], the results suggest that, within our specific context, the *x-vectors* outperform them in all recorded speech tasks, for both intelligibility and severity. This aspect not only continues to validate the usage of *x-vectors* for pathological speech, but also shows that not all types of speaker embeddings are suited for this type of analysis, despite showing better EER performance (see section 2.1). A comparative, deeper study on the reliability of different speaker embeddings for this type of assessment as well as searching a better metric to analyze their performance on pathological speech becomes an interesting lead for future work.

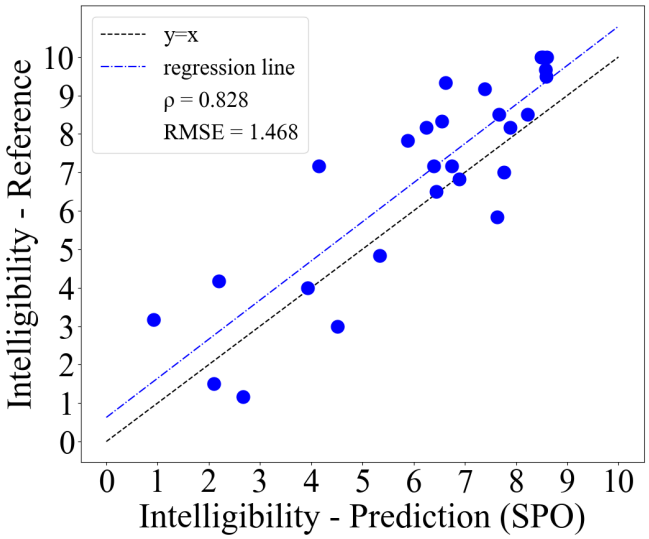

Fig. 2. Results of the intelligibility prediction on the SpeeCOmco corpus using the SPO task (*x-vectors*).

As far as the speech tasks are concerned, the *x-vectors* displayed interesting and reliable results in the majority, with the exception of the CSN and DAP tasks that illustrated larger errors. This was somehow expected since the full pseudo-word task file contains some noise artifacts (recording sounds between consecutive words), and no prosody and coarticulation between consecutive words are present. Furthermore, since the system was trained on the passage reading task, it was expected that during test it would perform the best on the same LEC task. This was evident for both intelligibility and speech disorder severity, that presented the best metrics when compared to the remaining tasks. The PHR task also displayed reliable results, especially when taking in consideration that the audio files are much shorter when compared to the entire passage reading task (i.e. two single sentences as opposed to a text). Finally, the results obtained on spontaneous speech presented a surprisingly interesting correlation and low error, which remains highly comparable to the other tasks. The assessment on this type of spontaneous speech

Table 1. Correlation and error values obtained when testing the system on the different speech tasks of the SpeeCOmco corpus. Bold values mark the tasks with the best correlation/error pairs. All correlations achieved a p-value < 0.05, making them statistically significant.

Perceptual Measures		**Speech Disorder Severity**				**Speech Intelligibility**			
Embeddings		**Ecapa TDNN**		***X-vectors***		**Ecapa TDNN**		***X-vectors***	
Evaluation Metrics		ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE
Speech Tasks	**LEC**	0.783	1.873	**0.866**	**1.384**	0.826	2.070	**0.891**	**1.322**
	PHR	0.784	1.782	0.805	1.772	0.807	1.854	**0.842**	**1.460**
	CSN	0.643	2.060	0.861	2.124	0.673	2.147	0.859	1.971
	DAP	0.296	2.673	0.724	2.219	0.371	2.805	0.731	1.881
	SPO	0.657	2.124	0.818	1.820	0.695	2.252	**0.828**	**1.468**

material becomes highly relevant due to the fact of being an under-explored medium [17] that more closely matches day-to-day communication, presenting an unbiased view on the patient's ability to communicate. This automatic assessment on spontaneous speech covers a gap in the literature concerning the test of automatic approaches on this type of recorded speech, and can be seen as a stepping stone towards more robust and reliable intelligibility predictions.

5. PERSPECTIVES

Despite speech intelligibility being the main subjective measure to be analyzed and predicted during the course of this work, the results displayed similar metrics on speech disorder severity. This aspect showcases that the multi-task paradigm can be effective when handling this type of measure, and leaves the possibility of further learning other perceptual measures together, such as prosody, resonance and phonemic distortions. Moreover, an intelligibility measure that can be regressed as a combination of these other parameters [35] can be seen as more interpretable, with an added value to a clinical environment.

The developed system was trained on a segmented and augmented reading passage task. Despite the system generalizing well on a variety of new patients and speech tasks, further training on other tasks, namely spontaneous speech material, as well as other languages and diseases (e.g. Parkinson's, Amyotrophic Lateral Sclerosis, etc.) could not only increase performance but also make the system even more robust. The development of a universal, multi-pathology automatic intelligibility model is an interesting perspective for future work. However, it should be devised carefully, since a working solution for speech intelligibility in head and neck cancers may not necessarily correspond to the best approach for neurological diseases. This is mainly due to the type of speech affecting problems that greatly differ between the two sets of diseases, making the devised systems not yet fully generalized to all pathologies.

Given the generalization ability of the proposed system on different types of hospital data in the context of head and neck cancer, the foreseeable future work will envision the direct implementation of the present system clinically, through the means of a mobile device application to be used by doctors and speech therapists alike.

6. CONCLUSIONS

This paper investigated the reliability of an automatic predictor of speech intelligibility based on speaker embeddings on hospital data recorded in ecological conditions, obtained from patient interviews in clinical appointments. Different assessments were conducted using different recorded speech tasks in a multi-task learning methodology. The results suggested a good generalization ability, illustrated by correlations (Spearman) as high as 0.891 and errors as low as 1.322 of the system on the different tasks. The metrics obtained on the spontaneous speech task are not only comparable to the other tasks, but also open up the possibility of a more deliberate use of this type of task in automatic assessments, a topic currently under-explored.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Alexander de Graeff, Rob J. de Leeuw, Wynand J.G. Ros, Gert-Jan Hordijk, Geert H. Blijham, and

Jacques A.M. Winnubst, "Long-term quality of life of patients with head and neck cancer," *The Laryngoscope, Volume 110, Issue 1*, 2000.

[2] Soren Fex, "Perceptual evaluation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 6, no. 2, pp. 155–158, 1992.

[3] Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain, "A comparative study of adaptive, automatic recognition of disordered speech," *Proceedings of Interspeech*, 2012.

[4] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, Julien Tardieu, and Cynthia Magnen, "Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss," *Journal of Speech, Language and Hearing Research, Volume 50(1)*, vol. 60(9), pp. 2394–2405, 2017.

[5] Sebastião Quintas, Julie Mauclair, Virginie Woisard, and Julien Pinquier, "Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer," *Proceedings of Interspeech*, 2022.

[6] Li Bin, Matthew C. Kelley, Daniel Aalto, and Benjamin V. Tucker, "Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks," *International Congress of Phonetic Sciences (ICPHs')*, 2019.

[7] Imed Laaridh, Corinne Fredouille, Alain Ghio, Muriel Lalain, and Virginie Woisard, "Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers," *Proceedings of Interspeech*, 2018.

[8] Sebastião Quintas, Julie Mauclair, Virginie Woisard, and Julien Pinquier, "Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer," *Proceedings of Interspeech*, 2020.

[9] Soroush Zargarbashi and Bagher Babaali, "A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language," *arXiv:1910.00330*, 2019.

[10] Juan M. Perero Codosero, Fernando Espinoza-Cuadros, Javier Antón-Martín, Miguel A Barbero-Alvarez, and Luis A. Hernández Gómez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14(2), pp. 240–250, 2019.

[11] Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen, "Non-intrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26(10), pp. 1925–1939, 2018.

[12] Mathias B. Pedersen, Morten Kolbæk1, Asger H. Andersen, Søren H. Jensen, and Jesper Jensen, "End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks," *Proceedings of Interspeech*, 2020.

[13] R.P. Clapham, L. van der Molen, R.J.J.H. van Son, M. van den Brekel, and F.J.M. Hilgers, "Nki-ccrt corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy," *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[14] Virginie Woisard, Corine Astésano, Mathieu Balaguer, Jérome Farinas, Corinne Fredouille, Pascal Gaillard, Alain Ghio, Laurence Giusti, Imed Laaridh, Muriel Lalain, Benoit Lepage, Julie Mauclair, Olivier Nocaudie, Julien Pinquier, Gilles Pouchoulin, Michèle Puech, Danièle Robert, and Vincent 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*, vol. 55, pp. 173–190, 2020.

[15] Richard Dufour, Yannick Estève, and Paul Deléglise, "Characterizing and detecting spontaneous speech: Application to speaker role recognition," *Speech Communication*, vol. 56, pp. 1–18, 2014.

[16] Corinne Fredouille, Alain Ghio, Imed Laaridh, Muriel Lalain, and Virginie Woisard, "Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers," *International Congress of Phonetic Sciences (ICPhS)*, 2019.

[17] Mathieu Balaguer, Timothy Pommée, Jérôme Farinas, Julien Pinquier, Virginie Woisard, and Renée 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*, vol. 42(1), pp. 111–130, 2019.

[18] Tove B. Lagerberg, Katarina Holm, Anita McAllister, and Sofia Strömbergsson, "Measuring intelligibility in spontaneous speech using syllables perceived as understood," *Journal of Communication Disorders, Volume 92*, vol. 92, 2021.

[19] Nathalie Boonen, Hanne Kloots, Pietro Nurzia, and Steven Gillis, "Spontaneous speech intelligibility: Early cochlear implanted children versus their normally hearing peers at seven years of age," *Journal of Child Language*, vol. 50(1), pp. 78–103, 2023.

[20] Mathieu Balaguer, Anth Boisguérin, Aanth Galtier, Nanth Gaillard, Manth Puech, and Virginie Woisard, "Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer," *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 136(5), pp. 355–359, 2019.

[21] Anthony Larcher, Ambuj Mehrish, Marie Tahon, Sylvain Meignier, Jean Carrive, David Doukhan, Olivier Galibert, and Nicholas Evans, "Speaker embeddings for diarization of broadcast data in the allies challenge," *Proceedings of ICASSP*, 2021.

[22] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv:2106.04624*, 2021.

[23] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embedings for speaker recognition," *Proceedings of ICASSP*, 2018.

[24] "X-Vector speaker embeddings extractor," `https://huggingface.co/speechbrain/spkrec-xvect-voxceleb`, Accessed: 20-02-2023.

[25] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A largescale speaker identification dataset," *Proceedings of Interspeech*, 2017.

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proceedings of Interspeech*, 2018.

[27] Jyh-Min Cheng and Hsiao-Chuan Wang, "A method of estimating the equal error rate for automatic speaker verification," *Proceedings of ISCSLP*, 2004.

[28] "ECAPA TDNN speaker embeddings extractor," `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`, Accessed: 20-02-2023.

[29] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proceedings of Interspeech*, 2020.

[30] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," *Proceedings of Interspeech*, 2018.

[31] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," *Proceedings of Interspeech*, 2015.

[32] Mathieu Balaguer, *Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé*, Ph.D. thesis, Université Paul Sabatier - Toulouse III, 2021.

[33] Alain Ghio, Gilles Pouchoulin, Bernard Teston, Serge Pinto, Corinne Fredouille, and et al, "How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?," *Speech Communication*, vol. 54, pp. 664–679, 2012.

[34] Muriel Lalain, Alain Ghio, Laurence Giusti, Danièle Robert, Corinne Fredouille, and Virginie Woisard, "Design and development of a speech intelligibility test based on pseudowords in french: Why and how?," *Journal of Speech, Language and Hearing Research*, vol. 63(7), pp. 2070–2083, 2020.

[35] Marc De Bodt, Maria E. Huici, and Paul Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, vol. 35(3), pp. 283–292, 2002.