INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Automatic Prediction of Speech Intelligibility based on X-vectors in the context of Head and Neck Cancer

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

[1]IRIT, Université de Toulouse, CNRS, Toulouse, France
[2]CHU Larrey, Oncopole, Toulouse, France

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

Abstract

In the context of pathological speech, perceptual evaluation is still the most widely used method for intelligibility estimation. Despite being considered a staple in clinical settings, it has a well-known subjectivity associated with it, which results in greater variances and low reproducibility. On the other hand, due to the increasing computing power and latest research, automatic evaluation has become a growing alternative to perceptual assessments. In this paper we investigate an automatic prediction of speech intelligibility using the *x-vector* paradigm, in the context of head and neck cancer. Experimental evaluation of the proposed model suggests a high correlation rate when applied to our corpus of HNC patients ($p = 0.85$). Our approach also displayed the possibility of achieving very high correlation values ($p = 0.95$) when adapting the evaluation to each individual speaker, displaying a significantly more accurate prediction whilst using smaller amounts of data. These results can also provide valuable insight to the redevelopment of test protocols, which typically tend to be substantial and effort-intensive for patients.

Index Terms: speech intelligibility, automatic speech processing, speaker embeddings, head and neck cancer

1. Introduction

Speech disorders, such as dysarthria or dysphonia, are usually associated with an underlying medical condition. These disorders can affect multiple components of speech (respiration, articulation, phonation, etc.) and can cause different sorts of speech impairment. Depending on the underlying condition, there are several methods and protocols that assess the overall speech ability of the patient. These protocols can be associated to a specific condition, such as Parkinson or amiotrophic lateral scleriosis [1], or can involve a more generalist approach [2]. Head and neck cancer (HNC) has major functional repercussions on the upper aerodigestive tract (breathing, swallowing, and phonation/speech). Due to this, a functional impairment at the level of communication is likely to appear, impacting the speech-related quality of life. As a result, perceptual evaluation has long been the most used method of disordered speech assessment. On the other hand, perceptual evaluation is usually time-consuming, biased and variant, since the evaluation can depend, amongst other facts, on the previous assessments that the health practitioner performed, affecting reproducibility [3]. Hence, the reliability of perceptual evaluations is mostly listener-dependent [4].With the increasing rate of oropharyngeal cancer incidence and the interjudge/intrajudge variance, the development of an automatic assessment that is able to output unbiased intelligibility measures becomes relevant [5, 6].

Loss of intelligibility is commonly found in the post-treatment of conditions that affect the vocal tract, such as HNC, and also in neurodegenerative diseases with dysarthria symptoms. An early diagnosis is usually correlated to a better prognosis, as a result of a progressive and timed implementation of post-treatment measures [7]. To better address the subjectivity and bias of intelligibility scores, automatic assessments have also been seen as a more objective and reproducible alternative. In the literature, one can distinguish two different approaches concerning automatic prediction of intelligibility measures. The first is based on the extraction of an intelligibility score as the result of the word error rate achieved by automatic speech recognition [8]. The second approach aims to extract relevant features from pathological speech by using automatic speech processing technologies, and then output a predicted intelligibility score [9].

Speaker embedding representations, such as *i-vectors*, have proven to represent well speaker characteristics [10]. In [11], we can see approaches based on *i-vectors* that aim to predict dysarthric speech evaluation metrics like intelligibility, severity and articulation impairment. In [12], the speaker embedding paradigm was applied to the specific case of intelligibility in the same HNC corpus. The study used the recording of a vast list of pseudo-words and addressed intelligibility as the phonetic distance between the original and the perceived word. The automatic measures achieved high correlation values in terms of intelligibility prediction (between 0.7 and 0.9) when compared to the reference values. Proposed by Snyder, *x-vectors* [13] are discriminative DNN speaker embeddings that have outperformed *i-vectors* in tasks such as speaker and language recognition [14, 15]. Recent advances suggest that *x-vectors* have been successfully applied to paralinguistic tasks such as emotion recognition [16], and to the detection of diseases like Obstructive Sleep Apnea [17] and Alzheimer's [18]. Following the line of research present in [11] and [12], we investigate the reliability of using *x-vector* speaker embeddings as features for automatic intelligibility prediction in the context of HNC. We perform comparisons between the automatic predicted scores and the perceptual evaluation issued by the professional assessment of dysarthria and healthy speakers.

The rest of this paper is organized as follows. Section 2 explains the methodology used, emphasizing the *x-vector* extraction network as well as the shallow neural network used. Section 3 presents the experiments performed on the French Head and Neck Speech Corpus (C2SI) and respective results. Section 4 presents the discussion as well as some perspectives on the results achieved. Finally, section 5 displays our conclusions and provides a few suggestions for future work.

http://dx.doi.org/10.21437/Interspeech.2020-1431

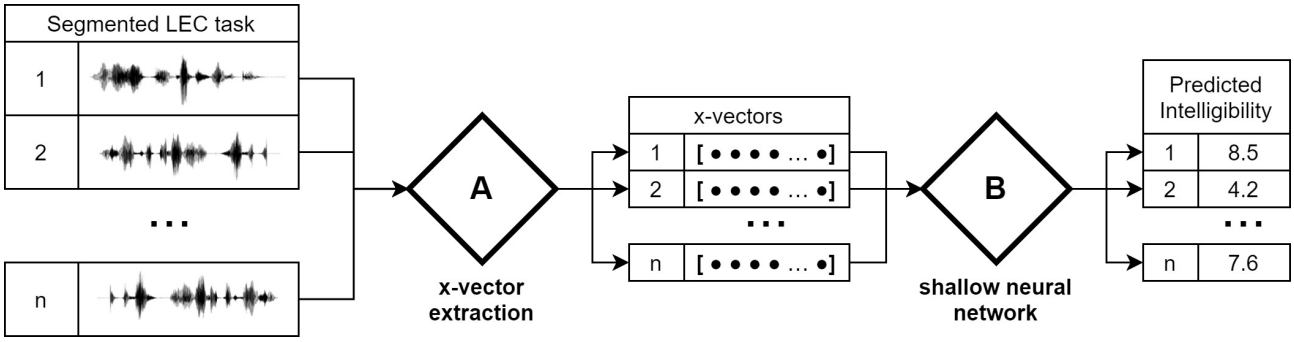

Figure 1: *Global Overview of the proposed system. The x-vectors are extracted from the segmented parts of a reading passage task (LEC), and then fed to a shallow neural network that regresses an intelligibility score.*

2. Methodology

The proposed methodology relies on two steps (see figure 1). The first one corresponds to the extraction of the *x-vector* speaker embeddings, in order to obtain a fixed-length representation of every speaker's utterance. In our specific case, we used the segmented recordings of a reading text task (LEC) which can be found fully described in 3.1. The second step relies on the regression task of predicting an intelligibility score based on the embedding representations. In order to do so, a shallow neural network was modeled to fit the data.

Both stages are fully explained in sections 2.1 and 2.2.

2.1. X-vector extraction

As mentioned before, *x-vectors* are DNN speaker embeddings that have seen a growing use in speaker recognition and paralinguistic tasks [16]. While *i-vectors* represent the total variability subspace of a channel or speaker, *x-vectors* aim to represent discriminative features between speakers. The comparison of both embeddings suggests that *x-vectors* require shorter temporal segments to achieve good results, and have been shown to be more robust to data variability and domain mismatches [13].

In order to extract *x-vectors*, we used the open source implementation present in the Kaldi toolkit[1]. The complete description of the extraction network can be found in [13]. Table 1 presents the outline of the DNN configuration used. Assuming a given speech signal has a total of *N* frames, the first five layers operate on speech frames with a small temporal context centered at the current frame *t*, building on temporal context of the previous layers. The statistics pooling layer aggregates all *N* frame-level outputs from layer *frame5* and computes its mean and standard deviation, which are then concatenated and propagated through segment-level layers, and finally to the softmax output layer. After training, embeddings are extracted from the affine component of layer *segment6*. The total dimension of each *x-vector* is 512. All long silences and noise bits were removed from the input audio files.

2.2. Shallow Neural Network

As previously stated, to predict an intelligibility score based on the embedding representations, a shallow neural network was modeled to fit our data. Only fully-connected layers (*fc-layers*) were used in our case. Figure 2 presents the proposed dimensions for the used network.

[1] https://github.com/kaldi-asr/kaldi

Table 1: *X-vector extraction DNN outline.*

Layer	Layer context	Input x Output
Frame 1	$\{t-2, t+2\}$	120×512
Frame 2	$\{t-2, t, t+2\}$	1536×512
Frame 3	$\{t-3, t, t+3\}$	1536×512
Frame 4	t	512×512
Frame 5	t	512×1500
Stats pooling	$[0, N]$	$1500T \times 3000$
Segment 6	0	3000×512
Segment 7	0	512×512
Softmax	0	$512 \times N$

Table 2: *Proposed shallow neural network outline.*

Layer	Input x Output
Input	512×128
fc-1	128×64
fc-2	64×1

3. Experiments and results

The present section displays the experiments performed and results achieved. On 3.1 we introduce the HNC corpus used, 3.2 presents the data augmentation scheme applied, 3.3 illustrates the training protocol, and finally sub-section 3.4 displays the evaluation scores.

3.1. C2SI Corpus

The present work is based on the French head and neck cancer speech corpus C2SI [19]. The corpus includes patients that suffer oral cavity or oropharyngeal cancer and also healthy speakers. All cancer patients have undergone at least one cancer treatment, such as surgery, radiotherapy and/or chemotherapy. All of the speakers were asked to record a different set of spoken tasks such as sustained vowels, picture description, spontaneous speech, passage reading and isolated pseudo-words.

In this study, the main focus of attention was set towards the passage reading task (LEC). In the context of the C2SI corpus, all speakers were asked to read the 1st paragraph of "La chèvre de M. Seguin", a tale by Alphonse Daudet that was chosen due to being long enough to include all French phonemes. This passage is also well known and widespread in French clinical phonetics [20].

For each speaker, the mean intelligibility and severity were computed based on the independent perceptual evaluation of 6 different health professionals. Each speaker was given a score between 0 and 10, the smaller the value, the less intelligible the speech is. The recordings were later segmented into 8 different segments of similar lengths, which can be found marked in the text. A total of 105 speakers, 84 patients and 21 controls, were used in this study.

The full LEC task is as follows, cut into segments: (S_1) *Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres.* (S_2) *Il les perdait toutes de la même façon.* (S_3) *Un beau matin, elles cassaient leur corde,* (S_4) *s'en allaient dans la montagne, et là-haut le loup les mangeait.* (S_5) *Ni les caresses de leur maître* (S_6) *ni la peur du loup rien ne les retenait.* (S_7) *C'était paraît-il des chèvres indépendantes* (S_8) *voulant à tout prix le grand air et la liberté.*

3.2. Data augmentation

In order to increase the training data available, a data augmentation scheme based on temporal distortion was implemented. Speed and tempo distortions have long been a reliable augmentation scheme applied to Automatic Speech recognition [21]. In our case, we employed a tempo distortion to all utterances of the training set. Tempo distortion is fairly similar to a speed perturbation, however, ensures that the pitch and spectral envelope of the signal remain the same [22]. We performed the augmentations by a factor of 0.9 and 1.1, where a factor of 1 corresponds to the original signal. A speech therapist listened to a subset of the augmented data in order to validate that there was no perceptual variation of intelligibility. Therefore, the same target values for intelligibility were used as labels for the augmented data.

3.3. 5-Fold Cross-Validation

A 5-fold cross-validation scheme was implemented in order to train the shallow neural network. At each fold, 84 speakers (patients and controls) were used for training and the remaining 21 unseen speakers were used for testing. Data augmentation was performed at every training fold. For each fold, the shallow neural network was trained during a total of 15 epochs using an exponential learning rate decay. Batch normalization and a dropout rate of 25% were applied on every layer.

3.4. Evaluation scores

In order to evaluate the resulting predictions, we evaluated our system on two metrics: Spearman's correlation (p), as the target intelligibility values were far from being normally distributed, and Root Mean Squared Error (RMSE). The scores were computed using the perceptual values mentioned in 3.1 as reference. The intelligibility perceptual assessment can be found fully described in [19].

The first intelligibility prediction experiment that we performed made use of all the 8 speaker's segments. In this case, the *x-vectors* were extracted, fed to the shallow neural network and finally paired to a predicted intelligibility value. The final score for each speaker was computed as the average score of each speaker's 8 segments. This analysis promoted more training data and a more granular analysis at sentence level, which will be explained further. Figure 2 depicts the predicted intelligibility values compared to the professional perceptual assessment. The correlation values achieved are consistent with the ones found in previous studies such as [12], which achieved correlation values between 0.75 and 0.84. However it is important to state that the perceptual intelligibility measures, in our case, are far more subjective due to being rated by health professionals instead of naive listeners.

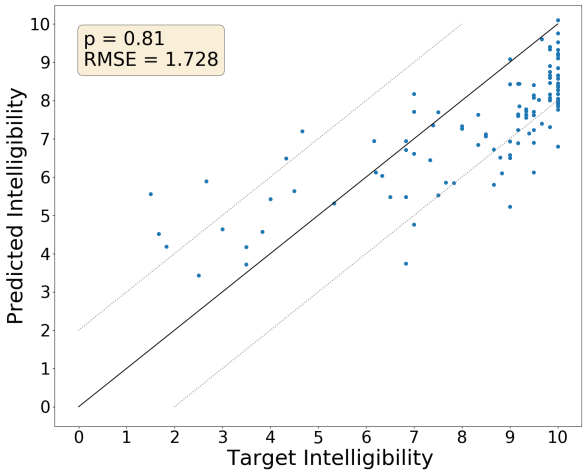

Figure 2: *Intelligibility prediction plot, using the average score of each speaker's segments.*

Outliers were accounted for speakers that had a predicted score outside a $[-2, T, 2]$ boundary, where T stands for the target value. From the 105 speakers, a total of 6 outliers were found above bounds and 21 below bounds, two of which were controls.

4. Discussion

4.1. Best segment scores

From the results achieved using the average score of 8 segments, we noticed that, in the majority of cases, there was a large variance within the individual scores of each speaker. Due to this, one can notice that there are sentences able to convey a much more precise intelligibility estimation. We further investigated this aspect by manually choosing, for each speaker, the segment that had the predicted value closest to target. The RMSE and p values were computed. The results can be found in table 3, paired with the results of choosing the worst segment as well.

The resulting values suggest that, for each speaker, there are segments that are able to convey a highly precise intelligibility measure, generally displaying a very high correlation value and a low root mean squared error. This points to a deeper analysis of those segments, showing that we are able to achieve very high correlation values by manually identifying utterances that fit each speaker best. The resulting best segments were further assessed. Although no clear preference was found towards a specific sentence, from the sub-list of best segments, number 2 and number 6 were the ones with larger representativeness, accounting for 22% and 15% of all cases respectively.

4.2. Choosing the best segment

As we have seen before, some of the segments were able to convey a much more accurate intelligibility measure than others. Due to this, it becomes relevant to devise a way to detect each speaker's most relevant sentence. While the subset of speaker's best segments displayed no clear preference towards a specific

one, we implemented a simple decision criterion based on external features, which points to a more accurate segment choice.

Table 3: *Results achieved by manually choosing, for each speaker, the best/worst segment scores.*

	p	RMSE
Best segment	0.95	0.900
Worst Segment	0.53	2.224

From the subset of speakers that had segments 2 and 6 as the most accurate ones, we analyzed the number of recognized phonemes that the speakers from each subset had. Word and phoneme error rates have long been used as an automatic way to assess speech intelligibility [8]. From this analysis, we found that segment 2 was associated with a clear above average recognition rate, while segment 6 was clearly below average. We believe this may be due to the larger presence of nasalised vowels in segment 6, which typically presents articulation issues for oropharyngeal cancer patients [23]. Afterwards, we correlated all speaker predictions from each segment group with the target values, and confirmed that segments 2 and 6 presented the highest correlation values when compared to the remaining ones. This was expected since from the subset of best segments, those two were the ones with larger representativeness. Interestingly, by simply using segment 2, the correlation/RMSE pair obtained (p: 0.82, RMSE: 1.434) was slightly better than the average score displayed in 3.4. Furthermore, we devised a simple decision tree that chooses the best speaker's segment based on the number of recognized phonemes by a kaldi-based phoneme recognizer. For speakers with or more than 167 overall recognized phonemes, segment 2 was assigned, while for the counterpart segment 6 was chosen. The split value mentioned corresponds to the mean value of recognized phonemes for the C2SI patients. The results using this criterion, present in table 4, suggest an improvement in correlation and a total decrease of 0.34 on the root mean squared error. *I-vectors* were obtained through the pretrained model [24], which served as baseline.

Table 4: *Comparison between the scores previously obtained and the decision tree criterion implemented.*

		p	RMSE
i-vectors	Averaged Scores	0.72	2.121
x-vectors	Averaged Scores	0.81	1.728
	Only Segment 2	0.82	1.434
	Decision Tree	0.85	1.389

4.3. Perspectives

The results presented suggest that by using the *x-vector* paradigm, we are able to obtain reliable intelligibility predictions with a given combination of individual segment scores. Moreover, when identifying the best segment for each speaker, a very high correlation value can be achieved, and the RMSE decreases to almost half of the value achieved in the averaged approach.

Concerning the results of the perceptual assessment, the intelligibility measures used suffered, in some cases, from very high variance within the same speaker, reaching standard deviations of up to 3 (on a 0-10 scale), pinpointing the large inter-class variance present in this type of clinical assessments [4]. This aspect points out the subjectivity of the intelligibility scores used, when compared to the more objective ones found previously in the literature [12], which were rated by naive listeners instead of health professionals. When comparing the usage of *i-vectors* with the *x-vector* paradigm, we can conclude that the latter does not rely on larger amounts of data to output better results [13]. This aspect was evident by analyzing the scores of only segment 2, present in 4.2, which were slightly better than the averaged approach described in 3.4. This can provide interesting cues to the development of less extensive and more precise batteries of exams, as the majority of the assessments are substantial and require much effort from both patient and therapist. A more precise and targeted assessment would strongly diminish the battery of exams required.

Regarding the outliers mentioned in 3.4, since only 12.4% of the total number of speakers had a target intelligibility score below 5, it was expected that the system would perform with larger margins of error in this specific context. This was the case for the 6 outliers predicted above bounds. Concerning the 23 outliers that were below bounds, half of them were found to have a tumor in the amygdala region. Tumors in this location are typically associated with changes in articulation of fricatives and stop consonants [25]. While perceptually, a larger vowel presence is usually correlated to an improvement in speech understandability [26], we introduce the hypothesis that in the context of an automatic assessment, a larger consonant presence may be related to a more accurate intelligibility score. When the decision tree criterion was applied, the number of lower outliers drastically reduced from 23 to 5. The usage of this simple decision method in the system promoted a correlation increase and a decrease in error. However it still leaves room for improvement. A deeper analysis on speaker individual features and speaker-specific phonetic content could provide a valuable insight to detect specific words and phonemes that are able to convey a more accurate intelligibility estimation.

5. Conclusions

This paper investigated an automatic approach for intelligibility estimation based on *x-vectors* and shallow neural networks. This approach was devised for the segmented parts of a phonetically rich passage, in the context of HNC. When using the average of all passage segments, a high correlation value of 0.81 was achieved, showing that *x-vectors* can indeed convey intelligibility measures, similarly to the *i-vector* paradigm. However they require smaller amounts of speech data, as it was evident in the single segment analysis devised. When choosing the segments that are closest to target, we achieved a very high correlation value of 0.95, pointing out the importance of selecting the sentences used in this automatic assessment to each speaker. Moreover, we devised a simple criterion to choose the speaker's best segment based on the statistics obtained and the number of recognized phonemes. The results suggest a correlation value of 0.85 and a total decrease of 0.34 on the root mean squared error. While this criterion promotes a better correlation value than the average score, it displays the relevance of an affecting pathology based way to detect the best sentence for each speaker. Future work will investigate this aspect, and a more granular phonetic analysis of the textual content used.

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] S. Zargarbashi and B. Babaali, "Dysarthria in amyotrophic lateral sclerosis: A systematic review of characteristics, speech treatment, and augmentative and alternative communication options," *Journal of Medical Speech-Language Pathology*, vol. 3, no. 19, pp. 12–30, 2011.

[2] A. Lowit and R. D. Kent, "Assessment of motor speech disorders," *Plural Publishing*, vol. 1, 2010.

[3] 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.

[4] L. Plisson, C. Pillot-Loiseau, and L. Crevier-Buchman, "Intelligibilité de la parole après le traitement d'uncancer de l'oropharynx: étude descriptive chez sept patients en pré-traitement et en post-traitement précoce," *7èmes Journées de phonétique clinique (JPC), Laboratoire de Phonétiqueet Phonologie, hôpital Européen G. Pompidou, Paris, France*, 2017.

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

[6] C. Middag, *Automatic analysis of pathological speech*. Doctoral Dissertation: Ghent University, Department of Electronics and information systems, Ghent, Belgium, 2012.

[7] B. P. Leifer, "Early diagnosis of alzheimer's disease: Clinical and economic benefits," *Journal of the American Geriatrics Society*, 2003.

[8] 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.

[9] 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.

[10] P. Verma and P. Das, "i-vectors in speech processing applications: A survey," *International Journal of Speech Technology*, 2015.

[11] I. Laaridh, W. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," *Proceedings of Interspeech*, 2017.

[12] 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.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embedings for speaker recognition," *Proceedings of ICASSP*, 2018.

[14] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proceedings of Interspeech*, 2017.

[15] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," *Proceedings of Interspeech*, 2018.

[16] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *Proceedings of ICASSP*, 2020.

[17] J. M. P. Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Alvarez, and L. A. H. Gómez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

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

[19] C. Astésano, M. Balaguer, J. Farinas, C. Fredouillec, A. Ghio, P. Gaillard, L. G. I. Laaridh, M. Lalain, B. Lepage, and et al., "Carcinologic speech severity index project: A database of speech disorder productions to assess quality of life related to speech after cancer," *Language Resources and Evaluation Conference*, 2018.

[20] A. Ghio, G. Pouchoulin, B. Teston, S. Pinto, C. 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.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *Proceedings of Interspeech*, 2015.

[22] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," *Proceedings of Interspeech*, 2018.

[23] I. Jacobi, M. A. van Rossum, and L. van der Molen, "Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy," *The annals of Otology, Rhinology and Laryngology*, 2013.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011. [Online]. Available: https://kaldi-asr.org/models/m7

[25] J. A. Logemann, B. R. Pauloski, A. W. Rademaker, and J. Johnson, "Speech and swallow function after tonsil/base of tongue resection with primary closure," *Journal of speech and hearing research*, 1993.

[26] D. Kewley-Portand, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *The Journal of the Acoustical Society of America*, 2007.