Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Relationship between objective and subjective perceptual measures of speech in individuals with head and neck cancer

Bence Mark Halpern[1], *Thomas Tienkamp*[2,3], *Teja Rebernik*[4,5], *Rob J.J.H. van Son*[6], *Martijn Wieling*[2], *Defne Abur*[2], *Tomoki Toda*[1]

[1]Nagoya University, Japan, [2]CLCG, University of Groningen, the Netherlands, [3]University Medical Center Groningen, the Netherlands, [4]LPP, CNRS/Sorbonne Nouvelle, France, [5]BCLS, Vrije Universiteit Brussel, Belgium, [6]Department of HNO, Netherlands Cancer Institute, the Netherlands

halpern.bence.e8@f.mail.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Meaningful speech assessment is vital in clinical phonetics and therapy monitoring. This study examined the link between perceptual speech assessments and objective acoustic measures in a large head and neck cancer (HNC) dataset. Trained listeners provided ratings of intelligibility, articulation, voice quality, phonation, speech rate, nasality, and background noise on speech. Strong correlations were found between subjective intelligibility, articulation, and voice quality, likely due to a shared underlying cause of speech symptoms in our speaker population. Objective measures of intelligibility and speech rate aligned with their subjective counterpart. Our results suggest that a single intelligibility measure may be sufficient for the clinical monitoring of speakers treated for HNC using concomitant chemoradiation.

Index Terms: pathological speech, perceptual measure, intelligibility, accent

1. Introduction

Meaningful assessment of speech through acoustic measures is important both in medical decision making and in clinical phonetics. In the decision making, the measures obtained through speech assessment directly influences speech therapy monitoring and planning. In clinical phonetics, these measures are crucial for reproducible research.

Assessing speech, whether in clinical or research settings, typically involves two types of measures: subjective (perceptual) evaluation and objective (computational) evaluation. On the one hand, subjective evaluations rely on trained listeners rating various aspects of speech, for example, intelligibility (here: the degree to which speech is understood) or phonation (here: accurate voicing distinction). However, these evaluations can be time-consuming, require trained raters, and may be influenced by biases such as listener familiarity [1] and professional experience [2]. On the other hand, objective evaluations use computational methods and algorithms to analyse speech signals and derive quantitative measures. These methods offer the potential for automated, consistent, and rapid assessment. However, a persistent challenge in objective evaluation, besides the lack of interpretability of most methods, is ensuring that the chosen metrics mimic human perception and are clinically relevant.

A common practice in developing objective speech measures is to focus on a single perceptual dimension, such as intelligibility. The objective speech measure is then validated by showing a strong correlation between this measure and subjective perceptual ratings [3, 4]. The goal is often to develop measures specific to either the articulatory subsystem (e.g., articulation) or the laryngeal subsystem (e.g., phonation, voice quality). The articulatory subsystem contains the structures and processes that shape the vocal tract for speech (e.g., tongue, lips, jaw), while the laryngeal subsystem contains the structures and processes that generate the voice source (e.g., vocal fold vibration). However, a high correlation between an objective speech measure and a perceptual rating does not guarantee that the measure specifically captures the intended percept. Perceptual dimensions corresponding to these different subsystems, for example, articulatory clarity and voice quality, are often correlated due to a common underlying factor, such as overall speech disorder severity [5]. These interdependencies between perceptual dimensions can lead to correlations between their respective ratings, even though voicing and articulation problems are attributed to distinct parts of the speech production apparatus. Furthermore, other factors not related to the produced speech can also play a role. For example, noise [6] has also been shown to have an impact on intelligibility ratings.

To this end, our present study investigates a range of perceptual measures across the articulatory and laryngeal subsystems: intelligibility, articulation (clarity of pronunciation), voice quality (overall vocal characteristics), phonation, rate (rate of speech), nasality (resonance in the nasal cavity), and the presence of background noise (extraneous sounds). Specifically, we investigated whether these perceptual measures correlate with objective measures and each other on longitudinal audio recordings from 53 Dutch individuals with head and neck cancer (HNC). This represents 2% of the annual Dutch HNC population [7]. We investigated the following research questions:

RQ1 What is the correlation between commonly used perceptual measures of speech with each other?

RQ2 How well can these subjective evaluation measures be predicted by objective evaluation measures? (as measured by Pearson's correlation)

2. Related works

This section reviews previous studies that investigated the relationships between multiple perceptual measures, as well as those that focused on the impact of one perceptual measure to another. However, note that most of these are studies on non-HNC populations. Due to limited space, we are not able to give a full account of the objective evaluation measures, and the reader is referred to the review of [8].

2.1. Interrelationships between perceptual measures

Tu and colleagues conducted a study comparing a number of perceptual measures (articulatory precision, nasality, vocal quality, severity, and prosody) [5]. Their study included 32 speakers with dysarthria rated by 15 second-year speech-

10.21437/Interspeech.2025-1127

language therapy master students. The lowest correlation was between vocal quality and nasality ($r = 0.69$), the highest correlation was between vocal quality, articulation precision, and severity (all of them $r = 0.91$), showing strong interrelatedness between measures overall. In another study, the De Bodt et al. [9] showed a strong correlation between articulatory precision and intelligibility ($r = 0.82$), and a moderate correlation between voice quality and intelligibility ($r = 0.46$), and weak correlation between nasality and intelligibility ($r = 0.32$) in speakers with dysarthria ($n = 79$).

2.2. Impact of perceptual measures on intelligibility

Speech rate: In the case of typical speech, faster speech is usually more difficult to understand [10]. For pathological speech, the relationship is a bit different. For example, for speakers with dysarthria, an atypical (either too fast or too slow) speech rate is often used as a diagnostic criterion [11]. For certain speakers with dysarthria who speak too fast, teaching them to speak slowly improves intelligibility [12].

Articulatory precision: Articulatory precision and intelligibility have recurrently been positively associated with each other. Apart from De Bodt et al. [9] results mentioned above, Thompson and Kim [13] also reported strong correlations ($r = 0.9$) between intelligibility and articulatory precision in 40 speakers with and without dysarthria.

Nasality: The literature reports varying correlations between intelligibility and nasality ratings. McWilliams at al. [14] found a high correlation between nasality and intelligibility ratings ($r = 0.72$) of 48 cleft-palate patients rated by seven listeners. In Cantonese children with cleft palate, no significant correlation was established neither for nasal ($r = -0.38$), nor for non-nasal sentences ($r = -0.41$)[15].

Noise: A common finding is that extraneous noise negatively affects intelligibility. Depending on the type of noise, there are some intricacies to this effect, i.e., noise with linguistic component in the same language (e.g., babble noise) is more detrimental to intelligibility compared to noise without linguistic content (e.g., white noise) [16]. Studies on speakers with dysarthria [17] or HNC [6] have both shown that babble noise impacts pathological speech more than typical speech.

3. Dataset and Experimental Percepts

We use the NKI-SpeechRT dataset, introduced in [18]. The dataset contains speakers with mild-to-mid severity pre and post-treatment for HNC with concomitant chemoradiotherapy (CCRT). The dataset includes 55 speakers (45 male, 10 female; mean age = 57 years, range = 32-79 years), of whom 47 are native speakers of Dutch. The remaining 8 speakers are non-native speakers. The dataset includes recordings from a speaker at a maximum of five time points: before CCRT, ten weeks post-CCRT, and 12 months post-CCRT. The other two stages are unknown. The total number of combination of speakers and stages (e.g., `speaker1_pre`, `speaker1_post_12`, from now on: speaker-stages) are 141 (pre-CCRT 54, post-CCRT 87). There are no typical speakers in the dataset. For the current data analysis, 5 speaker-stages (2 pre-CCRT, 3 post-CCRT), and two speakers (both male) are excluded due to the inability to get forced alignments. In total 136 speaker-stages, and 53 speakers are included. Speaker-stages are used for the correlation experiments. The total recorded audio is ca. 4 hours.

Participants were asked to read the Dutch text 'De vijvervrouw' by Godfried Bomans. Recordings were made with a Sennheiser MD421 Dynamic Microphone and portable 24-bit digital wave recorder (Edirol Roland R-1). The audio samples were cut for surrounding silences during manual annotation, therefore no additional voice activity detection was applied. The speech samples were energy normalised to -10 dB, downsampled to 16 kHz and quantized to 16-bit PCM for the analysis.

3.1. Subjective measures

In a 70-minute online listening test, 14 Dutch recent speech language pathology graduates without any self-reported hearing difficulties rated the entire speech text cut into three segments of approximately equal lengths. The audio was presented at 70 dB using Sennheiser HD418 headphones. Each segment received a single rating from the 14 listeners, and the mean scores are used for the correlation experiments. All listeners rated all the stimuli/speakers.

Several dimensions, listed below, were rated simultaneously. The relevant experimental details here are reproduced based on [19, 20]. Ratings statistics and interrater correlations (`ICC2, K`) are in Table 1.

Intelligibility (INT): Listeners were asked to rate the speech intelligibility on a 7-point scale (1 = completely unintelligible, 7 = good). The listeners were able to check the text with the ability to replay the stimuli. This allowed them to more accurately judge the intelligibility.

Phonation (PHO): Listeners were asked to rate the degree to which phonation deviated from what they considered normal on a 5-point scale (1 = very deviant, 5 = normal).

Articulatory precision (AP): Listeners were asked to evaluate the general precision of vowel and consonant production as compared to normal running speech on a 5-point scale (1 = extremely imprecise, 5 = normal/precise). Precise articulation was defined as correct manner and place of production and clear coordination between sounds.

Perceived speed (SPEED): Listeners were asked to evaluate the speech rate on a 9-point scale (1 = slow, 5 = normal, 9 = fast).

Voice quality (VQ): Listeners rated the overall impression of voice quality on a 5-point scale (1 = severely deviated, 5 = normal). Listeners were explicitly asked to not rate pleasantness but rather the degree of voice deviation compared to normal voice.

Nasality (NAS): Listeners were asked to evaluate nasality on a 5-point scale (1 = very nasal, 5 = normal).

NOISE: A separate study was done with one expert phonetician (R.v.S.) who rated the noisiness of the recordings on a 3-point scale. Zero meant no or barely any audible noise, one meant audible noise, and two meant noisy, including sometimes other voices or ringing of the telephone. We decided to keep all the recordings even those rated very noisy for the further experimentations.

4. Objective measures

In this section, we introduce the objective methods that we compare to the perceptual measures. We categorised the objective measures based on what they are intended to measure. For the objective analysis, the texts were manually cut into 23 utterances. Individual ratings for the utterances were obtained with each method, and averaged to obtain a speaker-stage level score for the correlation analysis.

	NAS (1-5)	PHO (1-5)	SPEED (1-9)	AP (1-5)	INT (1-7)	VQ (1-5)	NOISE (0-2)
Mean ± Std	4.42 ± 0.29	3.89 ± 0.58	5.14 ± 0.80	3.97 ± 0.66	5.51 ± 0.97	4.32 ± 0.42	0.44 ± 0.46
Range [Min, Max]	$[2.63, 4.77]$	$[1.44, 4.69]$	$[3.36, 7.48]$	$[2.11, 4.86]$	$[2.31, 6.73]$	$[2.93, 4.85]$	$[0.00, 1.67]$
IQR	$[4.33, 4.58]$	$[3.67, 4.24]$	$[4.74, 5.63]$	$[3.66, 4.44]$	$[5.07, 6.19]$	$[4.19, 4.58]$	$[0, 0.67]$
ICC2,K	0.58	0.91	0.90	0.91	0.92	0.78	N/A

Table 1: *Summary statistics for different speech attributes. Note that the statistics are calculated on the averaged ratings, hence the decimals in min/max. IQR = interquartile range, ICC2,k = intra-class correlation. NAS = nasality, PHO = phonation, SPEED = speech rate, AP = articulatory precision, INT = intelligibility, VQ = voice quality, NOISE = recording noisiness.*

4.1. Intelligibility estimation methods

In the intelligibility estimation methods, we aimed to compare both reference-based and reference-free methods. The phoneme error rate needs written transcription of the audio (written reference), the neural acoustic distance needs audio reference and transcriptions (written and speech reference), and the XPPG-PCA does not need any reference.

Phoneme error rate (PER): To obtain prediction for the phonemes in the utterances, we used a Dutch phoneme recogniser[1] pre-trained on the Dutch Common Voice dataset [21]. All of our datasets had word-level transcriptions, which we converted to phoneme-level transcriptions using the Dutch `espeak` frontend of *phonemizer* [22].

Neural acoustic distance (NAD): NAD was initially proposed as a pronunciation evaluation distance measure [23]. As the *wav2vec-large* feature used by this distance measure has shown to be sensitive to pathological speech, too [24], we think it can work as an intelligibility measure. First, for each of utterance, we obtained word boundaries using the pre-trained `dutch_cv` acoustic model from the Montreal Forced Aligner (MFA) [25]. Then, speech features were extracted from utterances using layer 10 of the *wav2vec2-large* model, as this provided the best result in [23]. The segmentation on the MFA was applied on the wav2vec features to extract feature sequences for individual words in an utterance. For the distance calculation, each word (from now: target word) in an utterance was systematically compared against the same word from all the other speakers in the dataset (from now: reference words). These comparisons were performed using dynamic time warping to match the naturally varying word durations. The scores from the target words were then first averaged across all reference words, then the word-level scores in an utterance. Open source code available[2].

XPPG-PCA (PCX): XPPG-PCA is a novel method for evaluating speech severity, combining x-vectors and phonetic posteriorgrams. These features are extracted from each utterance, normalized, and concatenated. Principal component analysis (PCA) is then applied to this combined feature set, estimated on NKI-OC-VC [26], to identify dominant variations related to speech severity. The first principal component is then used to calculate a reference-free score for each utterance in a new dataset, reflecting the degree of deviation from typical speech patterns [27]. Open source code available[3].

[1] `https://huggingface.co/Clementapa/wav2vec2-base-960h-phoneme-reco-dutch`

[2] `https://github.com/Bartelds/neural-acoustic-distance`

[3] `https://github.com/karkirowle/xppg-pca`

4.2. Speed estimation methods

Speech rate $\mathbf{RATE_S}$: To calculate speech rate, we divided the total number of words in the transcription by the duration of the recording.

Articulation rate $\mathbf{RATE_A}$: Compared to speech rate, articulation rate excludes pauses. To estimate the total duration of the speech without pauses we use an energy-based voice activity detection, and consider all speech samples less than 20 dB under the peak as speech frames. The total duration of these speech frames is used as a proxy for the duration excluding pauses. The articulation rate is calculated as number of words in the transcription divided by the duration excluding pauses.

4.3. Noise estimation methods

$\mathbf{SNR_N}$: The NIST (National Institute of Standards and Technology) signal-to-noise ratio (SNR) estimation method [28] uses sequential Gaussian mixture estimation to model noise. It generates a short-time energy histogram, which is then used to determine the energy distributions of both the signal and noise, from which the SNR is calculated.

$\mathbf{SNR_W}$: Waveform Amplitude Distribution Analysis signal-to-noise-ratio (WADA-SNR) is a reference-free SNR estimation method. This method assumes that clean speech has a Gamma distribution while the additive noise is Gaussian [29].

5. Results

Please note that we have inverted the sign for the NAD, PER and the NOISE to make all measures the same directionality.

5.1. RQ1: Correlations between subjective measures

As there are a large number of correlations in Figure 1, we will only comment on the perceptual measure correlations with intelligibility (INT), as INT showed overall the strongest correlations. There was a very strong correlation of INT with VQ ($r = 0.92$) and AP ($r = 0.95$). The correlation between SPEED and INT was moderate and positive ($r = 0.38$), with faster speech being more understandable for the raters. Phonation had a weak correlation with intelligibility ($r = 0.25$). Noise showed only a weak correlation with intelligibility ($r = 0.21$). Nasality had a none-to-weak correlation with intelligibility ($r = 0.14$).

5.2. RQ2: How well objective measures predict subjective measures?

The objective intelligibility measures correlated well with INT. NAD achieved the best performance ($r = 0.9$), followed by PCX ($r = 0.83$), and finally PER ($r = 0.82$). SPEED showed a strong positive correlation with RATE_S ($r = 0.83$), and a moderate positive correlation with RATE_A ($r = 0.42$). The objective noise measures were moderately correlated with their

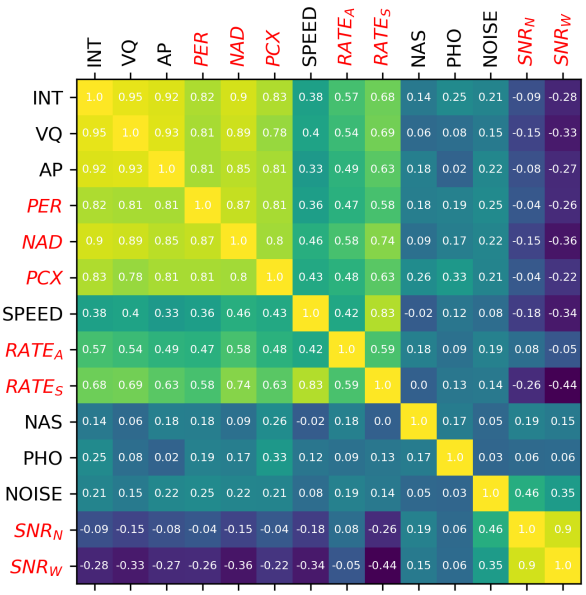


Figure 1: *Correlation matrix of the perceptual (subjective, in black) and objective measures (in red/italic).*

subjective counterpart, with the SNR_N showing a higher correlation ($r = 0.46$) than SNR_W ($r = 0.35$).

6. Discussion

6.1. RQ1: Correlations between subjective measures

Regarding RQ1, our results indicated strong correlations between the measures of intelligibility (INT), articulatory precision (AP), and voice quality (VQ). This indicates that, within this HNC speaker population, these aspects of speech, although originating from distinct speech motor subsystems, tend to deteriorate concurrently, most likely due to radiation treatment's effect on both the articulatory [30] and laryngeal [31] subsystems. This finding has several implications. First, for clinical practice, the strong correlation suggests that a sole intelligibility measure may be sufficient for clinical tracking in many speakers with HNC, and as shown by Tu et al. [5], in many speakers with dysarthria, too. Second, the strong correlations raise the risk of the 'common cause fallacy' during measure development. Developing a targeted measure, e.g., for articulation, needs validation on a population where the subjective measures are not correlated due to this common cause fallacy.

The moderate correlation of SPEED and INT was partially surprising as slower speech of typical speakers is generally easier to understand compared to faster speech [10]. However, sometimes individuals with speech difficulties have trouble reaching some articulatory targets, resulting in slower speech. Therefore, it could be that speakers who are more severely affected have to slow their speech to a greater extent compared to speakers who are less severely affected [32].

No strong correlations were found between INT and phonation (PHO), INT and nasality (NAS), and INT and NOISE. With nasality, the poor rater agreement could be one reason why correlation could not be established. The overall typical scores on nasality with low variances may also explain why no robust correlations were observed, i.e., likely the cohort did not have nasality issues. The phonation result also suggests that these voicing distinctions do not seem to influence intelligibility too much. For noise, the moderate correlation is due to a single listener rating the noise.

6.2. RQ2: How well objective do measures predict subjective measures?

Turning to RQ2, our results show that objective measures showed strong correlations with their subjective counterpart, with the exception of noise. For intelligibility, we found that the performance depended on the reference type used. The lower correlation of PCX and INT compared to NAD and INT can be attributed to the fact that PCX does not need a reference; the lower correlation between PER and INT than between NAD and INT indicates that acoustic references are likely better than written ones. The speed measures were different in their correlation, with RATE_S having a higher correlation. We would have expected RATE_A to have a higher correlation than the RATE_S due to the pauses influence on the perception but this was not the case.

In general, the strong correlations show that objective measures are promising for clinical use, offering a potentially more consistent and less subjective way to assess speech of individuals with HNC. This holds true even considering the inclusion of non-native speakers and the presence of some noisy samples in the dataset, which shows robustness of the objective methods. However, subjective measures of nasality (NAS) and phonation (PHO) still lack reliable correlations with objective methods. With nasality, the poor rater agreement could be one reason why it is challenging to develop such measures. In contrast, phonation showed excellent inter-rater agreement yet still no correlation was found between phonation and any objective method. This suggests that an objective measure is attainable with focused research effort.

6.3. Limitations and future work

Limitations of our study include only assessing individuals with HNC, and the lack of nasality and phonation measures. We could not include the nasal severity index for nasality in the current work as it requires sustained vowels, which was not available to us [33]. For the phonation measure, we are not aware of any existing method specifically looking at voicing distinctions. Developing nasality and phonation measures on running speech should be part of future aims.

Despite a great performance of the objective methods several key challenges remain. The most important is interpretability, i.e., both NAD and XPPG-PCA use neural network-based features that are not transparent enough for clinical practice. Another limitation is that all of our models are Dutch. It would be desirable to transition to language-independent models. Finally, all methods used read instead of spontaneous speech, which may be not representative of everyday speech use.

7. Conclusion

Our study found strong correlation between intelligibility, voice quality, and articulation in individuals with HNC which is consistent with previous findings in speakers with dysarthria. Objective measures showed promising predictive capabilities, particularly the NAD and XPPG-PCA methods, which effectively estimated intelligibility, voice quality, and articulatory precision. Future work should focus making current measures language independent and interpretable.

8. Acknowledgements

The data collection in the paper received ethical approval. The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden). This work was further supported by the Research School of Behavioral and Cognitive Neurosciences of the University of Groningen. This work is partly financed by the NWO under project number 019.232SG.011, and partly supported by a project, JPNP25006, commissioned by NEDO.

9. References

[1] S. Landa *et al.*, "Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding," *International Journal of Speech-Language Pathology*, vol. 16, no. 4, pp. 408–416, 2014.

[2] M. S. De Bodt *et al.*, "Test-retest study of the grbas scale: influence of experience and professional background on perceptual rating of voice quality," *Journal of Voice*, vol. 11, no. 1, pp. 74–80, 1997.

[3] P. Janbakhsi *et al.*, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6405–6409.

[4] C. Middag *et al.*, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–9, 2009.

[5] M. Tu *et al.*, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL416–EL422, 2016.

[6] T. L. Eadie *et al.*, "Effect of noise on speech intelligibility and perceived listening effort in head and neck cancer," *American Journal of Speech-Language Pathology*, vol. 30, no. 3S, pp. 1329–1342, 2021.

[7] IKNL, "Kankerregistratie," https://iknl.nl/kankersoorten/hoofd-halskanker/registratie/incidentie, 2023, accessed: 2025-01-25.

[8] B. M. Halper *et al.*, "Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners," *Speech Communication*, vol. 149, pp. 84–97, 2023.

[9] M. S. De Bodt *et al.*, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–292, 2002.

[10] A. Du *et al.*, "Effect of speech rate for sentences on speech intelligibility," in *2014 IEEE International Conference on Communiction Problem-solving*. IEEE, 2014, pp. 233–236.

[11] P. G. Blanchet and G. J. Snyder, "Speech rate deficits in individuals with parkinson's disease: A review of the literature," *Journal of Medical Speech-Language Pathology*, vol. 17, no. 1, pp. 1–7, 2009.

[12] G. Van Nuffelen, M. De Bodt, F. Wuyts, and P. Van de Heyning, "The effect of rate control on speech rate and intelligibility of dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 2, pp. 69–75, 2009.

[13] A. Thompson and Y. Kim, "Acoustic and kinematic predictors of intelligibility and articulatory precision in parkinson's disease," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 10, pp. 3595–3611, 2024.

[14] B. J. McWilliams, "Some factors in the intelligibility of cleft-palate speech," *Journal of Speech and Hearing Disorders*, vol. 19, no. 4, pp. 524–527, 1954.

[15] J. C. Chun and T. L. Whitehill, "The relationship between nasalance and nasality in cantonese children with cleft palate," *Asia Pacific Journal of Speech, Language and Hearing*, vol. 6, no. 3, pp. 135–147, 2001.

[16] S. H. Lee *et al.*, "Effects of various background noises on speech intelligibility of normal hearing subjects," *Korean Journal of Otorhinolaryngology-Head and Neck Surgery*, vol. 52, no. 4, pp. 307–311, 2009.

[17] M. J. McAuliffe, M. Schaefer, G. A. O'Beirne, and L. L. LaPointe, "Effect of noise upon the perception of speech intelligibility in dysarthria," in *Poster at the American Speech-Language-Hearing Association (ASHA) Convention*. University of Canterbury, Communication Disorders, 2009.

[18] B. M. Halpern and T. Toda, "Reference-free automatic speech severity evaluation using acoustic unit language modelling," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, 2024, pp. 1–5.

[19] R. P. Clapham, L. van der Molen, R. van Son, M. W. van den Brekel, F. J. Hilgers *et al.*, "Nki-ccrt corpus-speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy." in *LREC*, vol. 4. Citeseer, 2012, pp. 3350–3355.

[20] R. Clapham, C. Middag, F. Hilgers, J.-P. Martens, M. Van Den Brekel, and R. Van Son, "Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer," *Speech Communication*, vol. 59, pp. 44–54, 2014.

[21] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[22] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021. [Online]. Available: https://doi.org/10.21105/joss.03958

[23] M. Bartelds *et al.*, "Neural representations for modeling variation in speech," *Journal of Phonetics*, vol. 92, p. 101137, 2022.

[24] J. Cai, Y. Song, J. Wu, and X. Chen, "Voice disorder classification using wav2vec 2.0 feature extraction," *Journal of Voice*, 2024.

[25] M. McAuliffe *et al.*, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech 2017*, 2017, pp. 498–502.

[26] B. M. Halpern *et al.*, "Improving severity preservation of healthy-to-pathological voice conversion with global style tokens," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.

[27] ——, "XPPG-PCA: Reference-free automatic speech severity evaluation with principal components," *Under Review*, 2024.

[28] "The NIST Speech SNR Measurement," http://www.nist.gov/smartspace/nist_speech_snr_measurement.htm, accessed: 2025-01-25.

[29] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech 2008*, 2008, pp. 2598–2601.

[30] I. Jacobi *et al.*, "Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy," *Annals of Otology, Rhinology & Laryngology*, vol. 122, no. 12, pp. 754–762, 2013.

[31] S. Kraaijenga *et al.*, "Assessment of voice, speech, and related quality of life in advanced head and neck cancer patients 10-years+ after chemoradiotherapy," *Oral oncology*, vol. 55, pp. 24–30, 2016.

[32] T. B. Tienkamp *et al.*, "The effect of speaking style on the articulatory-acoustic vowel space in individuals with tongue cancer before and after surgical treatment," in *Proceedings of the 13th International Seminar on Speech Production*, 2024, pp. 65–68.

[33] K. Bettens, M. De Bodt, Y. Maryn, A. Luyten, F. L. Wuyts, and K. M. Van Lierde, "The relationship between the nasality severity index 2.0 and perceptual judgments of hypernasality," *Journal of Communication Disorders*, vol. 62, pp. 67–81, 2016.