Characterizing Sustained Phonation in Text-To-Speech Models

Amelie Daum[a], Nina Goes[a], Andreas M. Kist[a]

[a]*Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU),*
Nürnberger Str. 74, Erlangen, 91052, Bavaria, Germany

Abstract

Sustained phonation is a central task in clinical voice assessment and provides a controlled setting to quantify acoustic voice characteristics. In contrast, the evaluation of modern text-to-speech (TTS) systems still relies predominantly on perceptual ratings such as the Mean Opinion Score (MOS), leaving open whether these systems can reliably generate sustained phonation and how their acoustic properties compare to human voices. The capability of TTS models to reproduce clinically relevant voice features remains insufficiently characterized.

Here, we systematically examine sustained phonation in contemporary TTS systems and compare synthetic and human voice samples using common acoustic measures. Multiple TTS models were screened for their ability to generate sustained vowels, such as /a/. One model, namely Eleven v3 by ElevenLabs, was subsequently analyzed in detail with respect to the distribution of phonation durations, the relationship between prompt length and generated duration, and differences between vowels and speaker types. Finally, TTS-generated sustained phonations were compared with human recordings from two independent cohorts using established clinical voice parameters.

We found that TTS systems were able to produce sustained phonation, although reliability varied between models. For the selected Eleven v3 model, phonation durations showed non-normal distributions and were partially predicted by prompt length. Most acoustic measures of synthetic samples overlapped with the ranges observed in human voices, while selected parameters showed statistically significant but inconsistent differences across vowels. These findings indicate that current TTS models can approximate key acoustic characteristics of sustained phonation, while also exhibiting systematic deviations that should be considered in applications involving clinical voice metrics and in further development of realistic TTS systems.

Keywords: acoustic voice analysis, sustained phonation, text-to-Speech, synthetic voice

1. Introduction

Speech quality can be assessed with a variety of methods, as described by Lemmetty (1999). With segmental evaluation methods such as the Modified Rhyme Test (MRT), the intelligibility of single phonemes and syllables can be tested. Here, the error rate of humans identifying single-syllable words (e.g., bus, but, bug, buff, bun, buck) that were synthesized by a Text-to-Speech (TTS) model is used for quantifying its quality. To assess quality on a higher level, sentences were developed to test the understanding of spoken sentences. For example, the Semantically Unpredictable Sentences (SUS) test uses one-syllable words to build nonsensical sentences using varying grammatical structures. This ensures that words that were not understood cannot be derived from context. An example of a SUS test sentence is "The table walked through the blue truth." Jekosch (1993). Furthermore, comprehension tests are used to quantify the quality of a TTS system, where subjects have to answer questions on the previously heard texts.

Nowadays, segmental evaluation methods, as well as sentence-level and comprehension tests lost relevance, since the quality of TTS heavily evolved in recent years. To evaluate overall quality, and naturalness of TTS-synthesized speech in particular, the Mean Opinion Score (MOS) is often employed now Stan (2022); Le Maguer et al. (2024). For this, listeners have to assign a score on a five-level Likert scale (bad - poor - fair - good - excellent) to the synthesized speech they hear. The mean of the scores is then called MOS.

The MOS is oftentimes used, even though the calculation process is time-consuming and costly, since it relies on human listening tests. For this reason many models to predict MOS for assessing speech quality have been developed recently Lo et al. (2019); Mittag and Möller (2020); Tseng et al. (2022); Saeki et al. (2022); Hajal et al. (2022); Yang et al. (2022); Choi et al. (2022); Qi et al. (2023); Sellam et al. (2023); Baba et al. (2024); Shi et al. (2024). The development was often motivated by the VoiceMOS Challenges that were introduced in 2022 Huang et al. (2022); Cooper et al. (2023); Huang et al. (2024).

While the MOS provides an overall insight into the perceived quality, it is not very specific in terms of transparency, concerning why listeners felt that way and, thus, does not provide any information which parts of the model should be improved. Le Maguer et al. (2024) concluded after evaluating MOS for speech synthesis that new measurements should be developed to analyze the quality of modern speech synthesis technologies.

When assessing speech quality in the medical field, there are objective parameters that can be assessed to determine the condition of the voice. One widely used method is the analysis of sustained phonation (SP), as it gives insight into respiratory control and vocal fold integrity. The duration and quality of sustained phonation depend directly on the efficiency of respiratory control and the ability of the vocal folds to vibrate continuously and evenly Gilman (2021); Harden and Looney (1984); Maslan et al. (2011). Studies found differences between clinical parameters calculated from pathologic and healthy SP samples Teixeira et al. (2013); Vashkevich et al. (2019); Wertzner et al. (2005). Acoustic parameters extracted from SP signals are commonly used to evaluate the health of the phonatory system.

The goal of this paper is to bridge the gap between subjective TTS evaluation methods and objective clinical voice analysis. The primary objective is to determine whether TTS models are capable of producing sustained phonation. If so, the study will further investigate whether measurable acoustic differences exist between human and TTS-generated sustained phonation.

2. Materials and Methods

In a first experiment, the general ability of TTS models to produce SP is reviewed (Experiment 1). Afterwards, the behavior of well-performing models is further analyzed. It is investigated whether the durations of the SP samples follow a Normal distribution (Experiment 2), whether longer prompts actually result in longer SP samples (Experiment 3), and whether human and TTS-generated SP differ in common voice metrics (Experiment 4).

2.1. Recording of Human SP

As a baseline for the comparisons in Experiment 4, a small population of human speakers were recorded.

A total of 20 participants were recorded, 11 of them identified as female, 9 as male. All stated to have no known voice disorders. The average age was 25.5 years (SD = 3.3 years). Participants were recorded in a small, acoustically untreated room. They were seated on a chair and instructed to hold a high-quality microphone (Rhøde NT1) 10 cm from their mouth, as implemented in other studies (Gerratt et al., 2016; Vashkevich et al., 2019; Wertzner et al., 2005). Each participant was asked to produce an SP of the vowels /a/, /i/, and /u/ for 10 seconds each. Participants were shown an instructional video that guided them through the task. The sampling rate was set to 44.1 kHz.

In addition to the recordings described above, the recordings of a healthy control group from another study were used as a baseline (Vashkevich et al., 2019). The recordings were collected by the Republican Research and Clinical Center of Neurology and Neurosurgery in Minsk, Belarus. The recordings of 33 healthy participants were published on GitHub in *WAV* format. Participants were on average 53.8 years old (SD = 11.6 years). The sustained vowels /a/ and /i/ of 20 women and 13 men were recorded. The average duration of the recordings was 3.69 s (SD = 1.52 s), and the sampling rate set to 44.1 kHz.

2.2. Generation of SP Signals Using TTS Models

Mirroring SP studies with human participants, SP samples were generated with TTS models. Models from five companies, such as ElevenLabs and Google, were chosen for the experiments. The model names were *P1* (Papla Media), *Eleven v3* (ElevenLabs), *speech-02-hd* (MiniMax), *Gemini 2.5 Pro TTS* (Google) and *Octave* (Hume).

All chosen models were hosted via a web interface and presented the user with a selection of speakers. Some models offered more in-depth settings, such as a temperature controller between "creative" and "robust" (ElevenLabs), buttons with emotions like "disgust" or "awe" (Hume), or controllers for speed, pitch, and volume (MiniMax). For all experiments, the default settings were retained. Furthermore, all models presented different speakers with their own descriptions. Names that are connoted with a certain sex were often given to the speakers, and in the following, there will be spoken of female and male speakers according to their given names.

As a first approach to generating SP samples, slashes were used to indicate phonetic content, following the standard notation for phonemic transcription (e.g., /aaaaaaaaaaaaa/ to represent a sustained /a/ vowel). However, some TTS models did not interpret the slashes as phonetic markers and instead read them out loud or misinterpreted the input. In such cases, the slashes were removed (Hume Octave, Papla P1).

For assessing the general ability to generate SP (Experiment 1), the models are systematically prompted and the output is graded afterwards. To be concise, for every model two arbitrary speakers are picked. With every speaker, the models are prompted 5 times with a chain of 10, 15, and 20 /a/s, each. In total, each model is prompted 30 times. A generation attempt is considered as failed if the TTS response takes longer than 30 s.

As only the ElevenLabs model produced SP consistently, it is the only model that is used to further analyze the model behavior regarding SP in Experiments 2-4.

To carry out Experiment 2, which tests whether output durations follow a statistical distribution, the prompt /aaaaaaaaaaaaaaa/ is submitted until 100 SP samples were generated that pass the exclusion criteria defined in Section 2.3. This is repeated for the other two investigated vowels as well.

In Experiment 3, the impact of the prompt length is analyzed. Hence, each prompt in the range of one vowel (/a/), to 20 vowels (/aaaaaaaaaaaaaaaaaaaa/), is taken as an input for the selected TTS model and the output duration is measured. This approach is used for each vowel /a/, /i/ and /u/ and two speakers (Alice and Liam) each. Each prompt is used for generation until an SP, which passes the exclusion criteria for SP is synthesized.

For the analyses of common voice metrics (Experiment 4), *Eleven v3* was used to produce roughly 20 SP samples per vowel. Each sample was produced with another "voice". Since *Eleven v3* allowed further instructions, this was used like this:

[holding a vowel for 10 seconds at a comfortable pitch]
/uuuuuuuuuuuuuuu/

This is how 23 /a/ phonations (m: 11, f: 12), 20 /i/ phonations (m: 9, f: 11) and 20 /u/ phonations (m: 9, f: 11) were generated.

2.3. SP Quality Classes

To evaluate the ability of current TTS models to produce SP in Experiment 1, a structured assessment metric is required.

Given the wide variety of outputs and the need to provide a fine-grain quality measurement, the assessment is split into two stages. In the first stage, a set of non-negotiable exclusion criteria is looked at. If any of these criteria are met, the corresponding audio sample is disqualified from further evaluation, as it cannot be considered a valid instance of sustained phonation. These criteria are as follows:

- The model fails to generate an audio file from the prompt within 30 s.
- The generated audio file is entirely silent.
- The audio contains elements unrelated to sustained phonation, such as music, background sounds, or arbitrary spoken text.
- The model generated an audio file with the wrong pronunciation of the intended vowel.
- The model generated an audio file in which the vowels are chopped instead of sustained. No part of the audio includes sustained phonation.

Only audio files that do not meet any of these exclusion criteria proceed to the second stage, where they are evaluated using a more fine-grained quality assessment. Duration, continuity, pitch height, and pitch movement are evaluated. Each category can be assigned between 0 (unnatural) and 1 (natural) points - except for pitch movement, where 2 points can be assigned at maximum.

This means that audio files that pass the first evaluation stage can receive a score between 0 and 5 points, whereas more points mean more naturalness.

For the criterion duration, the point is given if, between the onset and offset, the vowel is held and can be perceived as such by a human rater. The continuity point is given if the SP is in one piece and not split into multiple segments. Pitch height is deemed unnatural if the voice seems strained or unnaturally high or low, as one would not expect from a participant holding a vowel at "a comfortable pitch".

2.4. Audio Processing

After recording and generating SP samples, the audio files are processed to extract further information. The audio files preprocessing, i.e. extracting the parts with sustained phonation, and computing the desired metrics, is described in the following.

All files were converted to the *WAV* file format, using the python library `pydub`. Praat, a speech analysis software developed by Paul Boersma and David Weenink (Boersma and Weenink, 2007), was utilized for analyzing and manually trimming the samples. The recordings of the human study participants, as well as the SP generations using prompts that did not purely consist of concatenated vowels, were trimmed by extracting the parts with SP.

For the TTS quality assessment (Experiment 1), no audio processing, besides file format conversion, was performed. For the experiments that investigated the output length of SP samples (Experiments 2 and 3), the durations were calculated with the Python library `librosa`.

For the comparison of synthetic and human SP samples (Experiment 4), the segment between the voicing onset and offset is analyzed, as it is considered relatively stable. To account for differences in sample duration, equal-length excerpts are extracted. To increase statistical robustness, multiple excerpts are randomly selected from each sample, thereby reducing the influence of local fluctuations. Specifically, for each signal, three randomly chosen 500 ms segments are taken from the SP samples. A safety margin of 200 ms at the beginning and end of the signal was excluded to minimize the effects of onset and offset.

The next step is calculating quantitative parameters, such as HNR, as seen in Gerratt et al. (2016) and originally described by Kreiman et al. (2010). Moreover, jitter and shimmer will be calculated as in Vashkevich et al. (2019); Wertzner et al. (2005). Furthermore, Patel et al. (2018) recommends looking at CPP, as well. Finally, formants are compared as well.

The code for calculating jitter, shimmer, HNR, and formants was modified from an existing script [GitHub] (Feinberg, 2022; Puts et al., 2012). CPP was calculated with another script, published by Satvik Dixit [GitHub]. Note that the definitions of these parameters are ambiguous. For example, for jitter $J_{loc,abs}$, J_{loc}, J_{rap}, J_{ppq5} and J_{ddp} were calculated.

2.5. Statistics

In Experiment 2, it is analyzed whether the duration of generated SP samples with the same prompt vary and specifically whether they are Normal distributed. Whether this assumption of Normality holds, is investigated visually with Q-Q-plots and confirmed with the Shapiro–Wilk test, as recommended previously (Ghasemi and Zahediasl, 2012). In Experiment 3, linear regression is used to quantify the relationship between the prompt length and the generated output duration.

In Experiment 4, common voice metrics are calculated for three groups of SP samples (human samples from Erlangen and Minsk, TTS-generated samples). Comparisons are made between two groups each. The variables under investigation are continuous metrics derived from short audio segments. Prior to hypothesis testing, the distributions of these metrics were assessed. Since the data was found to deviate from normality, the non-parametric Mann–Whitney U test is employed to evaluate whether the distributions of the two groups differ significantly. The significance threshold was set to $\alpha = 0.05$. Furthermore, the Benjamini-Hochberg method was used to correct for multiple hypothesis testing. Alongside the test statistic U and adjusted p-values p_{adj}, the rank-biserial correlation r is reported as a measure of effect size.

3. Results

3.1. Experiment 1: Ability to Produce Sustained Phonation

In the first experiment, models from various providers were tested with prompts of different lengths and with different speakers. To judge the ability of SP models to generate SP, a quality assessment system was developed (see Section 2.3). Of 150 SP generation attempts, only 41 passed the first assessment. ElevenLabs' model *Eleven v3* performs best, by far, with a passing rate of 70.0 %. Gemini's model is second with only 40.0 % of audio generations that did not fall into one or more of the exclusion criteria. The other models have passing rates below 15 %. The reasons for failing the definition of SP vary per model. *Eleven v3* failed mostly due to wrong pronunciation of the intended sound or generation of music or random text. 83 % of failed attempts of Gemini's model were caused by taking more than 30 s to generate a SP sample. Almost all of Hume *Octave*'s failed attempts were flagged as such because the generations were silent audio files. The model of MiniMax produced mostly concatenated, not sustained vowels, which were often pronounced near to the American "a", not the intended /a/, additionally. Papla's model *P1*, however, also struggled mostly with the intended pronunciation, see Table 3.1 for detailed results.

The average scores for audios that pass the exclusion criteria are between 3.25 (MiniMax) and 4.57 (ElevenLabs) of 5 possible points. ElevenLabs' model performs best, as before which means that of the many audios that pass the criteria, the quality is high as well. With ElevenLabs, the female voice has slightly higher quality (4.80 vs. 4.36 points), and the different lengths are almost equal in quality (4.50 / 4.50 / 4.71). The Gemini model is the only model besides *Eleven v3* where more than 4 audios passed the first assessment stage. The average quality score is 0.4 points below ElevenLabs; differences in sex and length are marginal.

In conclusion, all tested models are capable of producing SP, however, with varying degrees of reliability. As the ElevenLabs model *Eleven v3* is the best model by far, it is analyzed further.

3.2. Experiment 2: Output Duration Distribution Analysis

The hypothesis that SP sample durations follow a Normal distribution can be rejected, since the Shapiro-Wilk test yields a p-value < 0.001, indicating highly significant results (see Table 3.2). This non-normality is also evident in the Q-Q plots (see Figure 1), where the sample quantiles deviate substantially from the theoretical quantiles. The lower-end deviation can be attributed to the impossibility of negative durations, but the upper-end deviation indicates that longer durations are not well captured by a Normal distribution.

3.3. Experiment 3: Impact of Prompt Length on Output Duration

Next, we investigated the impact of prompt length on output length. We continued using *Eleven v3* due to its performance in Experiment 1. Figure 2 shows that the duration of the voice signals was rising with longer prompts. Prompts with 1 - 4 vowels seem to produce substantially shorter sustained phonations. However, with 5 vowels and more, the durations were oscillating around 2 seconds, with a peak at around 16 vowels in the prompt.

A linear relationship between the prompt length and the output length is expected. This can be confirmed, see Figure 2.

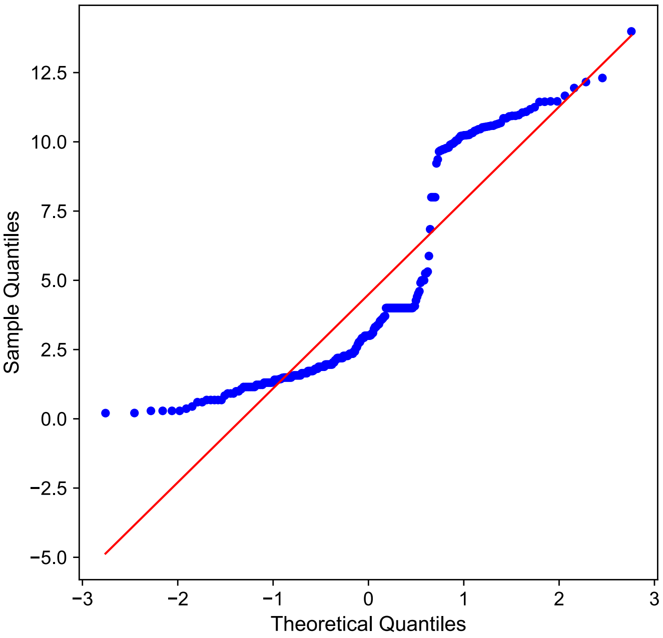

Figure 1: Q-Q-plot of audio duration distribution. 100x /aaaaaaaaaaaaaaa/, /iiiiiiiiiiiiii/ and /uuuuuuuuuuuuuuu/ each, generated with *Eleven v3*

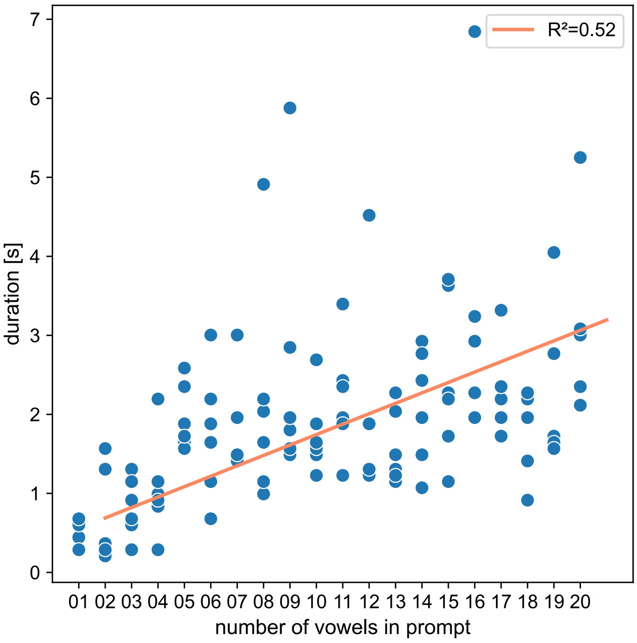

Figure 2: Impact of prompt length on duration of generated output. A linear regression line was fitted after excluding outliers.

Table 1: Results of Experiment 1. Pass Rate gives the percentage of generated audios that did not fail the exclusion criteria for SP, for each model, a total of 30 generations were made, see Section 2.3. Under quality scores are the scores that were assigned on average. Audios that were failing the inclusion criteria were assigned −1. The quality scores, including the failing audios, are given in parentheses.

Model	Pass Rate	Quality (overall)	Quality (by sex)	Quality (by prompt length)
ElevenLabs	**21**/30 (**70.0%**)	**4.57** (**2.90**)	F=4.80, M=4.36	10=4.50, 15=4.50, 20=4.71
Gemini	12/30 (40.0%)	4.17 (1.07)	F=4.12, M=4.25	10=4.25, 15=4.50, 20=4.00
Hume	2/30 (6.7%)	4.00 (-0.67)	F=4.00, M=4.00	10=4.00
MiniMax	4/30 (13.3%)	3.25 (-0.43)	F=3.00, M=3.33	10=3.00, 15=3.50, 20=3.00
Papla	2/30 (6.7%)	3.50 (-0.70)	M=3.50	15=4.00, 20=3.00

Table 2: Results of duration distribution of 300 equal-length prompts (100 per vowel). Shapiro-Wilk tests were performed to test whether generated audio durations are Normally distributed.

vowel	duration [s]	Shapiro-Wilk p
a	3.045 ± 1.197	0.0137
i	2.724 ± 1.217	0.0006
u	2.500 ± 0.931	0.0270
all	2.760 ± 1.144	<0.0001

52 % of the variation in audio length is explained by prompt length. The regression line can be described with

$$t = 0.13n + 0.56, \tag{1}$$

whereas t is the duration of the SP in seconds and n is the number of vowels. This shows that the impact of the number of vowels is small since almost 8 more vowels are needed to increase the duration by 1 second.

When looking at each vowel individually, audio duration has the highest correlation with the prompt length for /u/ with 69% variation explained, then /i/ with 56 % explained and lastly, /a/ with 41 % variation explained.

3.4. Experiment 4: Comparing Human and TTS-generated Sustained Phonation

In this experiment, clinical metrics are calculated for human voices from Erlangen (N = 20 for all vowels), and Minsk (N = 31 (/a/), N = 33 (/i/)), and for synthetic voices (N = 23 (/a/), N = 20 (/i/), N = 20 (/u/)) generated with ElevenLabs' *Eleven v3*. Pairwise Mann-Whitney U tests were performed. All tests were corrected for multiple hypothesis testing using Benjamini-Hochberg method.

3.4.1. Jitter

Jitter describes the perturbation of fundamental frequency. High jitter values are significantly correlated to hoarseness ($p =< 0.001, r = 0.712$) (Yumoto et al., 1984; Jones et al., 2001). In general, jitter can be used to tell pathological and healthy voices apart (Guimarães, 2007). In a meta study, jitter values from healthy control groups were compared. Mean jitter values between 0.3 % and 1.2 % were reported across various studies. However, the value variance between studies and in studies is high, with standard deviations of up to 0.5 % reported. (Guimarães, 2007)

The measured samples show jitter values that align with these reports. For example, J_{rap} is 0.29 % ± 0.32 % for human audios (Erlangen), 0.28% ± 0.43% for the human jitter of the audios of Minsk. Jitter for TTS-generated samples are also within the reported ranges (0.47 % ± 1.0 %).

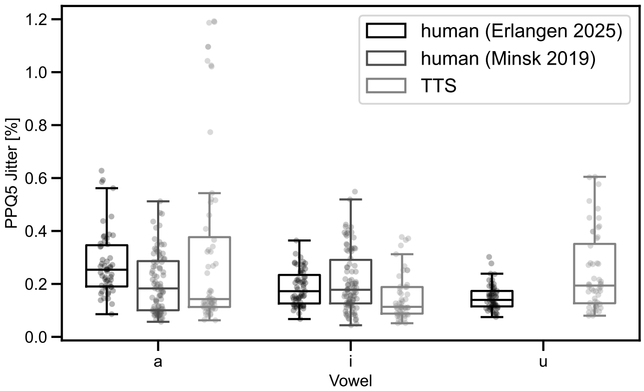

Figure 3: J_{ppq5} of SP samples, by vowel and source

Generally speaking, jitter values are lower for TTS voices than for human ones; however, the effect is mostly marginal, see Figure 3. While some significant differences are found ($p < 0.05$), they are mostly not of relevance due to their small effect sizes.

However, between the Minsk human signals and the synthetic voices, a relevant difference was found: jitter values for /i/ are significantly lower when comparing TTS voices with the ones from Minsk. 3 out of 5 jitter measurement calculations found significant results with a medium effect size ($J_{ddp} : p = 0.003, r = 0.33; J_{ppq5} : p = 0.004, r = 0.32; J_{rap} : p = 0.003, r = 0.33$). Those differences are not confirmed by the Erlangen data. ($J_{ddp} : p = 0.019, r = 0.29; J_{ppq5} : p = 0.054, r = 0.25; J_{rap} : p = 0.019, r = 0.29$)

In contrast, all jitter calculation methods show significant and medium-to-large sized differences between the two human groups for the vowel /a/ (for all jitter measures: $p =< 0.001$; $J_{ddp} : r = 0.50; J_{loc} : r = 0.49; J_{loc,abs} : r = 0.42; J_{ppq5} : r = 0.49; J_{rap} : r = 0.50$).

3.4.2. Shimmer

Shimmer is the amplitude equivalent to jitter, and perceptually correlates with breathiness (Petrovic-Lazic et al., 2015). In a study on the "Standardization of acoustic measures for normal voice patterns", $S_{loc,dB}$ shimmer values for healthy voices between 0.132 dB and 2.37 dB were mentioned (De Felippe

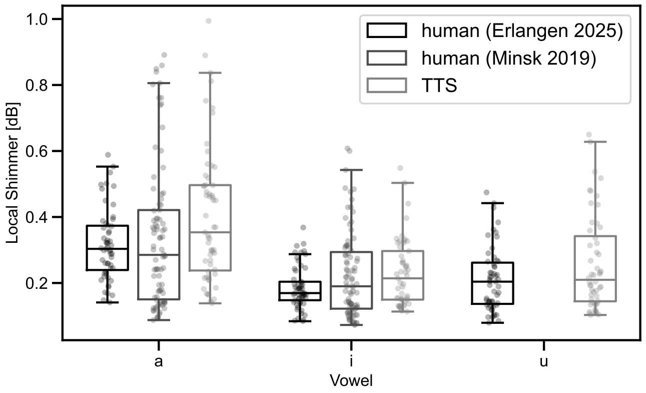

Figure 4: Local Shimmer of SP samples, by vowel and source

et al., 2006). All $S_{loc,dB}$ shimmer values calculated from the analyzed signals lie well within that range, with 0.32 dB ± 0.32 dB (Erlangen) and 0.33 dB ± 0.25 dB (Minsk), as well as 0.44 dB ± 0.45 dB (TTS).

Synthetic voices have significantly higher shimmer values for the vowel /a/, see Figure 4. With all 6 shimmer calculation methods, pairwise Mann-Whitney U tests with human SP samples from Minsk yield significant results. The effect sizes range from $r = -0.22$ to $r = -0.38$. While differences are larger between human voices from Minsk and synthetic voices, signals from Erlangen show significant differences for /a/ as well ($S_{loc,dB} : p = 0.008, r = -0.32; S_{loc} : p = 0.027, r = -0.27$).

3.4.3. Formants

Formants are a property of each individual, like a vocal fingerprint, and show different patterns per vowel.

The first formant frequencies are significantly higher for synthetic voices when sustaining the vowel /a/, see Figure 5. Both Erlangen and Minsk human signals show decent effects (Erlangen: $p =< 0.001, r = -0.39$; Minsk: $p =< 0.001, r = -0.60$). Furthermore, the frequency of the first formant seems to be lower for /i/ when comparing data from Minsk with TTS-generated signals ($p =< 0.001, r = 0.43$). This is not the case for signals from Erlangen compared to TTS voices. In fact, voices from Erlangen have significantly higher first formant frequencies for /i/ than voices of the Minsk group ($p =< 0.001, r = -0.42$).

Generally, the second formant frequencies are higher across all vowels in synthetic voices. This effect is significant and medium-sized for the vowel /a/ spoken by voices from both Erlangen ($p = 0.002, r = -0.36$) and Minsk ($p = 0.006, r = -0.30$). Largely higher F_2 values have synthetic voices compared to human voices concerning /u/ (Erlangen: $p =< 0.001, r = -0.54$).

Looking at F_3, human voices from Minsk have significantly lower formant frequencies for /a/ ($p = 0.002, r = -0.33$). This is not the case for the comparison of Erlangen voices with TTS voices ($p = 0.327, r = 0.13$). However, both human groups are showing significantly, disaligning F_3 value distributions ($p =< 0.001, r = 0.61$).

For F_4, no significant differences between groups were found.

Figure 5 shows the relationship between formant frequencies F_1 and F_2. The groups cannot be distinguished from the first two formants alone for all vowels, as differences between the two human groups seem as strong as differences between synthetic and human SP samples.

3.4.4. Harmonics-To-Noise Ratio (HNR)

High Harmonics-To-Noise Ratio (HNR) values are correlated with hoarseness ($p =< 0.001, r = .809$) (Yumoto et al., 1984). De Felippe compared HNR values across studies and found mean values between 7.82 dB and 10.98 dB (De Felippe et al., 2006). The analyzed signals show much higher mean HNR values: 23.7 dB ± 5.06 dB (Erlangen), 24.8 dB ± 6.29 (Minsk) and 20.6 dB ± 6.86 dB (TTS), see Figure 6.

The HNR is significantly different ($p =< 0.001$) between human (Minsk) and synthetic voice signals for /a/ with $r = 0.52$ and for /i/ with $r = 0.41$. HNR is lower for synthetic voices. This is not supported by the data from Erlangen, though. In fact, significant differences are found between both human groups (/a/: $p =< 0.001, r = -0.39$, /i/: $p = 0.002, r = -0.33$).

3.4.5. Cepstral Peak Prominence (CPP)

The prominence of the strongest peak of the cepstrum (CPP) is significantly higher in the synthetic recordings for the vowel /i/. Both in Erlangen and Minsk, originating data confirms this (Erlangen: $p = 0.002, r = -0.38$, Minsk: $p =< 0.001, r = -0.53$), see Figure 7.

4. Discussion

4.1. Summary

The generation of audio files that pass the definition of SP proved to be challenging for most of the examined models. ElevenLabs' model *Eleven v3* outperforms the other models by far with a criteria passing rate of 70.0 % and the highest average quality of generated SP samples. Common problems were, on the one hand, processing the prompt. Models failed to produce an audio file in under 30 seconds or produced silent audio or ones with random text, or even music. On the other hand, it proved to be challenging to ensure the correct pronunciation as well as that the vowels should be sustained for several seconds.

All following TTS samples were produced with *Eleven v3*. The output length correlated with the prompt length ($R = 0.72$), however, the effect is rather small as the prompt length must increase drastically for longer audio outputs (circa 8 vowels / s).

Then, sustained vowel productions from two human datasets (from Erlangen and Minsk) were compared with synthetic voices generated by a TTS system. Acoustic analyses focused on jitter, shimmer, the first four formants, HNR, and CPP. Overall, all measures fell within normal ranges reported for healthy speakers, with the exception of HNR, which was consistently higher than values reported in previous studies. To account for statistical robustness, only results that were both significant

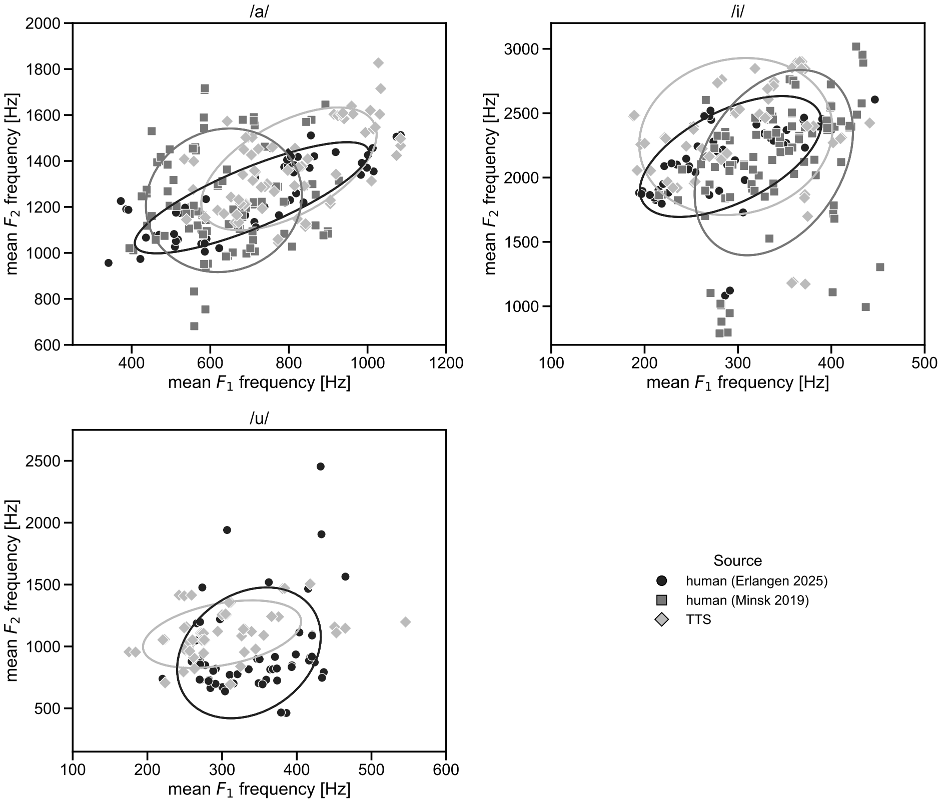

Figure 5: F_1 vs. F_2 of SP samples, by vowel and source

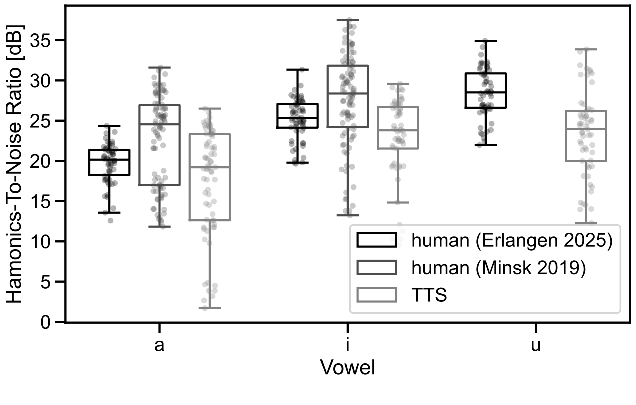

Figure 6: Harmonics-To-Noise Ratio values of SP samples, by vowel and source

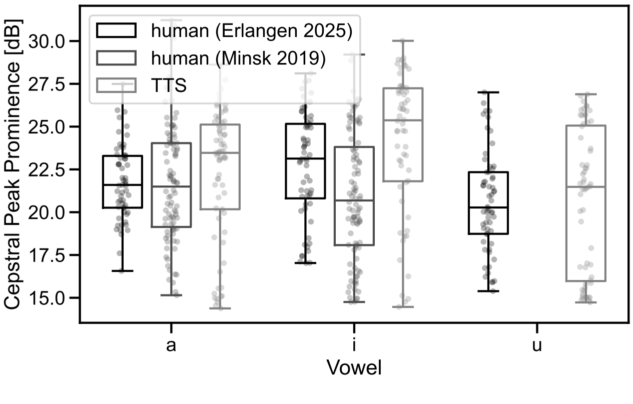

Figure 7: Cepstral Peak Prominence of SP samples, by vowel and source

($p < 0.05$) and of at least moderate effect size ($r > 0.3$) were considered relevant.

Across both human groups, three consistent effects emerged when comparing natural to synthetic voices. First, shimmer values were higher in synthetic voices for the vowel /a/. This finding is noteworthy given that shimmer is caused by irregularities in vocal fold vibration. Since TTS voices are not trained on sustained vowel production, instabilities may arise when the model attempts to generate long, steady phonation, resulting in artificially elevated shimmer. Second, formant frequencies (F_1 and F_2) of /a/ were consistently higher in synthetic voices. While regional language differences could account for some variability in the human groups, the greater magnitude of difference between TTS and both human groups suggests that the discrepancy reflects systematic properties of the synthetic voices. Third, CPP values for /i/ were significantly higher in the synthetic voices, indicating smoother and less noisy spectral energy distribution. This aligns with the expectation that TTS voices, which are trained to sound pleasant and clear, lack the breathiness, hoarseness, or microphone artifacts present in natural recordings.

In addition to these robust findings, several effects were observed only in the comparison between the Minsk dataset and the synthetic voices. Specifically, higher F_1 values for /i/, higher F_3 values for /a/, and lower HNR values for /a/ and /i/ were found. These effects were not replicated in the Erlangen dataset and, importantly, significant differences were also observed between the Erlangen and Minsk human groups themselves. The most plausible explanation is that these effects reflect regional or language-specific influences on vowel production as well as methodological differences in recording conditions. For instance, cross-linguistic variation in vowel quality could explain the formant discrepancies, while differences in microphone equipment, room acoustics, or preprocessing are likely sources of the divergent HNR results. Furthermore, the age of study participants varied. While participants from Erlangen were quite young (25.5 years ± 3.3 years), speakers from Minsk were older (53.8 years ± 11.6 years).

4.2. Limitations and Future Directions

While our study provides useful insights, some limitations should be acknowledged. Due to the overall success of the *Eleven v3*, we focused on the analysis of this TTS system. Future research should investigate whether the observed deviations in acoustic parameters are *Eleven v3*-specific or also occur in other models. Moreover, a closer matching of regional origins of voices in future studies could prevent confounding effects, as many of the observed differences may have resulted from variations in the speakers' native languages. Both fundamental frequency (Natour and Wingate, 2009) and formants (Fox and Jacewicz, 2009) are influenced by the speaker's ethnicity or region of growing up.

In a follow-up study, we would suggest testing the SP features being able to distinguish to a larger extend synthetic and real voices using Machine Learning. Our results suggest that there are several distinctions across TTS systems and real human voices, but future research would be able to use explainable AI methods to further uncover distinctions in a more latent parameter space.

Ultimately, the overall goal of future work should be to adapt TTS models in ways that more closely reflect physiological constraints and natural voice variability, using the information gained from differences between human and TTS-generated SP samples. At the same time, analyses should be extended beyond SP to more complex speech material, such as connected speech, in order to capture aspects like prosody that are ignored in sustained phonation.

5. Conclusions

The present study set out to evaluate state-of-the-art TTS systems using objective measures, rather than conventional MOS ratings. While, to the best of our knowledge, MOS remains the predominant benchmark in current literature, it is inherently limited by its subjective nature. By adopting measures commonly used in medical voice assessment, this work provides a more fine-grained image of where synthetic speech systematically diverges from human phonation.

The generation of sustained phonation samples posed a considerable challenge for most TTS models. Among the systems tested, only ElevenLabs' model *Eleven v3* was consistently able to produce acceptable samples. Even so, deviations from human voices were observed. These deviations suggest that while synthetic speech can approximate human-like sustained phonation in certain respects, it lacks the physiological constraints that anchor variability in real human voices.

Systematic mismatches are revealed that are invisible to MOS-based scores, but which may nevertheless influence perceptual judgments at a more implicit level. Whether such mismatches directly affect listener perception remains an open question, but they provide concrete starting points for future improvements.

Ultimately, this study shows that objective, clinical voice measures are a powerful tool for complementing traditional evaluations. By grounding the assessment of TTS voices in measurable deviations from human physiology, they not only enable more transparent comparisons across models but also point toward targeted strategies for making synthetic voices more natural and physiologically plausible in the future.

References

Baba, K., Nakata, W., Saito, Y., Saruwatari, H., 2024. The T05 System for The VoiceMOS Challenge 2024: Transfer Learning from Deep Image Classifier to Naturalness MOS Prediction of High-Quality Synthetic Speech. doi:10.48550/arXiv.2409.09305.

Boersma, P., Weenink, D., 2007. Praat: Doing phonetics by computer (version 5.3.51) URL: https://www.researchgate.net/publication/259810776_PRAAT_Doing_phonetics_by_computer_Version_5351.

Choi, Y., Jung, Y., Suh, Y., Kim, H., 2022. Learning to Maximize Speech Quality Directly Using MOS Prediction for Neural Text-to-Speech. IEEE Access 10, 1–1. doi:10.1109/ACCESS.2022.3175810.

Cooper, E., Huang, W.C., Tsao, Y., Wang, H.M., Toda, T., Yamagishi, J., 2023. The VoiceMOS Challenge 2023: Zero-shot Subjective Speech Quality Prediction for Multiple Domains. doi:10.48550/arXiv.2310.02640, arXiv:2310.02640.

De Felippe, A.C.N., Grillo, M.H.M.M., Grechi, T.H., 2006. Standardization of acoustic measures for normal voice patterns. Brazilian Journal of Otorhinolaryngology 72, 659–664. URL: https://linkinghub.elsevier.com/retrieve/pii/S1808869415310235, doi:10.1016/S1808-8694(15)31023-5.

Feinberg, D.R., 2022. Parselmouth praat scripts in python. URL: osf.io/6dwr3, doi:10.17605/OSF.IO/6DWR3.

Fox, R.A., Jacewicz, E., 2009. Cross-dialectal variation in formant dynamics of American English vowels. The Journal of the Acoustical Society of America 126, 2603–2618. URL: https://pubs.aip.org/jasa/article/126/5/2603/901156/Cross-dialectal-variation-in-formant-dynamics-of, doi:10.1121/1.3212921.

Gerratt, B.R., Kreiman, J., Garellek, M., 2016. Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech. Journal of Speech, Language, and Hearing Research 59, 994–1001. doi:10.1044/2016_JSLHR-S-15-0307.

Ghasemi, A., Zahediasl, S., 2012. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. International Journal of Endocrinology and Metabolism 10, 486–489. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/, doi:10.5812/ijem.3505, arXiv:23843808.

Gilman, M., 2021. Revisiting Sustained Phonation Time of /s/, /z/, and /α/. Journal of Voice 35, 935.e13–935.e18. URL: https://www.jvoice.org/article/S0892-1997(20)30104-1/abstract, doi:10.1016/j.jvoice.2020.03.012, arXiv:32345503.

Guimarães, I., 2007. A Ciência e a Arte da Voz Humana. ESSA – Escola Superior de Saúde do Alcoitão, Alcabideche, Portugal. Depósito Legal n.º 255774/07, il.

Hajal, K.E., Wu, Z., Scheidwasser-Clow, N., Elbanna, G., Cernak, M., 2022. Efficient Speech Quality Assessment using Self-supervised Framewise Embeddings. doi:10.48550/arXiv.2211.06646, arXiv:2211.06646.

Harden, J.R., Looney, N.A., 1984. Duration of sustained phonation in kindergarten children. International Journal of Pediatric Otorhinolaryngology 7, 11–19. URL: https://linkinghub.elsevier.com/retrieve/pii/S016558768480049X, doi:10.1016/S0165-5876(84)80049-X.

Huang, W.C., Cooper, E., Tsao, Y., Wang, H.M., Toda, T., Yamagishi, J., 2022. The VoiceMOS Challenge 2022. doi:10.48550/arXiv.2203.11389, arXiv:2203.11389.

Huang, W.C., Fu, S.W., Cooper, E., Zezario, R.E., Toda, T., Wang, H.M., Yamagishi, J., Tsao, Y., 2024. The VoiceMOS Challenge 2024: Beyond Speech Quality Prediction. doi:10.48550/arXiv.2409.07001, arXiv:2409.07001.

Jekosch, U., 1993. Speech quality assessment and evaluation, in: 3rd European Conference on Speech Communication and Technology (Eurospeech 1993), ISCA. pp. 1387–1394. doi:10.21437/Eurospeech.1993-11.

Jones, T.M., Trabold, M., Plante, F., Cheetham, B., Earis, J., 2001. Objective assessment of hoarseness by measuring jitter. Clinical Otolaryngology and Allied Sciences 26, 29–32. URL: http://doi.wiley.com/10.1046/j.1365-2273.2001.00413.x, doi:10.1046/j.1365-2273.2001.00413.x.

Kreiman, J., Antoñanzas-Barroso, N., Gerratt, B.R., 2010. Integrated software for analysis and synthesis of voice quality. Behavior Research Methods 42, 1030–1041. doi:10.3758/BRM.42.4.1030.

Le Maguer, S., King, S., Harte, N., 2024. The limits of the Mean Opinion Score for speech synthesis evaluation. Computer Speech & Language 84, 101577. URL: https://linkinghub.elsevier.com/retrieve/pii/S0885230823000967, doi:10.1016/j.csl.2023.101577.

Lemmetty, S., 1999. Review of Speech Synthesis Technology. Helsinki University of Technology URL: http://research.spa.aalto.fi/publications/theses/lemmetty_mst/thesis.pdf.

Lo, C.C., Fu, S.W., Huang, W.C., Wang, X., Yamagishi, J., Tsao, Y., Wang, H.M., 2019. MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion, in: Interspeech 2019, ISCA. pp. 1541–1545. URL: https://www.isca-archive.org/interspeech_2019/lo19_interspeech.html, doi:10.21437/Interspeech.2019-2003.

Maslan, J., Leng, X., Rees, C., Blalock, D., Butler, S.G., 2011. Maximum Phonation Time in Healthy Older Adults. Journal of Voice 25, 709–713. URL: https://linkinghub.elsevier.com/retrieve/pii/S0892199710001724, doi:10.1016/j.jvoice.2010.10.002.

Mittag, G., Möller, S., 2020. Deep Learning Based Assessment of Synthetic Speech Naturalness, in: Interspeech 2020, pp. 1748–1752. doi:10.21437/Interspeech.2020-2382, arXiv:2104.11673.

Natour, Y.S., Wingate, J.M., 2009. Fundamental frequency characteristics of jordanian arabic speakers. Journal of Voice 23, 560–566. URL: https://www.sciencedirect.com/science/article/pii/S0892199708000064, doi:10.1016/j.jvoice.2008.01.005.

Patel, R.R., Awan, S.N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Švec, J.G., Hillman, R., 2018. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. American Journal of Speech-Language Pathology 27, 887–905. doi:10.1044/2018_AJSLP-17-0009.

Petrovic-Lazic, M., Jovanovic, N., Kulic, M., Babac, S., Jurisic, V., 2015. Acoustic and Perceptual Characteristics of the Voice in Patients With Vocal Polyps After Surgery and Voice Therapy. Journal of Voice 29, 241–246. doi:10.1016/j.jvoice.2014.07.009.

Puts, D.A., Apicella, C.L., Cárdenas, R.A., 2012. Masculine voices signal men's threat potential in forager and industrial societies. Proceedings of the Royal Society B: Biological Sciences 279, 601–609. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234546/, doi:10.1098/rspb.2011.0829, arXiv:21752821.

Qi, Z., Hu, X., Zhou, W., Li, S., Wu, H., Lu, J., Xu, X., 2023. LE-SSL-MOS: Self-Supervised Learning MOS Prediction with Listener Enhancement. doi:10.48550/arXiv.2311.10656, arXiv:2311.10656.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., Saruwatari, H., 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. doi:10.48550/arXiv.2204.02152, arXiv:2204.02152.

Sellam, T., Bapna, A., Camp, J., Mackinnon, D., Parikh, A.P., Riesa, J., 2023. SQuId: Measuring Speech Naturalness in Many Languages, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. doi:10.1109/ICASSP49357.2023.10094909, arXiv:2210.06324.

Shi, Y.F., Ai, Y., Lu, Y.X., Du, H.P., Ling, Z.H., 2024. SAMOS: A Neural MOS Prediction Model Leveraging Semantic Representations and Acoustic Features. doi:10.48550/arXiv.2411.11232, arXiv:2411.11232.

Stan, A., 2022. The ZevoMOS entry to VoiceMOS Challenge 2022. doi:10.48550/arXiv.2206.07448, arXiv:2206.07448.

Teixeira, J., Oliveira, C., Lopes, C., 2013. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. Procedia Technology 9, 1112–1122. doi:10.1016/j.protcy.2013.12.124.

Tseng, W.C., Kao, W.T., Lee, H.y., 2022. DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores. doi:10.48550/arXiv.2204.03219, arXiv:2204.03219.

Vashkevich, M., Petrovsky, A., Rushkevich, Y., 2019. Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato, in: 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), IEEE, Poznan, Poland. pp. 267–272. doi:10.23919/SPA.2019.8936657.

Wertzner, H.F., Schreiber, S., Amaro, L., 2005. Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. Brazilian Journal of Otorhinolaryngology 71, 582–588. doi:10.1016/S1808-8694(15)31261-1.

Yang, Z., Zhou, W., Chu, C., Li, S., Dabre, R., Rubino, R., Zhao, Y., 2022. Fusion of Self-supervised Learned Models for MOS Prediction. doi:10.48550/arXiv.2204.04855, arXiv:2204.04855.

Yumoto, E., Sasaki, Y., Okamura, H., 1984. Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. Journal of Speech, Language, and Hearing Research 27, 2–6. doi:10.1044/jshr.2701.02, arXiv:https://pubs.asha.org/doi/pdf/10.1044/jshr.2701