Speech Communication

Voice quality modification using glottal pulse parameters in the WORLD vocoder

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Section/Category:	Regular Paper
Keywords:	Voice Quality; Vocoder
Corresponding Author:	Matías Zañartu Valparaiso, Chile
First Author:	Benjamin Opazo
Order of Authors:	Benjamin Opazo
	Matías Zañartu, Ph.D.
Abstract:	Self-perception of voice quality is a poorly understood phenomenon that is hypothesized to play a critical role in the development of hyperfunctional voice disorders. By means of voice quality perturbation experiments altering auditory feedback, researchers could further investigate the underlying laryngeal motor control mechanisms. However, there are no synthesizers capable of altering voice quality in a physiologically relevant and controlled manner, even less so in real-time. This study introduces a Vocoder capable of altering voice quality at the glottal source level with very low latency and high quality resynthesized signals. The proposed Vocoder is built upon the well-known WORLD synthesizer by adding a Rosenberg++ pulse as glottal excitation signal with a special shape control parameter that simultaneously alters the instants of maximum excitation, maximum flow, and the return phase. In addition, methods for altering fundamental frequency, spectral envelope, and voice aperiodicity are considered. Both perceptual and objective experiments were carried out to assess the ability of the proposed method to control voice quality. Perceptual experiments were based on a CAPE-V assessment for six simulated qualities (modal, breathy, rough, vocal fry, dysphonia) and objective experiments related the control parameters with standard measures of voice quality such as spectral tilt, cepstral peak prominence, etc. Both evaluations illustrate that the resynthesized voice is natural, physiologically accurate and that it can mimic different pathological voice qualities for vocally healthy subjects. Further efforts are needed to assess the performance of the proposed system in the opposite direction, i.e., mimicking normal voices for patients. Computational performance results show sufficiently low latency to enable future real-time implementation of the proposed Vocoder in an embedded system.

Universidad Técnica Federico Santa María,
Avenida España 1680, Valparaíso, CL 2390123

Editor-in-Chief
Speech Communication
Dear Sir/Madam,
We wish to submit an original research article titled *Voice quality modification using glottal pulse parameters in the WORLD vocoder* for consideration by Speech Communication. We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

In this paper, we present an extension of the well-known WORLD Vocoder adding a Rosenberg++ pulse as glottal excitation signal with a special shape control parameter that simultaneously alters the instants of maximum excitation, maximum flow, and the return phase. The resulting Vocoder is capable of modifying voice quality in a physiologically accurate way, it can mimic different pathological voice qualities for vocally healthy subjects and the synthetic voice is natural.

Please address all correspondence concerning this manuscript to me at matias.zanartu@usm.cl.

Thank you for the consideration of this manuscript.
Sincerely,

Corresponding author
Matías Zañartu, PhD, Assistant Professor
Department of Electronic Engineering, Universidad Técnica Federico Santa María., Valparaíso, Chile

Co-authors
Benjamín Opazo, MS Student
Department of Electronic Engineering, Universidad Técnica Federico Santa María., Valparaíso, Chile

Highlights:

- The WORLD Vocoder is extended by adding a Rosenberg++ pulse as glottal excitation signal
- Vocal Quality is modified in real-time by using the proposed extended WORLD Vocoder
- Six simulated qualities are achieved: modal, breathy, rough, vocal fry and dysphonia
- A perceptual experiment shows that the synthethic modified voice is natural
- An objective assesment shows that voice quality is successfully modified

Voice quality modification using glottal pulse parameters in the WORLD vocoder

Benjamín Opazo[a], Matías Zañartu[a]

[a] *Universidad Tecnica Federico Santa Maria, Electronics Department, Avenida Espana 1680, Valparaiso, 2390123, Chile*

Abstract

Self-perception of voice quality is a poorly understood phenomenon that is hypothesized to play a critical role in the development of hyperfunctional voice disorders. By means of voice quality perturbation experiments altering auditory feedback, researchers could further investigate the underlying laryngeal motor control mechanisms. However, there are no synthesizers capable of altering voice quality in a physiologically relevant and controlled manner, even less so in real-time. This study introduces a Vocoder capable of altering voice quality at the glottal source level with very low latency and high quality resynthesized signals. The proposed Vocoder is built upon the well-known WORLD synthesizer by adding a Rosenberg++ pulse as glottal excitation signal with a special shape control parameter that simultaneously alters the instants of maximum excitation, maximum flow, and the return phase. In addition, methods for altering fundamental frequency, spectral envelope, and voice aperiodicity are considered. Both perceptual and objective experiments were carried out to assess the ability of the proposed method to control voice quality. Perceptual experiments were based on a CAPE-V assessment for six simulated qualities (modal, breathy, rough, vocal fry, dysphonia) and objective experiments related the control parameters with standard measures of voice quality such as spectral tilt, cepstral peak prominence, etc. Both evaluations illustrate that the resynthesized voice is natural, physiologically accurate and that it can mimic different pathological voice qualities for vocally healthy subjects. Further efforts are needed to assess the performance of the proposed system in the opposite direction, i.e., mimicking normal voices for patients. Computational performance results show sufficiently low latency to enable future real-time implementation of the proposed Vocoder in an embedded system.

Keywords: Voice Quality, Vocoder
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

Auditory feedback plays an important role in speech production [1], and by modifying auditory feedback, it is possible to study the underlying mechanisms of speech motor control. A well-known example of perturbed auditory feedback is the Delayed Auditory Feedback (DAF), where a delay of 200ms may induce an increase in speech errors and a decreased speech rate[2]. Auditory feedback perturbation including formant changes [3] and pitch perturbation [4][5] has been used to study neurological mechanisms of voice production. The role of self-perception of voice quality has not been explored in literature and it is believed to play a critical role in the development of hyperfunctional voice disorders [6].

In order to synthesize and modify voice in real-time, a specialized Vocoder is needed. Usually, Vocoders are not designed with real-time capabilities, or perform poorly in real-time conditions. An early example of a high quality real-time Vocoder is STRAIGHT[7], but resulting voices degrade in quality if the Vocoder is implemented in real-time, hence STRAIGHT has been phased out by more modern implementations of similar ideas such as TANDEM-STRAIGHT[8][9], which attempts to simplify the STRAIGHT Vocoder but its real-time performance is degraded with respect to the original Vocoder. The WORLD Vocoder[10] is a modern Vocoder that aims to synthesize high-quality real-time voice. It is an order of magnitude faster than the STRAIGHT Vocoder and perceptually performs better than the STRAIGHT Vocoder. The Vocoder decomposes the voice signal into its fundamental frequency, spectral envelope and aperiodicity, which allows for the modification of different characteristics of voice. Using machine learning to synthesize high-quality voice in real-time is possible using Vocoders such as Vocaine [11], which uses traditional Vocoders for the analysis and machine learning for the synthesis. An important shortcoming with this approach is the need for a dataset with pathological voices, and also that the subsequent trained model will probably be limited by dataset biases, such as language.

This work proposes an extension of the state-of-the-art WORLD voice Vocoder to modify voice quality of a speaker with low latency, and it also assesses the performance of the extended Vocoder in terms of voice quality, using objective and perceptual measures of voice quality.

2. Methods

To modify voice quality, three main tasks were performed. First, the WORLD Vocoder was extended so that it can be used to modify voice quality, then, synthetic voices were generated

Email address: matias.zanartu@usm.cl (Matías Zañartu)

with modified voice quality to be evaluated both subjectively and objectively.

2.1. Overview

The WORLD Vocoder consists of three estimation blocks; DIO or HARVEST for the fundamental frequency estimation [12][13], Cheaptrick for the spectral envelope estimation [14], and D4C for the aperiodic parameter estimation[15]. Then, a synthesis block uses an impulse train along with the estimated parameters to synthesize natural sounding voice. To modify voice quality using the WORLD Vocoder, three modifications were done to the Vocoder; 1) The excitation signal of the synthesis part was changed from an impulse train to a Rosenberg++ (R++) [16][17] pulse train, 2) the spectral envelope estimation was modified so that the estimated spectrum is consistent with the new excitation signal, 3) and several modifications were applied to the parameters and variables of the Vocoder during the synthesis time to modify voice quality. Figure 1 shows the proposed WORLD Vocoder extension, where the green blocks are modified with respect to the original implementation, and purple blocks are new proposed blocks. For synthesis, a set of parameters was created that are used to modify voice quality.

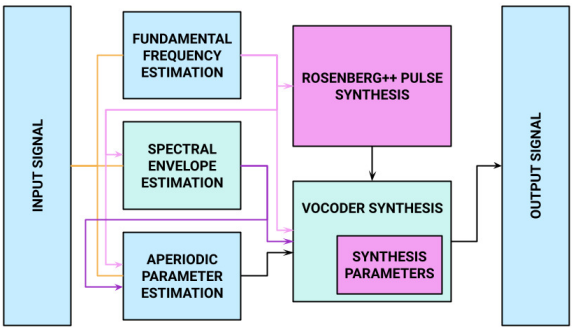


Figure 1: Proposed Extended WORLD Vocoder. Blue blocks are left unmodified, green blocks are modified with respect to the original implementation, and purple blocks are new proposed blocks. The Vocoder Synthesis block has an internal section where new parameters are defined.

The Rosenberg++ pulse is controlled using a wave shape parameter known as the R_d parameter [18], which is related to the quantification of the covariance of the parameters of the Liljencrants-Fant Model, although it also works with the R++ model, where lower values of the R_d parameter are related to relaxed glottal pulses and higher values are related to tighter phonation as it can be seen in Figure 2. This parameter has been successfully used to modify the breathiness of synthesized voices [19] [20] [21], but this work presents a framework capable to do it in real-time if implemented in an embedded system.

Modal voice was also resynthesized using the R++ as an excitation signal with mixed results; while the synthesized voice still sounds natural and similar to the input signal, the resulting voices are rougher, leaving place for improvements of the modal voice synthesis. Jitter and shimmer were synthesized by

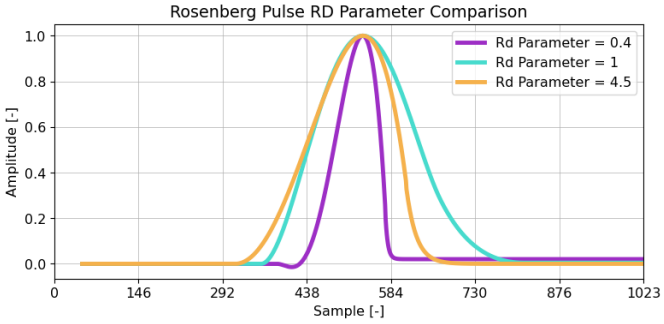


Figure 2: Comparison of Rosenberg++ Pulse Synthesized with different values of the R_d Parameter. Lower values of the R_d Parameters show a tighter pulse. When the R_d parameter is lower than 0.4, a small negative section starts appearing at the beginning of the pulse.

modulating the frequency and amplitude of the excitation signals. To synthesize vocal fry, the fundamental frequency was lowered to the range of $50 - 75$ [Hz] with good results. Dysphonic voices were synthesized by lowering the amplitude of the excitation signal while maintaining lower values of the R_d parameter with respect to breathy voices. Rough voice was synthesized with by slightly lowering the fundamental frequency and adding jitter and shimmer, but results were not satisfactory.

2.2. WORLD Vocoder extension

The original implementation of the WORLD Vocoder uses an impulse train as an excitation signal, which means that the design of the Vocoder considers it in the implementation of the spectral envelope estimation (Cheaptrick). If no modifications are made to the Cheaptrick after modifying the synthesis module, then the synthesized voice sounds muffled with the lower frequencies exacerbated. This is due to the fact that the spectral envelope that extracts Cheaptrick includes the effect of the glottal folds, which is similar to a low-pass filter, and the excitation signal using a R++ pulse works as a second cascading low-pass filter applied to the spectrum. To solve this problem, a solution is proposed where the estimated spectral envelope is modified using the following function:

$$f[n] = \exp(n\frac{p}{m} - p) \tag{1}$$

Where $n \leqslant m$ and p is related to the slope of the exponential. This modifier is applied to the first m samples. Note that $f[m] = 1$, so that the change between the modified spectrum and the unmodified part is smooth. The proposed modifier is applied directly to the estimated spectral envelope array frame by frame. Results shows that by tuning the m and p parameters it is possible to synthesize natural sounding voice that resembles the input signal.

The R++ pulse is used as an excitation signal for the synthesis module. The implementation of the R++ does not differ from the specification of the pulse [16][17], except for 1) if $t_p = 2t_e/3$ or $|t_p - t_e| < 0.001$ the pulse is not synthesized, and 2) the amplitude is defined as the maximum value of the pulse at $f(t_p)$, which differs to the original implementation where the amplitude is defined as the maximum value of the derivative of

the pulse. To calculate the amplitude, an internal parameter K_{int} is defined as follows. If $t_a > 0$

$$K_{int} = \begin{cases} 3K/(t_p^3(2t_x - t_p)) & t_p < \frac{(4D(t_0,t_e,t_a)t_a t_e + t_e^2)}{(2D(t_0,t_e,t_a)t_a + t_e)} \\ -3K/(t_p^3(2t_x - t_p)) & \text{otherwise} \end{cases} \quad (2)$$

If $t_a = 0$, then K_{int} is defined as

$$K_{int} = \begin{cases} 3K/(t_p^3(2t_x - t_p)) & t_p < t_e \\ -3K/(t_p^3(2t_x - t_p)) & \text{otherwise} \end{cases} \quad (3)$$

To avoid a division by 0, the restriction $|t_p - t_e| < 0.001$ is applied. With this implementation of the R++ pulse, it is possible to determine the amplitude of the pulse with the parameter K.

The synthesis module was extended so that it can use different parameters that allow for the modification of voice quality. The parameters that are implemented in the synthesis block are the following:

i) R_d Parameter: This parameter controls the R_d value used to synthesize a R++ Pulse. The R_d parameter is used to obtain the values of the t_e, t_p and t_a parameters of the pulse using the (COVAREP) toolbox [22](v1.4.2), which corresponds to the instant of maximum excitation, the instant of the maximum glottal flow, and the return phase constant, respectively.

ii) F_0 filter frequency and order: These parameters apply a low-pass Butterworth filter to the F_0 array used by the Vocoder, controlling the frequency and the order of the filter.

iii) F_0 multiplier: This parameter is used to multiply the values of the F_0 array with a fixed value.

iv) Jitter Amplitude and Frequency: These parameters are used to add jitter to the voice, by modulating the array of the F_0 estimated values with multiplicative Brownian noise. The noise is generated by low-pass filtering white Gaussian noise with a Butterworth filter, where the Amplitude parameter refers to the power of the Gaussian noise and the Frequency parameter refers to the cutoff frequency of the low-pass filter.

v) Shimmer Amplitude and Frequency: These parameters are used to add shimmer to the voice, by modulating the amplitude of the synthesized pulses used for the excitation signal. Like the jitter parameters, the amplitude is modulated with Brownian noise generated with the low-pass filtering of a Butterworth filter, but in this case, the noise is added. The amplitude controls the power of the white Gaussian noise and the Frequency controls the cutoff frequency of the low-pass filter.

vi) Vibrato Amplitude and Frequency: These parameters add vibrato to the F_0 estimated array by multiplying it with a sine wave, with an amplitude defined by the Amplitude parameter and the frequency determined by the Frequency parameter.

vii) Spectrum Filtering p and m: This parameter controls the values of the p and m variables of Eq. 1

viii) R++ t_e, t_p and t_a multiplier: These multipliers allow for a fine tuning of the t_e, t_p and t_a parameters obtained with the R_d parameter. It multiplies each value with a number, modifying the values used for the synthesis of the pulse.

ix) R++ K multiplier: This multiplier changes the amplitude of the synthetic R++ pulse.

x) Band Aperiodicity Multiplier: This parameter modifies the estimated aperiodic parameter obtained by D4C in the analysis phase of the Vocoder. It multiplies the values of the aperiodic parameter for each individual frequency band that the Vocoder uses.

2.3. Voice quality modification: perceptual assessment

In order to modify voice quality by adding different pathological sounds to the input signals, the Vocoder uses different parameters and values, which is highly dependant of the input voice. Table 1 shows the parameters used to synthesize different pathological voices, but not the specific values. The input signals were taken from the Perceptual Voice Quality Database [23], the male signal corresponds to audio file *LA_9011_ENSS* and the female signal to audio file *LA9023_ENSS*.

A perceptual assessment task was done on the resulting voices with three expert raters using the CAPE-V evaluation tool [24] and adding two extra parameters to be evaluated: Vocal Fry and Aperiodicity.

2.4. Voice quality modification: objective assessment

The same two input files that are used in section 2.3 are used in this experiment, with the sole difference that only two sections of the audio files are used, one where the /ɑː/ utterance is said and one where the subject says "*Peter will keep at the peak*".

The signals are then synthesized while modifying a single parameter in a specified range, as seen in Table 2. For each synthesized signal a set of objective parameters are calculated which are shown and defined in Table 3.

3. Results

Results indicate that both the perceptual and objective assessment of the synthesized signals are satisfactory. The worst performing synthetic signal in the perceptual assessment was the rough voice both for male and female subject. The objective assessment shows that the synthesis parameters are physiologically relevant.

3.1. Voice quality modification: perceptual assessment

Tables 4 and 5 show the averaged results of the three perceptual evaluations. All synthetic voices were evaluated using the CAPE-V protocol with three expert raters.

Results show that the modal voice synthesis works as expected for the male voice, synthesizing a signal with similar CAPE-V values, whereas the female synthesis shows mildly deviant values of overall severity, roughness and aperiodicity. This means that the Vocoder needs more fine tuning for the synthesis of female voices. This suggests that the synthesized

Parameter	Modal	Breathy	Vocal fry	Rough Voice	Dysphonia
R_d	1	1	1	1	1
F_0 filter frequency and order	0	0	0	0	0
$F0$ multiplier	0	0	1	1	0
Jitter Amplitude and Frequency	0	0	1	1	0
Shimmer Amplitude and Frequency	0	0	1	1	0
Vibrato Amplitude and Frequency	0	0	0	0	0
Spectrum Filtering p and m	1	1	1	1	1
R++ t_e, t_p and t_a multiplier	1	1	1	1	1
R++ K multiplier	1	1	1	1	1
Band Aperiodicity Multiplier	0	1	0	0	1

Table 1: Parameters used to synthesize modified voice quality. A 1 means that the parameter is used and a 0 means that the parameter is not used

Parameter	**Min Value**	**Step Size**	**Max Value**
R_d	0.35	0.05	4
$F0$ multiplier	0.5	0.1	4
Jitter Amplitude	0	2	50
Jitter Frequency	0	100	22000
Shimmer Amplitude	0	2	50
Shimmer Frequency	0	100	22000
R++ K multiplier	0	0.2	5

Table 2: Minimum, maximum and step values of the parameters

female voice using the R++ Pulse and the proposed WORLD Vocoder is not synthesizing a completely authentic voice, although in terms of the overall severity, it is the best rated voice for male and female voice, which serves as a baseline for comparison when modifying voice quality. It is worth noting that, however, the rough synthesized voice has a comparable value of overall severity for the case of female voice.

The synthesized breathy voice shows that the male voice is severely deviant, although the female voice needs more work, with a moderately deviant score for overall severity and a lower than expected value of perceived breathiness. For the dysphonic voice, results show that the synthetic voice has high values of breathiness (severely deviant), but roughness and aperiodicity are also present with moderately deviant values, and mildly deviant values of strained voice. The breathiness value for the female dysphonic synthetic voice show higher values than its breathy synthetic counterpart, and the male synthetic voice shows similar values, this indicates that breathy voice needs more fine tuning, and some characteristic of the dysphonic synthesis may be useful like lower values of the pulse amplitude. Vocal fry was successfully synthesized for both male and female voice, showing severely deviant values for the male voice and moderately deviant values for the female voice. It was not possible to synthesize severely deviant vocal fry voices for the female voice, and another method is probably needed to create a stronger effect. Rough voice was unsatisfactory, and it shows the mixed results. It was possible to synthesize mildly deviant values of roughness, but these values appear in the other synthesized voices, with higher values in the dysphonic voice, and in the case of the female voice, the dysphonic voice is severely rough, whereas the rough voice is not, which probably means that the implementation of both rough and dysphonic voices needs more tuning. In general, all voices, except for the male rough voice, show higher values of the overall severity with respect to the original voice, and even with respect to the synthetic modal voice. According to one of the expert raters, the synthesizer is working for all the evaluated voices except for the rough voice, and the modal voice needs more work.

3.2. Voice quality modification: objective assessment

In accordance with the perceptual results and literature [26][29][28][30], Figure 3a shows that CPP is $6[dB]$ lower when the R_d parameter has high values, Figure 3b shows that the spectral envelope varies from 10 to $40[dB]$, and in Figure 3c, the peak slope varies from -0.5 to -0.4. All these values indicate that the voice is changing from tense to breathy. Results of measured PESQ show that in general, higher values of the R_d parameter synthesize a more different voice with respect to the original signal. It can be noted that for lower values of the R_d parameter, the female sustained signal has values of over 4 in the PESQ measurements. As expected, the other parameters do not show correlations to the R_d parameter.

It is clear from Figure 4a that the peak value of the PESQ is when the F0 multiplier is 1, which is consistent to the idea that not modifying the fundamental frequency allows for a voice that is perceptually similar to the original voice. It is important to note that this happens with all four signals, and that the peak values are similar to the peak values of Figure 3d. This is an important result because when the F0 multiplier is 1, the Vocoder is not modifying in any aspect the voice, which means that the synthesized voice with the R_d parameter can synthesize modal voice like the input signal. The other measurements are not necessarily related to the parameter, even though the results of Figure 4b are attention grabbing, in the sense that there is a strong correlation with the PS, this could be a spurious correlation. Figure 4c shows that HNR is correlated with the multiplier. This is probably related to the specific implementation of the HNR, although this could not be checked because the details of the implementation of this parameter are not publicly available. The parameter is not related to the CPP as expected,

Objective Measurement	Description
Cepstral Peak Prominence (CPP)	Measure of the cepstral peak amplitude normalized for overall amplitude [25]. It is known to be a robust measure of dysphonia severity and breathy voices [26]
Harmonic-to-Noise Ratio (HNR)	Measure of the Ratio of the estimated harmonic component of the signal and the noise component of the signal. Note that, in order to estimate the harmonic component, an F_0 estimation is needed, which is not done using DIO or Harvest from the WORLD Vocoder
Mean Jitter	Corresponds to the pitch period deviations, calculated in 0.2 [ms] frames, and averaged across the signal[27].
Mean Shimmer	Corresponds to the pitch period amplitude deviations, calculated in 0.2 [ms] frames, and averaged across the signal[27].
Perceptual Evaluation of Speech Quality (PESQ)	Measure used for the automated assessment of speech quality. To measure the PESQ, a python implementation was used, corresponding to the pip package PESQ V0.0.3
Peak Slope (PS)	Corresponds to the regression line that fits the log10 of the peaks of the time domain maxima of the signal that was previously decomposed in octaves. This parameter is used to identify breathy to tense voice qualities [28]
Spectral Envelope H1-H2 (SE)	Corresponds to the difference in decibels (or energy ratio) of the first harmonic (H1) and second harmonic (H2).

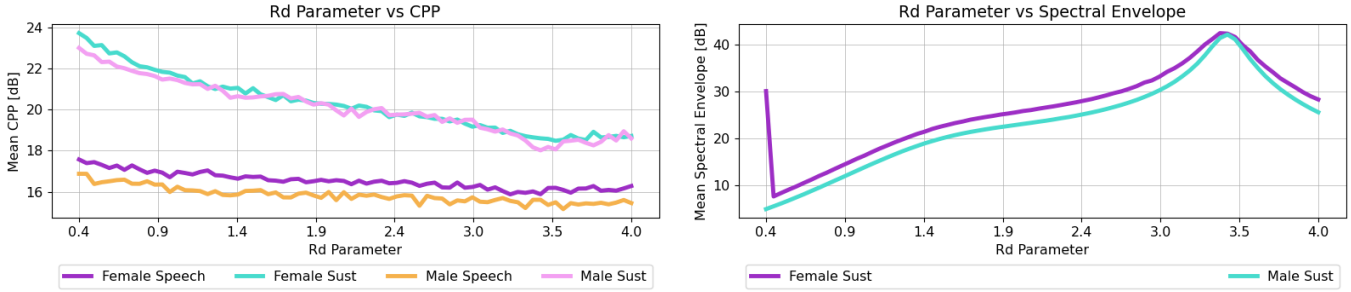
Table 3: Objective measurements definitions

Label	Original	Modal	Breathy	Rough	Vocal Fry	Dysphonia
Overall Severity	12	7	53	17	33	73
Roughness	0	0	17	27	7	58
Breathiness	0	3	57	3	3	53
Strain	8	0	12	17	25	38
Pitch	0	0	0	0	20	0
Loudness	0	0	10	0	10	17
Vocal Fry*	12	10	3	23	70	0
Aperiodicity*	12	10	13	30	30	43

Table 4: Averaged CAPE-V results of the original and synthesized signals of the male subject

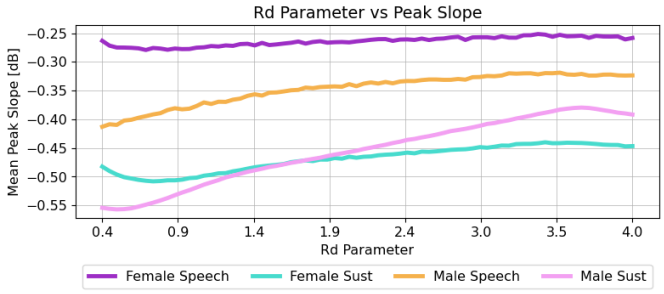
Label	Original	Modal	Breathy	Rough	Vocal Fry	Dysphonia
Overall Severity	3	20	40	33	43	60
Roughness	0	10	30	30	20	60
Breathiness	0	3	27	20	23	50
Strain	0	7	13	17	20	23
Pitch	0	0	0	0	33	0
Loudness	0	0	7	0	17	0
Vocal Fry*	0	0	0	17	40	0
Aperiodicity*	7	20	13	33	23	37

Table 5: Averaged CAPE-V results of the original and synthesized signals of the female subject

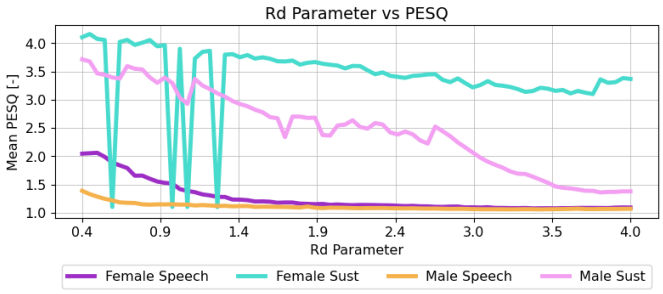


(a) R_d Parameter vs Mean CPP

(b) R_d Parameter vs Spectral Envelope



(c) R_d Parameter vs Peak Slope



(d) R_d Parameter vs Peak Slope

Figure 3: R_d Parameter results

however, there is a steep drop in the value in the female sustained vowel when the parameter is over 2.2.

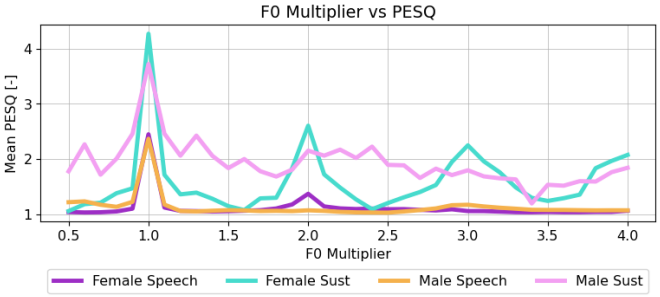
In Figure 5b it can be noted the drastic change of the CPP of $6[dB]$ in the case of the sustained vowels when the pulse amplitude changes from 0 to a higher value. This change is due to the fact that when this parameter is 0, there is no excitation pulse, and the synthesized signal is excited only by noise, emulating a breathy or dysphonic voice. It can be seen that this drastic change is repeated in Figure 5a, showing that when an excitation signal is present, the synthesized signal is closer to the input signal. The other parameters show a quick convergence to their maximum when R++ K multiplier parameter is greater than 0, with the exception of the measured jitter and shimmer, which is expected.

As expected, PESQ lowers when either the amplitude or the frequency of the jitter increases, being more susceptible to changes in frequency. The measured value of the jitter increases from 0.35% to 31% for changes in frequency (Figure 6b) and from 1.2% to 25% for changes in amplitude (Figure 6a), which is consistent with the values of the variation of parameters. The measured shimmer is also affected in a similar manner which is unexpected, although it is probably related to how the algorithm was implemented, where it shows mixed sensitivities with respect to jitter and shimmer. The amplitude and frequency of jitter are inversely correlated with the HNR, which is what is expected, as higher values of jitter are related with a lower harmonic component of the signal. On the other hand, the Shimmer Amplitude parameter and Shimmer Frequency parameter affect the measured shimmer as it is shown in Figures 6c and 6d, changing from 2.5% to 30% for changes in frequency and from 1.4% to 12% for changes in amplitude, which is consistent with the values of the parameters. Following a similar behaviour with respect to the 'Jitter Frequency parameter' and 'Jitter Amplitude parameter' parameters, the measured jitter for both amplitude and frequency of shimmer is affected. As expected, the PESQ values drop from near 4 to 1 when the 'Jitter Amplitude parameter' and 'Jitter Frequency parameter' parameters are higher.

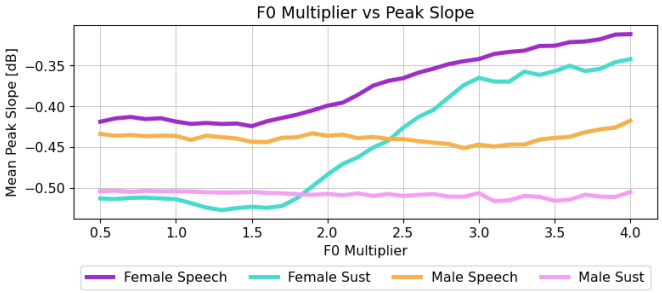
4. Conclusion

In this study, it was shown that it is possible to extend the WORLD Vocoder to modify voice quality in a physiologically relevant way. The R++ pulse was introduced as an physiologically relevant excitation signal with a special shape control parameter that simultaneously altered the instants of maximum excitation, maximum flow, and the return phase. In order to introduce this new and physiologically relevant excitation signal, core components of the WORLD Vocoder were adapted. In addition to the pulse shape, several source control parameters were included to account for other temporal variations. The resulting method combining these parameters allowed for altering voice quality to mimic breathiness, dysphonia, and vocal fry.

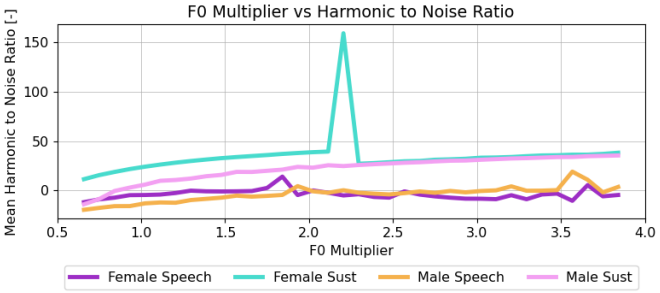
The perceptual relevance of the resulting synthesized signals was assessed using the CAPE-V framework. The proposed scheme was able to produce modal, breathy, dysphonia, and vocal fry qualities with good perceptual quality. However, the rough voice quality sounded less natural and its results are not perceptually satisfactory. Further work is needed to describe rough voice quality and to extend the vocoder to improve this performance. In general, the Vocoder exhibits a lower perceptual quality with female voices. Considering the multidimensional aspect of Voice Quality, it is difficult to completely iso-



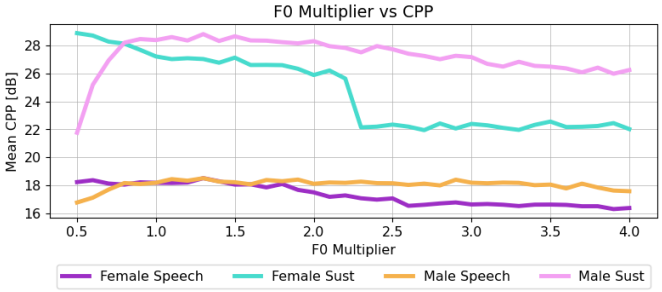
(a) F0 Multiplier vs PESQ



(b) F0 Multiplier vs Peak Slope

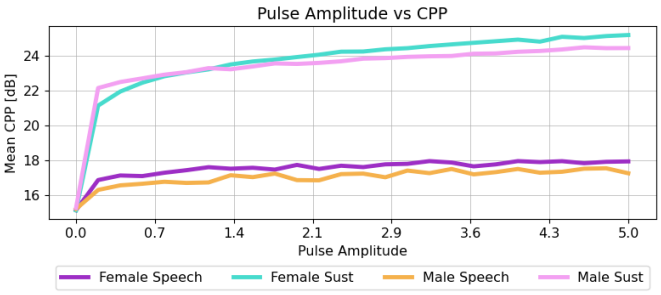


(c) F0 Multiplier vs HNR

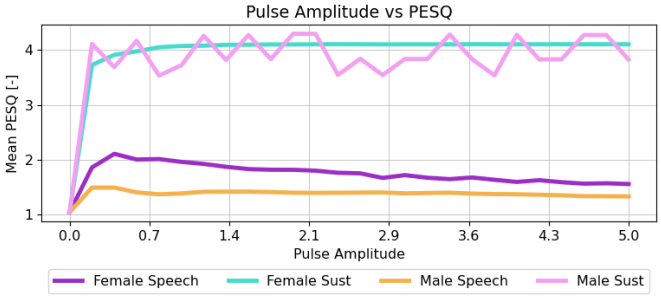


(d) F0 Multiplier vs CPP

Figure 4: F0 Multiplier Parameter results

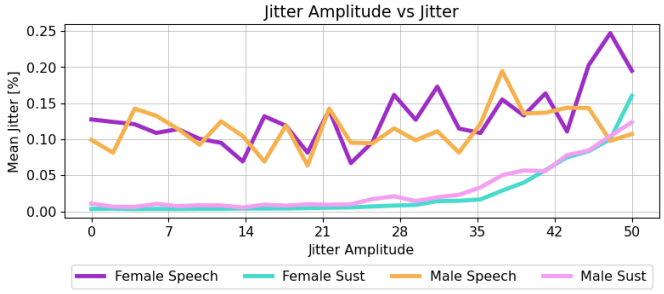


(a) F0 Multiplier vs CPP

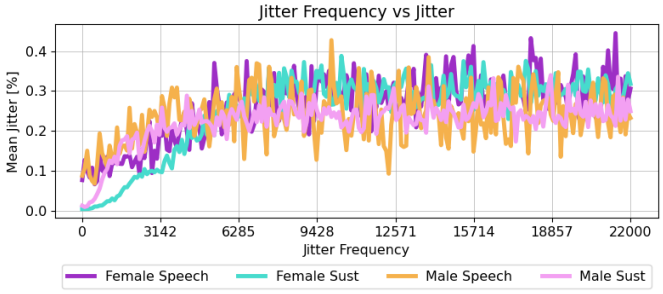


(b) F0 Multiplier vs PESQ

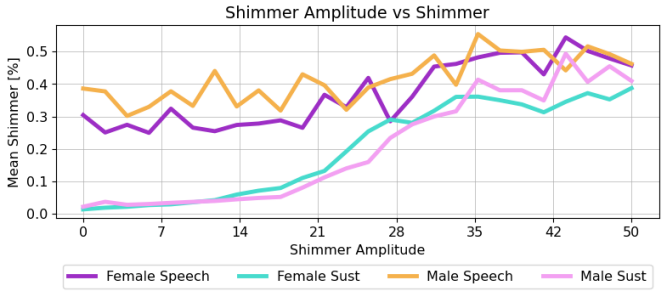
Figure 5: R++ K Multiplier Parameter results



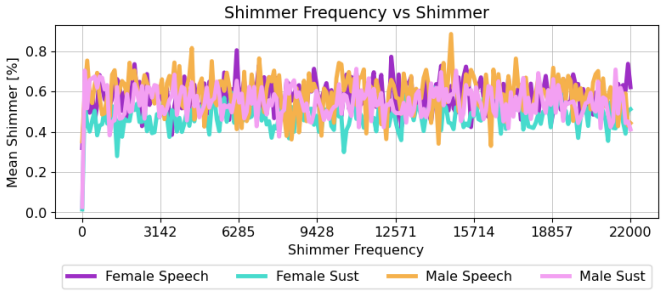
(a) Jitter Amplitude vs Mean Jitter



(b) Jitter Frequency vs Mean Jitter



(c) Shimmer Amplitude vs Mean Shimmer



(d) Shimmer Frequency vs Mean Shimmer

Figure 6: Jitter and Shimmer (Amplitude and Frequency) Parameter results

late the effect of a given label of voice quality with a unique transformation. Our results show that the proposed methods affect more than one component of Voice Quality, even when the objective is to modify only one. For example, dysphonic synthethic voices are associated with higher values of roughness, strain and aperiodicity, even when only the amplitude of the excitation pulse was modified.

The objective evaluation of the synthesized voices provided insights on the performance of the proposed parameters. The pulse shape parameter R_d yielded results that are in agreement with the literature, following the expected trends from CPP, PS, SE and PESQ. The F_0 multiplier was the main factor when synthesizing vocal fry and rough voices. The pulse amplitude had a clear effect on CPP and PESQ and was a key parameter to synthesize dysphonic voices. Changes in voice quality using jitter and shimmer showed relatively good performance in terms of the objective measures (mean jitter, mean shimmer), although its perceptual relevance was lower. In general, all parameters affect the resulting PESQ, which means that modifications are being made to the voice, although, each parameter perturbs differently the synthesized voice, affecting with different degrees the value of PESQ.

Note that experiments with sustained vowels showed clear trends of voice quality changes that are well aligned with the expected outcomes in the objective measures according to the literatyre. Even though they tend to have a more natural sound, experiments with running speech do not always agree with these trends. This is rather expected because the objective measurements are designed for sustained vowels, and in running speech the parameters are averaged between (sometimes drastically) different frames. In the case of the perceptual setup, the opposite effect is seen: When listening to running speech, the small imperfections of the synthesis remain unnoticed, whereas with sustained vowels, the imperfections of the synthesis are more easily noticed.

The modification of the Cheaptrick algorithm in the WORLD Vocoder allows to use different excitation signals. In this work only the R++ signal was used, but other glottal pulses could be considered, although with further modifications to the Cheaptrick algorithm. Another shortcoming of this work is that, at the moment, it is only possible to degrade voice quality rather than improve it. Future work could explore training a deep learning framework to change the estimated spectral envelope, and thus fixing vocal quality, as seen in Text-To-Speech synthesizers[31] and voice separation schemes[32] that use the WORLD Vocoder.

The proposed parameters are probably insufficient for the modification of all kinds of voice quality, although the proposed framework allow for further development in this area. Fine tuning the algorithms associated with the parameters and proposing new parameters can lead to better results of voice quality modification. Nevertheless, the proposed vocoder provides a framework for future investigation of real-time modification of voice quality in an embedded system and laryngeal motor control experiments.

5. Acknowledgments

This study was funded by ANID through grants BASAL FB0008 and FONDECYT 1191369. We acknowledge the Speech Language Pathologists Christian Castro, Diego Romero and Alejandro Herrera, who performed the perceptual CAPE-V assessments.

References

[1] J. L. Elman, Effects of frequency-shifted feedback on the pitch of vocal productions, The Journal of the Acoustical Society of America 70 (1981) 45–50. doi:10.1121/1.386580.

[2] J. Chesters, L. Baghai-Ravary, R. Möttönen, The effects of delayed auditory and visual feedback on speech production, The Journal of the Acoustical Society of America 137 (2015) 873–883. doi:10.1121/1.4906266.

[3] J. A. Tourville, K. J. Reilly, F. H. Guenther, Neural mechanisms underlying auditory feedback control of speech, NeuroImage 39 (2008) 1429–1443. doi:10.1016/j.neuroimage.2007.09.054.

[4] D. Abur, A. Subaciute, M. Kapsner-Smith, R. K. Segina, L. F. Tracy, J. P. Noordzij, C. E. Stepp, Impaired auditory discrimination and auditory-motor integration in hyperfunctional voice disorders, Scientific Reports 11 (2021). doi:10.1038/s41598-021-92250-8.

[5] D. Abur, C. E. Stepp, Acuity to changes in self-generated vocal pitch in parkinson's disease, Journal of Speech, Language, and Hearing Research 63 (2020) 3208–3214. doi:10.1044/2020_jslhr-20-00003.

[6] G. E. Galindo, S. D. Peterson, B. D. Erath, C. Castro, R. E. Hillman, M. Zañartu, Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds, Journal of Speech, Language, and Hearing Research 60 (2017) 2452–2471. doi:10.1044/2017_jslhr-s-16-0412.

[7] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné, Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds, Speech Communication 27 (1999) 187–207. doi:10.1016/s0167-6393(98)00085-5.

[8] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation, IEEE, 2008. doi:10.1109/icassp.2008.4518514.

[9] H. KAWAHARA, M. MORISE, Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework 36 (2011) 713–727. doi:10.1007/s12046-011-0043-3.

[10] M. MORISE, F. YOKOMORI, K. OZAWA, WORLD: A vocoder-based high-quality speech synthesis system for real-time applications, IEICE Trans. Inf. & Syst. Transactions on Information and Systems E99.D (2016) 1877–1884. doi:10.1587/transinf.2015edp7457.

[11] Y. Agiomyrgiannakis, Vocaine the vocoder and applications in speech synthesis, in: ICASSP, 2015.

[12] M. Morise, H. Kawahara, H. Katayose, Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech, in: Audio Engineering Society Conference: 35th International Conference: Audio for Games, 2009. URL: http://www.aes.org/e-lib/browse.cfm?elib=15165.

[13] M. Morise, Harvest: A high-performance fundamental frequency estimator from speech signals, in: Interspeech 2017, ISCA, 2017. doi:10.21437/interspeech.2017-68.

[14] M. Morise, CheapTrick, a spectral envelope estimator for high-quality speech synthesis, Speech Communication 67 (2015) 1–7. doi:10.1016/j.specom.2014.09.003.

[15] M. Morise, D4c, a band-aperiodicity estimator for high-quality speech synthesis, Speech Communication 84 (2016) 57–65. doi:10.1016/j.specom.2016.09.001.

[16] R. Veldhuis, A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation, The Journal of the Acoustical Society of America 103 (1998) 566–571. doi:10.1121/1.421103.

[17] B. Doval, C. d'Alessandro, N. Henrich, The spectrum of glottal flow models, Acta acustica united with acustica 92 (2006) 1026–1046.

[18] G. Fant, The lf-model revisited. transformations and frequency domain analysis, Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm 2 (1995) 40.

[19] G. Degottex, A. Roebel, X. Rodet, Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011. doi:10.1109/icassp.2011.5947511.

[20] G. Degottex, P. Lanchantin, A. Roebel, X. Rodet, Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis, Speech Communication 55 (2013) 278–294. doi:10.1016/j.specom.2012.08.010.

[21] A. Murphy, I. Yanushevskaya, A. N. Chasaide, C. Gobl, The role of voice quality in the perception of prominence in synthetic speech, in: Interspeech 2019, ISCA, 2019. doi:10.21437/interspeech.2019-2761.

[22] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP: A collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014. doi:10.1109/icassp.2014.6853739.

[23] P. R. Walden, Perceptual voice qualities database (pvqd), 2020. doi:10.17632/9DZ247GNYB.1.

[24] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, R. E. Hillman, Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol, American Journal of Speech-Language Pathology 18 (2009) 124–132. doi:10.1044/1058-0360(2008/08-0017).

[25] J. Hillenbrand, R. A. Cleveland, R. L. Erickson, Acoustic correlates of breathy vocal quality, Journal of Speech, Language, and Hearing Research 37 (1994) 769–778. doi:10.1044/jshr.3704.769.

[26] R. Fraile, J. I. Godino-Llorente, Cepstral peak prominence: A comprehensive analysis, Biomedical Signal Processing and Control 14 (2014) 42–54. doi:10.1016/j.bspc.2014.07.001.

[27] X. Na, opensmile, 2019. URL: https://github.com/naxingyu/opensmile.

[28] J. Kane, C. Gobl, Identifying regions of non-modal phonation using features of the wavelet transform., 2011, pp. 177–180.

[29] Y. D. Heman-Ackah, D. D. Michael, G. S. Goding, The relationship between cepstral peak prominence and selected parameters of dysphonia, Journal of Voice 16 (2002) 20–27. doi:10.1016/s0892-1997(02)00067-x.

[30] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, Z. Zhang, Toward a unified theory of voice production and perception, Loquens 1 (2014) e009. doi:10.3989/loquens.2014.009.

[31] Mozilla, Tts: Text-to-speech for all, 2021. URL: https://github.com/mozilla/TTS.

[32] M. Blaauw, J. Bonada, A neural parametric singing synthesizer modeling timbre and expression from natural songs, Applied Sciences 7 (2017) 1313. doi:10.3390/app7121313.


Click here to access/download
LaTeX Source File
bibliography.bib

Click here to access/download

LaTeX Source File

elsarticle.dtx

Click here to access/download

LaTeX Source File

elsarticle.ins

Figure 1 (Color)

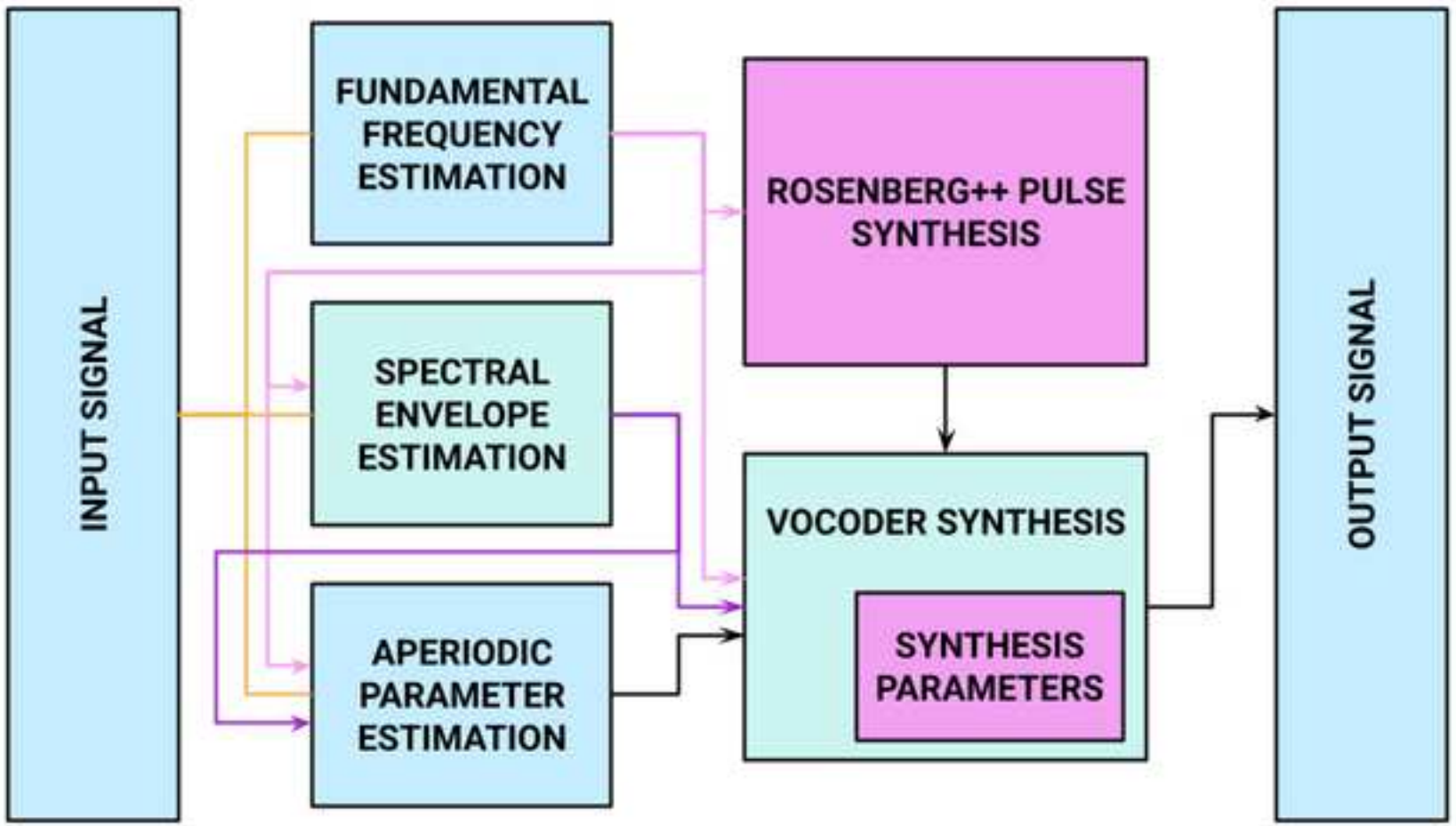


Figure 2 (Color) Click here to access/download;Figure;rosenberg_rd.png

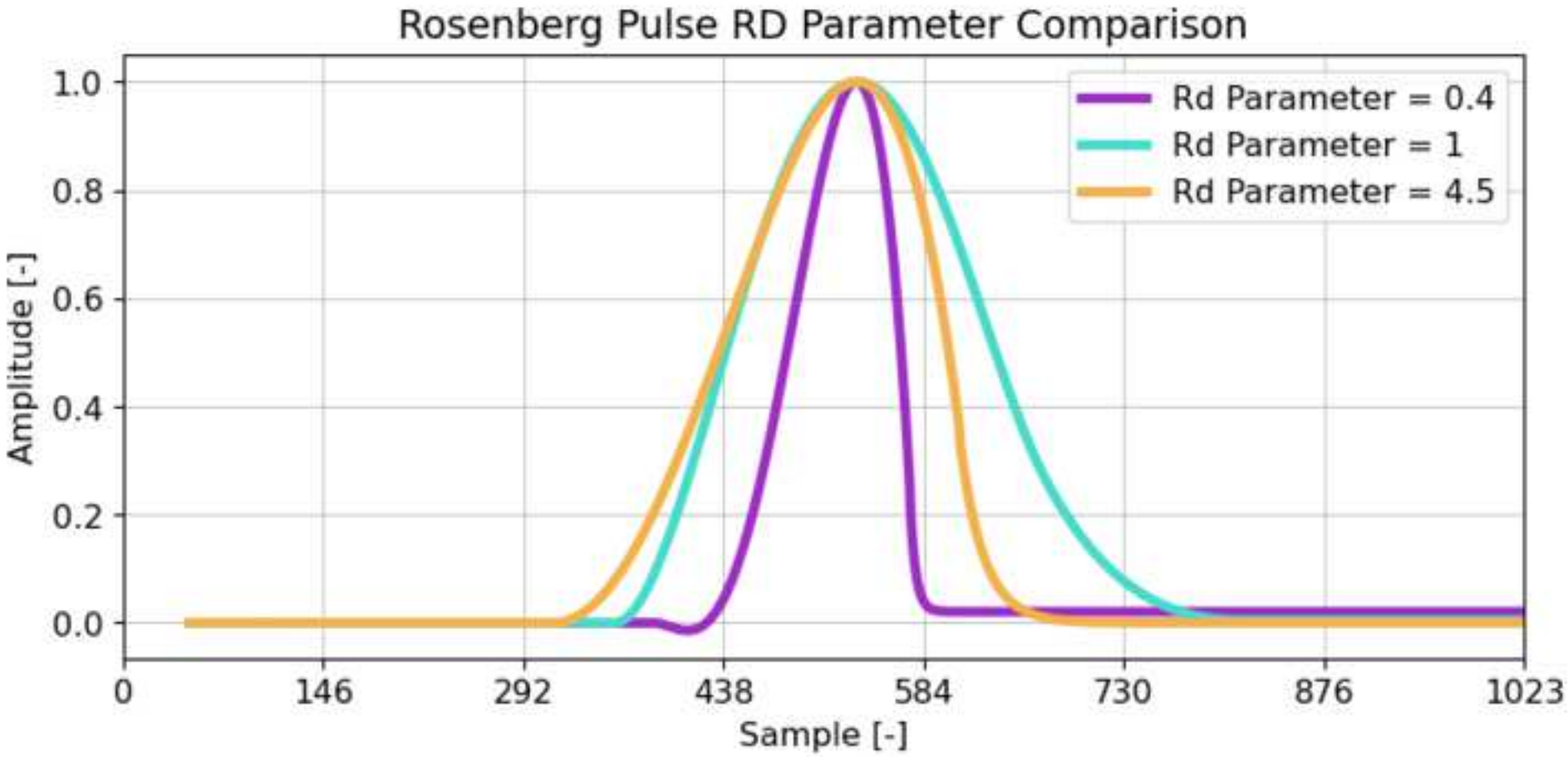


Figure 3a (Color)

Click here to access/download;Figure;rd_cpp.png

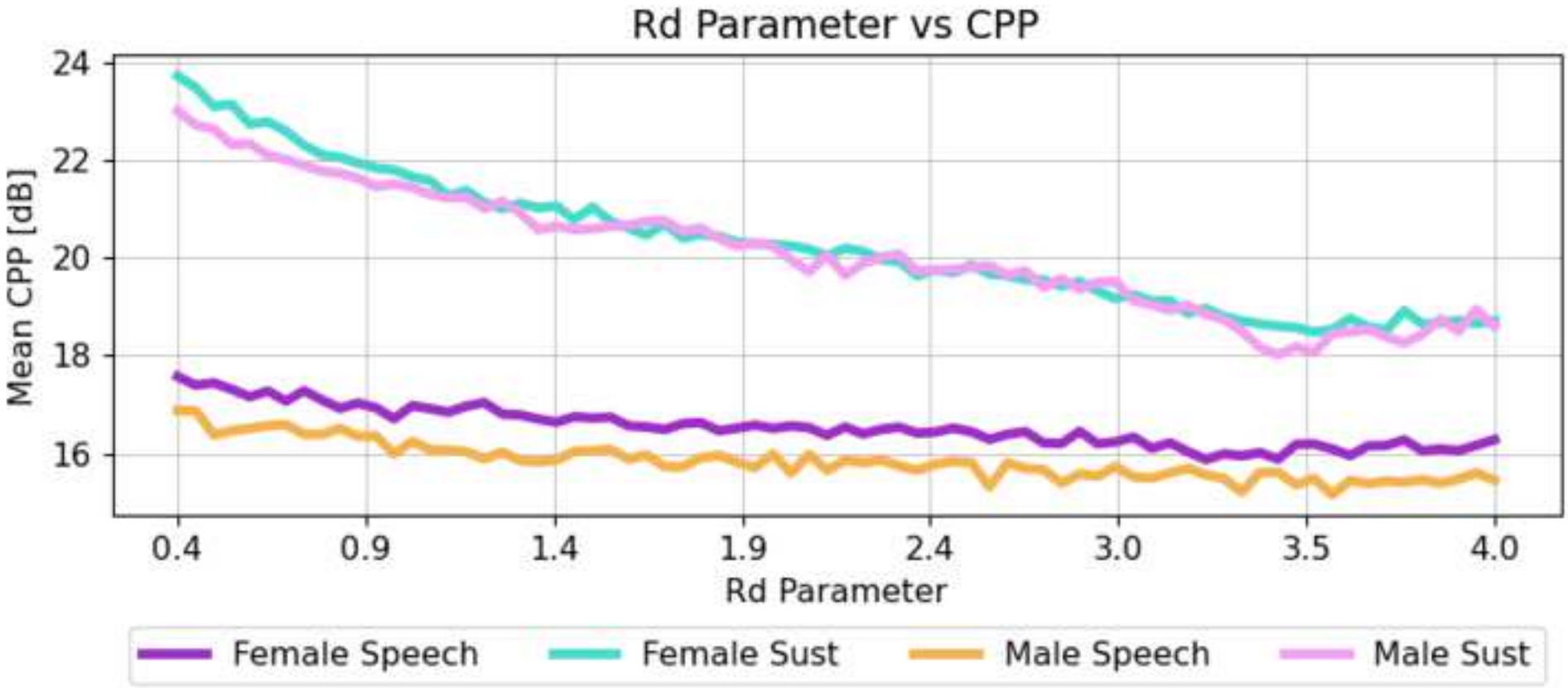


Figure 3b (Color)

Click here to access/download;Figure;rd_se.png

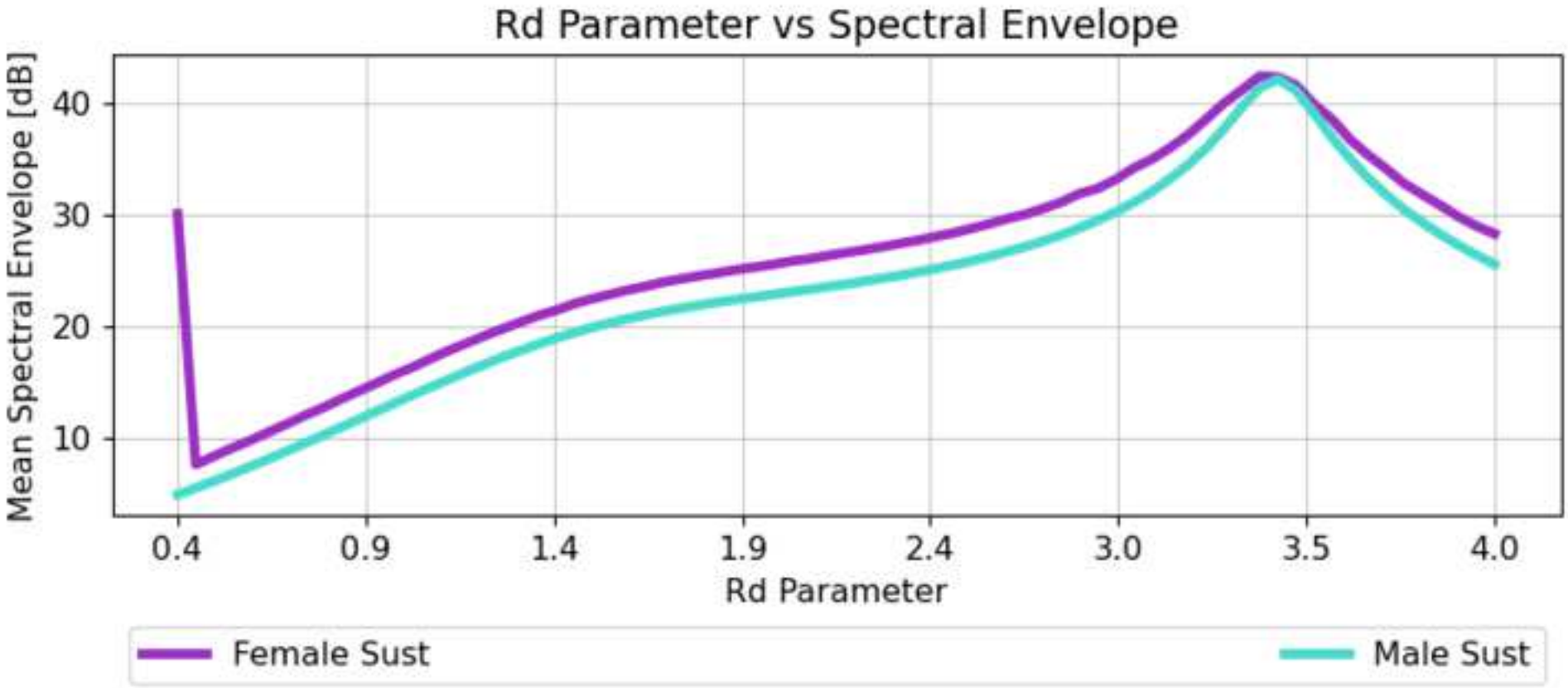


Figure 3c (Color)

Click here to access/download;Figure;rd_ps.png

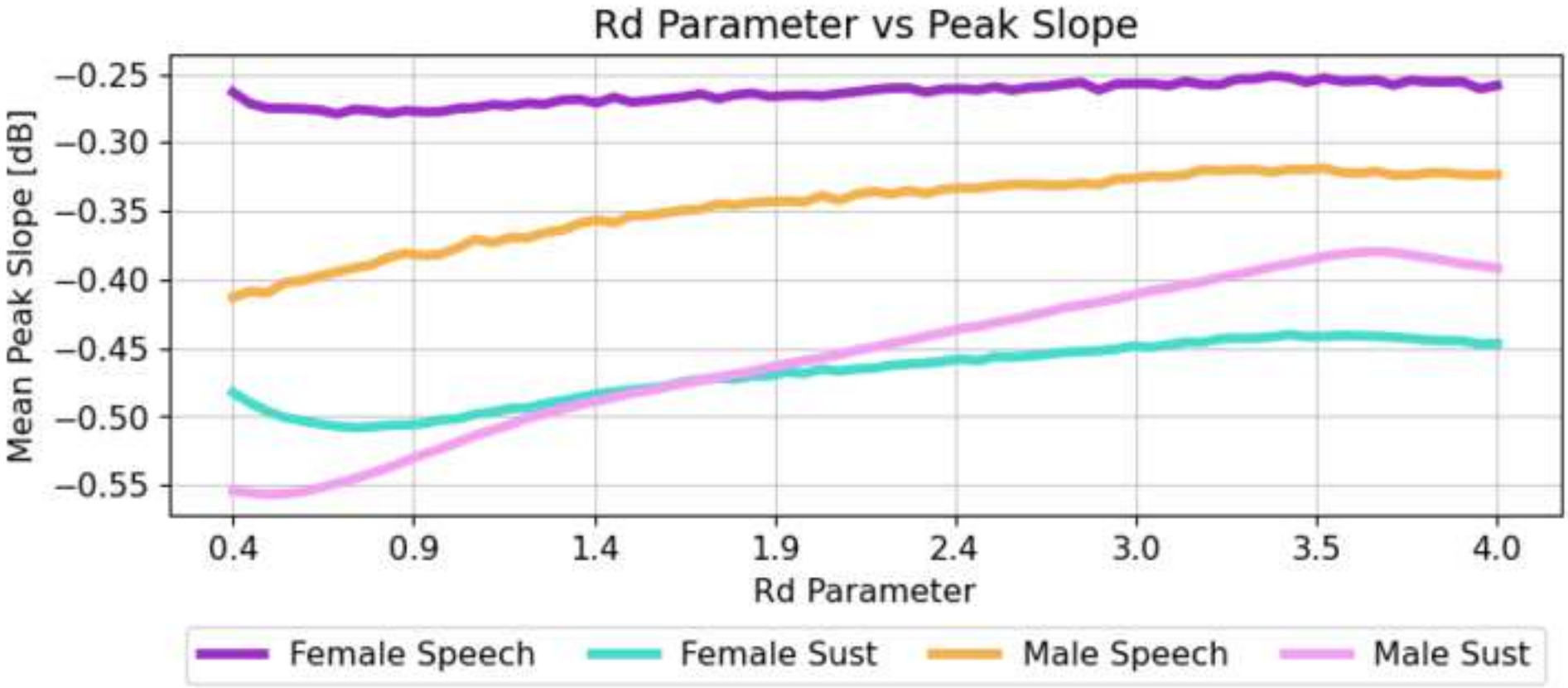


Figure 3d (Color)

Click here to access/download;Figure;rd_pesq.png

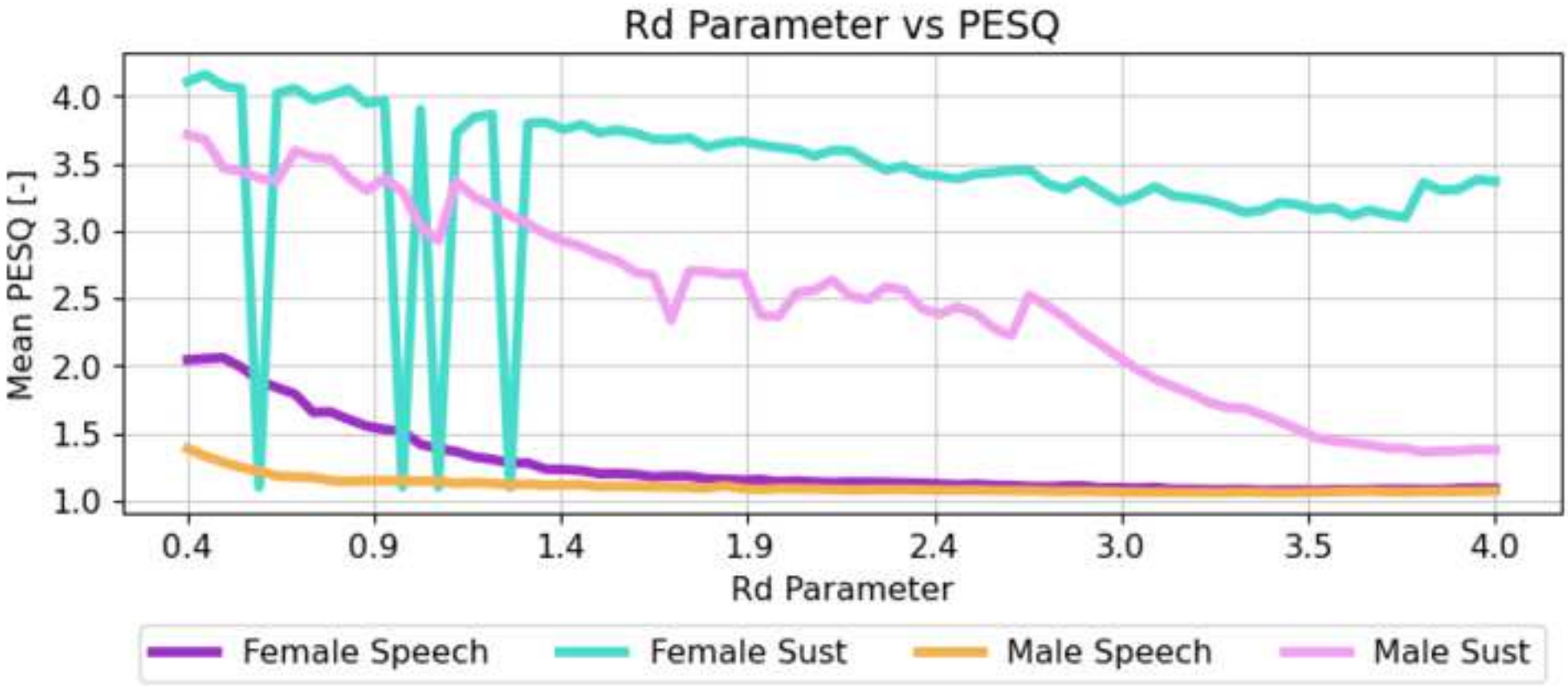


Figure 4a (Color)

Click here to access/download;Figure;f0_multiplier_pesq.png

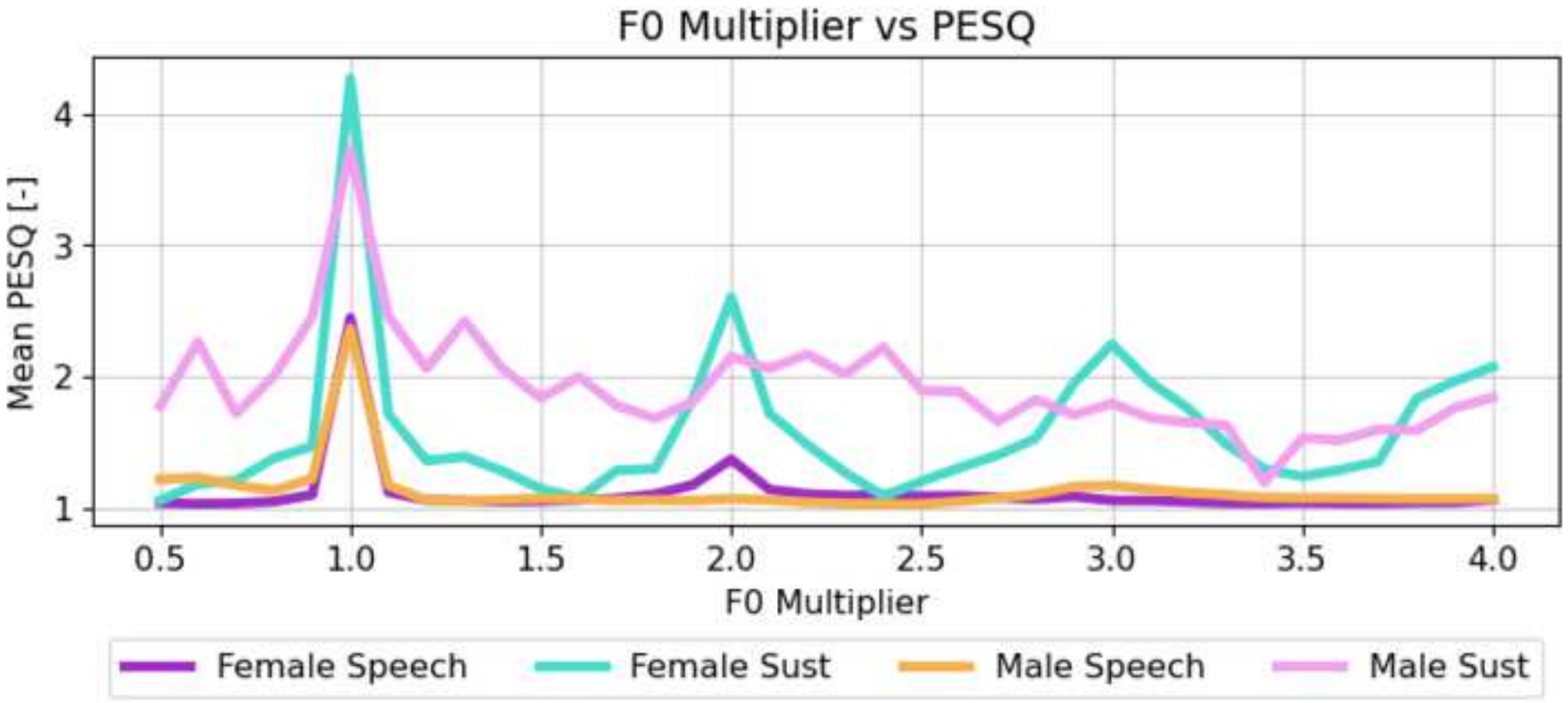


Figure 4b (Color)

Click here to access/download;Figure;f0_multiplier_ps.png

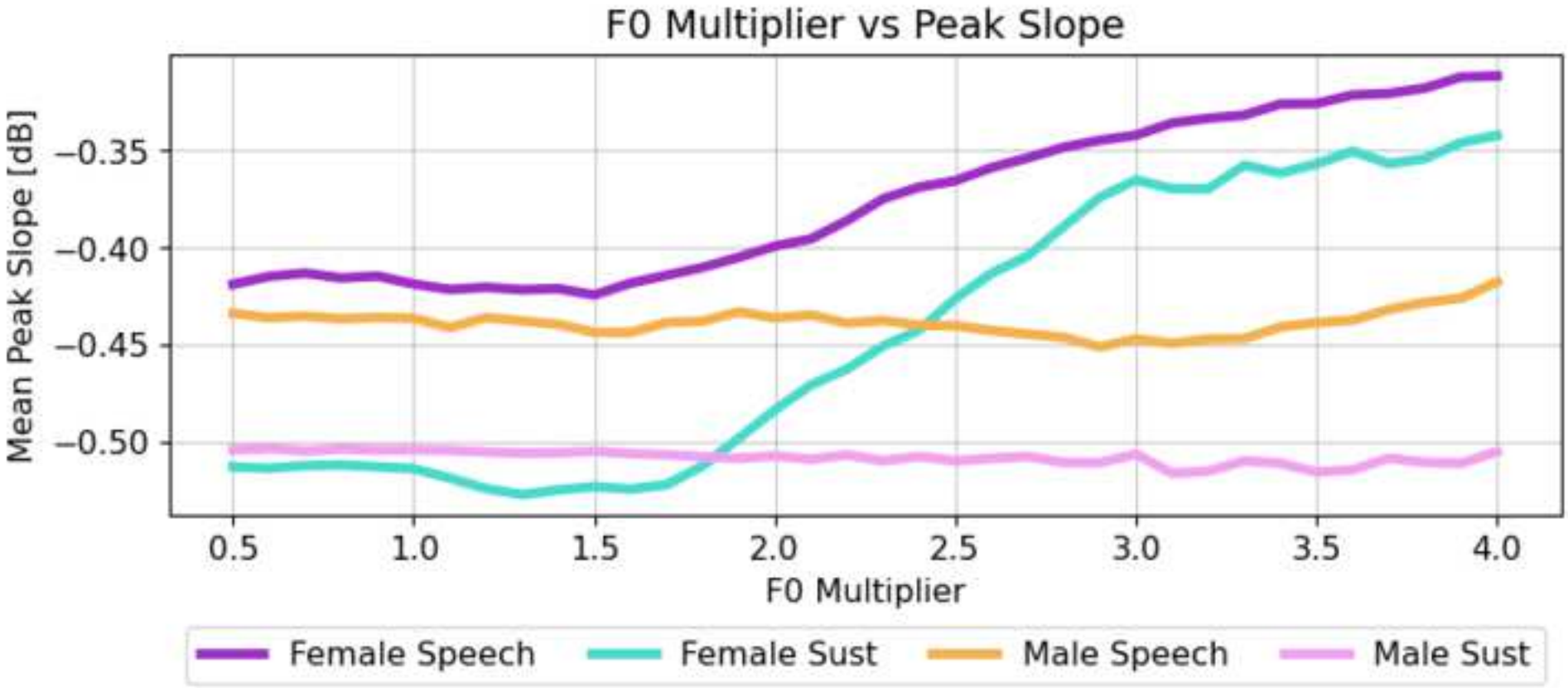


Figure 4c (Color)

Click here to access/download;Figure;f0_multiplier_hnr.png

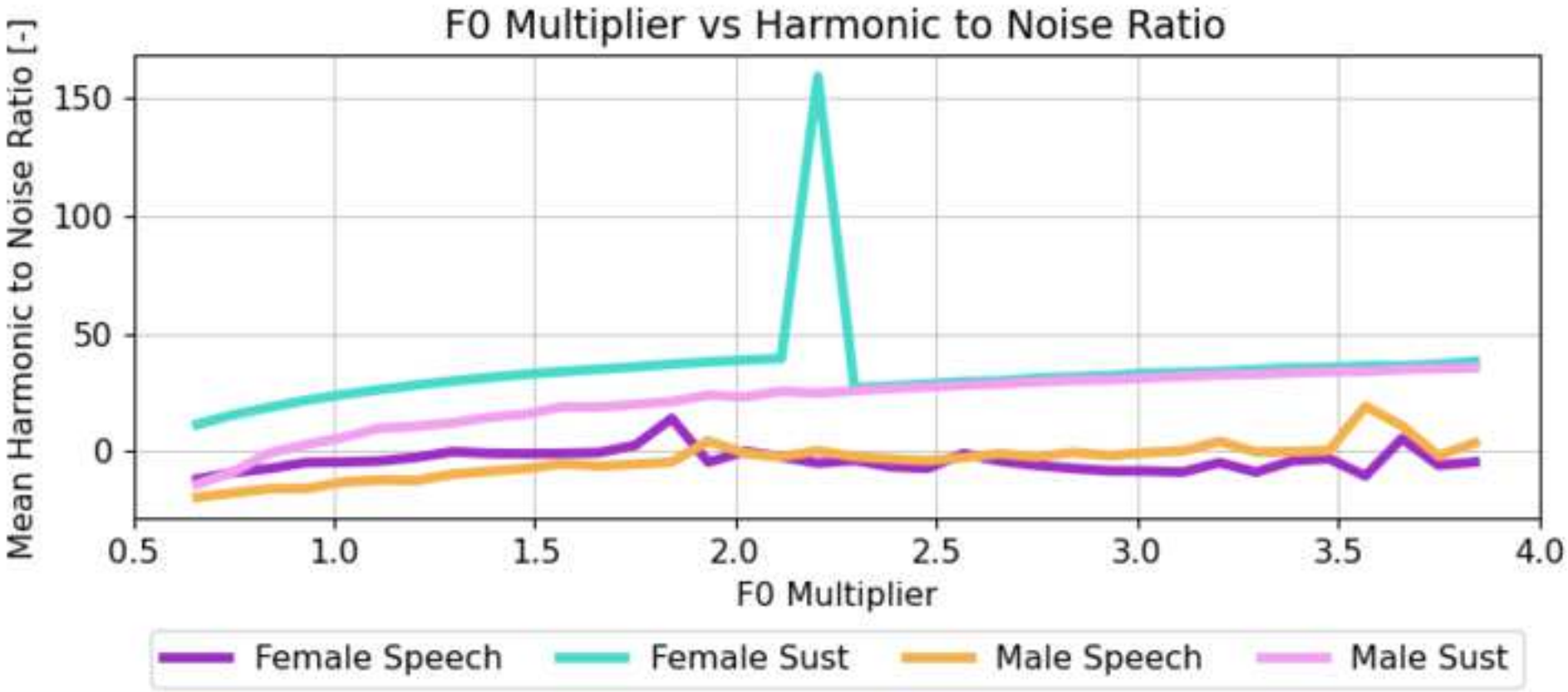


Figure 4d (Color)

Click here to access/download;Figure;f0_multiplier_cpp.png

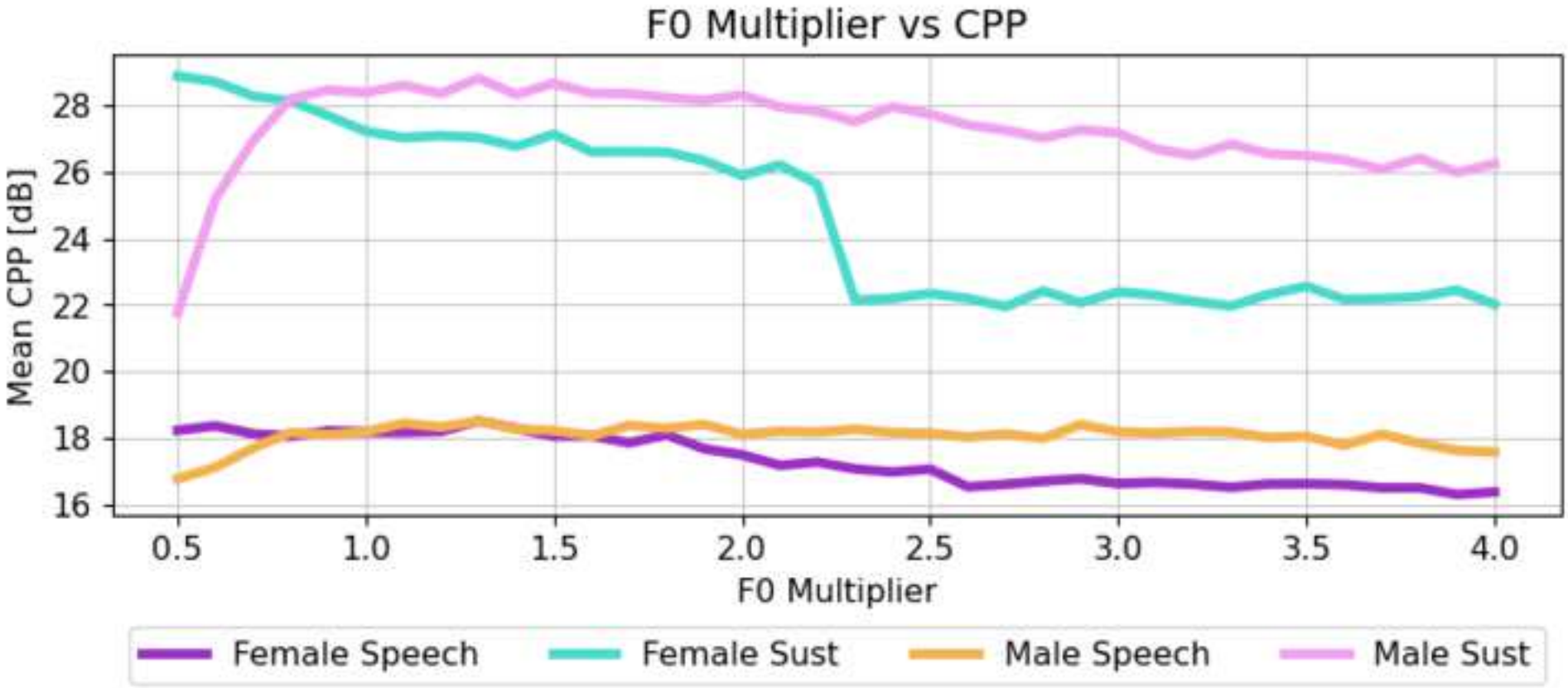


Figure 5a (Color)

Click here to access/download;Figure;rpp_k_cpp.png

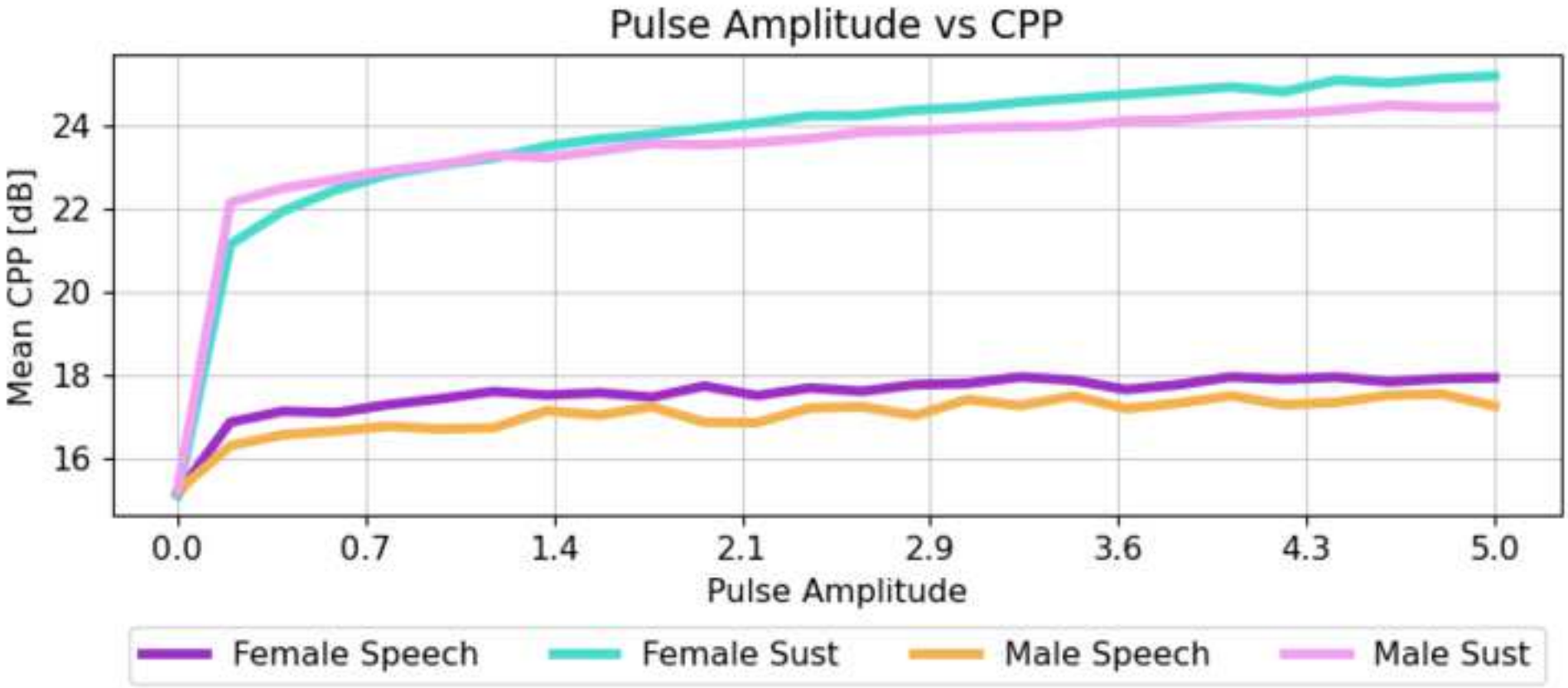


Figure 5b (Color)

Click here to access/download;Figure;rpp_k_pesq.png

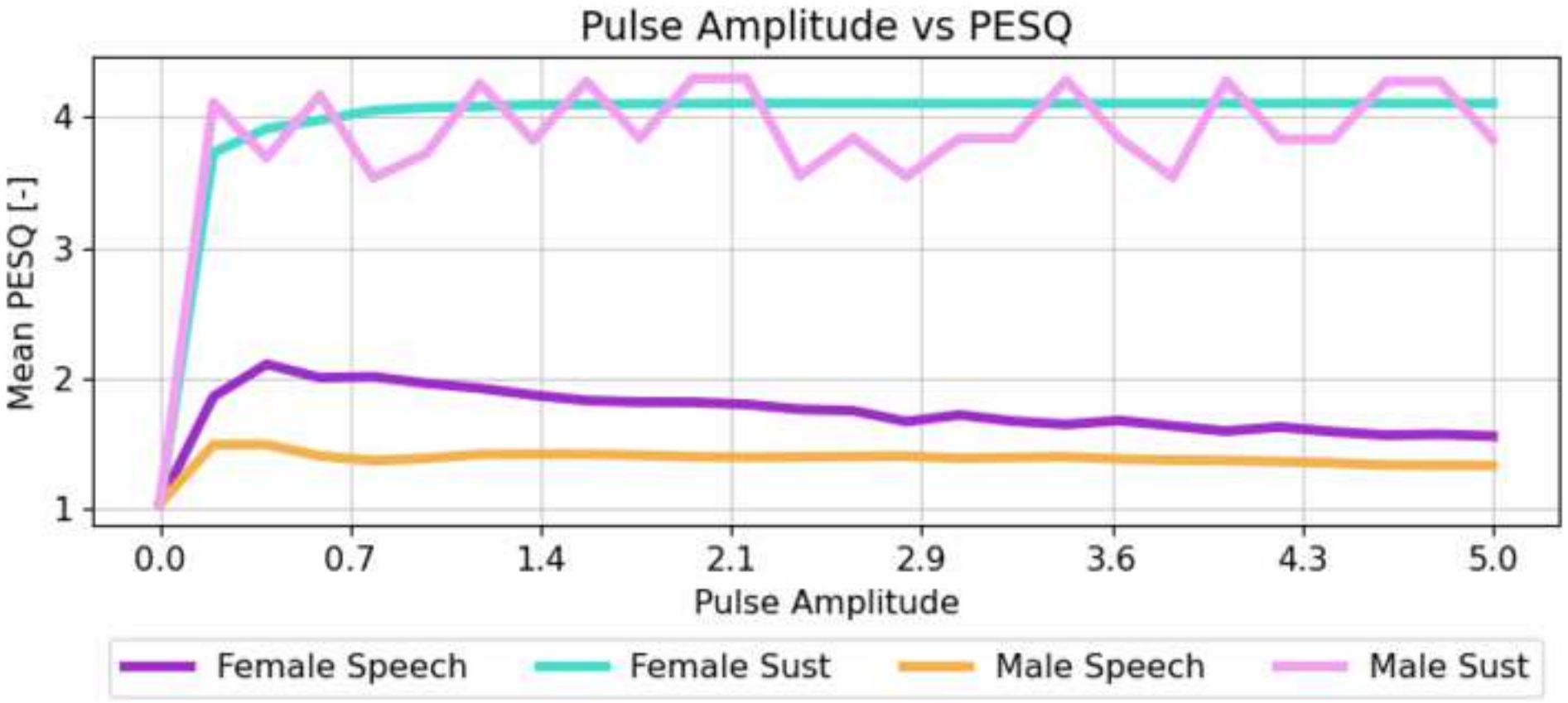


Figure 6a (Color)

Click here to access/download;Figure;jitter_amplitude_jitter.png

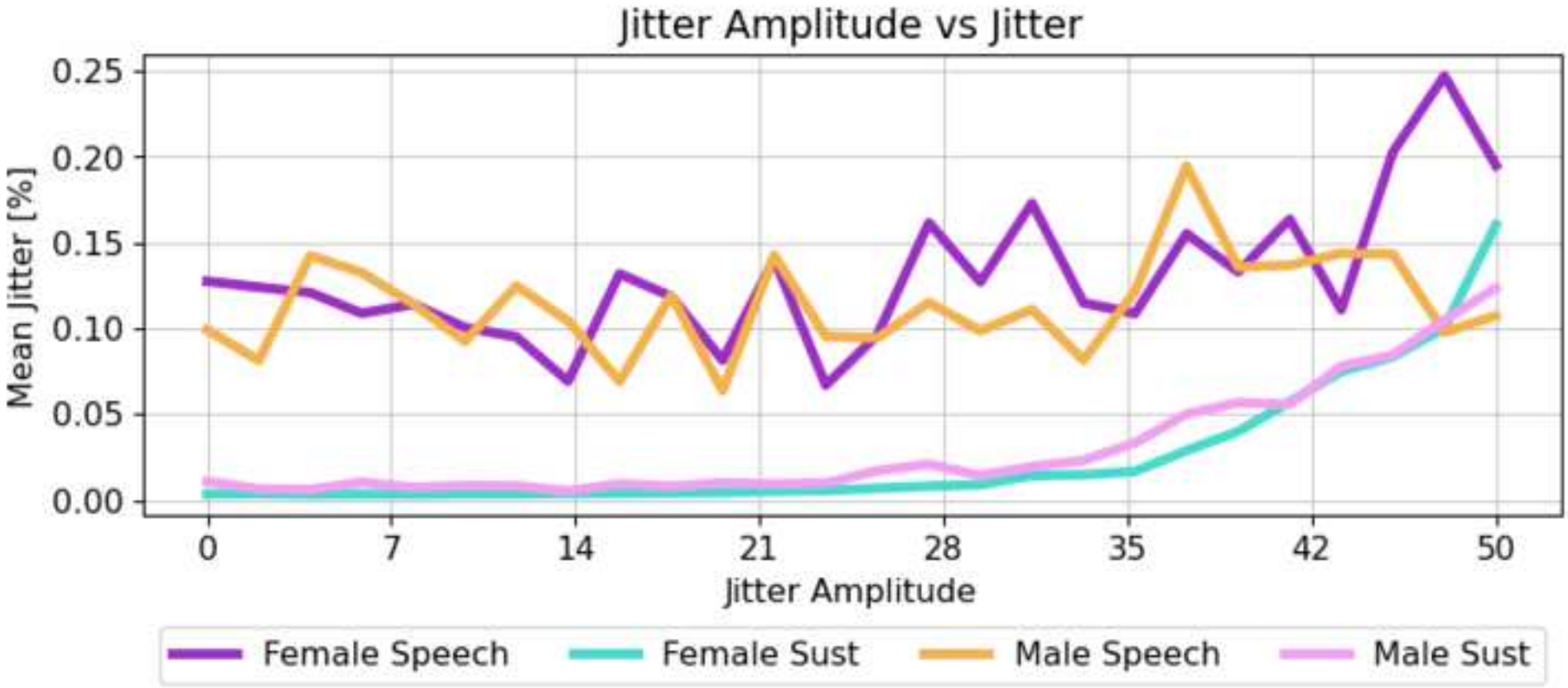


Figure 6b (Color)

Click here to access/download;Figure;jitter_frequency_jitter.png

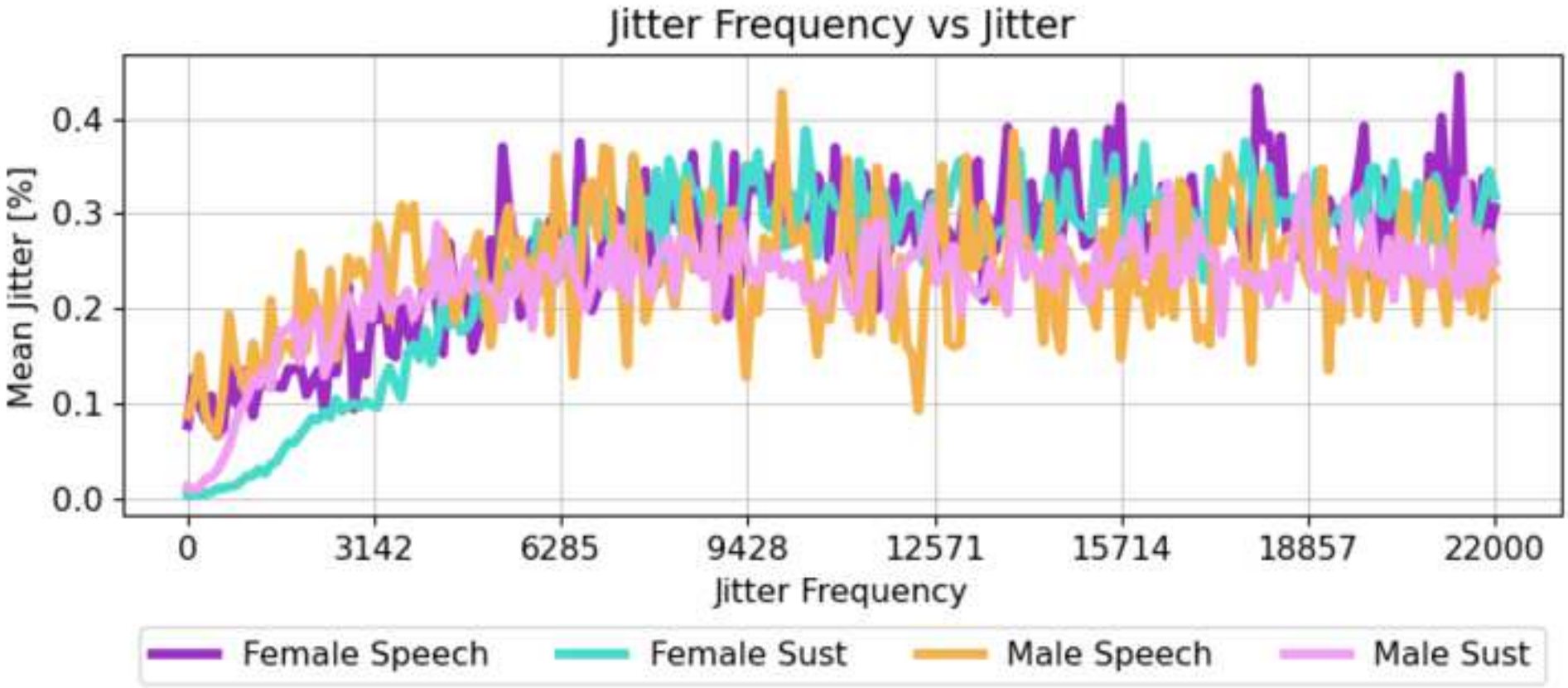


Figure 6c (Color)

Click here to access/download;Figure;shimmer_amplitude_shimmer.png

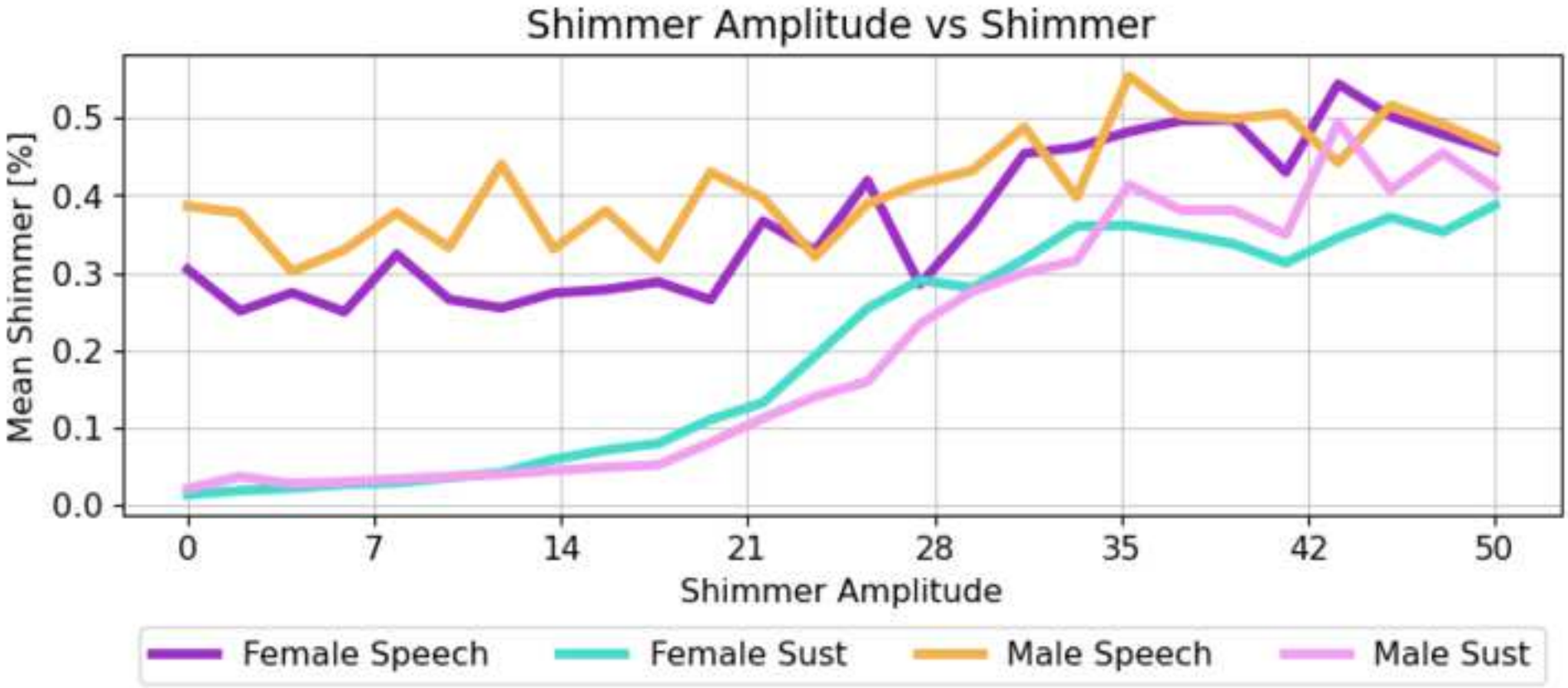
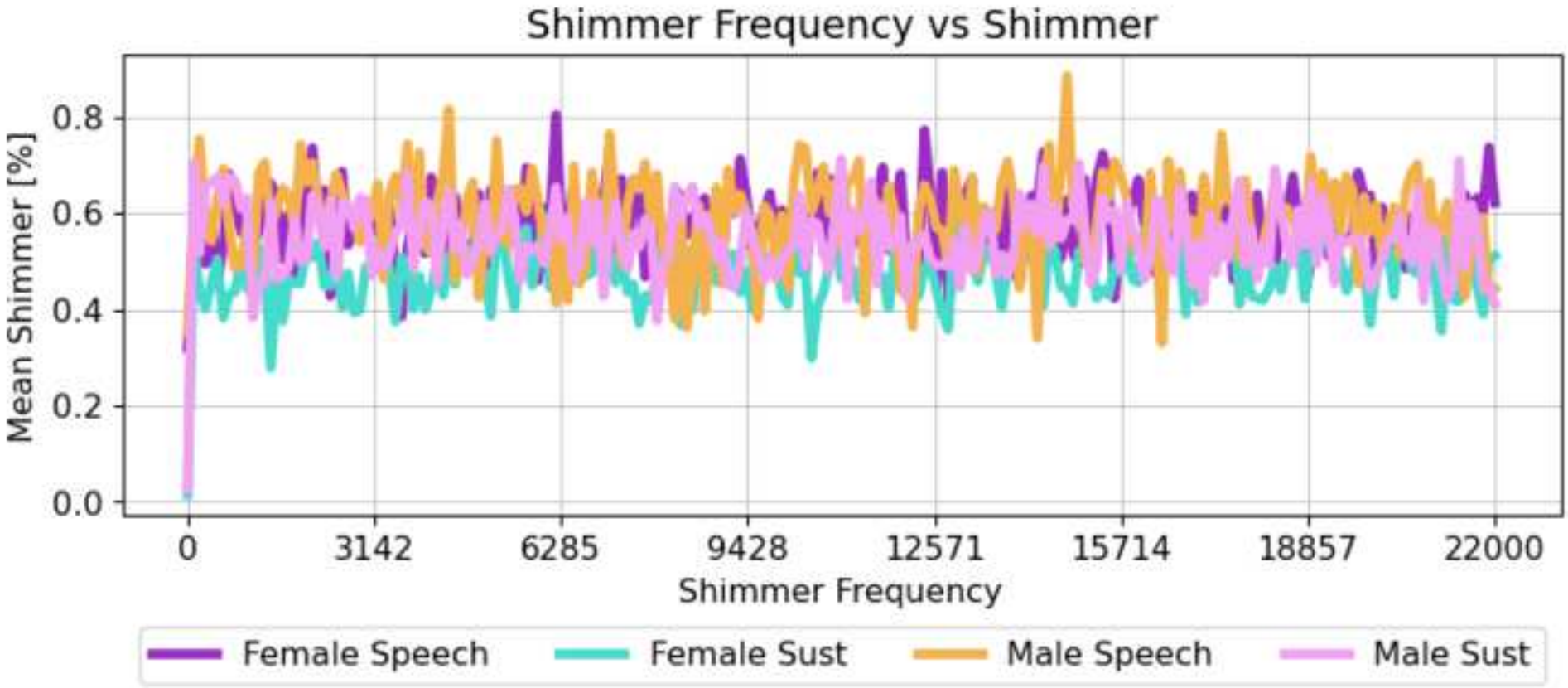


Figure 6d (Color)

Click here to access/download;Figure;shimmer_frequency_shimmer.png



Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Benjamin Opazo reports financial support was provided by National Fund For Scientific Technological and Technological Innovation Development. Matias Zanartu reports financial support was provided by National Fund For Scientific Technological and Technological Innovation Development. Matias Zanartu reports a relationship with Lanek SPA that includes: equity or stocks.