Voice Quality Modification using the WORLD Vocoder

Benjamín Opazo

Motivación

- Se estudia la modificación de Calidad Vocal para hacer experimentos de neurociencia
  - De la misma forma que se hacen experimentos de perturbación de tono pero con Calidad Vocal
- Se cree que el rol de la autopercepción juega un rol crítico en la hiperfunción vocal

Objetivos de la tesis

- Extender un Vocoder del estado del arte con la capacidad de incorporar y modificar la fuente glotal de manera de modificar la calidad vocal
- Evaluar el rendimiento del vocoder extendido usando medidas objetivas y perceptuales de la calidad vocal

Puntos Previos

Producción de Voz

Se puede aproximar el proceso de producción de voz con la siguiente ecuación

$$y(t) = h(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT_0) \tag{1}$$

O en el dominio de la frecuencia

$$Y(\omega) = \frac{2\pi}{T_0} H(\omega) \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \tag{2}$$

Calidad Vocal

- Dificil de definir objetivamente
- Perceptual en su origen
- Multidimensional
- Depende tanto del hablante como de quien percibe el habla

Según la US National Library of Medicine se puede definir como:

> *Las diferentes propiedades del habla que entregan la distinción primaria de la voz de un hablante cuando el tono y el volumen son excluidos*

Evaluación de la Calidad Vocal

- Evaluar la calidad vocal es dificil
- Algunos métodos de evaluación de Calidad Vocal
  - Buffalo Voice Profile (BVP)
  - Vocal Profile Analysis (VPA)
  - Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)
  - Grade, Rough, Breathy, Asthenic, Strained Scale (GRBAS)

Modelos Glotales: Rosenberg++ (R++)

- Basado en el pulso de Rosenberg-B
- Modelo de Flujo Glottal simple de calcular
- Parámetros de forma: (T_e, T_p, T_a)
- Rendimiento similar al modelo LF

Modelos Glotales: Rosenberg++ (R++)

$$g(t) = \begin{cases} At^2(t_e - t) & 0 \leq t \leq t_e \\ 4At(t_p - t)(t_x - t) & t_e \leq t \leq t_0 \end{cases} \tag{3}$$

Donde

$$t_x = t_e \left(1 - \frac{\frac{1}{2}t_e^2 - t_e t_p}{2t_e^2 - 3t_e t_p + 6t_a(t_e - t_p)D(t_0, t_e, t_a)}\right) \tag{4}$$

$$D(t_0, t_e, t_a) = 1 - \frac{(t_0 - t_e)/t_a}{\exp((t_0 - t_e)/t_a) - 1} \tag{5}$$

Síntesis de Voz: WORLD Vocoder

Separa la señal de voz en

- Frecuencia Fundamental
- Envolvente Espectral
- Parámetro Aperiódico

WORLD Vocoder

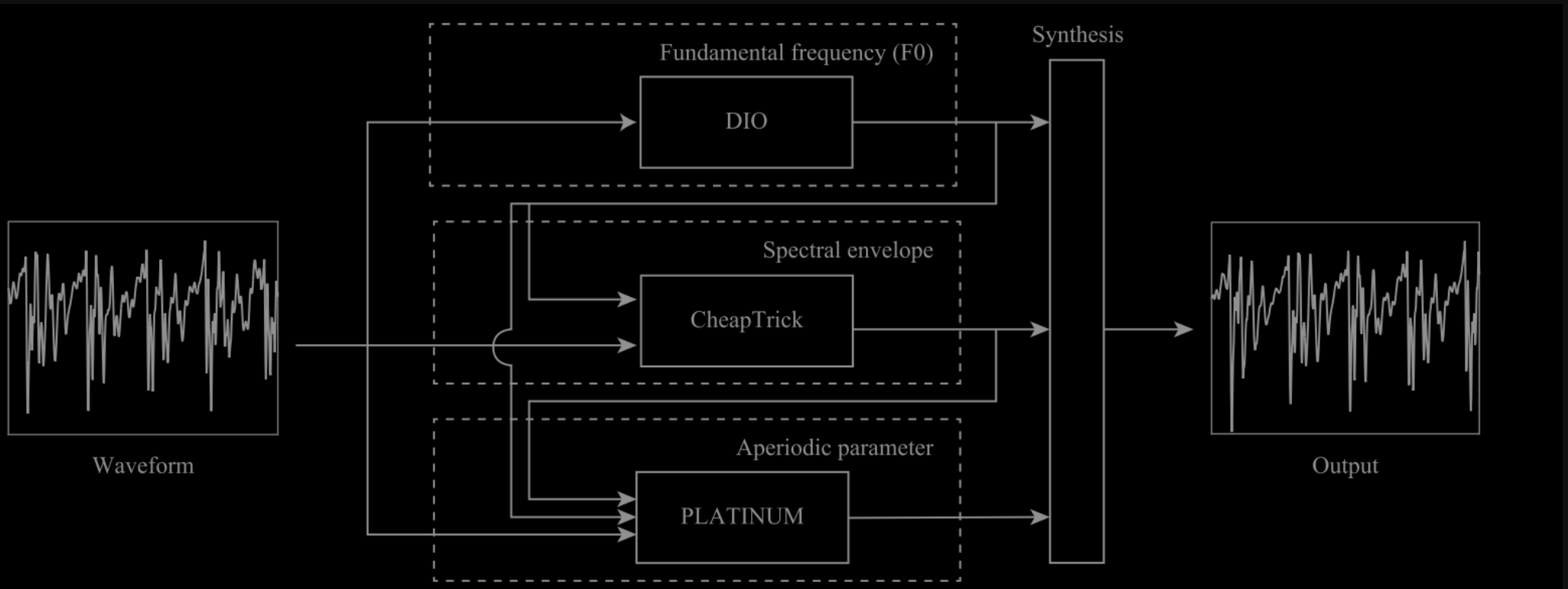


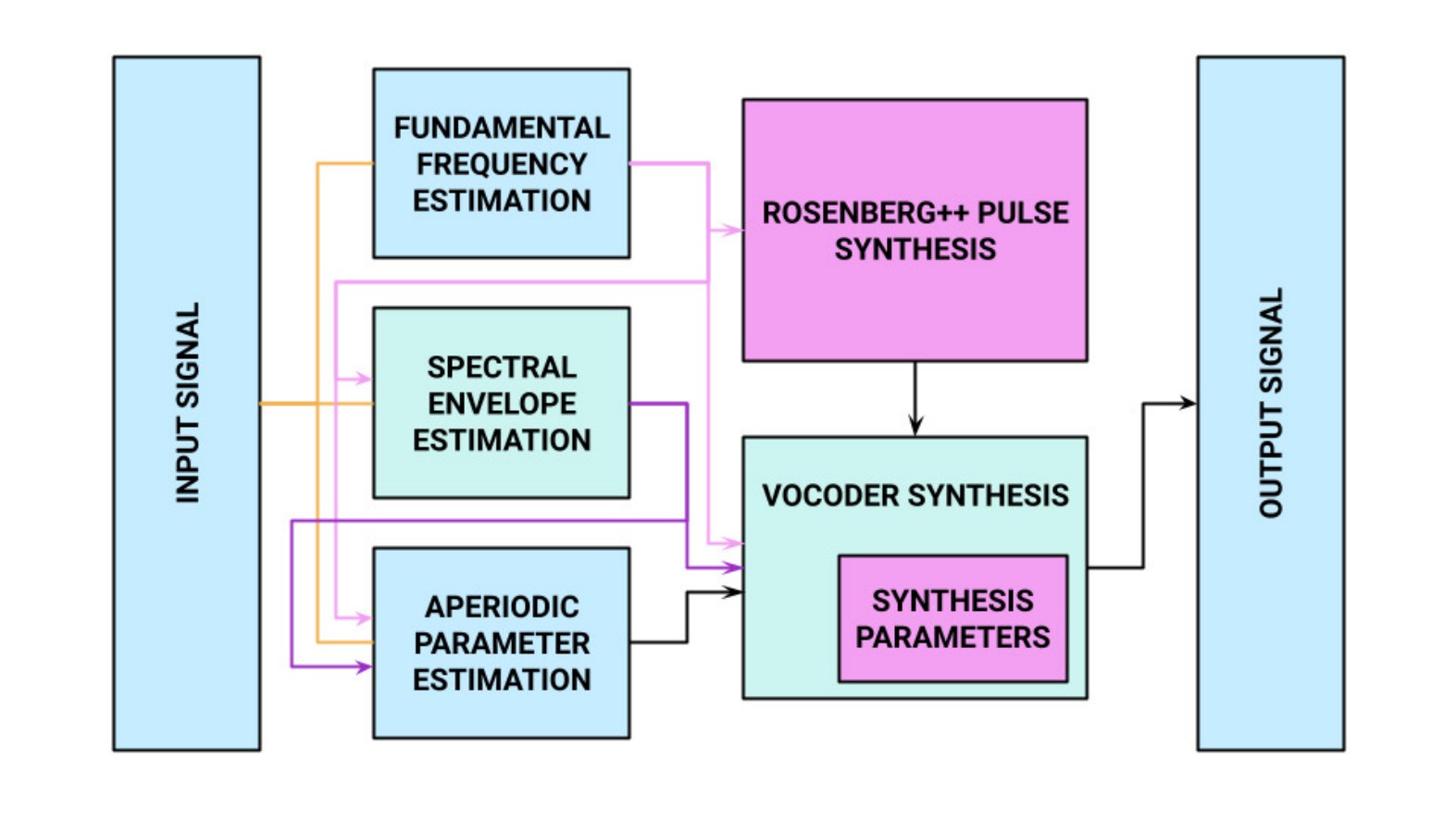
Fig. 1 Overview of the developed system. WORLD consists of three analysis algorithms for determining the F0, spectral envelope, and aperiodic parameters and a synthesis algorithm incorporating these parameters.

WORLD: CheapTrick

- Calcula la envolvente espectral en tres pasos:
  - Usa una ventana Hanning para calcular una densidad espectral que no depende de la variable temporal
  - Suaviza la densidad espectral con un filtro (para evitar $\log 0$)
  - Recuperación Espectral usando Liftering

Métodos

Extensión del WORLD Vocoder



Rosenberg ++

- La implementación restrinje la síntesis del pulso dadas ciertas condiciones
- Se modifica la definición de la amplitud K

Rosenberg ++

La variable de amplitud K se redefine de la siguiente forma. Si $t_a > 0$

$$K_{int} = \begin{cases} 3K/(t_p^3(2t_x - t_p)) & t_p < \frac{(4D(t_0,t_e,t_a)t_at_e+t_e^2)}{(2D(t_0,t_e,t_a)t_a+t_e)} \\ -3K/(t_p^3(2t_x - t_p)) & \text{e.o.c} \end{cases} \tag{6}$$

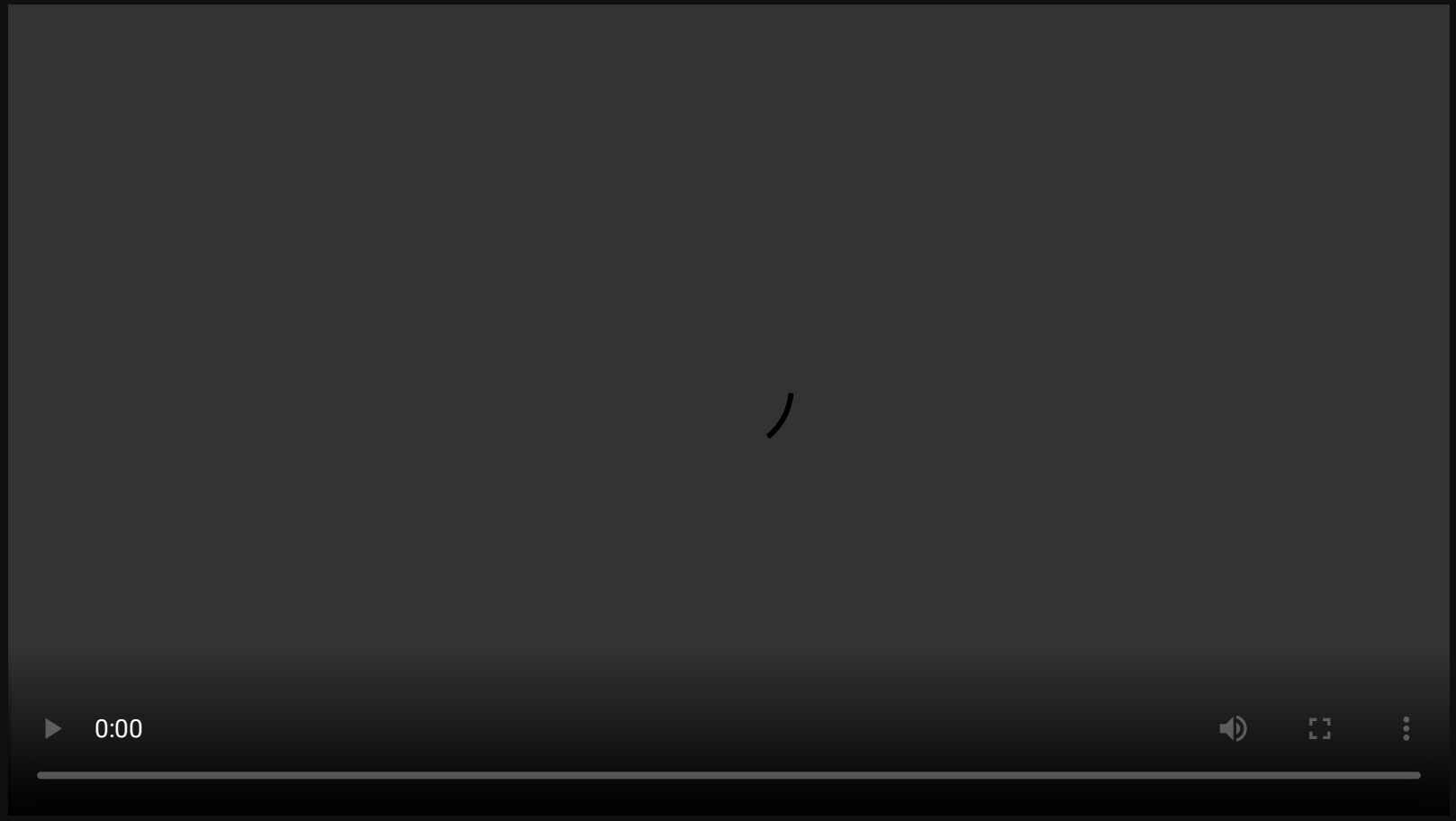
Si $t_a = 0$, entonces K_{int} se define como

$$K_{int} = \begin{cases} 3K/(t_p^3(2t_x - t_p)) & t_p < t_e \\ -3K/(t_p^3(2t_x - t_p)) & \text{e.o.c} \end{cases} \tag{7}$$

Parámetro R_d

- Este parámetro fue definido basado en el modelo LF
- Se obtiene al hacer un análisis funcional y estadístico de la covariación de los parámetros del modelo LF
- Puede generar desde una fonación apretada y aducida hasta una fonación aspirada y abducida

Parámetro R_d



Parámetro R_d

Se pueden definir los siguientes parámetros:

$$R_a = t_a/t_0 \tag{8}$$
$$R_g = t_0/(2t_p) \tag{9}$$
$$R_k = (t_e - t_p)/t_p \tag{10}$$

Parámetro R_d

Los valores de R_a y R_k se pueden predecir usando el parámetro R_d

$$R_a^* = (-1 + 4.8R_d)/100 \quad (11)$$
$$R_k^* = (22.4 + 11.8R_d)/100 \quad (12)$$

Además, se tiene la siguiente relación

$$R_d = (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a)) \quad (13)$$

Modificación de CheapTrick

- WORLD Vocoder asume que la señal de excitación es un tren de impulsos
- CheapTrick extrae la envolvente espectral

Por lo tanto es necesario modificar CheapTrick

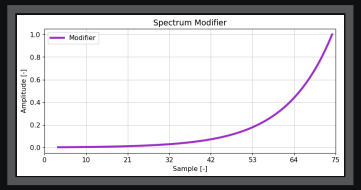
Modificación de CheapTrick

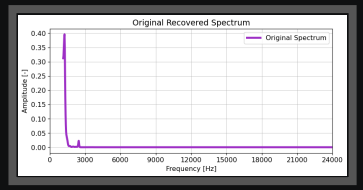
- El efecto de las cuerdas vocales es principalmente en las bajas frecuencias
- Se propone aplicar el siguiente modificador a la envolvente espectral

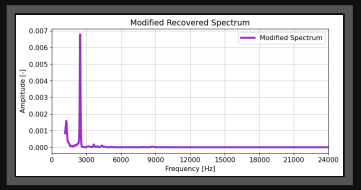
$$f[n] = \exp(n\frac{p}{m} - p) \tag{14}$$

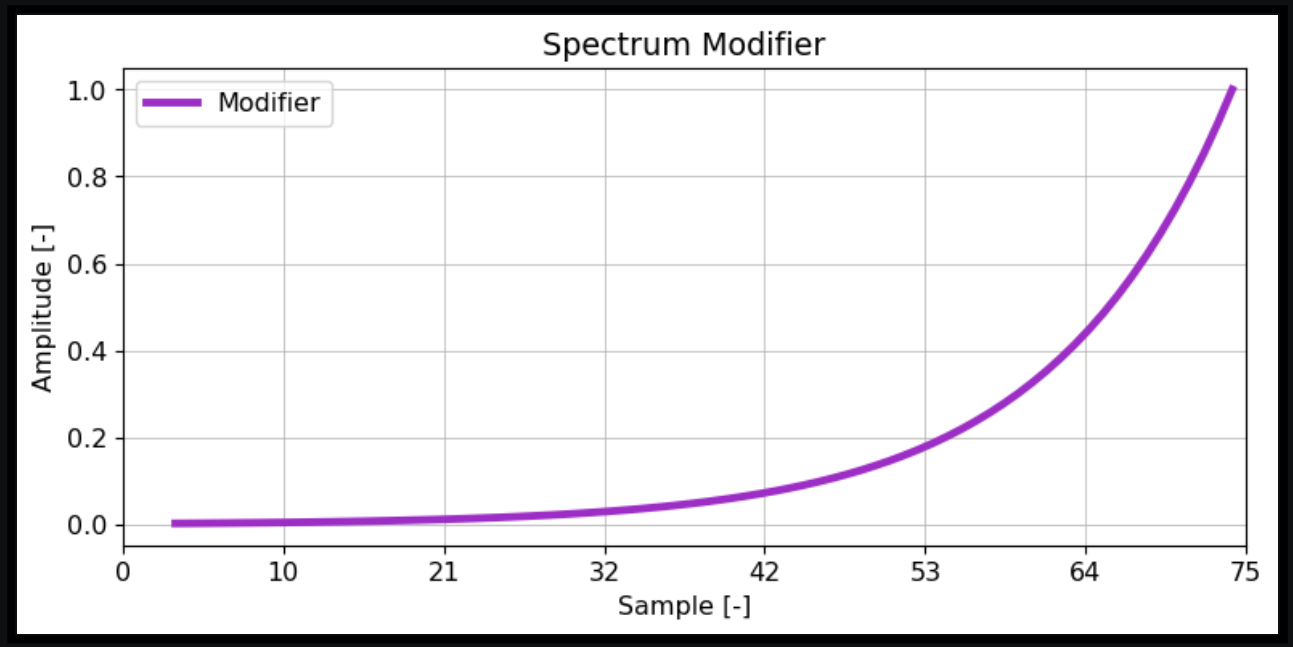
Donde $n \leq m$ y p está relacionado con la pendiente del modificador

Modificación de CheapTrick



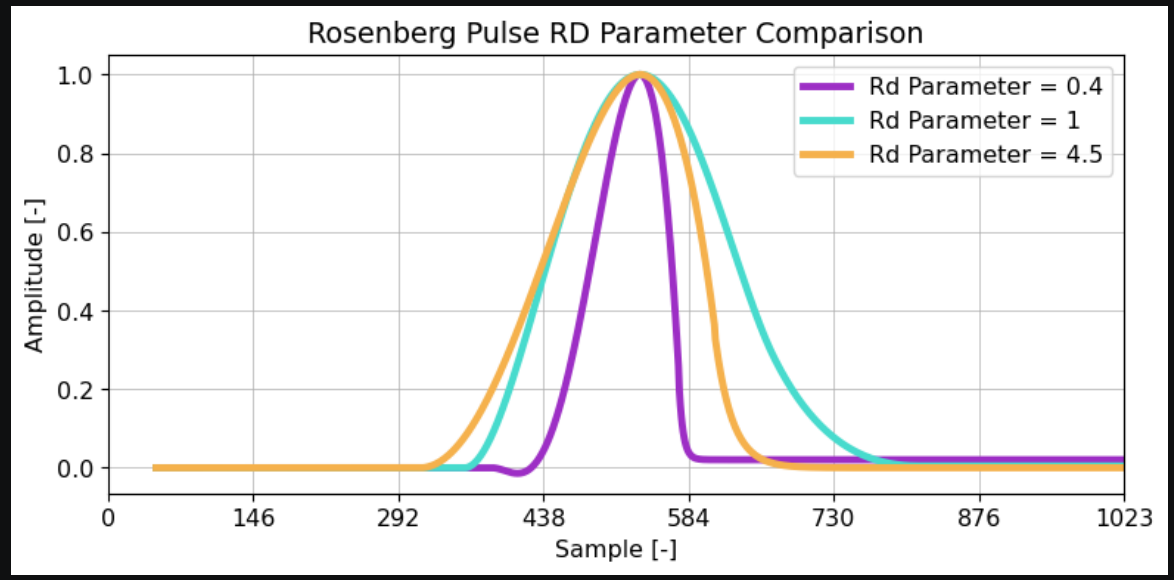






Parámetros Propuestos: *rd_param*

- Controla el parámetro R_d del tren de impulsos usados como señal de excitación. Los parámetros t_e, t_p y t_a se obtienen utilizando el toolbox COVAREP. Varía entre 0.35 y 4



Parámetros Propuestos: *f0_filter_frequency* y *f0_filter_order*

- Filtran las variaciones de F_0 o frecuencia fundamental.
- El filtrado se hace a partir de una filtro pasabajos Butterworth de orden *f0_filter_order* y frecuencia de corte frecuencia *f0_filter_frequency*.
- Se puede utilizar para reducir jitter.

Parámetros Propuestos: *jitter_amplitude* y *jitter_frequency*

- Estos parámetros se utilizan para agregar jitter a la voz sintetizada. *jitter_amplitude* controla la amplitud del jitter y *jitter_frequency* controla la frecuencia del jitter.
- El jitter se genera a sumando ruido Browniano sintetizado a partir de un ruido blanco Gaussiano de amplitud *jitter_amplitude* filtrado con un filtro Butterwroth pasabajos con frecuencia de corte *jitter_frequency*.

Parámetros Propuestos: *shimmer_amplitude* y s*himmer_frequency*

- Estos parámetros agregan shimmer a la voz sintetizada
- Multiplican la amplitud de la señal de excitación por $(1 + BN)$ donde BN corresponde a ruido Browniano

Parámetros Propuestos: *vibrato_amplitude* y *vibrato_frequency*

- Agrega vibrato a la frecuencia fundamental
- la ecuación de vibrato corresponde a

$$F_0^*[n] = F_0[n]VA * \sin(n * VF) \tag{15}$$

Parámetros Propuestos: *spectrum_filtering_exponential* y *spectrum_filtering_samples*

- Implementación de la modificación de CheapTrick
- Aplica la siguiente función

$$f[n] = \exp(n\frac{p}{m} - p) \tag{16}$$

Donde p corresponde a *spectrum_filtering_exponential* y m corresponde a *spectrum_filtering_samples*

Parámetros Propuestos: *rpp_multiplier_te*, *rpp_multiplier_tp*, *rpp_multiplier_ta* y *rpp_k*

- Multiplicadores de los parámetros t_e, t_p, t_a y k

Parámetros Propuestos: *band_aperiodicity_multiplier*

- Matriz que multiplica la variable de aperiodicidad obtenida por D4C Lovetrain

Experimentos: Metodología

Rendimiento en Tiempo Real

Se sintetiza el siguiente audio:

- Muestras de 16bit@48KHz
- 120000 Muestras (o 2.5 [s])
- Hombre de mediana edad diciendo "Esta es una grabación de prueba"

En el siguiente computador

- Dell XPS 13 Modelo 9343
- 5th Gen Intel Core i5-5200 @2.2GHz
- 8 Gb de Ram @1600MHz
- Ubuntu 20.04

Rendimiento en Tiempo Real

- Implementación en C del Vocoder WORLD
- Se utilizó DIO + Stonemask o Harvest, CheapTrick, D4C
- Síntesis con 3 métodos distintos
- 3 Ventanas de Análisis: 5 ms, 3 ms, 1 ms.

Experimento 1: Evaluación Objetiva

El objetivo de este experimento es generar medidas objetivas de los parámetros mas relevantes de forma que pueden ser comparados y evaluados con medidas de Calidad Vocal

Experimento 1: Evaluación Objetiva

Se sintetizan los siguientes audios:

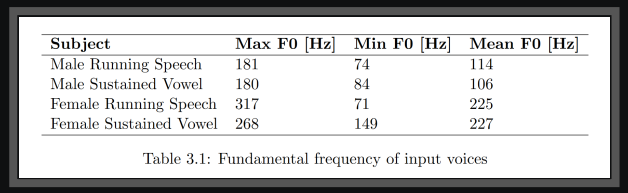
- Señales obtenidas de la Perceptual Voice Quality Database: LA9023_ENSS y LA9011_ENSS
- Las voces son una masculina y una femenina con buena calidad vocal
- Se sintetizan dos secciones, la vocal "a" y la frase "Peter will keep at the peak"

Experimento 1: Evaluación Objetiva

Se utilizan las siguientes medidas objetivas:

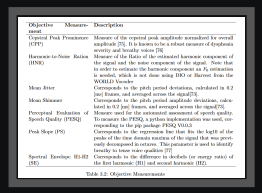
- Cepstral Peak Prominence (CPP)
- Harmonic-to-Noise Ratio (HNR)
- Mean Jitter, Mean Shimmer
- Perceptual Evaluacion of Voice Quality (PESQ)
- Peak Slope (PS)
- Spectral Envelope (H1-H2) (SE)

Experimento 1: Evaluación Objetiva



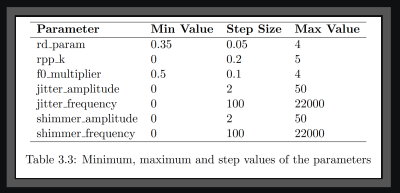
Subject	Max F0 [Hz]	Min F0 [Hz]	Mean F0 [Hz]
Male Running Speech	181	74	114
Male Sustained Vowel	180	84	106
Female Running Speech	317	71	225
Female Sustained Vowel	268	149	227

Table 3.1: Fundamental frequency of input voices



Objective Measurement	Description
Cepstral Peak Prominence (CPP)	Measure of the cepstral peak amplitude normalized for overall amplitude [75]. It is known to be a robust measure of dysphonia severity in connected speech [76].
Harmonic-to-Noise Ratios (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[73].
Mean Shimmer	Corresponds to the pitch period amplitude deviations, calculated in 0.2 [ms] frames, and averaged across the signal[73].
Perceptual Evaluation of Speech Quality (PESQ)	Measure used to objectively assess the quality of speech, according to ITU-T P.862, a standard implementation was used, corresponding to the pip package PESQ V0.0.4
Peak Slope (PS)	Corresponds to the regression line that fits the log10 of the peaks of the time domain cepstrum of the signal that was previously decomposed in octaves. This parameter is used to identify breathy to tense voice qualities [77]
Spectral Envelope H1-H2 (SE)	Corresponds to the difference in decibels (or energy ratio) of the first harmonic (H1) and second harmonic (H2).

Table 3.2: Objective Measurements



Parameter	Min Value	Step Size	Max Value
rd_param	0.35	0.05	4
rpp_k	0	0.2	5
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

Table 3.3: Minimum, maximum and step values of the parameters

Experimento 2: Evaluación Subjetiva

- Se utilizan las mismas señales de audio que en el Experimento 1
- Se sintetizan voces modales, aspiradas, *vocal fry*, disfonía y voz áspera
- La voz modal sintetizada funciona como una base para evaluar las otras voces
- Se utilizan las descripciones de distintas fuentes de Calidad Vocal para sintetizar las voces
- La evaluación se hace usando CAPE-V con 3 evaluadores expertos (fonoaudiólogos)

Experimento 2: Evaluación Subjetiva

Voz Modal

Parameter	Male Voice	Female Voice
rd_param	0.35	1
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	1
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	1
rpp_k	0.9	2.5

Table 3.4: Parameter values for modal voice

Experimento 2: Evaluación Subjetiva

Voz Aspirada

Parameter	Male Voice	Female Voice
rd_param	2.8	4
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	0.8
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	1
rpp_k	0.5	0.8
band_aperiodicity_multiplier	0.5*[1 1 1 1 1 1 1]	0.5.*[1 1 1 1 1 1 1]

Table 3.5: Parameter values for breathy voice

Experimento 2: Evaluación Subjetiva

Vocal Fry

Parameter	Male Voice	Female Voice
rd_param	0.35	1
f0_multiplier	0.5	0.35
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	-
rpp_multiplier_tp	0.94	-
rpp_multiplier_ta	1	-
rpp_k	0.5	-
jitter_amplitude	12	18
jitter_frequency	50*0.7	50*0.7
shimmer_amplitude	12	15
shimmer_frequency	50*0.7	50*0.7

Table 3.6: Parameter values for vocal fry

Experimento 2: Evaluación Subjetiva

Voz Disfónica

Parameter	Male Voice	Female Voice
rd_param	1.35	1.35
spectrum_filtering_exponential	8.5	10
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	0.45
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	0.1
rpp_k	0.5	0.5
band_aperiodicity_multiplier	0.05*[0.1 1 1 1 1 1 1]	0.05*[0.5 1 1 1 1 1 1]

Table 3.7: Parameter values for dysphonia

Experimento 2: Evaluación Subjetiva

Voz Áspera

Parameter	Male Voice	Female Voice
rd_param	0.35	1.35
f0_multiplier	0.95	0.95
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	0.45
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	0.1
rpp_k	0.9	0.8
jitter_amplitude	14	17
jitter_frequency	50*0.5	0.8
shimmer_amplitude	15	17
shimmer_frequency	50*0.8	0.8

Table 3.8: Parameter values for rough voice

Experimentos: Resultados

Rendimiento en Tiempo Real

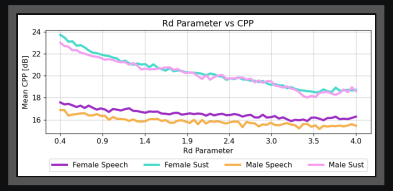
Window [ms]	Harvest [ms]	Cheaptrick [ms]	D4C [ms]	Synthesis [ms]	Total [ms]	RTF
1	1426	501	2807	111	4845	1.94
3	1422	169	942	109	2642	1.06
5	1477	106	580	113	2276	**0.91**

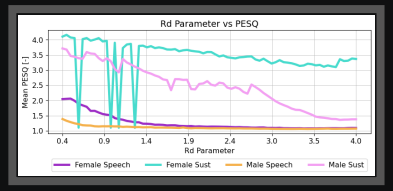
Table 4.1: RTF of WORLD Vocoder analysis-synthesis using Harvest

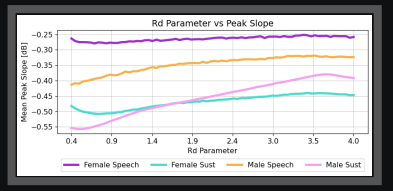
Window [ms]	DIO [ms]	StoneMask [ms]	Cheaptrick [ms]	D4C [ms]	Synthesis [ms]	Total [ms]	RTF
1	69	437	541	3387	89	4523	1.81
3	65	145	179	1108	92	1589	**0.64**
5	70	56	104	445	124	799	**0.32**

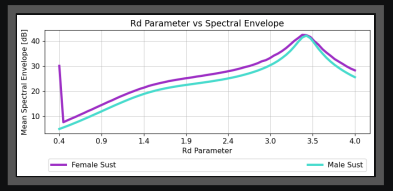
Table 4.2: RTF of WORLD Vocoder analysis-synthesis using DIO

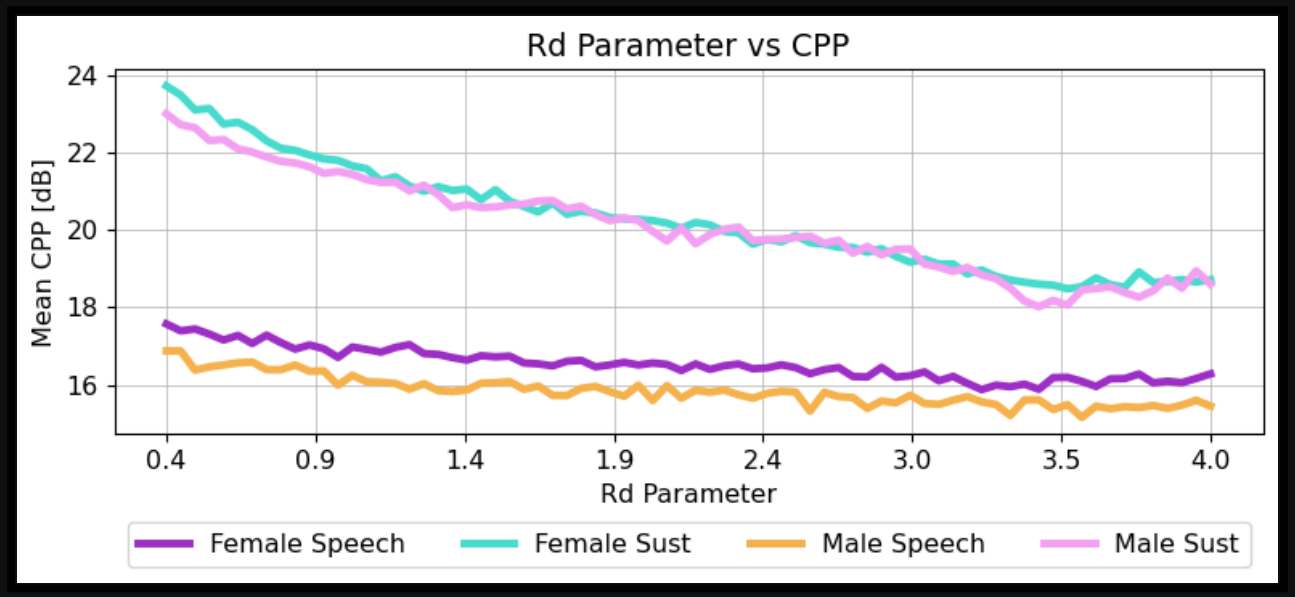
Experimento 1: Evaluación Objetiva - rd_param



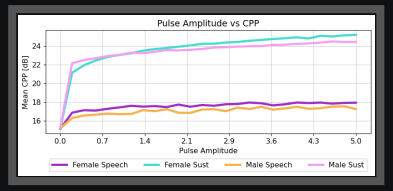


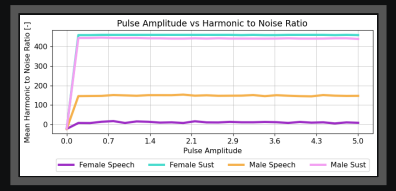


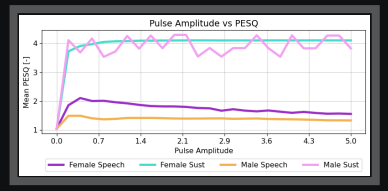


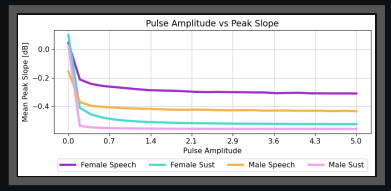


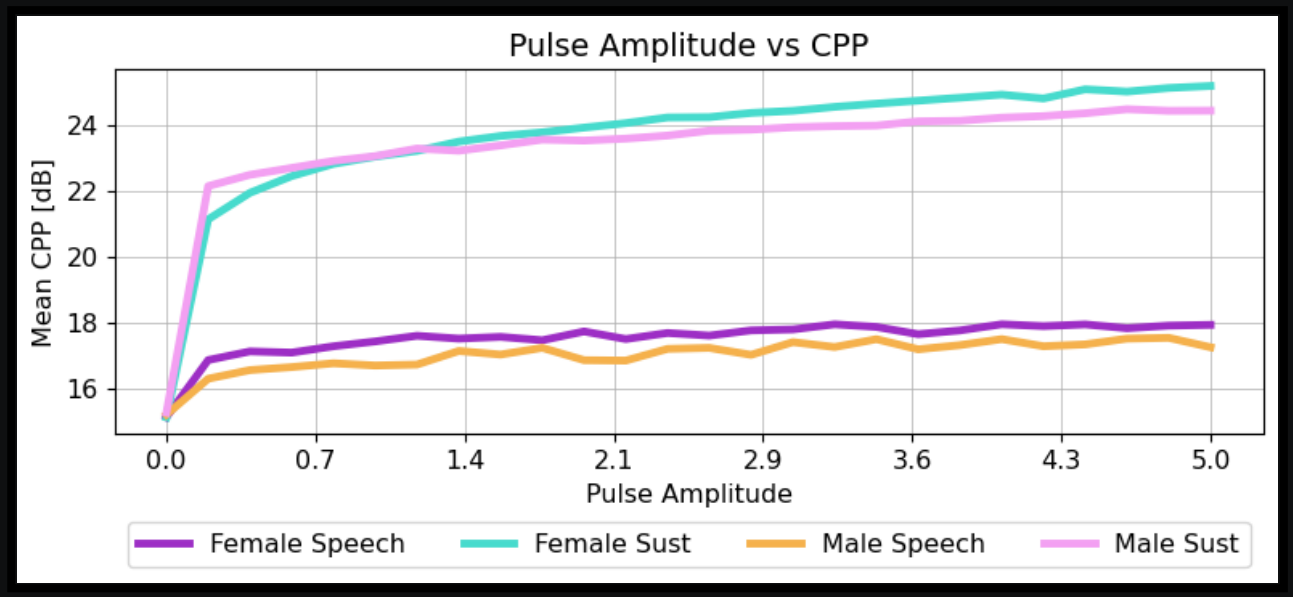
Experimento 1: Evaluación Objetiva - rpp_k



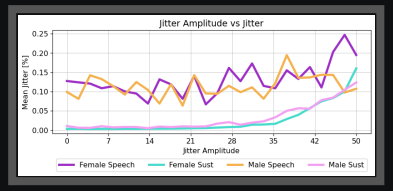


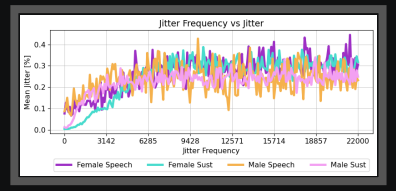


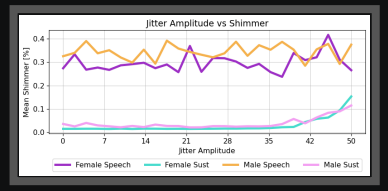


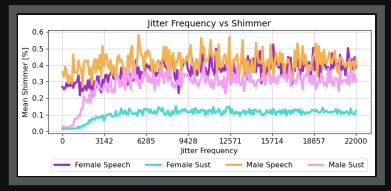


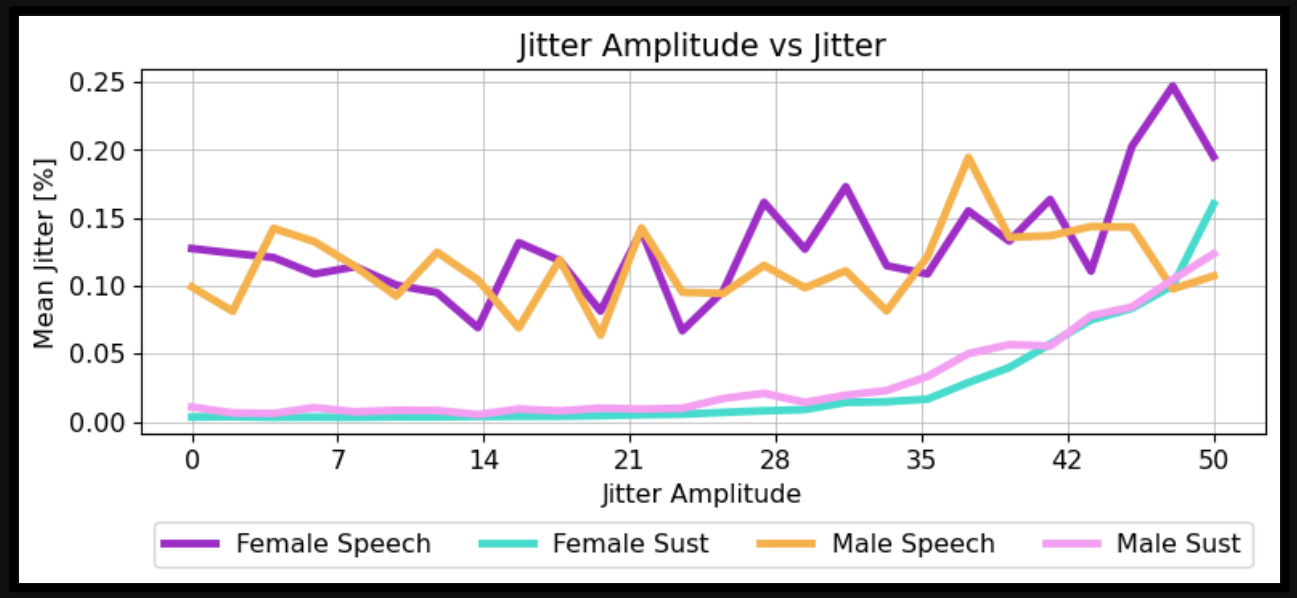
Experimento 1: Evaluación Objetiva - jitter_amplitude y jitter_frequency



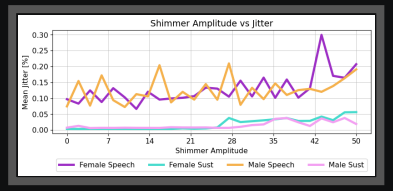


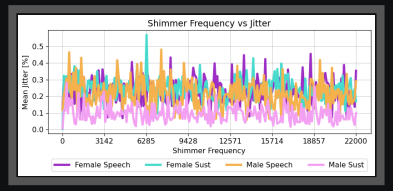


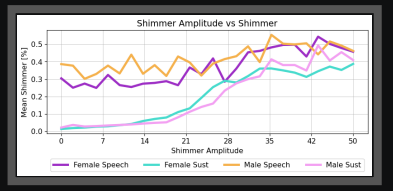


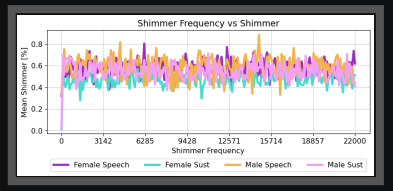


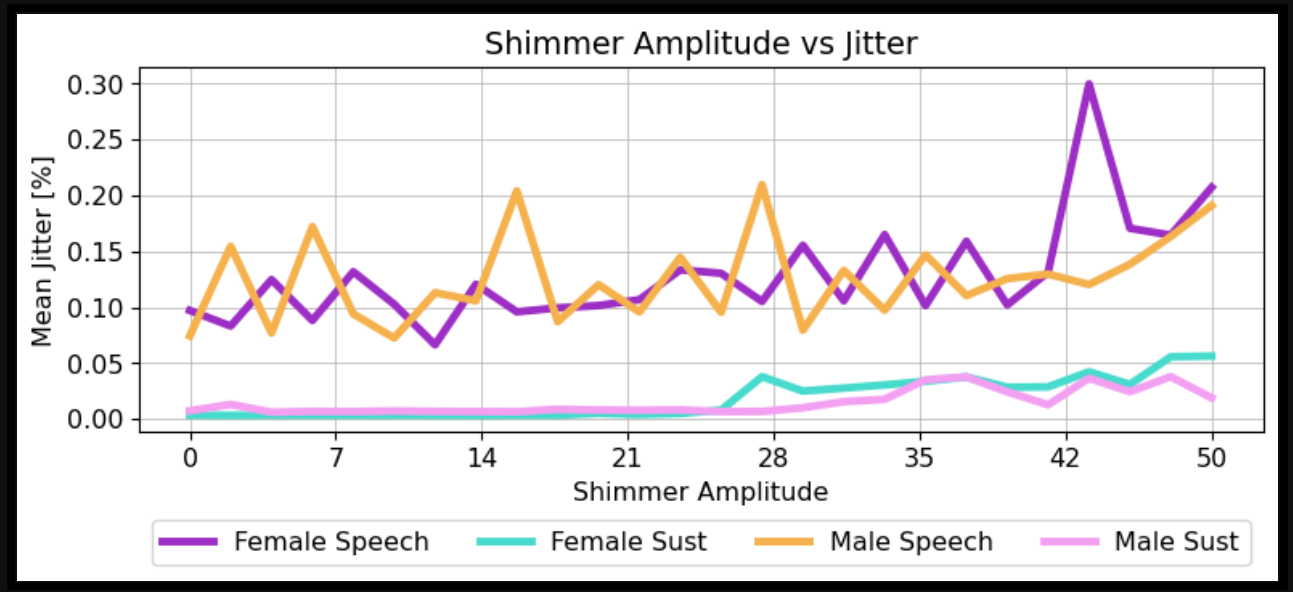
Experimento 1: Evaluación Objetiva - shimmer_amplitude y shimmer_frequency



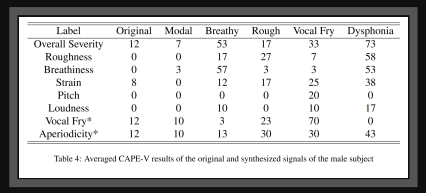






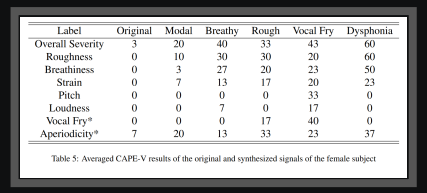


Experimento 2: Evaluación Subjetiva



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

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

Conclusiones

Muchas gracias por su atención