



UNIVERSIDAD DE BUENOS AIRES  
Facultad de Ciencias Exactas y Naturales  
Departamento de Computación

# Evaluación automática de la calidad del habla artificial

Tesis presentada para optar al título de  
Doctor de la Universidad de Buenos Aires  
en el área Computación

**Christian Gustavo Cossio Mercado**

Director de tesis: Dr. Jorge Gurlekian

Consejero de Estudios: Dr. Agustín Gravano

Lugar de trabajo: Laboratorio de Investigaciones Sensoriales, INIGEM, UBA-CONICET

Buenos Aires, 2023



## RESUMEN

---

El español es la principal lengua del continente americano y la cuarta más hablada en el mundo, además de la segunda con más hablantes nativos. Aún así, existen pocos sistemas con voces artificiales que soportan variantes locales, con sus diferencias fonéticas y de entonación, entre otras, como el español de Argentina.

El desarrollo de un sistema de conversión de texto a habla (TTS) necesita de buenas bases de datos, y que estén procesadas y etiquetadas adecuadamente, lo que requiere trabajo intensivo de recursos humanos, en muchos casos, con tareas manuales. Así, luego de que se completó el desarrollo de una voz artificial se tiene que probar que su calidad es adecuada para las necesidades de sus futuros usuarios.

Normalmente, se siguen varias iteraciones de evaluación y mejora de un sistema, de acuerdo al tiempo y los recursos disponibles. Este proceso suele ser largo, entre otras cosas, por el tiempo que toma la realización de las evaluaciones perceptuales con humanos.

En una evaluación subjetiva del habla una persona emite juicios sobre distintas elocuciones, tanto artificiales como naturales, y expresa directa o indirectamente, cuán aceptables y agradables le son, además de realizar otras evaluaciones sobre las mismas, como, por ejemplo, qué bien articuladas están o si tienen algún tipo de defecto sonoro. Uno de los objetivos de estas evaluaciones es determinar qué características del habla se asocian con buenos puntajes de los evaluadores, de forma de identificar atributos del habla que permitirían la evaluación automática de los sistemas, ‘copiando’ los criterios humanos.

Las pruebas clásicas para la evaluación perceptual de voces artificiales no evalúan completamente la experiencia del usuario, ya que no consideran totalmente el contexto en el cual se realizan las pruebas, y sólo se analizan en un contexto de laboratorio. Esto se plantea como el dilema principal de la evaluación de la calidad del habla.

Este trabajo buscó diseñar métodos de evaluación automática de la calidad del habla artificial generada a través de Sistemas TTS para el español de Buenos Aires. Los métodos incluyen nuevas métricas y otras ya existentes, y tienen como base las características de la percepción humana de la voz, así como el procesamiento automático de los parámetros acústicos de la señal de habla.

## PALABRAS CLAVE

---

voces artificiales • procesamiento automático del habla • sistemas de conversión de texto a habla • evaluación de calidad de la voz • percepción del habla

## ABSTRACT

---

Spanish is the main language in the Americas, and the fourth most spoken and the second with native speakers in the world. However, there are few systems that have artificial voices for local variants of this language, with their phonetical and intonational differences, among others, as Argentine Spanish.

The development of a text-to-speech (TTS) system requires high-quality databases, what is a human resource intensive goal, mainly due to it includes manual tasks as tagging and audio editing. Once a first version of a voice is available, it is necessary to evaluate if it fulfils the needs of its future users.

It is common practice to complete several iterations of evaluation and error correction cycles, as it is possible according to time and other available resources. This full process takes a long time to be completed, as human perceptual evaluations are highly time consuming.

In speech subjective evaluation, a person has to make judgements about several natural and artificial utterances, and they have to answer explicitly or implicitly how acceptable and likeable they are, additionally to other assessments related to speech articulation and signal artifacts, among others. One of the main objectives of this type of evaluations is to determine which speech characteristics are associated to better perceptual evaluations from the listeners. Thus, it will be possible to 'copy' the human criteria, in order to automatically identify relevant features in order to evaluate the quality of a system.

Standard test designs for the evaluation of artificial voices do not cover all the aspects related to user quality of experience, given they not present an ecological context for the tests, as they are only evaluated within laboratory setups. This might be the main issue in the evaluation of the quality of speech.

This thesis was aimed at the design of methods for the automatic evaluation of the quality of artificial speech generated through TTS systems with Argentine or similar variants of Spanish. In this work, new metrics as well as other known indicators are explored, mainly based on features related to human perception of voice, based on automatic acoustic processing of speech signals.

## KEYWORDS

---

artificial voices • automatic speech processing • text-to-speech systems  
• voice quality evaluation • speech perception

## AGRADECIMIENTOS

---

A mi Director, Jorge Gurlekian, y a todo el equipo del Laboratorio de Investigaciones Sensoriales (LIS), de/con quienes aprendí muchísimo y, más importante, pasé momentos muy lindos. Entre otras personas, muchas gracias, Diego, Miguel, Mercedes, Humberto y, especialmente, a la Dra. Miguelina Guirao, fundadora y alma máter del laboratorio que inició y marcó el camino de los sentidos y la percepción en Argentina y la región. A Hansjörg Mixdorff, por todo el apoyo durante el viaje a Alemania, cuyos resultados marcaron totalmente el trabajo de mi tesis.

Para toda la gente del Departamento de Computación (DC) de Exactas-UBA, donde encontré no sólo un lugar de trabajo, sino un hogar y una gran inspiración. En especial quiero agradecer a Agustín Gravano, mi consejero de estudios y referente en el área de procesamiento automático del habla en el DC, al equipo de divulgadores (divus), tanto actuales como anteriores, el cual dirijo desde hace más de seis años, y que me ayudaron a cumplir un montón de objetivos y proyectos en estos años, a Pablo, apoyo indispensable para llevar a cabo todos los proyectos en los que trabajé y seguiré trabajando, a Viviana, que me brindó su amistad y con quien compartimos muchas horas de trabajo y de charlas, y para toda la gente de Secretaría, Aída, Mónica, Lara, Paula y Alejandro. Para los equipos docentes de Algoritmos II y La Programación y su Didáctica, que me apoyaron para que pudiera cerrar este trabajo.

Para todo el equipo de Popularización de la Ciencia de Exactas-UBA, en especial a Valeria, Guillermo, Romina, Florencia y Vanesa, que no sólo hacen cosas fantásticas y con mucho compromiso, sino que me brindaron su cariño y amistad. A todas las integrantes de Dirección de Orientación Vocacional (DOV) de la Facultad, Luciana, Diana, Vanina y Claudia.

Para todo el equipo de Profesorados y CEFIEC de Exactas, en especial a Mariví y Gastón, que me sumaron como un integrante más de su equipo y me brindaron su afecto, y con quienes seguramente vendrán muchos proyectos interesantes.

A todos y todas los y las docentes que pasaron en el trabajo de formación docente de todos estos años, que realicé a la par de la tesis, y que terminaron de ajustar la orientación de mis próximas temas de investigación, además de la Fundación Sadosky, por la buena onda y el trabajo que hacen, en especial a Herman, Alfredo, Fidel, Fernando, Julián, Virginia, Fernando Sch. y Mara.

A todos mis amigos, y, en especial, a Fernando, Juan, Vanesa y Carmen, que siempre me brindaron todo su apoyo y cariño. A Claudia y Laura, que me apoyaron en distintos momentos del doctorado, y que me impulsaron a avanzar y terminar el trabajo.



Para mis padres, Jorge y Julieta, que, con mucho amor, esfuerzo y sacrificios, y enseñando con el ejemplo, me mostraron el valor del trabajo y, por sobre todo, que nunca hay que bajar los brazos.

Para mis hermanos, Jorge y Heber, y mis sobrinos, Irene, Sandra y Jorgito, que me llenaron de amor en estos años de trabajo.

Para Ágata, este trabajo y todo de mí.





## ÍNDICE GENERAL

---

1	El problema de la evaluación de la calidad del habla	1
1.1	Contexto	1
1.2	¿Por qué no se hacen más voces del español de diferentes regiones de América?	1
1.3	¿Por qué sería importante contar con voces en nuestra variante de la lengua?	3
1.4	El problema de la evaluación subjetiva del habla	3
1.5	La calidad de la experiencia como medida global de la calidad	5
1.6	Propuesta general de la tesis	5
1.7	Organización de la tesis	6
2	El habla artificial	7
2.1	Proceso general de los sistemas de conversión de texto a habla	7
2.1.1	Procesamiento Lingüístico	8
2.1.2	Síntesis de Habla	9
2.2	Escenarios de uso para el habla artificial	9
2.3	Tecnologías para la síntesis de habla	14
2.3.1	Síntesis por formantes	14
2.3.2	Síntesis por concatenación	14
2.3.3	Síntesis por articulación	16
2.3.4	Síntesis por Modelos Ocultos de Markov	16
2.3.5	Síntesis por Redes Neuronales Profundas	17
2.4	El sistema de conversión de texto a habla 'Aromo'	18
2.4.1	Corpus de oraciones en diarios	18
2.4.2	Evaluación del sistema	19
2.4.3	Diseño de extensión de base de datos	19
2.5	Conclusiones	19
3	Percepción del Habla	21
3.1	La comunicación humana	21
3.2	Modelos de percepción del habla	22
3.2.1	Modelos Pasivos	23
3.2.2	Modelos Activos	23
3.2.3	Otros modelos	24
3.3	Percepción del habla artificial	24
3.4	Información que se comunica con la voz	28
3.5	Conclusiones	29
4	Evaluación de la Calidad del Habla	31
4.1	Evaluación de Calidad de un Sistema TTS	32
4.2	Tipos de evaluación	33

4.3	Limitaciones de las pruebas usadas habitualmente . . . . .	36
4.4	Evaluación de sistemas de conversión de texto a habla . .	36
4.4.1	Sujetos . . . . .	37
4.4.2	Diseño del experimento . . . . .	37
4.4.3	Resultados . . . . .	39
4.4.4	Análisis de los resultados . . . . .	42
4.5	Propuesta de evaluación perceptual del habla . . . . .	45
4.5.1	El <i>corpus</i> EVALPERCEP2023 . . . . .	46
4.5.2	Diseño propuesto . . . . .	48
4.5.3	Diseño de la evaluación de la prueba . . . . .	50
4.5.4	Prueba básica para la evaluación general del habla artificial	52
4.6	Conclusiones . . . . .	54
5	Evaluación automática de la calidad del habla . . . . .	55
5.1	Contexto actual de la evaluación automática . . . . .	55
5.2	Características para la evaluación automática de la calidad del habla . . . . .	56
5.3	Creación de los modelos y prueba . . . . .	59
5.4	Resultados y Discusión . . . . .	60
5.4.1	Desempeño por dimensión de evaluación . . . . .	60
5.4.2	Modelos elegidos para la predicción de la calidad general	62
5.4.3	Validación de los modelos . . . . .	72
5.5	Conclusiones . . . . .	72
5.6	Trabajo futuro . . . . .	73
6	Prominencia en el habla . . . . .	75
6.1	Foco en el habla . . . . .	75
6.2	Perspectivas en el estudio de la prominencia prosódica . .	77
6.3	Evaluación del foco y prominencias para el español de Buenos Aires . . . . .	78
6.3.1	Diseño experimental . . . . .	81
6.4	Detección automática de la prominencia . . . . .	83
6.5	Resultados y Discusión . . . . .	84
6.5.1	Modalidad y Foco . . . . .	84
6.5.2	Detección de prominencias con y sin contexto . . . . .	86
6.6	Conclusiones . . . . .	90
7	Caracterización de habla alegre y enojada . . . . .	93
7.1	Corpus de habla alegre y enojada . . . . .	93
7.1.1	Diseño del Experimento . . . . .	94
7.1.2	Recolección de datos . . . . .	94
7.2	Diseño de la Entrevista . . . . .	95
7.3	Análisis preliminar de diferencias entre habla alegre y enojada	97
7.3.1	Caracterización del habla con MFCC . . . . .	98
7.3.2	Materiales y métodos . . . . .	99
7.3.3	Corpus utilizado . . . . .	99
7.3.4	Análisis realizados . . . . .	99

7.4	Resultados . . . . .	100
7.5	Discusión . . . . .	100
7.6	Conclusiones . . . . .	107
8	Cierre . . . . .	109
8.1	Resumen de resultados . . . . .	109
8.2	Conclusiones . . . . .	109
8.3	Posible trabajo futuro . . . . .	111
A	Diseño de la prueba de evaluación de calidad del habla . . . . .	113
A.1	Oraciones utilizadas . . . . .	113
A.1.1	Prueba ITU . . . . .	113
A.1.2	Oraciones prueba SUS . . . . .	114
B	Diseño de la prueba de evaluación de prominencia . . . . .	117
B.1	Oraciones utilizadas . . . . .	117
B.2	Instrucciones para los sujetos . . . . .	117
C	Diseño de la prueba de evaluación de calidad del habla . . . . .	119
C.1	Oraciones utilizadas . . . . .	119
C.2	Prueba básica: Instrucciones para los participantes . . . . .	121
D	Diseño de corpus para la evaluación de habla alegre y enojada . . . . .	123
D.1	Formulario a completar previo a la grabación . . . . .	123
D.1.1	Datos del Entrevistado . . . . .	123
D.2	Frases de connotación alegre, enojada y ambigua . . . . .	124
D.2.1	Grupo 1: Frases con connotación ambigua . . . . .	124
D.2.2	Grupo 2: Frases con connotación alegre . . . . .	124
D.2.3	Grupo 3: Frases con connotación de enojo . . . . .	125
	Bibliografía . . . . .	127

## ÍNDICE DE FIGURAS

---

Figura 2.1	Esquema de un Sistema de Conversión de Texto a Habla . . . . .	8
Figura 2.2	Máquina de von Kempelen, versión de C. Wheatstone [57] . . . . .	14
Figura 2.3	Esquema del sintentizador por formantes de Klaat [106] . . . . .	15
Figura 2.4	Esquema del aparato fonador humano [1] . . . . .	17
Figura 3.1	La cadena del habla [45] . . . . .	22
Figura 4.1	Ciclo de aprendizaje para un sistema TTS manual vs. uno con evaluación automática . . . . .	36
Figura 4.2	Promedio de las respuestas de la Inteligibilidad ITU, por sistema y grupo de oyentes . . . . .	39
Figura 4.3	Promedio de las respuestas de la Calidad ITU, por sistema y grupo de oyentes . . . . .	40
Figura 4.4	Promedio de las respuestas de la Aceptabilidad ITU, por sistema y grupo de oyentes . . . . .	40
Figura 4.5	Promedio del % de palabras correctas por oración y % de oraciones correctas para la Prueba ITU . . . . .	41
Figura 4.6	Palabras correctas en la evaluación SUS, para los 20 oyentes . . . . .	42
Figura 4.7	Promedio de respuestas de la evaluación MOS por grupo de oyentes . . . . .	43
Figura 4.8	Evaluaciones MOS por sistema para los 20 oyentes	43
Figura 4.9	Ejemplo de la escala para la evaluación de <i>Naturalidad</i> [87] . . . . .	49
Figura 6.1	Estructura de las oraciones y sílabas acentuadas	81
Figura 6.2	Pantalla del sitio web para la obtención de datos de foco y prominencia . . . . .	82
Figura 6.3	Prominencia normalizada con z-score por evaluador, de acuerdo a la posición en la oración, por modalidad y foco percibido . . . . .	86
Figura 6.4	Medida F (Prod y Desv. Est) para cada clasificador, atributos y sílabas de contexto. a) Duración b) Energía c) Atributos de Fo . . . . .	91
Figura 7.1	Relación de Escala Mel con respecto a las frecuencias . . . . .	99
Figura 7.2	Diferencia de Fo promedio para los 4 sujetos . . . . .	101
Figura 7.3	Promedio de Fo sobre todas las frases para los 4 sujetos . . . . .	102
Figura 7.4	Jitter y Shimmer para los sujetos 4 y 6 . . . . .	103

Figura 7.5	Coefficientes de Mel cepstrum para todos los sujetos . . . . .	104
Figura 7.6	Promedio de duración sobre todas las frases para los 4 sujetos . . . . .	105

## ÍNDICE DE CUADROS

---

Cuadro 2.1	Escenarios de uso de voces artificiales y sus características principales . . . . .	13
Cuadro 4.1	Longitud de las oraciones, en fonemas del alfabeto SAMPA [69] . . . . .	47
Cuadro 4.2	Evaluaciones promedio recibidas por cada voz y sistema para cada dimensión, con su desvío estándar . . . . .	53
Cuadro 5.1	$R^2$ y NMSE por cada dimensión usando regresión R.Forest con atributos básicos . . . . .	60
Cuadro 5.2	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Claridad' . . . . .	63
Cuadro 5.3	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Defectos' . . . . .	64
Cuadro 5.4	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Entonación' . . . . .	65
Cuadro 5.5	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Familiaridad' . . . . .	66
Cuadro 5.6	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Naturalidad' . . . . .	67
Cuadro 5.7	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Velocidad del habla' . . . . .	68
Cuadro 5.8	Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Evaluación General' . . . . .	69
Cuadro 5.9	Resumen de mejores modelos por cada dimensión de evaluación de acuerdo a $R^2$ y NMSE . . . . .	70
Cuadro 5.10	Evaluación de modelos con los datos de validación . . . . .	73
Cuadro 6.1	Prominencias. Identificación de modalidad de la oración . . . . .	85
Cuadro 6.2	Prominencias. Identificación de las condiciones de foco . . . . .	85

Cuadro 6.3	Tasas de detección de prominencias silábicas promedio para tres clasificadores . . . . .	85
Cuadro 6.4	Resultados de la detección de prominencia de acuerdo los conjuntos de atributos utilizados, en promedio y desv. est. de la medida F . . . .	87
Cuadro 6.5	Resultados de la detección de prominencia de acuerdo a la información de contexto silábica, en promedio y desv. est. de la medida F . . . .	87

## ACRÓNIMOS

---

ACR	Evaluación por categorías absolutas, del inglés <i>Absolute Category Rating</i>
TTS	Conversión de Texto a Habla, del inglés <i>Text-to-Speech</i>
HMM	Modelos Ocultos de Markov, del inglés <i>Hidden Markov Models</i>
CART	Árboles de Regresión y Clasificación, del inglés <i>Classification And Regression Trees</i>
SAMPA	Métodos para la evaluación del habla: Alfabeto Fonético, del inglés <i>Speech Assessment Methods: Phonetic Alphabet</i>
ASCII	Código estándar americano para el intercambio de información, del inglés <i>American Standard Code for Information Interchange</i>
RRNN	Redes Neuronales
PSOLA	Suma solapada sincrónica con la frecuencia fundamental, del inglés <i>Pitch Synchronous OverLap Add</i>
MBROLA	Suma solapada multibanda de resíntesis, del inglés <i>Multi-Band Resynthesis OverLap Add</i>
SMO	Algoritmo de optimización mínima secuencial, del inglés <i>Sequential Minimum Optimization</i>
POS	Parte de oración, del inglés <i>Part-of-Speech</i>
MOS	Nota media de opinión, del inglés <i>Mean Opinion Score</i>

## EL PROBLEMA DE LA EVALUACIÓN DE LA CALIDAD DEL HABLA

---

### 1.1 CONTEXTO

El español es la principal lengua del continente americano y la cuarta más hablada en el mundo, además de la segunda con más hablantes nativos del mundo —sólo después del Chino Mandarín— [41]. Consecuentemente, está entre las opciones de idioma a elegir en los principales asistentes personales de celulares, tablets y otros dispositivos, como *Google Assistant*, *Apple Siri* y *Amazon Alexa*, aunque para este último caso sólo recientemente se incorporó la variante del español americano [2].

Por otro lado, en los sistemas más usados de Conversión de Texto a Habla, del inglés *Text-to-Speech (TTS)*, que son los más reconocidos en el mercado, en general sólo aparece la variante americana neutra. Existen algunos pocos sistemas que tienen voces de variantes locales, como el español de Argentina en *Nuance Vocalizer*, pero en general no poseen la calidad de las variantes neutras, que tienen mucho más tiempo de desarrollo y uso.

### 1.2 ¿POR QUÉ NO SE HACEN MÁS VOCES DEL ESPAÑOL DE DIFERENTES REGIONES DE AMÉRICA?

El desarrollo de un sistema *TTS* requiere tener buenas bases de datos, y que estén procesadas y etiquetadas adecuadamente, lo que implica el trabajo intensivo de recursos humanos, en muchos casos, con tareas manuales, con su consecuente alto costo. Sin embargo, hay muchos proyectos que pudieron llegar a resultados con voces de calidad para distintas variantes del español, como el de Buenos Aires [205], distintas regiones de España [21, 59, 123], y Colombia [175], entre varias otras, así como en otras lenguas que se hablan en regiones donde impera la lengua española, como el Vasco [146]. Adicionalmente, hay esfuerzos como el realizado por Guevara-Rukoz y col. [67], quienes recopilaron y pusieron a disposición un *corpus* con diferentes tipos de habla en español de las variantes de Argentina, Chile, Perú y Venezuela, recolectado por medio de *crowdsourcing*.

Una vez que se completó el desarrollo de una voz artificial se tiene probar que, efectivamente, su calidad es adecuada para las necesidades de sus futuros usuarios —esto es, otros seres humanos—. Así, lo habitual es realizar una prueba integral de los sistemas, por medio de la realización de evaluaciones perceptuales, donde a muchas personas

se le hace escuchar distintas elocuciones generadas por varias fuentes artificiales e, inclusive, hechas por otros seres humanos, y éstas deben evaluar cada una de ellas. Así, con los resultados de estas evaluaciones, el equipo de desarrollo de un sistema debe analizar los resultados, y determinar posibles ajustes de acuerdo a cuál fuera el posible origen de cada problema detectado. Un ejemplo de las evaluaciones que deben realizar los oyentes es la que está descrita en la Norma P.85 de la ITU [211], sobre la cual hablaremos en el [Capítulo 4](#).

En algunos casos, las falencias detectadas en las voces tienen origen en defectos en las bases de datos desde donde se entrenó o se tomaron las muestras como parte del desarrollo de los sistemas, mientras que en otros se deben a cuestiones relacionadas a los algoritmos utilizados para la generación de las elocuciones —e. g., la elección de los sonidos a utilizar—. En particular, dentro del último escenario se encuentran los problemas asociados a la mala definición de los parámetros de los algoritmos, como ocurre con los coeficientes o pesos, entre otras cosas que se pueden ajustar en un sistema, que determinan en mayor o menor medida la calidad final de los audios generados.

Así, normalmente, se siguen varias iteraciones de evaluación y mejora, de acuerdo al tiempo y los recursos disponibles por cada equipo de desarrollo. Este proceso suele ser largo, entre otras cosas, debido al tiempo que toma la realización de las evaluaciones con humanos, además de que lo habitual es buscar la mayor cantidad posible de personas participantes, de manera de cubrir cierta variedad en las preferencias personales, y poder realizar una generalización robusta sobre la calidad percibida de las voces generadas por los sistemas.

En este proceso se pueden identificar algunas falencias comunes, que en muchos casos tienen que ver con lo artificial de las pruebas realizadas [221], sobre las que profundizaremos en próximas secciones de este trabajo. A continuación los principales problemas identificados y posibles acciones que se pueden tomar para resolverlos:

1. **¿Cómo bajar el costo en tiempo y dinero de las evaluaciones perceptuales?**

Las evaluaciones perceptuales insumen mucho tiempo y recursos. Es necesario poder hacer más cantidad de evaluaciones, y en menos tiempo. Para esta tarea se puede entrenar a una computadora para que realice las evaluaciones en forma automática, tomando como referencia evaluaciones humanas sobre una variedad de elocuciones, tanto artificiales como naturales.

2. **¿Cómo ajustar los parámetros de los sistemas de conversión de texto a habla a partir de las evaluaciones recibidas, de forma de maximizar la satisfacción de los usuarios?**

A partir de las evaluaciones realizadas en forma automática, se puede realizar un proceso de optimización, por ejemplo, por



medio de algoritmos de aprendizaje por refuerzos, con el fin de maximizar las evaluaciones recibidas.

3. **¿Cómo saber si las bases de datos utilizadas son adecuadas para el uso que se le va a dar?**

Además de los métodos conocidos de evaluación lingüística de la cobertura de sonidos (por ejemplo, el algoritmo descrito en [12]), también se puede sacar información de las evaluaciones recibidas por algoritmos automáticos. Así, no sólo podría ser utilizada para saber si el conjunto de sonidos es el adecuado, sino también servir de guía para eliminar sonidos redundantes de las bases.

4. **¿Cómo facilitar y acelerar los tiempos de puesta a punto para que un sistema TTS esté disponible más rápidamente?**

El trabajar con evaluaciones largas y costosas como las que se realizan habitualmente implica una demora en el desarrollo inicial de los sistemas TTS. Si se comenzara con evaluaciones automáticas, se estaría reduciendo el tiempo hasta obtener un sistema de calidad aceptable. Recién luego de esa etapa de ajustes preliminares, se harían las pruebas con seres humanos, pero sólo para completar la puesta a punto final.

1.3 ¿POR QUÉ SERÍA IMPORTANTE CONTAR CON VOCES EN NUESTRA VARIANTE DE LA LENGUA?

Un sistema que no satisface las demandas y necesidades de los usuarios, finalmente, no será aceptado y utilizado de manera extensiva. Así, de acuerdo a investigaciones sobre habla natural [63], así como en agentes virtuales y voces artificiales [110], los usuarios prefieren, evalúan mejor, y confían más en voces que tienen características similares a las propias, y esto, además, podría tener impacto en escenarios donde se usan esas voces, como ocurre en los entornos virtuales de aprendizaje [29]. Adicionalmente, se encontró que la variante de la lengua que expresa un hablante tiene efecto en la percepción de confiabilidad e inteligencia que se tiene del mismo [19], lo que puede ser especialmente importante para el uso de voces artificiales en venta y promoción de productos comerciales.

1.4 EL PROBLEMA DE LA EVALUACIÓN SUBJETIVA DEL HABLA

La evaluación subjetiva del habla ocurre cuando una persona emite juicios sobre distintas elocuciones, tanto artificiales como naturales, y expresa directa o indirectamente, cuán aceptables y agradables le resultan, además de realizar otras evaluaciones sobre las mismas, como por ejemplo, qué bien articuladas están o si tienen algún tipo de defecto sonoro. Uno de los objetivos de estas evaluaciones es, finalmente,

poder determinar qué características del habla se asocian con mejores puntajes por parte de los evaluadores. De esta manera, se pueden identificar atributos del habla que permitirían una evaluación automática de los sistemas, de manera de ‘copiar’ los criterios humanos.

Hay varios problemas que van inherentemente asociados a las evaluaciones perceptuales en este contexto:

1. **¿Cómo preguntar?:** Se debe determinar cómo preguntar a un oyente acerca de cómo evalúa una elocución. Esto tiene varios problemas, pero el principal es que una componente de la evaluación tiene que ver con las emociones que suscita el habla en el oyente, y se sabe que el reporte personal no es una fuente confiable [229, 230]. Existen varias formas para mitigar este riesgo, como el realizar evaluaciones por comparación y determinar órdenes más que *rankings* absolutos [228], pero es un problema que no termina por resolverse.
2. **¿Qué hace que algo sea agradable para todos los oyentes?:** Aún en elocuciones naturales hechas por locutores profesionales, que no tienen ‘errores’ o defectos evidentes, no hay acuerdo total sobre lo que es ‘agradable’. Esto es, no existen referencias o formas de predecir totalmente cuán agradable será percibido un estímulo por parte de cualquier oyente.
3. **¿Cómo generalizar?:** Si una voz es evaluada como ‘de buena calidad’ de acuerdo a un conjunto de oyentes, es relevante saber hasta qué punto esas evaluaciones pueden generalizarse para una población más grande, en muchos casos totalmente indefinida<sup>1</sup>, de forma de estar seguros de que pueden ser tomados como referencia.
4. **La componente *hedónica* en la evaluación:** Al evaluar una voz se pueden encontrar características que tienen que ver con cosas ‘medibles’ acústicamente, tanto en pequeña como en gran escala, mientras que hay otras características asociadas a cosas más difusas, generalmente relacionadas al placer que puede dar el escuchar una voz. Estas componentes que aparecen al momento de evaluar una voz representan un desafío para quienes necesitan abstraerse, al menos parcialmente, de los gustos particulares de ciertos oyentes.

---

<sup>1</sup> No alcanza con pensar en hablantes nativos de la lengua, ya que nuestras voces podrían ser usadas por personas que tienen al español como segunda lengua, u otras que están aprendiendo el idioma o, simplemente, están interesadas en él.

## 1.5 LA CALIDAD DE LA EXPERIENCIA COMO MEDIDA GLOBAL DE LA CALIDAD

Las pruebas clásicas para la evaluación perceptual de voces artificiales no evalúan completamente la experiencia del usuario (QoE, del inglés *Quality of Experience*) [89], ya que, entre otros aspectos, no consideran totalmente el contexto en el cual se realizan las pruebas—o, directamente, lo obvian por completo—, y sólo se analizan en un contexto de laboratorio. Esto se plantea como el dilema principal de la evaluación de la calidad del habla [221].

La posibilidad de contar con escenarios concretos de prueba, más cercanos al uso esperado de las voces artificiales, son determinantes para lograr pruebas que realmente predigan la aceptación de un sistema y puedan evaluar adecuadamente cómo será el desempeño de un sistema en el campo [34, 223]. Sin embargo, para el caso de las voces artificiales de uso general, no existe previamente una definición de cuáles serán los escenarios de uso que permitan definir pruebas ‘ecológicamente válidas’. Así, se hace necesario realizar diseños de pruebas que permitan evaluar la calidad del habla en diversas situaciones, incorporando criterios relacionados a la experiencia del usuario, entre otros, como parte de su elaboración.

De acuerdo a lo anterior, se puede considerar que el problema de la evaluación de la calidad del habla está aún abierto, y presenta muchas áreas de desarrollo que deben ser exploradas para no sólo mejorar su poder predictivo, sino también aportar a una mayor y mejor comprensión del funcionamiento de la percepción humana del habla, así como de la relación que tiene esta con la calidad de la experiencia de los usuarios de esas voces.

## 1.6 PROPUESTA GENERAL DE LA TESIS

A partir de la problemática planteada anteriormente, este trabajo buscó diseñar métodos de evaluación automática de la calidad del habla artificial generada a través de Sistemas TTS para el español de Buenos Aires. Los métodos integran nuevas métricas con otras ya existentes, y tendrán como base las características de la percepción humana de la voz, así como el tratamiento automático de los parámetros acústicos de la señal.

El trabajo tiene como uno de sus aportes principales la integración, en un mismo modelo, de métricas originadas en la percepción humana con otras basadas en atributos físicos acústicos del habla. Así, sus resultados podrán ser utilizados dentro de ciclos de aprendizaje automático, debido a que brinda información relevante a diseñadores y desarrolladores de sistemas TTS.

En este trabajo de tesis también se profundizará en los problemas actuales de la evaluación de calidad, además de en las propuestas diseñadas para poder aportar al desarrollo de esta área.

### 1.7 ORGANIZACIÓN DE LA TESIS

Este trabajo de tesis se organizó de la siguiente manera:

A lo largo de este [Capítulo 1](#) se realizó un resumen del problema de la evaluación de la calidad del habla, y se mostró la importancia de contar con evaluaciones adecuadas, que no sólo tengan un buen diseño desde el punto de vista técnico, sino que también contemplen particularidades de nuestra lengua para permitir el desarrollo de las tecnologías del habla en español. A continuación, en el [Capítulo 2](#) se hará una presentación de distintos escenarios en los cuales se utilizan o podrían utilizarse voces artificiales, además de los distintos procesos y técnicas utilizadas para generarlas. En el [Capítulo 3](#) se hará una revisión de los distintos modelos de percepción humana del habla, además de profundizar sobre la percepción del habla artificial, en tanto ambos aspectos son el sustento para poder tener evaluaciones de calidad que sean representativas de lo humano. Luego, en el [Capítulo 4](#) se desarrollará acerca de las distintas metodologías y técnicas utilizadas para la evaluación de la calidad del habla, de forma de evaluarlas críticamente en tanto ellas sirven, o no, para evaluar los distintos escenarios de uso de las voces artificiales. Complementariamente, en este capítulo se mostrarán los resultados de una evaluación de la calidad del habla artificial, donde se compararon dos voces de sistemas muy reconocidos y la voz en la que se trabajó como parte del trabajo de doctorado. Además, en este capítulo se propone el diseño de una prueba para la evaluación perceptual del habla, con el fin de mejorar las principales falencias detectadas en las pruebas del estado del arte. Por otro lado, en el [Capítulo 5](#) se proponen características para evaluar el habla artificial, así como un modelo para la predicción de distintas dimensiones asociadas a la calidad percibida, y en el [Capítulo 6](#) se mostrarán los avances en la descripción y reconocimiento automático de la prominencia en el habla para el español y el alemán, la cual se avisa como un factor relevante para la predicción de la calidad percibida del habla. Adicionalmente, en el [Capítulo 7](#) se resumirá el trabajo para la creación de un *corpus* de habla natural expresiva, y los resultados preliminares obtenidos de él para la clasificación de habla alegre y enojada. Por último, en el [Capítulo 8](#) se hará un resumen de los principales resultados obtenidos durante este trabajo de tesis. Además, quedarán definidas las principales conclusiones, así como las posibles líneas de trabajo futuro e interrogantes abiertos que pueden abrir la puerta para otros trabajos de tesis de grado y posgrado.

En este capítulo haremos una revisión del habla artificial, a través de los distintos métodos utilizados para generarla y las situaciones en las que se las utiliza. Para ello, comenzaremos con un resumen de las distintas situaciones —a las que llamaremos *escenarios de uso*— en las cuales se usan voces artificiales o podría llegar a ser necesaria su utilización. Luego se presentará un proceso general para los sistemas de habla artificial, que abarca desde que se define qué es lo que se quiere comunicar hasta la salida de la elocución. Por último se hará una revisión de las técnicas más extendidas para la generación de habla artificial.

El **habla artificial** comprende toda manipulación y creación de habla semejante a la humana utilizando, por ejemplo, técnicas matemáticas [213] o algoritmos computacionales de varios tipos —e. g., Redes Neuronales Profundas [131]. En particular, esta tesis hace foco en el habla artificial generada por medio de Sistemas de TTS, como se mencionó anteriormente. Un sistema TTS se encarga de transformar un texto de entrada en una secuencia de sonidos equivalente a la que produciría una persona al leerlo en voz alta. Para ello se sigue un proceso para llegar desde el texto original hasta el audio generado, como se describe a continuación.

## 2.1 PROCESO GENERAL DE LOS SISTEMAS DE CONVERSIÓN DE TEXTO A HABLA

El proceso de conversión de texto a habla puede dividirse en dos partes esenciales [181, 207], como se muestra en la [Figura 2.1](#): 1) *Procesamiento Lingüístico*, también denominada ‘parte frontal’ o *frontend*, donde se recibe un texto a convertir y a partir de él se obtiene la secuencia de sonidos a pronunciar y sus características, y 2) *Síntesis de Habla*, también denominada ‘parte trasera’ o *backend*, donde se materializa esa secuencia de sonidos por medio de alguna técnica determinada y se obtiene el segmento de habla de salida —i. e., la onda de sonido.

A continuación resumiremos cada uno de los subprocesos de estos dos bloques de un sistema de conversión de texto a habla:

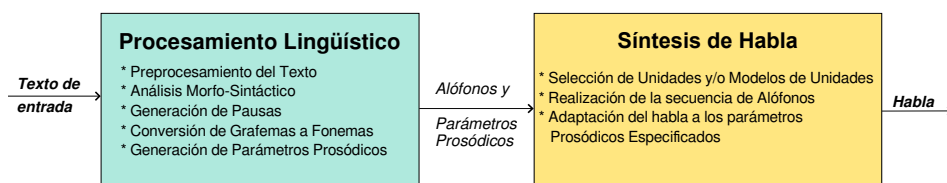


Figura 2.1: Esquema de un Sistema de Conversión de Texto a Habla

### 2.1.1 *Procesamiento Lingüístico*

#### *Preprocesamiento del texto*

Este subproceso se encarga de obtener un texto que pueda ser procesado por las subsiguientes etapas. Así, se debe segmentar el texto original en unidades más pequeñas, denominadas *tokens*, y se expanden aquellas palabras abreviadas o que tengan un pronunciación que no se obtenga unívocamente del texto. Por ejemplo, dado el texto original “Me dijo que para las 12:45 le tuviera listos los 1320.50 u\$s”, este se puede leer de varias formas como “me dijo que para las doce y cuarenta y cinco le tuviera listos los mil tres veinte con cincuenta dólares”, pero también como “. . . para las doce horas cuarenta y cinco minutos le tuviera listos los mil trescientos veinte dólares con cincuenta centavos”.

Dentro de esta parte del proceso se encuentran la *tokenización*, como se comentó anteriormente, la *expansión de abreviaturas*, *expansión de números*, *fechas y unidades de medida*, entre varias otras tareas, como describe Llisterri en [121].

#### *Análisis morfológico y pronunciación de palabras*

Aquí se analiza cómo están formadas las palabras para identificar correctamente la forma de pronunciarlas. Esto es un problema en idiomas como el inglés, entre muchos otros, donde la conversión de grafemas a fonemas no es directa, sino que requiere conocer el origen de las partes que la componen.

Dentro de este subproceso se incluyen el desarrollo y utilización de *lematizadores*, que se encargan de obtener la forma base de un palabra — el *lema*—. Por ejemplo, ‘jugar’ es el lema del verbo conjugado ‘jugaron’.

#### *Análisis Sintáctico, acentuación y fraseo*

En esta parte se analiza la estructura sintáctica del texto a convertir, de forma de encontrar los lugares donde deberían corresponder distintos tipos de pausas, de acuerdo a distintas técnicas como, por ejemplo, Árboles de Regresión y Clasificación, del inglés *Classification And Regression Trees (CART)* [207], o por medio de Modelos Ocultos de Markov, del inglés *Hidden Markov Models (HMM)* [201].

*Conversión de grafema a fonema*

Aquí se obtiene la secuencia de fonos y alófonos que se deberían generar a partir del texto original y las etapas de procesamiento posterior. Para el español existen reglas precisas que permiten realizar esta conversión, de acuerdo en el contexto en el que se encuentre cada grafema, como se describe en [207]. En esta tesis se utilizará la adaptación al español de Argentina de *SAMPA* [69], ya que, entre otras, tiene la ventaja de presentar a los sonidos en formato *ASCII*, lo que facilita su utilización sin la necesidad de utilizar caracteres especiales, como los que requiere, por ejemplo, el alfabeto *IPA* [90].

*Generación de parámetros prosódicos*

En esta etapa se añade información de cómo deben pronunciarse los fonos y alófonos definidos previamente. Así, como se describe en [207], entre otras se deben hacer las siguientes definiciones para poder completar la síntesis posteriormente:

**FRECUENCIA FUNDAMENTAL (FO)** : se debe determinar el contorno de  $F_0$ , que es el correlato físico de la entonación del texto que se está leyendo. Entre otros, se utiliza el Método de Fujisaki, como se describe en [71].

**DURACIÓN SEGMENTAL** : cada segmento acústico definido previamente debe tener una duración específica, la cual se puede obtener a través de una importante variedad de métodos, como *HMM*, *CART* o Redes Neuronales (*RRNN*).

**ENERGÍA** : así como varía la duración y la frecuencia fundamental de cada segmento del habla, también lo hace energía total.

2.1.2 *Síntesis de Habla*

Una vez que se obtuvo la secuencia de sonidos a generar —i. e., tanto la secuencia de fonos/alófonos y parámetros prosódicos, se debe materializar la elocución final. Para ello actualmente se utilizan diversas técnicas, como síntesis por formantes, síntesis por concatenación de unidades, Modelos Ocultos de Markov y Redes Neuronales, entre muchas otras técnicas, como se describe en la [Sección 2.3](#).

## 2.2 ESCENARIOS DE USO PARA EL HABLA ARTIFICIAL

A continuación se resumen algunos escenarios en los cuales se utilizan sistemas de generación de habla artificial, de forma de comprender cuáles son los requisitos que se tienen para este tipo de sistemas para, posteriormente, poder desarrollar pruebas que se asemejen a uno o varios de posibles los escenarios de uso. Así, tomando como base

la lista presentada por Wagner y col. [223], se pueden ver distintas características que pueden definirse para cada uno de estos casos, como se resume en el Cuadro 2.1.

- Información por llamada: incluye a aquellos servicios manejados por medio del teclado, o que funcionan como sistemas de diálogo, pero con poca interacción, ya sea por vía telefónica común como por banda ancha —e. g., llamadas vía servicios de llamadas como *Skype*.
- Lectores de pantalla para usuarios con visión disminuida: aquellas personas con visión disminuida cuentan con lectores de pantalla tanto para el uso de computadoras, como tablets y celulares [176].
- Audiolibros: lectura de libros para su uso por medio de teléfonos celulares, tanto en ámbitos silenciosos como en aquellos donde puede haber ruido, como mientras se maneja un auto.
- Locución en *off*: para acompañar videos y presentaciones.
- Voces para personajes de juegos: en esta categoría van las voces creadas para ser utilizadas en el contexto de juegos o en otros entornos interactivos [129], con los cuales el usuario deberá interactuar regularmente
- Asistentes en celulares o tabletas: aquí nos referimos a los asistentes de software que están disponibles en teléfonos celulares y tablets, como *Alexa* y *Google Assistant*, los cuales se caracterizan por tener que responder a todo tipo de situaciones, desde la búsqueda de una receta de cocina hasta el de una película, o distintos comandos para ejecutar funcionalidades de los dispositivos.
- Asistentes hogareños: en general, aparatos dedicados exclusivamente a ser asistentes hogareños, que en muchos casos tienen control sobre diferentes partes del hogar, como sistemas de domótica <sup>1</sup>.
- Voces para asistentes robóticos: en esta categoría se incluyen los robots asistentes, en general disponibles para adultos mayores o problemas que deben estar bajo seguimiento continuo —e. g., con cuadros en los cuales pueden tener episodios que deben ser detectados rápidamente, o que están pensados para mantener diálogos cotidianamente con sus usuarios.
- Voces para robots humanoides: en esta categoría se incluyen las voces para robots de tipo humanoide. Así, se destaca que

<sup>1</sup> e. g., *Amazon Echo Dot* o *Google Home*.



la voz a utilizar va a interactuar especialmente con el aspecto visual del robot, y puede ocurrir el efecto de ‘valle inquietante’ [126] (generalmente conocido por su nombre en inglés, *uncanny valley*).

- Voces para personas que no pueden hablar: entran en esta categoría aquellas voces desarrolladas para que sean utilizadas por personas que, o bien nunca pudieron hablar o actualmente tienen imposibilidad o dificultades para poder hacerlo —e.g., porque perdieron la voz por una enfermedad que afectó su motricidad, como la Esclerosis Lateral Amiotrófica (ELA).
- Personajes de ficción: entran en este tipo las voces que se crean para acompañar personajes de ficción, por ejemplo, para su uso en la industria cinematográfica.
- Voces para cantar (e.g., vocoders): entran en esta categoría las voces artificiales desarrolladas para cantar, como la descrita en [144].
- Sistemas de información en espacios públicos: este es el caso de los sistemas de voz utilizados en los espacios públicos como aeropuertos, terminales de ómnibus, parques y estadios, entre muchas otros contextos donde se debe emitir una voz por medio de altavoces a un público indefinido y en un espacio físico amplio.
- Voces para aparatos hogareños (e.g., heladeras): se trata de aquellos aparatos hogareños, como diferentes tipo de electrodomésticos, que incorporaron avisos por medio de voz como parte de sus funcionalidades. Se destacan por usarse esporádicamente, y en ambientes relativamente silenciosos.
- Voces para aparatos de uso público (e.g., ascensores): entran en esta categoría todos los dispositivos que están pensados para el espacio público y con un uso eventual —i. e., normalmente, en espacios donde no se puede asegurar silencio—, como ocurre con ascensores o terminales de consulta.
- Robots industriales: equipamientos robotizados utilizados en la industria, que tienen interfaces de voz, entre otras posibles. Se destacan por usarse en contextos ruidosos y a lo largo de varias horas.

Cada una de estas situaciones puede caracterizarse de acuerdo a las siguientes dimensiones:

- Motivación del uso: cuál es la principal motivación para la persona que escuchará la voz [entretenimiento, comunicación personal, información pública, interacción humano-máquina, integración].

- Canal: el tipo de canal en el cual se presentará el audio, no sólo tiene que ver con el ancho de banda disponible, sino con las características del espacio donde será presentada la voz [Telefónico, VoIP (banda ancha), altoparlante, parlante, auriculares].
- Valor más importante para la voz: cuál es la característica más importante que se espera de la voz, de acuerdo al contexto de uso [naturalidad (que suene similar lo realizado por un humano), inteligibilidad (que se comprendan las palabras comunicadas), agradabilidad (que sea agradable de escuchar), expresividad (que pueda demostrar distintos grados de expresividad), robustez (que pueda articular expresiones de diverso tipo, complejidad y hasta idioma)].
- Tipo de interacción más importante: en qué tipo de comunicación se utilizará la voz [diálogo (intercambio interactivo de mensajes), comunicación unidireccional (mayormente comunicación en un solo sentido), ráfagas (intercambio, de baja interacción)].
- Frecuencia y duración del uso: de acuerdo a cuándo un usuario estándar debe interactuar con la voz artificial, cuán frecuentemente lo hace y con qué duración por sesión. La frecuencia puede ser Alta (uso diario), Media (uso regular, semanal o mensual), Baja (uso eventual), mientras que la duración puede ser Extra Larga (60min+), Larga (21-59min), Media (6-20min), o Corta (<5min).
- Dominio: de qué tipo son los contenidos que se deben sintetizar [dominio cerrado (el vocabulario y las estructuras de las frases es específica), dominio acotable (de uso general, aunque podría acotarse de acuerdo al contexto de uso), dominio general (no acotado, que puede ser usado para cualquier dominio)].

ESCENARIO DE USO	MOTIVACIÓN	CANAL	VALOR	TIPO COM.	FREC/DUR	DOMINIO
Info. por llamada	Información	Tel/VoIP	intelig.	Diálogo	Baja/C	Cerrado
Lectores de pantalla	Integración	Auric/Parlante	robustez	Unidirec.	Alta/XL	General
Audiolibros	Entretenimiento	Auric/Parlante	Nat/Expr	Unidirec.	Alta/XL	Acotable
Locución 'en off'	Entretenimiento	Auric/Parlante	Nat, Agrad	Unidirec.	Media/L	General
Personajes de juegos	Entretenimiento	Auric/Parlante	Int/Expr	Unidirec.	Media/L	Acotable
Asist. celulares	Uso general	Auric/Parlante	intel/rob	Diálogo	Alta/Corta	General
Asist. hogareños	Int. Hum-Máq	Parlante	rob/agrad	Diálogo	Alta/Corta <sup>2</sup>	General
Asist. robóticos	Uso general	Parlante	rob/agrad	Diálogo	Alta/Corta	General
Robots humanoides	Entretenimiento	Parlante	todos	Diálogo	Alta/XL	General
Personas sin voz	Integración	Parlante	rob/nat/exp	Diálogo	Alta/M	General
Personajes de ficción	Entretenimiento	Auric/Parlante	Int/Expr	Unidirec	Media/Media	Acotable
Voces para cantar	Entretenimiento	Auric/Parlante	exp/agrad	Unidirec	Baja/Media	Acotable
Info espacios públicos	Información	Altoparlante	intel/rob	Unidirec	Media/Corta	Cerrado
Aparatos hogareños	Int. Hum-Máq	Parlante	intel/rob	Ráfagas	Media/Corta	Cerrado
Aparatos de uso público	Int. Hum-Máq	Parlante	intel/rob	Ráfagas	Media/Corta	Cerrado
Robots industriales	Int. Hum-Máq	Parlante	intel/rob	Ráfagas	Alta/Corta	Cerrado

Cuadro 2.1: Escenarios de uso de voces artificiales y sus características principales

<sup>2</sup> Aquí la consideración es sobre la persona que usa la voz, y no sobre quienes deben escucharla, como es en el resto de los casos

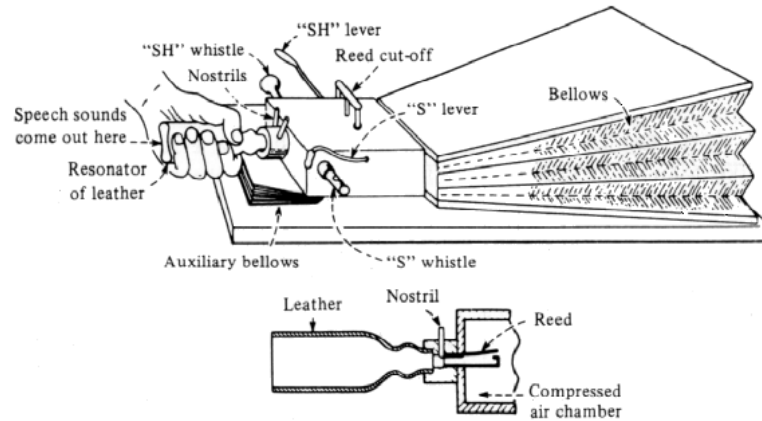


Figura 2.2: Máquina de von Kempelen, versión de C. Wheatstone [57]

### 2.3 TECNOLOGÍAS PARA LA SÍNTESIS DE HABLA

La generación de habla artificial se remonta al menos al año 1779, donde se registra la primera máquina documentada capaz de generar sonidos en forma mecánica. Esta máquina había sido desarrollada por Christian Gottlieb Kratzenstein, y fue ganadora del premio de la Academia Imperial San Petersburgo [57]. Algunos años después, en 1791, se registra la máquina desarrollada por Wolfgang von Kempelen [104], que es la hoy se conoce mayormente como el primer antecedente de una máquina parlante, cuya reconstrucción se puede apreciar en la Figura 2.2. Con la aparición de las computadoras se incorporó todo un conjunto de técnicas electrónicas, tanto analógicas como digitales, para la generación de habla, como las que se describen brevemente a continuación.

#### 2.3.1 Síntesis por formantes

También conocida como *síntesis por reglas*, es una técnica que tuvo éxito masivo hasta comienzos de los años ochenta [200]. En la síntesis por formantes se modela al aparato fonador como un filtro que actúa sobre una señal de excitación, modelando las cavidades del tracto vocal —i. e., los formantes. Uno de los sistemas por formantes más conocidos fue el desarrollado por Klatt [106], cuyo esquema se muestra en la Figura 2.3.

#### 2.3.2 Síntesis por concatenación

En este tipo de síntesis se tienen segmentos de voz pregrabados, de distintas longitudes de acuerdo al sistema, que luego se utilizan para poder construir las elocuciones requeridas según la secuencia de

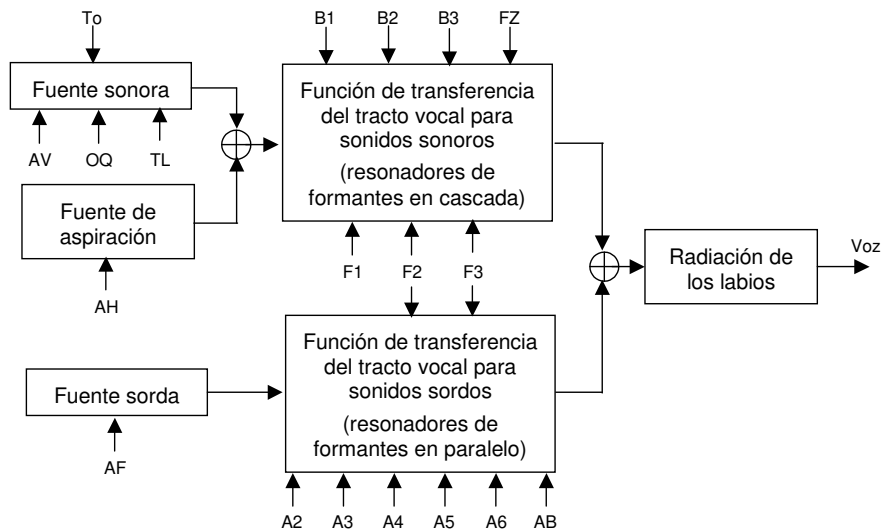


Figura 2.3: Esquema del sintetizador por formantes de Klaat [106]

alófonos y características prosódicas ya obtenidas. En los tamaños de las unidades se han desarrollado diferentes propuestas, tanto por fonos, difonos, trifonos, sílabas, demisílabas, palabras, frases, segmentos de longitud variable, o alguna versión que combina algunas de las anteriores.

Una de las ventajas de este método es que permite obtener habla inteligible y natural, de manera relativamente sencilla. Entre las desventajas principales, el sistema ocupa más espacio y requiere realizar un post-procesamiento de la señal para cubrir los atributos prosódicos definidos. Esto último se puede resolver a través de varias maneras, siendo las más extendidas la Suma solapada sincrónica con la frecuencia fundamental, del inglés *Pitch Synchronous OverLap Add (PSOLA)* y la Suma solapada multibanda de resíntesis, del inglés *Multi-Band Resynthesis OverLap Add (MBROLA)*, aunque tienen como desventaja que si el cambio a realizar es grande, la calidad final del audio generado no es buena, ya que agrega distorsiones [82]

Así, una alternativa es realizar la denominada *Síntesis por selección de unidades*, donde se tienen varias realizaciones acústicas de cada una de las unidades del repertorio, para distintas entonaciones, curvas de frecuencia fundamental y otros atributos acústicos. Como desventaja, requiere almacenar muchas variantes por cada sonido posible, con el consiguiente impacto en el tamaño de las bases de datos, lo que aumenta la complejidad del proceso de selección de las unidades a usar y, por consiguiente, el costo temporal del proceso completo de síntesis. Sin embargo, hay varios algoritmos que permiten reducir el tamaño de las bases sin una pérdida significativa de la calidad [189].

Por otro lado, el desarrollo completo de una base de datos de selección de unidades es largo y complejo, ya que no sólo deben contemplarse todas las variantes posibles de cada sonido, sino también

se debe verificar que la calidad del habla de todas las unidades sea similar, lo que se dificulta cuando se realizan extensiones de bases de datos para tener una mayor cobertura de sonidos, o si la base debe grabarse en varias sesiones debido a su extensión.

Así, para este tipo de sistemas el armado de un cuerpo de datos —comúnmente denominado *corpus*— requiere tener en cuenta varios aspectos para mantener una buena calidad, como los textos de las oraciones a usar, la precisión del etiquetado realizado, el número de instancias de cada unidad de concatenación y la cobertura del mismo. [206]

### 2.3.3 Síntesis por articulación

En este tipo de síntesis, también llamada *síntesis articuladora*, se trata de modelar la estructura y fisiología del aparato fonador humano (ver Figura 2.4) por medio de modelos matemáticos de la mecánica y acústica de la producción del habla. Así, se modela el comportamiento de los articuladores (e. g., lengua y labios), glotis y excitación del tracto vocal, tanto para sonidos sonoros, que no requieren la vibración de las cuerdas vocales, como para los no sonoros o sonidos fricativos (e. g., el sonido de la ‘z’ en ‘pez’).

En tanto parte del aporte de estas técnicas es el modelado del funcionamiento del aparato fonatorio, en la síntesis por articulación se utiliza la medición experimental del comportamiento de los articuladores para poner a prueba y ajustar sus modelos, como se reporta en [8].

### 2.3.4 Síntesis por Modelos Ocultos de Markov

La síntesis por HMM, es una técnica de síntesis estadística paramétrica [15, 232]. En ella se parte de datos para extraer características del habla, y se generan las ondas a partir de estos HMM basados en criterio de máxima verosimilitud, como se describe en [231]. Entre sus características se incluyen: espectro de frecuencias, que representa las características del tracto vocal; la frecuencia fundamental, relacionada a la fuente de sonido; y los atributos prosódicos, como las duraciones; entre otras.

Esta técnica es una de las más utilizadas en los últimos años para la generación de habla que expresa emociones [46], debido a la facilidad que tiene para ajustar a la expresividad deseada por medio de la modificación de los parámetros. Además, también se la utiliza ampliamente para la conversión de voces [139] —i. e., modificar una voz para denotar características distintas a las que tenía originalmente, pero sin cambiar las palabras pronunciadas—, entre otras aplicaciones

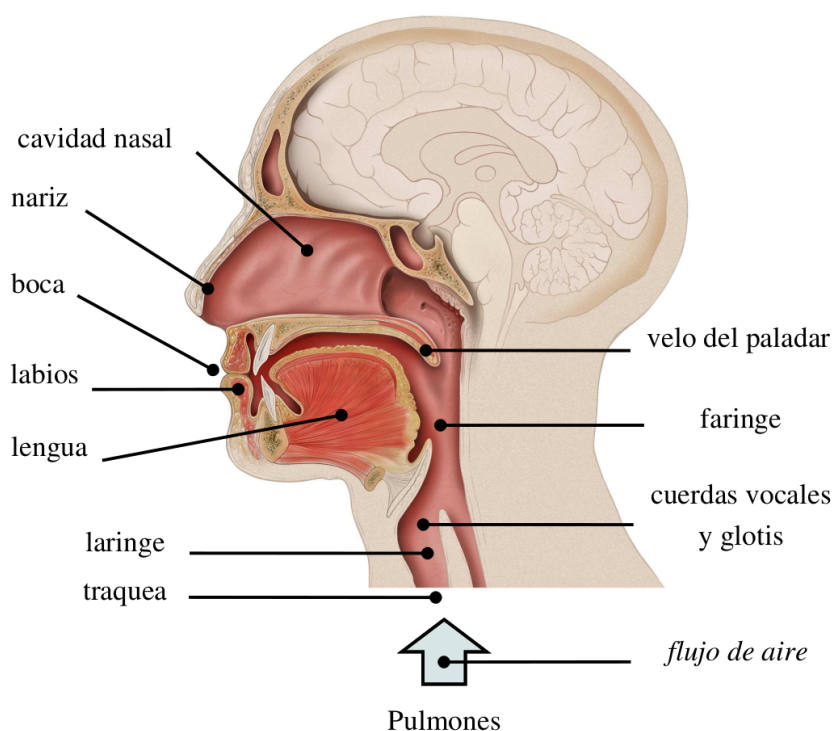


Figura 2.4: Esquema del aparato fonador humano [1]

### 2.3.5 Síntesis por Redes Neuronales Profundas

Las Redes Neuronales no empezaron a usarse intensivamente para la síntesis de habla sino hasta los últimos años, desde que se presentó con éxito *WaveNet*, desarrollado por el equipo de *Google* [152]. A partir de su éxito y de su facilidad para poder trabajar con pocos datos, tomando como base la transferencia de aprendizaje y a la adaptación de modelos [25], esta variedad de técnicas se convirtió en uno de los estándares del área.

A diferencia de lo visto anteriormente, este tipo de sistemas trabaja con la modalidad de extremo a extremo [224], del inglés *end-to-end*, ya que el entrenamiento de las redes se realiza con pares de texto-sonido, y no hay procesos intermedios como sí ocurre con los otros métodos vistos.

Así, podemos encontrar Redes Neuronales y sus diversas variantes en distintas aplicaciones de la síntesis de habla, como la generación de habla expresiva [196], sintetizadores multilingüaje [142, 237], generación de voces para canto [88], y síntesis de voces con pocos recursos disponibles [25], entre varias otras.

Además, en los últimos años se desarrollaron sistemas de conversión de texto a habla para el español usando RRNN profundas, como el de Bonafonte, Pascual y Dorca [18].

## 2.4 EL SISTEMA DE CONVERSIÓN DE TEXTO A HABLA ‘AROMO’

El sistema de conversión de texto a habla ‘Aromo’ [205] fue desarrollado en el Laboratorio de Investigaciones Sensoriales, con el fin de construir un sistema de conversión de texto a habla para el español de Buenos Aires, que fuera de alta calidad y naturalidad.

Como parte de su desarrollo se realizaron distintas tareas, que se resumen a continuación.

### 2.4.1 *Corpus de oraciones en diarios*

El sistema TTS requiere de la utilización de analizadores morfosintácticos — normalmente denominados etiquetadores Parte de oración, del inglés *Part-of-Speech* (POS)—, que permiten identificar, entre otras cosas, la función de una palabra en una oración para determinar los atributos prosódicos de la oración a sintetizar [207].

Para el sistema Aromo se compilaron oraciones obtenidas de diarios *online* de tirada nacional, que fueron etiquetados automáticamente con su clase de palabra en tres niveles, siguiendo la notación de Eagles para el español [93], que luego fueron corregidos en forma manual por lingüistas y expertos en procesamiento de lenguaje natural. El *corpus* contiene oraciones de noticias etiquetadas con clases de palabras, contenidas en 10 años de ediciones del diario Clarín entre los años 1995 y 2001, más un agregado de oraciones del diario La Nación obtenidas a fines de 2011.

Así, se compiló un total de 52100 oraciones, para un total de 80000 palabras etiquetadas, aproximadamente, siendo 17300 distintas, de las cuales un 52 % son sustantivos (dentro de ellos, un 34 % son nombres propios), 26 % verbos, 18 % adjetivos, 1 % de adverbios, y un 3 % de palabras de función (e. g., conjunciones, preposiciones, etc.)

Dentro del corpus realizado aparecen 59 clases de palabras utilizadas (2 adjetivos, 19 verbos, 2 sustantivos, 13 signos de puntuación, entre otras).

Para el entrenamiento del etiquetador se utilizó la herramienta Apache OpenNLP 1.5.1<sup>3</sup>, con la que se generó un modelo estadístico representativo del corpus utilizado. Dentro de la herramienta se utiliza la técnica de aprendizaje por Máxima Entropía, algoritmo GIS, tomando la información de cada palabra etiquetada y su contexto (palabras previas y siguientes). Con el modelo generado por Apache OpenNLP se obtiene el etiquetado para las nuevas oraciones. Por cada palabra se genera una lista de etiquetas posibles, con sus respectivas probabilidades, de la cual luego se selecciona la de mayor probabilidad para definir el etiquetado definitivo.

El etiquetador de clase de palabra obtenido alcanzó un 99 % de precisión en pruebas de validación cruzada de cinco partes, y poste-

<sup>3</sup> Apache OpenNLP project, [://opennlp.apache.org/](http://opennlp.apache.org/)



riormente fue incorporado como parte integral del sistema Aromo [206].

#### 2.4.2 Evaluación del sistema

Como parte del trabajo de desarrollo del sistema se realizó la prueba perceptual del mismo, y su comparación con otros sistemas comerciales conocidos, lo que se describe en el [Capítulo 4](#).

#### 2.4.3 Diseño de extensión de base de datos

Como se describe en [206], el sistema Aromo tuvo tres etapas en la creación de su base de oraciones. Como parte de este trabajo se realizó el diseño de las oraciones a utilizar a partir de los resultados obtenidos en las pruebas mencionadas anteriormente. Así, se agregaron 625 nuevas oraciones, de las cuales 36 fueron interrogativas, orientadas al uso en sistemas de diálogo, que abarcaban desde preguntas breves (e. g., “¿Qué tal?”), hasta preguntas utilizadas como parte de sistemas de información por teléfono (e. g., “¿Dónde desea recibir la correspondencia” y “¿Qué le pareció el servicio? ¿Volvería a utilizarlo?”). Adicionalmente, se incluyeron 195 oraciones declarativas extensas, diseñadas especialmente para incluir múltiples frases dentro de las mismas oraciones. Finalmente, se incorporaron 393 oraciones que permitían cubrir toda combinación posible de sonidos (difonos) del español, incluyendo combinaciones poco comunes, como las que surgen al pronunciar nombres extranjeros o nombres propios poco comunes (e. g., “Baruj Star apareció en el concierto con una campera de Asatej Turismo, donde junto con Baj cantaron sus temas favoritos”).

## 2.5 CONCLUSIONES

En este capítulo se presentó una visión general sobre el habla artificial y, en particular, sobre los Sistemas de TTS. Se hizo especial énfasis en la definición de los *Escenarios de Uso*, debido a que representan la base fundamental sobre la que, en futuras secciones, realizaremos un diseño de pruebas que permitan evaluar adecuadamente el desempeño de un sistema.

Con respecto a esto último, creemos que no es posible realizar una evaluación adecuada de una voz generada por medios automáticos si primero no se identifica y se describe claramente el contexto, y las restricciones y la forma en la cual sería utilizada. Así, el [Cuadro 2.1](#) representa un aporte para los futuros trabajos en el área, ya que resume el análisis de las características de distintos escenarios de uso de las voces artificiales.

Adicionalmente, en este capítulo se resumieron las partes principales de un proceso estándar para los Sistemas de Conversión de Texto

a Habla, desde la entrada del texto a convertir hasta la obtención de la onda que materializa el mensaje indicado, lo que será relevante en los siguientes capítulos. También se hizo un resumen de las principales técnicas utilizadas para la síntesis de habla, las cuales serán analizadas posteriormente en cuanto a qué características tienen con respecto a la calidad del habla generada, además de algunas ventajas y desventajas de cada una de ellas.

Por último, se resumió el trabajo realizado como parte del desarrollo del sistema Aromo [205] y, en especial, en distintos aspectos y herramientas lingüísticas del mismo [206], como un corpus y un etiquetador morfosintáctico para el español de Argentina, que alcanzó un 99 % de precisión en las pruebas sobre textos informativos.

En este capítulo se hará una revisión de los distintos modelos de percepción humana del habla, además de particularizar en la percepción del habla artificial. El conocimiento profundo acerca de cómo el ser humano percibe el habla, y, por consiguiente, sobre cómo procesa los estímulos acústicos y lingüísticos que recibe, es fundamental para crear evaluaciones del habla que sean representativas de la percepción humana.

A diferencia de otras señales que se pueden estudiar y analizar, el habla tiene la particularidad de ser generada por un ser humano, generalmente realizada con un fin de comunicación entre un hablante y un oyente [167].

La percepción del habla es un proceso dual, donde se integra la información de manera descendente (*top-down*) y ascendente (*bottom-up*). En su parte descendente, intervienen aspectos relacionados a la expectativa, generada tanto por atributos lingüísticos, como la estructura sintáctica, y, entre otros, por atributos acústicos suprasegmentales, como la entonación. En su parte ascendente, se trata de la información resultante del procesamiento de la señal recibida. Al momento no está definida totalmente cómo es la interrelación de estos dos tipos de procesos, pero sí está ampliamente estudiado que ambos tienen impacto en la percepción habla [35, 156, 194, 236]. En este sentido, son importantes las características segmentales del habla —por ejemplo, que no haya artefactos y las realizaciones de las unidades acústicas sean adecuadas—, pero también lo son los atributos suprasegmentales del habla, como las curvas de entonación, entre otras.

### 3.1 LA COMUNICACIÓN HUMANA

Para comenzar, recuperaremos el contexto en el cual se enmarca la percepción del habla, que es el de la comunicación humana.

Por otro lado, la percepción del habla está mediada por las características físicas del oído humano, que hacen que deba procesarse y analizarse de manera no estándar, si lo que se desea es tener en cuenta aquellos aspectos que son más relevantes. Así, por ejemplo, la percepción de las frecuencias se conoce que no es lineal, sino que sigue una escala logarítmica, como se plantea en la Escala Mel [36].

La percepción humana es multicanal y multisensorial. Esto es, recibimos información del mundo que nos rodea a través de nuestros sentidos, y dentro de cada sentido puede haber varios canales simultá-

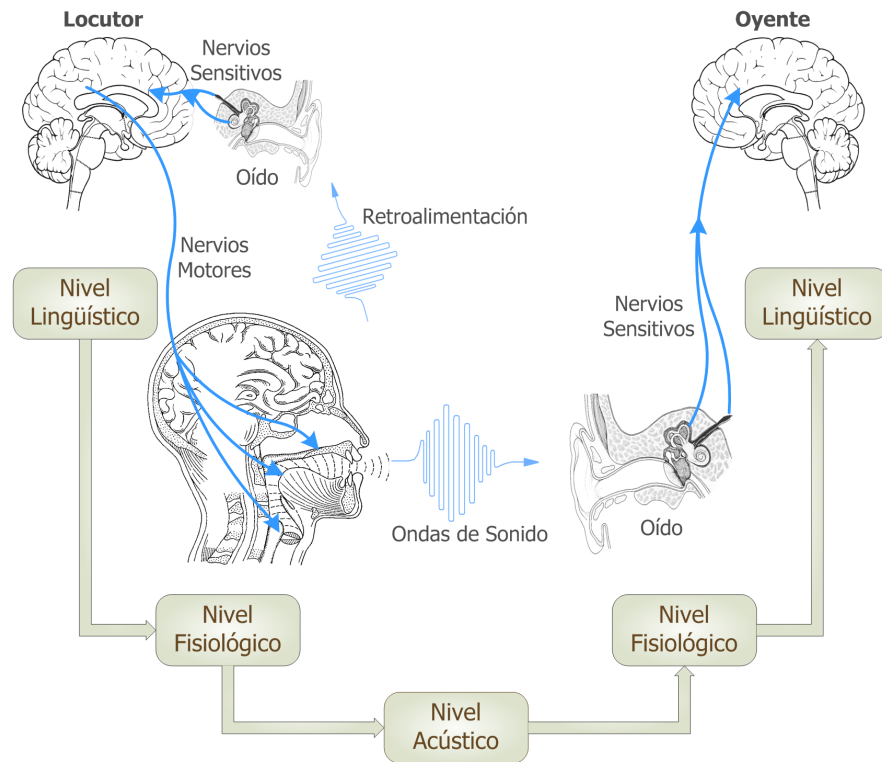


Figura 3.1: La cadena del habla [45]

neamente (e.g., escuchar un sonido de una alarma, el ruido de la calle, y una voz que nos habla).

Nuestro cuerpo es una 'máquina' que codifica y decodifica mensajes a través de múltiples canales (por ej., visual y auditivo). La voz juega un papel muy importante en nuestra vida cotidiana

En lo que respecta a este trabajo, nos interesa trabajar sobre la percepción humana del habla, en tanto es esencial para poder evaluar características relevantes de la voz artificial.

### 3.2 MODELOS DE PERCEPCIÓN DEL HABLA

Hay una separación básica entre los modelos para la percepción del habla, entre modelos pasivos y activos, y estos a su vez se dividen en más categorías [160].

**MODELOS PASIVOS** : Se entiende a la percepción como un proceso de clasificación directo, donde se reconocen los segmentos del habla a partir de una señal auditiva, considerada robusta y simple, además de lineal

**MODELOS ACTIVOS** : Se le asigna estructura a los segmentos del habla, para lo cual se requiere un esfuerzo activo del oyente, para obtener el mensaje a partir de una señal auditiva ruidosa y

compleja. Así, no se considera a la percepción como lineal, sino que es vista como un proceso con una estructura jerárquica

### 3.2.1 *Modelos Pasivos*

#### 3.2.1.1 *Modelos basados en plantillas*

Hay una cantidad de plantillas (*templates*) que se asocian a realizaciones acústicas. Se elige el que coincide exactamente (o el más parecido).

Debido al carácter ruidoso de la señal, sólo es aplicable si utiliza hipótesis o filtra *templates* a partir del contexto.

#### 3.2.1.2 *Modelos basados en filtros*

Se le aplican filtros fijos a la señal recibida. Con eso se obtienen las representaciones fonológicas en la mente de la persona.

#### 3.2.1.3 *Percepción directa*

El significado (representación fonológica) se obtiene en el oído por análisis espectral y temporal de la señal. El significado está presente en la señal acústica. Hay poca o nula mediación cognitiva en el procesamiento. Sólo se tiene en cuenta al oído, la señal, y su proyección en la corteza auditiva.

### 3.2.2 *Modelos Activos*

#### 3.2.2.1 *Teoría Motora*

Planteada por Liberman, en una primera versión en 1967 y, con actualizaciones, en 1985. Para esta teoría la percepción es un proceso activo que requiere cognición del oyente y tiene relación directa con el proceso de producción de habla. Así, el oyente evoca las propiedades coarticulatorias y motoras del habla escuchada para obtener sus etiquetas fonológicas.

Uno de sus puntos débiles es que no puede explicar la variabilidad en el habla (e.g., no se contempla la prosodia ni el contenido expresivo).

Adicionalmente, autores como Casserly y Pisoni [22], entre otros, ven que la percepción del habla no puede escindirse de la producción, ya que son parte un mismo proceso de comunicación humana, y deben ser estudiados juntos.

#### 3.2.2.2 *Teoría de Análisis por Síntesis*

Desarrollada por Stevens y Halle, en 1967. Es similar a la Teoría Motora, pero aquí es el conocimiento acústico y no el de la articula-

ción el que se tiene en cuenta. Así, el oyente evoca las propiedades acústicas del habla conocida y de la percibida, y obtiene sus etiquetas fonológicas.

Una de sus debilidades es que no puede explicar la variabilidad en el habla (e.g., no se contempla la prosodia ni el contenido expresivo).

### 3.2.2.3 *Otros modelos activos*

**TEORÍA DE LA PERCEPCIÓN CATEGÓRICA** (Liberman, 1958 y 1984): En este modelo se tienen representaciones simbólicas (categorías) para las señales físicas, y se realiza un esfuerzo cognitivo en asignar esas categorías.

**REALISMO DIRECTO** (Fowler, 1986): De la señal se puede obtener la representación cognitiva del sonido. Se usa una representación de la articulación.

**TEORÍA DE LA CANTIDAD** (Blumstein-Stevens 1979, Stevens 1989): La relación entre la señal acústica y la configuración física del tracto vocal no es lineal. Grandes cambios de configuración en el tracto pueden derivar en pequeños cambios acústicos.

**TEORÍA DEL ALMACENAMIENTO ASOCIATIVO** (1970): Se espera una señal que no es perfecta. Así, se la compara con representaciones ideales de señales. Para ello se define una hipótesis de una corrección, que luego podrá ser confirmada/descartada por el contexto. Si se descartó una hipótesis, se probará con otra.

### 3.2.3 *Otros modelos*

#### 3.2.3.1 *Modelos híbridos*

Se usan dos o más modelos de los anteriores. Por ejemplo, se empieza con un método pasivo (e.g., Percepción Directa), y, si se falla, se pasa a algún método activo. Un punto débil es que al momento tiene poca base experimental sobre los mecanismos de pasaje de uno a otro, pero aparece como más razonable.

## 3.3 PERCEPCIÓN DEL HABLA ARTIFICIAL

En lo que respecta al estudio de la percepción humana del habla artificial, tomando como base la revisión realizada por Winters y Pisoni [225], se pueden resumir las siguientes características que la diferencian de la percepción del habla natural [158, 159]:

1. **El habla sintética es menos inteligible que la natural:** Más allá de que en la actualidad el desarrollo de los sistemas de conversión de texto a habla permita tener buenos niveles de

inteligibilidad, anteriormente vimos que la percepción en entornos con ruido depende de información de más alto nivel, como la entonación. Este es uno de los aspectos que los sistemas TTS todavía no tienen resuelto por completo —en especial para oraciones de largas, de varias frases—.

2. **La percepción del habla sintética requiere de más recursos cognitivos:** Relacionado al punto anterior, la carga de trabajo cognitiva asociada a la percepción de habla artificial es mayor que la del habla natural, en tanto el oyente debe hacer un mayor esfuerzo para comprender lo que se está diciendo, ya que no recibe la misma cantidad y calidad de pistas que en el habla natural. Esto es especialmente notable al momento de utilizar los sistemas interactivos de voz (en inglés, IVR) por tiempos prolongados, o al escuchar audiolibros generados con voces artificiales.
3. **La percepción del habla sintética interactúa con el conocimiento lingüístico de alto nivel:** Como se mencionó anteriormente, la percepción de habla requiere de información que se obtiene de manera ascendente a partir de la señal, aunque también intervienen procesos descendentes, principalmente mediados por la expectativa. Así, debido a que el habla natural adolece de problemas en la inteligibilidad, especialmente en entornos ruidosos, se hace mucho más dependiente de que las oraciones que están siendo escuchadas sean consistentes con la información lingüística de mayor nivel, como la sintáctica y semántica. En particular, en los casos en que no hay incoherencias semánticas en las oraciones que están siendo escuchadas, se resiente la tasa de reconocimiento de palabras de las voces artificiales, mientras que el impacto es menor en aquellas naturales.
4. **El habla sintética es más difícil de comprender que la natural:** Relacionado a lo que se mencionó anteriormente de que el habla artificial es menos inteligible, también hay resultados que muestran que, al realizar preguntas luego de escuchar habla artificial, más allá de que se las responda correctamente, aumentan los tiempos de respuesta si se los compara con aquellos casos en donde se presenta habla natural. En este caso, la mayor demora en la respuesta se explicaría por el impacto de la calidad del habla artificial en el proceso ascendente de la señal acústico-fonética —i.e., aun cuando la respuesta fuera correcta, se tarda más tiempo en formularla—. Adicionalmente al mayor esfuerzo relacionado a la decodificación, también se suma que podría haber más dificultades al recuperar de la memoria aquellas palabras escuchadas si estas fueran artificiales. Por último, se puede asociar la mayor dificultad en la comprensión del habla artificial a que el oyente debe aplicar procesos de compensación, ya sea con-

ciente o inconcientemente, para poder procesar la información del habla artificial.

5. **La percepción del habla artificial mejora con la experiencia:**

Así como ocurre con otras tareas que debe realizar un ser humano, el desempeño con el reconocimiento y comprensión de las elocuciones con voces artificiales mejora con la experiencia. Esto es distinto a lo que pasa con el habla natural, donde, una vez que se adquirió la habilidad de procesar el lenguaje, la dificultad para decodificar el habla natural se mantiene relativamente constante a medida que se avanza en la realización de la tarea—descartando pequeñas mejoras relacionadas a la adaptación a la voces utilizadas—. Aún así, se espera que el desempeño con las voces artificiales nunca alcance el que se obtiene por medio de aquellas naturales, ya que se llega a un techo en la mejora que se puede tener, aunque la diferencia quedará determinada por el tipo de tarea realizada así como por la características del sistema utilizado. Por último, más allá de que el entrenamiento con una tarea determinada no puede extrapolarse a otras tareas, se encontró que la mejora general en el desempeño queda determinada por la variedad acústico-fonética de las muestras de elocuciones artificiales utilizadas.

6. **Diferentes grupos de oyentes procesan de manera distinta el habla artificial:**

De acuerdo al tipo de población que debe procesar el habla artificial se pueden obtener distintos resultados en las tareas asociadas a la percepción de esas elocuciones. Entre algunas de todas las variantes posibles están si la persona es hablante nativa de la lengua o no, si se trata de niños o adultos mayores, o si las personas tienen problemas auditivos o, por el contrario, son expertas en la percepción del habla y no poseen dificultades para escuchar. En las personas hablantes no nativas, entre otras situaciones, se vio que estas tenían un desempeño menor en tareas con oraciones comunes de la lengua, mientras que cuando se trataba de oraciones semánticamente impredecibles no había una diferencia notable con aquellos hablantes nativos. Esto último es esperable, en tanto el conocimiento de los hablantes no nativos de las estructuras de alto nivel del lenguaje es menor y, por ende, en las tareas donde estas no son determinantes para definir el desempeño de la tarea, las diferencias con los hablantes nativos se reduce considerablemente. Sin embargo, esto último queda acotado por el nivel del idioma que maneje la persona no nativa.

En los niños, una vez normalizado el grado del manejo del lenguaje, se vio que tienen una baja en el desempeño en las tareas al trabajar con habla artificial, más baja que en los adultos. En este caso, no se trataría de un problema específico con el habla



artificial, sino con la complejidad de las tareas que se deben realizar una vez que se escucharon las muestras de las voces. Del otro lado se encuentran los adultos mayores, en los cuales se encontró que, en líneas generales, al presentarles distintos tipos de estímulo tenían preferencia por velocidades de habla más bajas, debido a las diferencias en la percepción de estímulos temporales. Además, en este grupo tiene impacto la pérdida de audición general y, en particular, de las frecuencias altas, entre otros cambios, que hacen que ciertos estímulos no se escuchen en todo su espectro. Sin embargo, como algo favorable, podrían desaparecer de la percepción aquellos artefactos del audio que están en las bandas superiores. En lo que respecta a las tareas realizadas, podrían encontrarse diferencias en el desempeño alcanzado, de acuerdo al grado de deterioro cognitivo propio de la edad de los oyentes.

Para los oyentes expertos, entre quienes puede incluirse a fonaudiólogos, locutores y lingüistas especializados en habla, muestran un mejor desempeño en las tareas asociadas a la percepción de habla. En este caso, también obtienen una mejora en las tareas aquellas personas que recibieron entrenamiento analítico sobre habla artificial, aunque no fueran profesionales del lenguaje y la voz.

7. **Hay una interrelación entre los atributos prosódicos, la inteligibilidad y la naturalidad:** Más allá de que, a priori, pueda verse a la inteligibilidad y naturalidad como dimensiones ortogonales al momento de realizar mejoras a una voz artificial, ambas dimensiones están relacionadas. De hecho, en las pruebas de evaluación de calidad del habla se ve que muchas medidas definidas para evaluar la inteligibilidad correlacionan con aquellas utilizadas para analizar la naturalidad (e.g., ver diseño de pruebas ITU en el [Capítulo 4](#)). En particular, hay una relación entre la evaluación de los atributos prosódicos, la naturalidad y la aceptabilidad de una voz artificial. Así, las voces que tienen buenos atributos prosódicos (por ejemplo, pausas y entonación) y buena inteligibilidad también son vistas como de mayor naturalidad y, finalmente, son más preferidas y categorizadas como aceptables. En particular, se vio que una mala inteligibilidad afecta la percepción de los atributos prosódicos, bajando la diferencia en segmentos de habla donde, por ejemplo, se aplica una mala y buena curva de entonación. Así, la inteligibilidad se pone por delante de la diferencia en la entonación.

Sin embargo, para niveles de inteligibilidad aceptables, se hacen notables las diferencias en los atributos prosódicos, en especial para segmentos de habla largos, donde aparecen múltiples frases, lo que redundaría en una mejor percepción de la naturalidad, así como de la aceptabilidad en general.

Por otro lado, una mejor realización de los atributos prosódicos en una elocución también se asocia a mejores resultados en tareas donde, por ejemplo, se requiere recuperar datos para responder una pregunta. En esos casos, una buena prosodia permite un mejor recupero de la información requerida. Finalmente, en tanto la inteligibilidad se evalúa habitualmente por medio de tareas de escritura de los textos escuchados o respuesta de preguntas con datos específicos, los atributos prosódicos, generalmente asociados a la naturalidad, estarían afectando directamente a la evaluación de qué tan inteligible es una voz artificial.

### 3.4 INFORMACIÓN QUE SE COMUNICA CON LA VOZ

Una parte importante de la percepción de la voz, tanto artificial como natural, tiene que ver con la información del hablante que se puede obtener a partir del análisis de las elocuciones. Estas evaluaciones se enmarcan dentro de los atributos paralingüísticos del habla, como, por ejemplo, las emociones, afecto y personalidad que se expresan en el habla y pueden analizarse en forma automática [183].

Así, además de poder reconocer la secuencia de palabras de la elocución, a partir de la señal de la voz pueden obtener distintas informaciones sobre el hablante [179]:

- Lugar de origen (país, región) e idioma/dialecto [112]
- Edad, género [10]
- Nivel educativo [99]
- Contextura física, fuerza y altura [157, 169]
- Atributos de la personalidad [138, 145, 162, 164, 220]
- Conocer cuál es su evaluación sobre algo (i.e., si le cae mal o bien) [55]
- Actitud con respecto a algo (defensiva, colaborativa) [63, 119]
- Emociones y estado del ánimo (enojo, alegría, tristeza) [172, 178, 218]
- Atractivo sexual [11]
- Qué tanta confianza expresa [29, 62]
- Qué tan persuasiva es (en general, aplicado en contexto de venta) [40, 105]

### 3.5 CONCLUSIONES

En este capítulo se presentó una revisión general de la percepción humana del habla, ya que conocer su funcionamiento es clave para poder determinar qué aspectos de una voz —en este caso, artificial— podrían tener mayor impacto la forma en que se procesa y se evalúa una elocución.

La percepción del habla es un proceso dual, donde se integra la información de manera descendente (*top-down*) y ascendente (*bottom-up*). Así, son importantes las características segmentales del habla —por ejemplo, que no haya artefactos y las realización de las unidades acústicas sean adecuadas—, pero también lo son los atributos suprasegmentales del habla, como las curvas de entonación, entre otras.

Hay varias diferencias del habla natural y artificial. En particular, se vio que el habla sintética es menos inteligible que la natural, más allá de que las técnicas actuales de los sistemas TTS aseguran una buena inteligibilidad; que la percepción del habla sintética requiere de más recursos cognitivos para su procesamiento, en particular teniendo en cuenta que se la suele utilizar en sistemas de información o diálogo, donde suele haber una tarea asociada detrás (e.g., saber la hora de un vuelo); que la percepción del habla sintética interactúa con el conocimiento lingüístico de alto nivel, en particular con las curvas de entonación; que el habla sintética es más difícil de comprender que la natural, aunque se puede mejorar con la experiencia, pero con un límite que no permite llegar a la alcanzada con voces naturales; y, por último, vimos que la inteligibilidad y naturalidad no están disociadas, sino que ambas se ven afectadas por los atributos prosódicos.



Como presentamos en el [Capítulo 2](#), un sistema TTS debe generar una elocución a partir de un texto definido. De esta forma, como resume Campbell [20], dadas las palabras escritas debe pronunciarlas de manera *inteligible*, *natural* y *expresiva*.

El habla es *inteligible* si se comprenden todas las palabras dichas, es *natural* si es similar a como la pronunciaría un ser humano, mientras que se la considera *expresiva* si acompaña, refuerza y es consistente con el mensaje comunicado.

En la mayoría de los sistemas de generación de habla artificial actuales se alcanza una inteligibilidad a nivel de palabra próxima a la del habla natural. Sin embargo, un problema hasta ahora no resuelto satisfactoriamente es el de la naturalidad y expresividad del habla sintetizada [85, 180]. Para que alguien preste atención al habla artificial durante un tiempo prolongado es necesario que esta sea lo más natural posible dado que, de lo contrario, la persona perderá la concentración, se sentirá cansada y perderá gran parte de la información, aun cuando la voz sea inteligible.

Existen varias aplicaciones desafiantes dentro de los sistemas de diálogo hablado, como es el caso de los agentes conversacionales para asistir a adultos mayores o niños con trastornos de comunicación. En estos casos la expresividad es aún más necesaria, por cuanto se desea transmitir un mensaje con una intención específica, demostrar un estado de ánimo determinado o indicar un tipo de relación con el oyente (e.g., de amistad o de comprensión).

La falta de naturalidad y expresividad en los sistemas de diálogo automático ha restringido su empleo a ámbitos específicos como la lectura de mensajes escritos cortos, acceso por voz a información de servicios, y solicitud de reservas. Sin embargo, su uso aún no es aceptable, por ejemplo, para la lectura de libros o diálogos extensos [30, 82].

A los fines de este trabajo, la evaluación del habla artificial está orientada a determinar su *calidad*.

La *calidad del habla* es un concepto subjetivo asociado a la inteligibilidad, naturalidad, agradabilidad, entre otros aspectos, que poseen correlatos físicos que se evalúan en esta tesis. Así, la calidad es el resultado de la evaluación de todos los atributos reconocidos del habla en estudio, para determinar qué tan apropiada es para satisfacer las expectativas que se tienen sobre los valores de esos atributos. [98]

La evaluación de la calidad del habla es un proceso que requiere de la participación de un ser humano, que percibe y valora el habla

escuchada, por lo que se lo considera un proceso que es inherentemente subjetivo [65]. Sin embargo, se han desarrollado evaluaciones objetivas, o instrumentales, que permiten tener una aproximación a la evaluación humana [24, 31, 53, 140], y que presentan como ventajas principales el ser mucho más rápidas y tener un menor costo, aunque aún no se correlacionan totalmente con las evaluaciones subjetivas [53, 130].

Al no depender de un ser humano para realizar las pruebas, que en algunos casos requieren que decenas de personas evalúen múltiples oraciones de diferentes sistemas [20], las evaluaciones automáticas podrían utilizarse como entrada para algoritmos de aprendizaje automático (e.g., para utilizar la técnica de *aprendizaje por refuerzos* [198] u otras técnicas con base en el aprendizaje estadístico [78]).

En la figura 4.1 se pueden ver dos esquemas distintos de aprendizaje: sobre la izquierda se observa el ciclo habitual, donde personas realizan las evaluaciones y luego, a partir de ellas, se realizan los ajustes en los parámetros del sistema en forma manual, lo que en total podría tardar varios días en completarse; por otro lado, en el bucle que utiliza un método de evaluación automática, se genera una salida que luego es utilizada por un algoritmo de aprendizaje automático, con el que se podrían definir los ajustes en forma automática.

Lo anterior es congruente con lo que destacan Jurafsky y Martin en [101]<sup>1</sup>, donde indican que la definición de una métrica automática para la evaluación de la síntesis de habla resulta un tema de investigación abierto y fascinante.

#### 4.1 EVALUACIÓN DE CALIDAD DE UN SISTEMA TTS

Al momento de evaluar el habla artificial generada por un sistema TTS entre las metas se destacan:

1. comunicar el mensaje deseado (i.e., la secuencia de palabras de entrada), y
2. que el resultado se perciba de la manera más humana posible, de acuerdo a las necesidades y restricciones propias de cada dominio de aplicación.

Estas metas reciben el nombre de *inteligibilidad* y *naturalidad*, respectivamente [200]. Campbell [20] agrega una tercera meta, denominada *agradabilidad* (del inglés *likeability*), por la que se busca que la voz obtenida se adecue a las necesidades del ámbito de aplicación y a las expectativas de los usuarios. Así, se analiza si el habla es atractiva y presenta características de las voces humanas, lo que posteriormente determinará su aceptación y uso, como se registra en [111].

---

<sup>1</sup> pág. 280

## 4.2 TIPOS DE EVALUACIÓN

Las evaluaciones utilizadas para la calidad del habla se dividen en *subjetivas*, que requieren la participación de un humano, y *objetivas* —o *instrumentales*—, que se pueden realizar en forma automática.

A continuación se listan los distintos tipos de evaluaciones para determinar la calidad de un sistema:

- Evaluación explícita
  - Evaluación cuantitativa de un estímulo de un sistema, por ejemplo, por medio de puntajes de tipo Evaluación por categorías absolutas, del inglés *Absolute Category Rating (ACR)*
  - Comparación de a pares, donde se debe indicar cuál se escucha mejor de acuerdo a cierta medida, como, por ejemplo, la inteligibilidad de lo que se dice
- Evaluación implícita
  - Evaluación comportamental, donde se tiene en cuenta qué es lo que hace el sujeto con la información brindada por medio de la voz artificial
  - Evaluación del diálogo, de manera de detectar cambios en la voz del sujeto, que indiquen frustración u otras emociones, tanto positivas como negativas
  - Mediciones fisiológicas, como, por ejemplo, por medio de registros de electroencefalograma [5]

En los primeros años de la aplicación de los sistemas TTS las evaluaciones se focalizaron en las **evaluaciones subjetivas**, las que hoy en día se siguen realizando en forma general. En particular, se comenzó evaluando la inteligibilidad [20, 65], para lo que se pedía a varios sujetos que transcriban las secuencias de palabras escuchadas. Así, se utilizaron tests de rima (e.g., presentando palabras que sólo difieren en un fono al comienzo, como 'cima' y 'rima') o frases semánticamente impredecibles [14] (denominadas SUS, por el inglés *Semantically Unpredictable Sentences*), que siguen una estructura sintáctica válida pero no tienen sentido, como, por ejemplo, 'El avión pintaba los días de melón cocido'. La principal crítica a estas técnicas es que no se relacionan con el contexto de aplicación real de un sistema TTS [20], y por ende sus resultados no son del todo representativos.

Otra de las evaluaciones subjetivas es la calificación media de opinión (MOS, del inglés *Mean Opinion Score*) [212], donde se utiliza una escala de cinco valores para indicar la calidad general del habla escuchada: Excelente, Buena, Regular, Mediocre o Mala.

Este tipo de evaluaciones unidimensionales tiene la ventaja principal de ser muy simple de realizar [20, 65]. Sin embargo, para que la

información sea útil para mejorar distintos aspectos de un sistema hace falta tener más detalle, por medio de la utilización de varios criterios a la vez.

Así, existen métodos que piden que el sujeto evalúe varios aspectos del habla analizada, como la recomendación ITU P.85 [211]. En ella se evalúa el *esfuerzo de escucha* necesario para entender el mensaje, la *dificultad de comprensión* del mismo, así como su *nitidez y claridad*, y si *velocidad del habla* es adecuada, *agradable* y si es *aceptable* para ser utilizada. Esta recomendación no se ha aplicado extensivamente debido a que requiere un mayor esfuerzo, además de que las medidas que utiliza están fuertemente correlacionadas [217].

Las evaluaciones SUS, MOS y las definidas en la recomendación ITU P.85 pueden aplicarse en conjunto con buenos resultados [70, 192]. Según Sityaev, Knill y Burrows [192], al comparar al mismo tiempo la prueba ITU con las evaluaciones SUS, la primera no muestra el mismo poder discriminativo que las oraciones SUS, que presentan un contexto de prueba más estricto y útil para la evaluación de inteligibilidad. Al comparar con las pruebas MOS, se encuentran algunas diferencias entre la prueba ITU que usa una escala de cinco valores posibles, comparada con la escala estándar de MOS de diez valores, y, finalmente, se encuentran distintos resultados en la posición en la que quedan los sistemas evaluados en las dos pruebas. Finalmente, concluyen que el tipo de estructura prosódica de las oraciones utilizadas juega un rol importante, ya que pone a prueba los problemas ya existentes en los sistemas. Por ejemplo, para las oraciones cortas la agradabilidad de una voz puede terminar dominando los juicios de calidad, mientras que para las oraciones largas la velocidad del habla termina siendo el factor principal. Siguiendo con la evaluación ITU, Sluijter y col. [193] muestran que hay una correlación débil entre inteligibilidad y calidad, y lo mismo ocurre para aceptabilidad y preferencia, mientras que la aceptabilidad se relaciona con la calidad general, aunque la agradabilidad de una voz determina las respuestas sobre la preferencia de un sistema.

También hay técnicas que utilizan evaluaciones subjetivas con un enfoque multidimensional, como el escalamiento multidimensional (o MDS, del inglés *Multidimensional Scaling*) [130]. En ellas se realizan experimentos de percepción para conocer cuáles son los atributos del habla sintetizada —e.g., entonación, duración, uniones, presencia de plops y clicks— que mejor explican las diferencias en la *naturalidad* percibida por los sujetos en los segmentos de habla presentados. De esta manera, los segmentos de habla se ubican en un espacio n-dimensional que permite evidenciar cuáles son las características que tienen mayor significancia perceptual y deben ser tenidas en cuenta al momento de realizar una evaluación de calidad [97].

Las **evaluaciones objetivas** se ocupan de analizar la señal del habla para predecir en forma automática uno o más atributos de su cali-



dad (e.g., inteligibilidad). Inicialmente esto se logró comparando el habla sintetizada con la versión natural para la misma secuencia de palabras [24, 31, 140], lo que se denominó con la categoría de *métodos intrusivos* [65]. Sin embargo, estas comparaciones no son aplicables en forma general, debido a que no se tiene la versión *ideal* para todas las elocuciones posibles; sin contar que el habla es inherentemente variable y, por ende, dos segmentos de habla pueden decir exactamente lo mismo aun teniendo características acústicas diferentes [130]. En consecuencia, en los últimos trece años se han desarrollado métodos de evaluación objetiva *no-intrusivos* por medio de análisis de la envolvente [37, 38], mapas autoorganizados [155], modelos bayesianos [68], modelos de mezcla de gaussianas (GMM, del inglés *Gaussian Mixture Models*) [52], y otros atributos del habla (e.g., si suena robotizada, o si se detectan interrupciones) como se describe en la recomendación ITU-563 [214].

Entre los trabajos principales con evaluaciones objetivas podemos mencionar a los de Möller y colaboradores [53, 83, 86, 140, 149], que se enfocan en la obtención de atributos acústicos del habla, tanto sintetizada como natural, de forma de ubicarlos en un espacio y utilizar su distancia como medida de la calidad del habla relativa, tal como se había descrito en [128]. Adicionalmente, realizaron la evaluación de la calidad del habla a partir de atributos prosódicos [151], y de la predicción del contorno de entonación, por ejemplo, por medio del método de Fujisaki [84]. Más recientemente, el mismo equipo se dedicó a la utilización de redes neuronales de aprendizaje profundo para la evaluación de la calidad general del habla, a partir de datos evaluados por seres humanos [132-136].

En los últimos años se incorporaron la mediciones electrofisiológicas la caracterización y predicción de la calidad del habla [3-6, 165] por medio del uso de potenciales evocados (ERP, del inglés *Event-Related Potentials*) y la medición de la activación negativa por discordancia [141] (MMN, del inglés *Mismatch Negativity*). Posteriormente, debido a los resultados obtenidos, estos métodos se han utilizado en sistemas multimodales [7].

Por otra parte, en el área clínica también se utilizan métodos automáticos para la evaluación de la voz [9]. Así, se puede determinar su envejecimiento [58], su grado de disfonía [154], determinar si la voz suena aspirada o áspera [118], y detectar defectos propios de patologías en las cuerdas vocales [77], entre otros aspectos. Estas medidas podrían aplicarse en la evaluación de habla artificial para determinar su calidad, dado que al momento hay pocos trabajos relevantes al respecto.

En todos los trabajos indicados anteriormente se utilizan diferentes parametrizaciones [61] de la voz humana, a partir de las cuales se obtienen atributos que la caracterizan. Las principales técnicas utilizadas en actualidad se basan en el funcionamiento del oído humano,

como la de coeficientes cepstrales en escala Mel (MFCC, del inglés *Mel-Frequency Cepstrum Coefficients*) [36], y la predicción lineal perceptual (PLP, del inglés *Perceptual Linear Predictive*) [81]. Recientemente se han comenzado a utilizar otras parametrizaciones espectro-temporales, como las obtenidas por la aplicación de filtros de Gabor [49] o modelos jerárquicos [79], que se basan en resultados de la neurofisiología de la audición [27, 188].

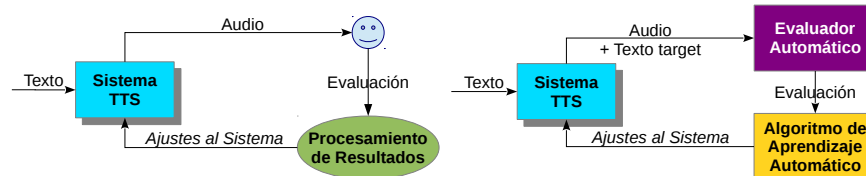


Figura 4.1: Ciclo de aprendizaje para un sistema TTS manual vs. uno con evaluación automática

#### 4.3 LIMITACIONES DE LAS PRUEBAS USADAS HABITUALMENTE

En general, las evaluaciones de las elocuciones generadas por sistemas TTS suelen evaluarse por medio de encuestas de opinión, por medio de la prueba MOS. Sin embargo, esta tiene el problema de que al hacer un promedio se pierde información, por lo que no termina siendo suficiente. Así, pierde de vista que hay gente que no está satisfecha con el sistema, más allá de que el promedio esté bien. Admite muchas curvas de calidad con el mismo valor de promedio, aunque eso puede mitigarse manteniendo la información de la distribución de las respuestas [89]. Más aún, teniendo en cuenta el punto de vista de alguien que está desarrollando una tecnología, para la cual debe tener el mayor detalle posible y no alcanza con tener una evaluación en promedio, sin considerar que una parte de los usuarios puede considerar inaceptable la calidad del sistema. Por lo tanto, es necesario poder contar con modelos multidimensionales para la evaluación de la calidad del habla, más allá de que luego se requiera armar un *ranking* entre distintos sistemas por medio de una sola medida general. Esto último se presentará como parte del trabajo de tesis en el [Capítulo 5](#).

#### 4.4 EVALUACIÓN DE SISTEMAS DE CONVERSIÓN DE TEXTO A HABLA

En [70] se resume la evaluación de tres sistemas de conversión de texto a habla en español con voces femeninas: el sistema *Aromo* (S<sub>1</sub>), descrito en la [Sección 2.4](#), el sistema *Nuance Vocalizer* (S<sub>2</sub>) y el sistema *AT&T Natural Voices* (S<sub>3</sub>). Todos los sistemas fueron creados utilizando concatenación por selección de unidades. Para el caso del S<sub>1</sub>, la voz es Emilia (español de Argentina), mientras para el caso el

S<sub>2</sub> se utilizó la voz Paulina (español mexicano) y para el S<sub>3</sub> se utilizó la voz Rosa (español latinoamericano), ya que las tres guardaban similitudes. Todos los audios generados se normalizaron en ganancia y frecuencia de muestreo (8Khz). Para evaluar las elocuciones se usaron evaluaciones MOS, SUS y lo indicado en la prueba ITU P.85, como se describió más arriba.

#### 4.4.1 *Sujetos*

Participaron veinte oyentes en el experimento, de entre 23 y 45 años. Diez de ellos eran especialistas en Fonoaudiología (de aquí en adelante, el grupo *SLT*), tanto profesionales experimentados como estudiantes avanzados de la Universidad de Buenos Aires, y los otros diez fueron no expertos sin problemas de audición ni neurológicos (de aquí en adelante, *non-SLT*).

#### 4.4.2 *Diseño del experimento*

Se le enviaron a los participantes las instrucciones y un documento para realizar las evaluaciones, así como los estímulos a evaluar —en el [Apéndice A](#) se encuentra la lista completa de oraciones utilizadas—. Luego, las personas debían enviar las respuestas de las tres pruebas para su posterior procesamiento automático.

#### *Prueba ITU*

Se crearon 27 elocuciones a partir de 9 diferentes textos, de entre 20 y 25 palabras de longitud, y con hasta 10 frases melódicas cortas. Los oyentes podían escuchar cada estímulo las veces que quisieran para hacer los juicios de inteligibilidad y calidad. Las oraciones utilizadas tenían una parte fija, específica de una tarea en particular, y una parte variable que cambia con la oración utilizada. Las tareas definidas tenían que ver con la venta telefónica, información de vuelos y el pago de servicios. Los sujetos tenían que escuchar 9 elocuciones por bloque, siendo 3 correspondientes a cada sistema en estudio, para un total de 18 estímulos por sujeto. En el primer bloque se realizó la evaluación de inteligibilidad (i), mientras que el segundo se trabajó con la evaluación de calidad (q). A continuación se detallan las evaluaciones solicitadas a los sujetos, y en cuál de los bloques aparecían:

**RECONOCIMIENTO DE PALABRAS (I)** Escriba el nombre, producto, características, código, precio y hora; empresa, país, número de vuelo, hora de salida, terminal y puerta de embarque; nombre, empresa, mes, monto y sucursal (el texto elegido depende de la tarea al cual corresponde el estímulo actual).

IMPRESIÓN GENERAL (Q) ¿Cómo es la calidad de lo que escuchó?: Excelente; Buena; Regular; Pobre; Mala.

ESFUERZO DE ESCUCHA (I) ¿Cómo describiría el esfuerzo realizado para entender el mensaje?: Ninguno; Bajo; Moderado; Alto; No se entendió el mensaje.

PROBLEMAS DE COMPRENSIÓN (I) ¿Hubo palabras difíciles de entender?: Nunca; Raramente; Ocasionalmente; Recurrentemente; Todo el tiempo.

ARTICULACIÓN (I) ¿Los sonidos se pueden distinguir?: Sí, claramente; Sí, suficientemente claro; Bastante claro; No muy claro; De ninguna manera.

PRONUNCIACIÓN (Q) ¿Notó anomalías en la pronunciación?: No; Sí, pero que no eran molestas; Sí, ligeramente molestas; Sí, molestas; Sí, muy irritantes.

VELOCIDAD DEL HABLA (Q) La velocidad del habla en promedio fue: Mucho más rápido de lo preferido; Más rápido de lo preferido; La preferida; Más lento de lo preferido; Mucho más lento de lo preferido.

AGRADABILIDAD DE LA VOZ (Q) ¿Cómo describiría la voz escuchada?: Muy agradable; Agradable; Suficientemente agradable; Desagradable; Muy desagradable.

ACEPTABILIDAD (I, Q) ¿Cree que esta voz se podría usar en un sistema de información por teléfono?: Sí; No.

#### *Prueba SUS*

Se utilizaron 50 textos de oraciones sin sentido semántico, aunque con estructura sintáctica correcta, de entre 6 y 10 palabras y con una o dos frases melódicas. Cada oyente recibió 45 oraciones seleccionadas en forma aleatoria, siendo 15 de cada sistema, sin repeticiones en los textos, y que sólo podían escuchar una vez. La instrucción dada a los sujetos fue "Escriba cada palabra que escuchó en la oración. No habrá posibilidad de escuchar la oración de nuevo, por lo que preste atención".

#### *Prueba MOS*

Se utilizaron 60 textos distintos, de entre 10 y 20 palabras de longitud, y con dos y tres frases melódicas. Los sujetos escucharon 15 elocuciones, 5 de cada sistema, con la posibilidad de que hubiera repeticiones en los textos entre un sistema y otro, pero sin repeticiones en los textos dentro del mismo sistema.

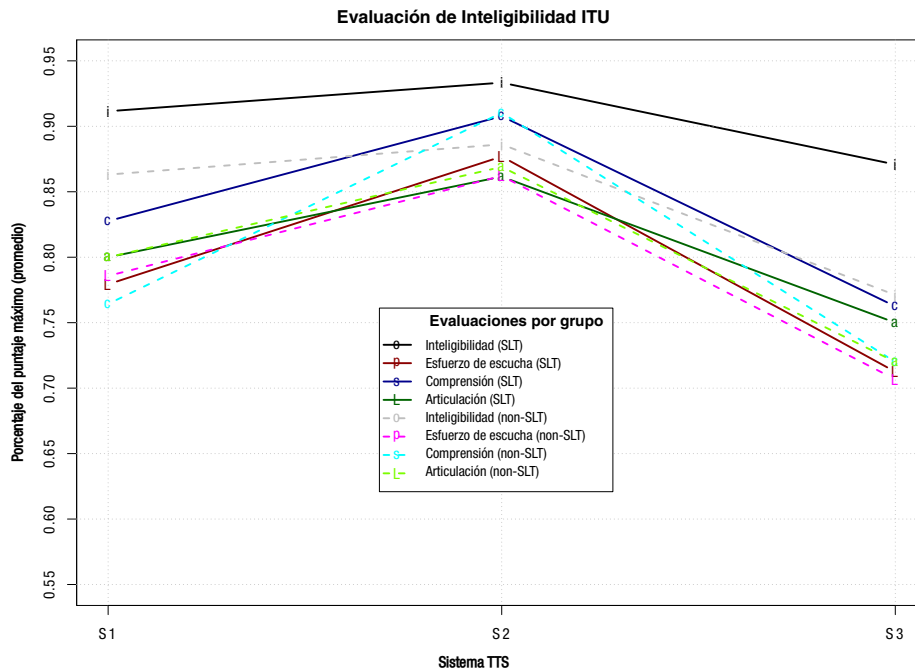


Figura 4.2: Promedio de las respuestas de la Inteligibilidad ITU, por sistema y grupo de oyentes

#### 4.4.3 Resultados

##### 4.4.3.1 Prueba ITU

Como se ve en la [Figura 4.2](#), la prueba de Inteligibilidad ITU muestra resultados similares para los tres sistemas, aunque el S3 presenta un peor desempeño para las personas no expertas (non-SLT). Por otro lado, de acuerdo a las evaluaciones de *Esfuerzo de escucha*, *Comprensión* y *Articulación* de ambos grupos, el orden de los sistemas queda en S2, luego el S1, y por último el S3.

En la prueba de Calidad ITU, como se aprecia en la [Figura 4.3](#), el S2 está por delante de los otros dos sistemas en todas las categorías. Adicionalmente, S1 y S3 alcanzan el mismo desempeño en la evaluación *General* y en *Pronunciación*, ambas detrás de S2. En lo respectivo a la *Velocidad del Habla*, S1 y S2 obtuvieron evaluaciones similares, y S3 se encuentra por debajo de ellos. Por otro lado, no se encontró acuerdo entre expertos y no expertos para la *Agradabilidad* de S1.

La *Aceptabilidad* de los sistemas aparecen en la [Figura 4.4](#). Así, S2 alcanzó un 83 % de aceptación, S1 un 62 %, y S3 llegó a un 49 %.

##### 4.4.3.2 Prueba SUS

Para la prueba SUS se tuvieron en cuenta el porcentaje de palabras correctas por oración, y también la cantidad de oraciones que fueron recordadas por completo por los sujetos, como se ve en la [Figura 4.5](#).

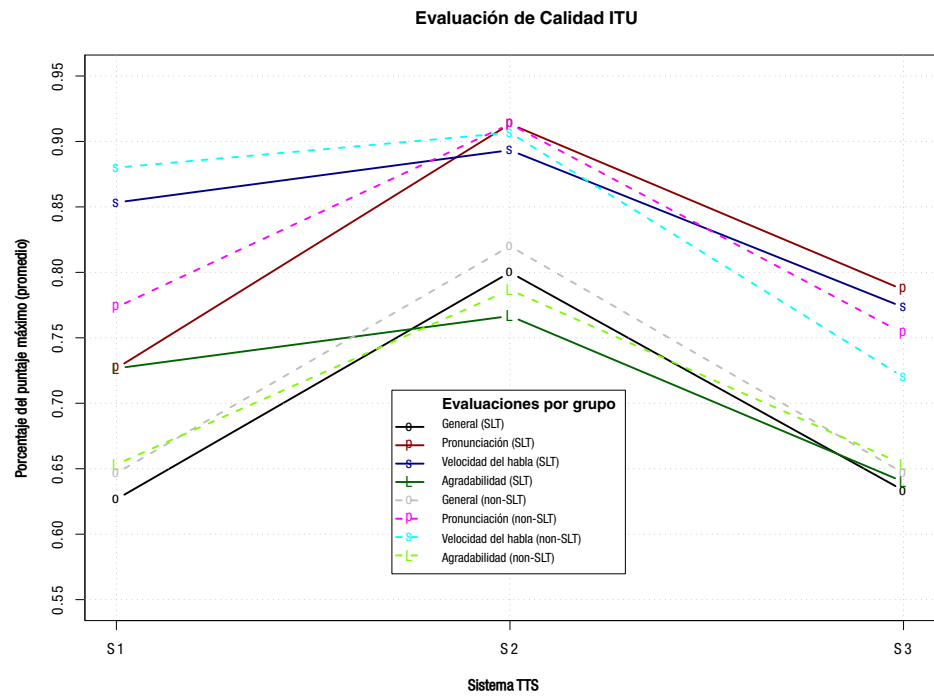


Figura 4.3: Promedio de las respuestas de la Calidad ITU, por sistema y grupo de oyentes

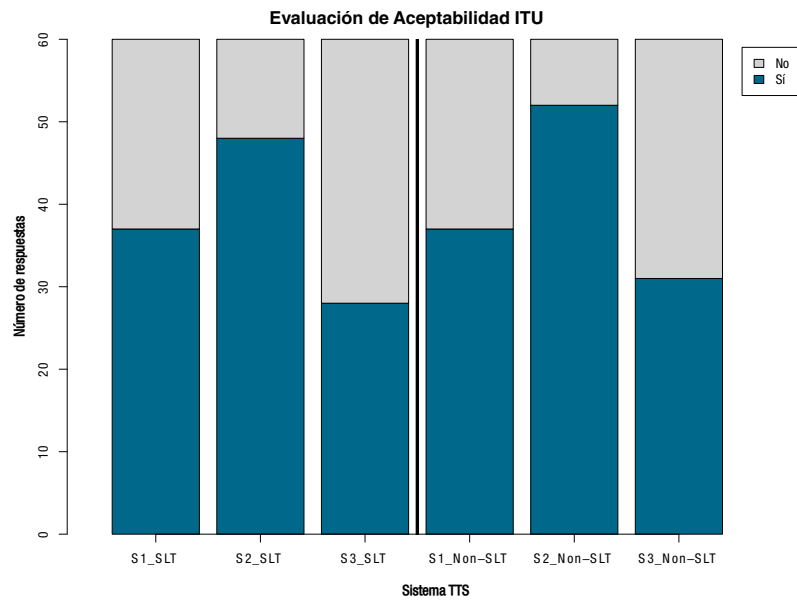


Figura 4.4: Promedio de las respuestas de la Aceptabilidad ITU, por sistema y grupo de oyentes

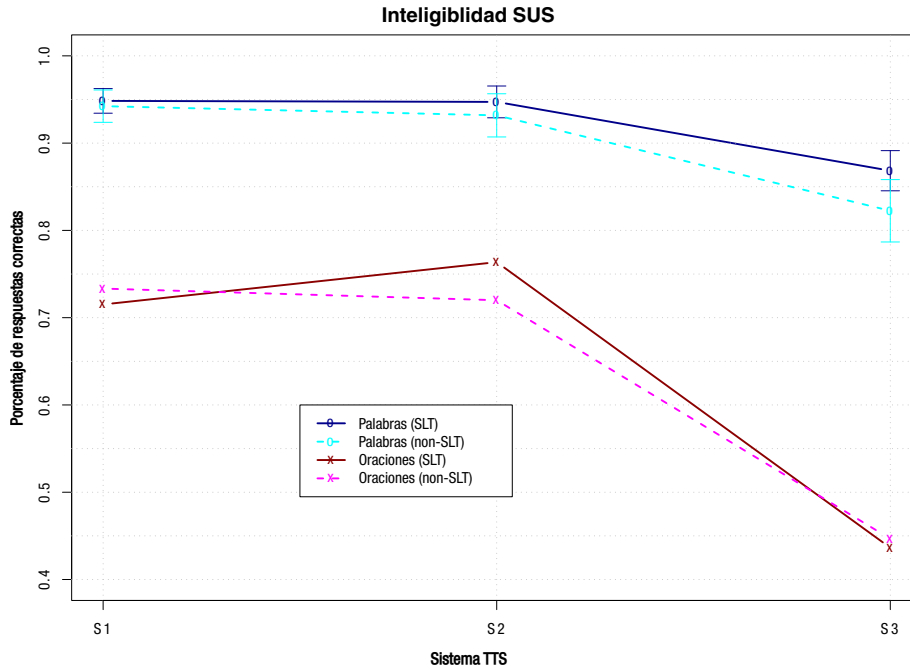


Figura 4.5: Promedio del % de palabras correctas por oración y % de oraciones correctas para la Prueba ITU

Así, en líneas generales, para los no expertos el desempeño de S1 fue mejor que el de S2 y el de S3, mientras que en los expertos se notó una igualdad entre S1 y S2.

En lo que respecta a las oraciones correctas, como puede verse en la Figura 4.5, S1 resultó mejor que S2 para los no expertos, mientras que S2 supera a S1 para los expertos. Sin embargo, ninguna de estas diferencias son significativas estadísticamente, usando una prueba de proporción de oraciones reconocidas entre par de sistemas (a dos colas, con 95 % de intervalo de confianza), aunque las diferencias sí son significativas entre ambos y el S3 (respectivamente,  $\chi^2 = 50.5457$ ,  $df = 1$ ,  $p < 0.00001$ ;  $\chi^2 = 58.0703$ ,  $df = 1$ ,  $p < 0.00001$ ).

El análisis de la varianza del porcentaje de oraciones correctas por cada sistema en la prueba SUS (usando una prueba F con 95 % de intervalo de confianza) muestra que hay diferencias significativas entre los tres sistemas (S1-S2,  $F = 0.5752$ , num  $df = 314$ , denom  $df = 314$ ,  $p \ll 0.001$ ; S1-S3,  $F = 0.3005$ , num  $df = 314$ , denom  $df = 314$ ,  $p \ll 0.001$ , S2-S3,  $F = 0.5225$ , num  $df = 314$ , denom  $df = 314$ ,  $p \ll 0.001$ ), y que S1 tiene la varianza más pequeña de los tres sistemas. No se encuentran diferencias significativas entre las medias de los sistemas 1 y 2, mientras que ambos difieren significativamente con respecto a S3 (respectivamente:  $t = 8.1123$ ,  $df = 487.093$ ,  $p < 0.00001$ ;  $t = 7.0721$ ,  $df = 571.753$ ,  $p < 0.00001$ ).

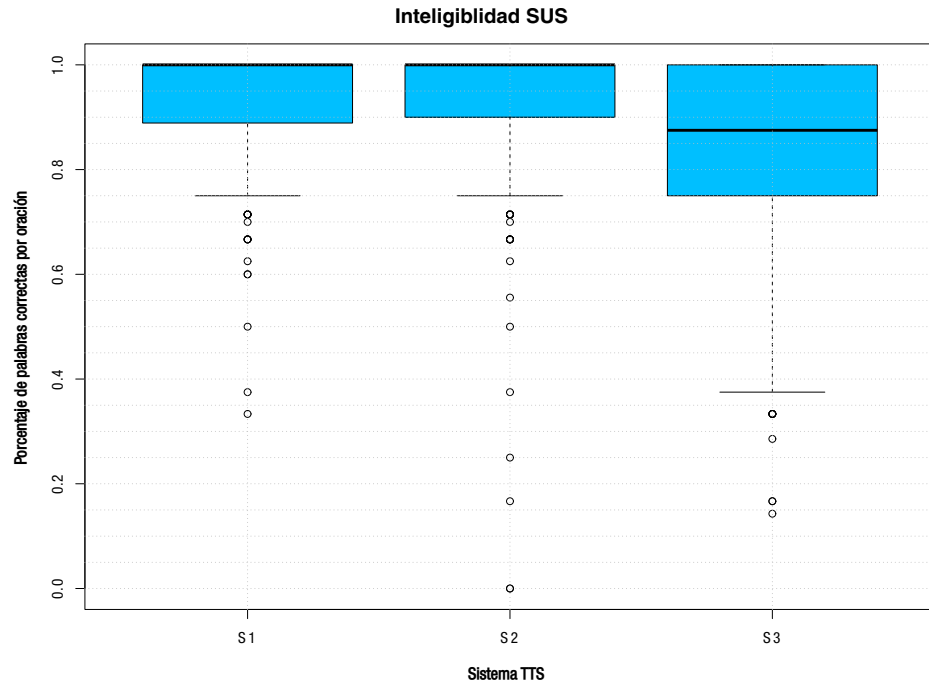


Figura 4.6: Palabras correctas en la evaluación SUS, para los 20 oyentes

#### 4.4.3.3 Prueba MOS

Como se aprecia en las Figuras 4.7 y 4.8, los resultados de la Prueba MOS de *Calidad General* muestran que el sistema S2 es el mejor para los no expertos, seguido por el S1, y luego el S3, mientras que para los oyentes expertos la evaluación para los sistemas 1 y 2 es similar. El análisis de la varianza indica que no hay diferencias significativas entre las evaluaciones para los tres sistemas, aunque no se encuentran diferencias significativas entre las medias de S1 y S2 (usando una prueba t de Welch de dos colas, con 95 % de intervalo de confianza), ambos sí se diferencian significativamente del S3 (respectivamente:  $t = 6.1381$ ,  $df = 207.015$ ,  $p < 0.00001$ ;  $t = 7.323$ ,  $df = 207.669$ ,  $p < 0.00001$ ).

#### 4.4.4 Análisis de los resultados

Los resultados de Inteligibilidad ITU muestran que S1 y S2 tienen un desempeño equivalente, mientras S2 está primero con 91 %, S1 sigue con 90 %, y S3 sólo llega al 82 %. La inteligibilidad obtenida por medio de la Prueba SUS muestra resultados similares, con S1 con un 95 %, S2 con 94 %, mientras que S3 llega a un 85 %. En ambos casos no se encontraron diferencias significativas entre los sistemas 1 y 2, tanto para expertos como para no expertos. Para el caso de la Prueba SUS sí se encontraron diferencias entre S1 y S2 con respecto a S3, como era de esperarse. Así, se encontró un acuerdo en ambas evaluaciones de inteligibilidad, aun cuando usan distintos tipos de oraciones. Adicio-



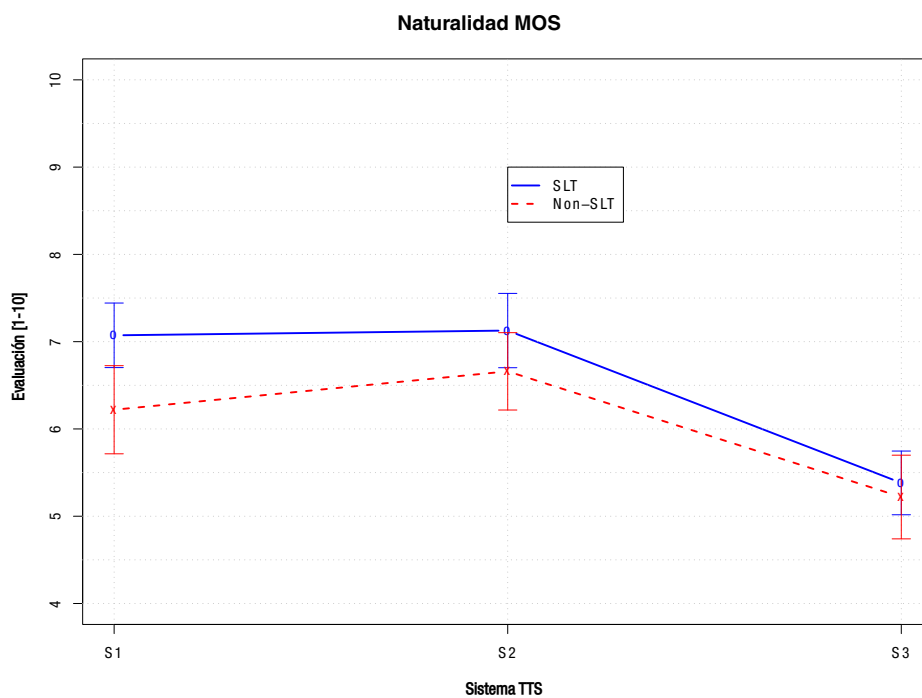


Figura 4.7: Promedio de respuestas de la evaluación MOS por grupo de oyentes

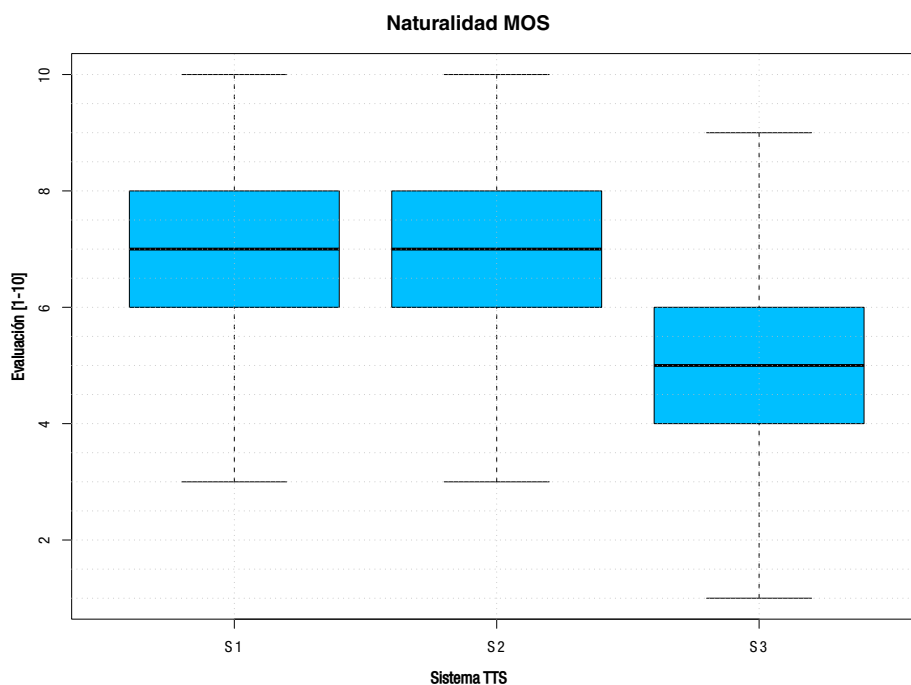


Figura 4.8: Evaluaciones MOS por sistema para los 20 oyentes

nalmente, se encontró una correlación alta (correlación de Pearson de dos colas, con 95 % de intervalo de confianza) entre los puntajes de *Esfuerzo de escucha* (E), *Comprensión* (C) y *Articulación* (A), con 0.768 ( $t = 14.2594$ ) entre E y C, 0.747 ( $t = 13.3538$ ) entre E y A, y 0.685 ( $t = 11.1785$ ) entre C y A ( $df = 141$ ,  $p \ll 0.001$ ), por lo que todas ellas podrían agruparse bajo el nombre *Pronunciación*.

También hay una alta correlación (correlación de Pearson de dos colas, con 95 % de intervalo de confianza) entre *Calidad General* y *Preferencia* en la prueba de Calidad ITU (0.699,  $t = 13.045$ ,  $df = 178$ ,  $p \ll 0.001$ ). Esto es aún más fuerte para los oyentes no expertos, lo que es razonable teniendo en cuenta que los fonoaudiólogos están entrenados para encontrar y analizar detalles finos en el habla.

Los resultados de calidad para la prueba MOS tomando una escala de 5 puntos ponen al sistema S2 como primero, seguido del S1 y, por último, el S3. Sin embargo, cuando se utilizan no se encuentran diferencias significativas entre S1 y S2, aunque las respuestas para S1 presentan menos varianza.

Se pueden explicar los resultados debido a que el Sistema 1 se creó con base en oraciones con pocos grupos melódicos, y el modelo de entonación está basado en esas oraciones, por lo que tiene problemas con oraciones más largas como las usadas en la prueba ITU. Esto deja claro que no alcanza con tener una buena cobertura fonética de la unidades a usar, sino que también hay que tener variedad en oraciones de distinta longitud y cantidad de frases, como se pone a prueba en la evaluación ITU. Así, se llegó a conclusión de que no se había podido modelar correctamente los contornos de entonación cortos que requerían las oraciones de la prueba ITU. De esta forma, la mala definición de las curvas generó una mala percepción general de parte de los oyentes.

Por otro lado, el Sistema 2 recibió mejores evaluaciones en todas las dimensiones. El S1 tuvo mejores resultados que S2 (usando una prueba t de Welch de dos colas, con 95 % de intervalo de confianza) en *Calidad General* y *Pronunciación* ( $t = -5.6318$ ,  $df = 114.739$ ,  $p \ll 0.0001$ ;  $t = -5.5743$ ,  $df = 108.474$ ,  $p \ll 0.0001$ ), pero no se encontraron diferencias significativas para *Velocidad del habla* y *Agradabilidad*. El S3 tuvo peores resultados que S1 y S2 en todas las dimensiones para *Velocidad del habla* y *Agradabilidad*. Los problemas de *Pronunciación* encontrados en S3 podrían deberse a la cantidad limitada de difonos disponibles o, directamente, a la falta de difonos para atender los nombres propios de origen extranjero.

Finalmente, encontramos que el sistema desarrollado (S1) obtuvo buenos resultados de inteligibilidad, tanto en las pruebas ITU como SUS. Más aún, si se miran aquellas oraciones con pocas frases melódicas, también se obtuvieron buenos resultados en las pruebas de calidad, comparables al de uno de los sistemas comerciales más reconocidos (S2). Finalmente, no se encontraron diferencias significativas

entre los juicios de expertos y no expertos para todas las dimensiones evaluadas.

Los resultados de S1 fueron peores en *Calidad* y *Pronunciación* para las tareas específicas incluidas en la Prueba ITU, donde se incluían varios grupos melódicos denotados por signos de puntuación —tanto por comas como por puntos—. Así, se cree que, para este caso en particular, se podrían haber obtenido mejores resultados de haber forzado el cierre de las frases utilizando los signos de puntuación como referencia, aunque esto no puede generalizarse para todo tipo de oraciones, donde no están tan marcadas las frases por medio de comas y puntos.

En síntesis, se encontró que para mejorar el desempeño del sistema deberían agregarse oraciones de variada longitud para el entrenamiento de los modelos de entonación. Por otro lado, en lo respectivo al problema con los nombres propios o que tiene difono no presentes en el español estándar, podría mejorarse si se agregan más oraciones con nombres propios de origen extranjero, pero son comunes en los países de habla hispana, aunque también podría resolverse por medio de la utilización de una base de datos multiidioma, de forma de cubrir todo tipo de nombres foráneos. Adicionalmente, para evitar artefactos propios de la concatenación de unidades pequeñas, podría utilizarse una base con unidades de diferente tamaño, desde fonos hasta trifonos, de forma de mejorar el resultado final.

#### 4.5 PROPUESTA DE EVALUACIÓN PERCEPTUAL DEL HABLA

Hay varios enfoques posibles para la evaluación subjetiva del habla sintetizada. El método más básico consiste en hacer una evaluación general de audios sintetizados, que pueden incluir o no a otros sistemas, por medio de ACR [212], y luego se obtiene la Nota media de opinión, del inglés *Mean Opinion Score* (MOS) [215], por cada estímulo/voz participante. Por otro lado, existen pruebas bien estructuradas, en las cuales se definen varias escalas, o dimensiones, las cuales deben ser evaluadas para cada estímulo y sistema, como se define en la recomendación ITU P85 [211] y su extensión para la evaluación de audiolibros [95].

En el [Capítulo 1](#) se habló de la importancia a nivel mundial del español, y, sin embargo, la poca presencia que tiene esta lengua en los sistemas de conversión de texto a habla más conocidos. Sin embargo, en los últimos años se hicieron varios proyectos de creaciones de sistemas TTS impulsados desde la academia, para el español de Argentina [205], de España [21, 59, 123], y Colombia [175], en otras variantes.

A partir del desarrollo de nuevas técnicas de síntesis, se espera que cada vez haya más sistemas multilingües [142, 237] o que se puedan adaptar rápidamente [26] para facilitar el desarrollo de voces para otras lenguas, además del soporte de variantes de las lenguas

existentes o la inclusión de lenguas o dialectos con pocos recursos disponibles [25].

De acuerdo a las investigaciones en lenguaje natural [63] además de en agentes virtuales y voces artificiales [110], el público prefiere, evalúa mejor y confía más en aquellas voces que se parecen más a la propia, lo que puede tener impacto en otros procesos donde media el lenguaje hablado, como el aprendizaje a través de entornos virtuales [29].

Las evaluaciones subjetivas comunes no contemplan la experiencia de los usuarios (QoE), ya que incluyen información del contexto y se realizan sólo en contexto de laboratorio. Así, se hace necesario un marco claro para el desarrollo de pruebas perceptuales [34, 223]. Sin embargo, para el caso de sistemas TTS de uso general no hay una definición clara de los escenarios de uso, de forma de poder diseñar procedimientos de evaluación más ecológicamente válidos. Con respecto a esto último, se sugiere ver el [Capítulo 2](#), donde se plantean en forma detallada estos posibles escenarios de uso.

En esta sección se describirá una propuesta de prueba perceptual para evaluar habla artificial, en especial para el Español de Buenos Aires, que fue diseñada pensando en sistemas TTS de uso general.

#### 4.5.1 El corpus *EVALPERCEP2023*

Como parte del trabajo de tesis se creó el *corpus* *EVALPERCEP2023* de habla en español de Buenos Aires, que posteriormente será publicado y puesto a disposición para la realización de pruebas perceptuales de calidad de sistemas de voces artificiales, incluyendo atributos acústicos de cada uno de los estímulos y las evaluaciones recibidas para cada una de las elocuciones.

Las oraciones se seleccionaron a partir de un listado de 133 oraciones en 6 diferentes categorías (sistemas de información, noticias, expresividad, preguntas, y comunicaciones personales), que extiende el diseñado en [70]. Se seleccionaron 24 oraciones de diferentes longitudes, con una o más frases, excluyendo aquellas que contenían palabras extranjeras (e. g., nombres propios de origen extranjero).

Las 24 oraciones seleccionadas corresponden a cuatro tipos, y hay 6 de cada categoría, de diferente longitud (ver [Cuadro 4.1](#)), y se seleccionaron cuatro más para entrenamiento (todas las oraciones se pueden ver en el [Apéndice C](#)):

- **Sistemas de Información (SI):** textos en el marco de un servicio de atención al cliente por voz, con información de un pedido de un producto que debe ser entregado, el recordatorio de la factura de un servicio que está impaga, y una oración de entrenamiento con el anuncio de información de un vuelo. Todas las oraciones incluyen expresiones vocativas, como “señor” y “señora”.

- Noticias (N): segmentos de noticias en medios de comunicación, donde la mitad de ellos son breves y la otra mitad son largos. Todos incluyen varias frases entonativas, verbos y contrucciones nominales complejas.
- Expresividad (E): textos que expresan varios estados de emociones del hablante, entre los que se incluyen oraciones inherentemente de enojo, otras inherentemente de alegría, y, por último, otras ambiguas (i. e., la misma oración podría ser evaluada como alegre o enojada, de acuerdo a cómo sea que sea la elocución finalmente generada).
- Oraciones Semánticamente Impredecibles (SUS): oraciones que son sintácticamente correctas para el español, pero que no tienen sentido en su interpretación semántica. Todas las oraciones tienen la estructura Sujeto-Verbo-Objeto.

Tipo	Mín.	Máx.	Prom.	Desv.Est.
Sistema de Información	126	179	157.0	20.9
Noticias (cortas/largas)	36/145	53/184	46.7/170.7	9.3/22.2
Expresividad	44	86	67.8	14.9
SUS	33	40	36.0	3.2

Cuadro 4.1: Longitud de las oraciones, en fonemas del alfabeto SAMPA [69]

Se generaron elocuciones para las oraciones indicadas previamente con catorce voces naturales y artificiales del español de Buenos Aires o diferentes variantes del español neutro:

- Dos voces naturales (ES-AR), una femenina (Lucía, AR-C) y otra masculina (Diego, AR-H), grabadas en un contexto de laboratorio, siguiendo las pautas definidas en [94]
- Sistema Aromo TTS, con su voz Emilia [Fem, ES-AR]
- Amazon Polly, con sus voces Lupe [Fem, ES-US] y Miguel [Masc, ES-US]
- Nuance Vocalizer, con sus voces Isabela [Fem, ES-AR] y Diego [Masc, ES-AR]
- Watson Text-to-speech system, con su voz SofíaV3 [Fem, ES-LA]
- AT&T Natural Voices, con sus voces Rosa [Fem, ES-US] y Alberto [Masc, ES-US]
- Microsoft Azure TTS con su voces Elena [Fem, ES-AR] y Tomás [Masc, ES-AR]

- Google Cloud TTS, con sus voces femenina (Wavenet A) y masculina (Wavenet B) [ES-US]

Para la creación del *corpus* de elocuciones y realización de las pruebas, las ondas se normalizaron en ganancia y muestreo y se convirtieron a formato mono, con 16khz de frecuencia de muestreo y 16 bits de profundidad.

#### 4.5.2 *Diseño propuesto*

##### 4.5.2.1 *Secuencia utilizada en la prueba*

La prueba propuesta consta de tres bloques, con 4 oraciones a modo de entrenamiento, 12 oraciones para el primer bloque de evaluación, y otras 12 oraciones para el segundo bloque, para un total de 28 evaluaciones por sujeto.

1. Bloque de Entrenamiento: En este bloque, los sujetos deben evaluar cuatro estímulos, presentados en orden aleatorio a partir de cada tipo de oración, en donde se les dan exactamente las mismas instrucciones que para los otros dos bloques, como se detalla más abajo. Se les indica que es sólo como entrenamiento, pero se les pide que respondan con cuidado. Los datos de este bloque no se guardan.
2. Bloque de Evaluación 1: Los sujetos deben evaluar 12 estímulos, tomados en forma aleatoria a partir de las 24 oraciones y sistemas disponibles, en las siguiente secciones:
  - a) Inteligibilidad:
 

Deben transcribir la oración que escucharon. La instrucción será “En un momento escuchará un audio que contiene un mensaje. Por favor, preste atención porque sólo se presentará una vez. Después de escuchar el audio, escriba lo que escuchó. Si no puede reconstruir la oración completa, trate de escribir cada palabra que recuerde en el orden en el que apareció.” Luego, debajo se mostrará el contexto de la oración antes de que se reproduzca el audio: “Va a escuchar esto porque. . .”

“compró un producto de un negocio en línea y está esperando la entrega, y está sonando el teléfono” (SI)

“está escuchando las noticias en la radio” (Noticias)

“está escuchando la lectura de una novela” (Expr.)

“está escuchando la lectura de un poema” (SUS)
  - b) Evaluación de la voz:
 

El sujeto vuelve a escuchar la misma oración, la cual puede ser reproducida todas las veces que quiera, y se le pide que evalúe la voz en dos grupos de escalas continuas, presentadas en orden aleatorio, por medio de selectores continuos

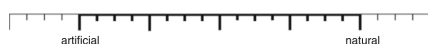


Figura 4.9: Ejemplo de la escala para la evaluación de *Naturalidad* [87]

como se describe en [109] and [87] (ver [Figura 4.9](#)). Estas escalas serán convertidas en números en el intervalo de 0 a 6, de acuerdo a la posición donde se hubiera ubicado el selector.

“¿Cómo evalúa la voz escuchada?”

- artificial vs. natural
- ruidosa vs. no ruidosa
- áspera vs. suave
- clara vs. no clara
- bien acentuada vs. mal acentuada
- con buen ritmo vs. con mal ritmo
- distorsionada vs. no distorsionada
- demasiado lenta vs. demasiado rápida

“En su opinión, la voz escuchada suena como de alguien. . .”

- masculino vs. femenino
- viejo vs. joven
- sano vs. enfermo
- formal vs. informal
- confiable vs. no confiable
- tenso vs. calmo
- activo vs. inactivo
- positivo vs. negativo
- de cuerpo pequeño vs. de cuerpo grande

Para el tercer grupo de evaluaciones el usuario tendrá que marcar con casillas de verificación los siguientes ítems, presentados en orden aleatorio:

“¿Escuchó algo de esto? Marque todas las que hayan ocurrido.”

- saltos en el sonido / discontinuidades
- clinks / jingles
- sonidos respirados
- sensación de voces múltiples
- otra clase de artefacto o ruido aleatorio

## c) Evaluación general:

“En resumen, ¿cuál es su evaluación general de la voz que escuchó?”

- Me gustó vs. No me gustó

“De todas las posibles aplicaciones de esta voz, ¿cuáles cree que serían las más adecuadas? Marque al menos dos opciones.”

- Ayuda para personas con problemas de visión
- Lectura de libros u otros textos largos
- Como narrador o voz en off de videos en YouTube u otras plataformas
- Sistemas de información por vía telefónica (e. g., para la venta de pasajes aéreos)
- Asistente personal en teléfono celular
- Asistente personal en un dispositivo externo (e. g., *Amazon Echo* o *Google Nest*)
- Asistentes de navegación (e. g., GPS)
- Voz para personas que no pueden o tienen dificultades para hablar
- Voz para un robot
- Voz para un personaje animado (e. g., dibujo animado)
- Otra (especificar)

“Por favor, deje otro comentario que tenga sobre la voz escuchada. Por ejemplo, algo que no hubiera estado incluido en las preguntas anteriores, pero que es relevante para usted.”

### 3. Bloque de Evaluación 2: idéntico al bloque 1, usando los 12 estímulos restantes.

#### 4.5.3 *Diseño de la evaluación de la prueba*

De forma de realizar la validación de la propuesta se diseñó un prueba comparativa con otras propuestas anteriores y la actual, donde sólo incluyen evaluaciones sin referencia:

- Recomendación ITU-T P.85 [94]
- Prueba AB modificada, utilizando varias escalas continuas para evaluar sistemas de a pares [190]
- La prueba propuesta en este trabajo



Posteriormente, a las personas participantes se les requerirá evaluar subjetivamente la prueba realizada de acuerdo las siguientes dimensiones, por medio a de escalas continuas con los valores extremos definidos:

- Cansancio generado por la prueba ('Bajo' a 'Muy alto'): se espera relevar si la persona se sintió cansada en la prueba, ya que el cansancio puede derivar en malos juicios acerca de las elocuciones, con el consiguiente impacto en la calidad de los resultados.
- Aburrimiento generado ('Ninguno' a 'Completo'): se le preguntará si la prueba le generó algún tipo de aburrimiento mientras la realizaba, ya que esto podría tener impacto en las respuestas recibidas, ya que los sujetos podrían responder de manera apurada para terminar la prueba o no prestar adecuada atención a los estímulos.
- Comodidad del diseño de la prueba ('Bajo' a 'Muy alto'): se buscará saber si el sujeto se sintió cómodo al realizar la prueba, ya que uno de los objetivos es generar un diseño que permita que la persona no se sienta incómoda, y pueda evaluar de una manera más cercana a como lo haría en un contexto más natural.
- Grado de adecuación de la prueba ('No adecuada' a 'Totalmente adecuada'): aquí se le pedirá su evaluación personal sobre qué tan adecuada cree que es la prueba para poder evaluar la calidad de voces artificiales, en tanto se buscar poder evaluar voces para un uso general, sin un ámbito de aplicación predefinido, entre otras cosas.

Además, los usuarios podrán dejar comentarios libres sobre la prueba realizada, que luego podrían tomarse en cuenta para hacer un análisis cualitativo de la percepción de cada prueba por parte de los sujetos experimentales.

#### 4.5.3.1 *Prueba AB Modificada*

Posteriormente a la prueba completa se realizará la prueba AB de comparación entre sistemas con elocuciones de los mismos textos.

Así, se realizarán comparaciones entre las 14 voces propuestas, para lo que existen 182 combinaciones posibles de a pares, contemplando el efecto del orden de presentación. A cada sujeto se le presentarán 4 estímulos de entrenamiento antes de comenzar, y luego recibirá en forma aleatoria 24 pares de estímulos de voces distintas, donde cada par tiene la misma secuencia de palabras, de forma de cubrir el corpus completos de textos.

#### 4.5.4 Prueba básica para la evaluación general del habla artificial

Como parte del trabajo de tesis se diseñó y realizó una prueba básica del habla artificial para el español, tomando como base el diseño descripto previamente y la lista de oraciones en la [Sección C.1](#). Las instrucciones dadas a los evaluadores se pueden ver en la [Sección C.2](#).

Así, las personas participantes debieron evaluar las siguientes aspectos del las oraciones presentadas del *corpus* EVALPERCEP2023:

- **Naturalidad:** qué tan natural parece la voz escuchada, en relación a la voz humana, entre (1) 'Artificial' y (5) 'Natural'.
- **Entonación:** evaluar si la voz escuchada tiene una buena entonación y usa bien las pausas al hablar, comparada con el habla humana común y de acuerdo texto de la oración, entre (1) 'Mala entonación' y (5) 'Buena entonación'.
- **Claridad:** qué tan bien pronunciado estaba el mensaje, entre (1) 'Poco claro' y (5) 'Muy claro'.
- **Velocidad:** evaluar la velocidad del mensaje, entre (1) 'Demasiado lento' y (5) 'Demasiado rápido'.
- **Defectos:** evaluar si se encontró algún defecto en el sonido, como ruidos extraños, chirridos, silbidos o interrupciones, entre (1) 'Ningún defecto' y (5) 'Muchos defectos'.
- **Familiaridad:** qué tan familiar le resulto la forma de hablar de la voz, desde (1) 'Nada familiar' a (5) 'Muy familiar'.

##### 4.5.4.1 Resultados

Nueve hablantes nativos de español de Buenos Aires, sin problemas de audición ni neurológicos, participaron de la prueba básica de evaluación de habla artificial: 7 hombres y 2 mujeres (edad media: 44.22 años, desvío estándar: 14.21 años), para un total 1322 juicios realizados para las 336 elocuciones generadas entre los 14 sistemas evaluados (6 textos  $\times$  4 grupos  $\times$  14 voces).

En el [Cuadro 4.2](#) se resumen las evaluaciones recibidas por cada una de las voces de la prueba, indicando el desvío estándar entre paréntesis, donde se marca en negrita la voz con mejor desempeño por cada una de las dimensiones.

Id	Nombre de voz	Sistema	Naturalidad	Entonacion	Claridad	Velocidad	Defectos	Familiaridad	GENERAL
4	Lucía	<i>natural</i>	4.80 (0.65)	<b>4.52 (0.77)</b>	4.68 (0.56)	3.14 (0.46)	1.43 (0.69)	<b>4.39 (1.10)</b>	4.41 (0.78)
5	Diego	<i>natural</i>	<b>4.86 (0.44)</b>	4.51 (0.68)	<b>4.75 (0.53)</b>	2.87 (0.40)	<b>1.32 (0.56)</b>	4.24 (0.99)	<b>4.56 (0.60)</b>
11	Emilia	Aromo	2.88 (1.31)	2.73 (1.11)	3.52 (1.10)	2.80 (0.48)	2.40 (1.21)	2.64 (1.22)	2.70 (1.16)
13	Lupe	Polly	2.61 (1.19)	3.20 (1.21)	3.88 (1.14)	<b>3.00 (0.46)</b>	1.85 (1.05)	2.50 (0.99)	2.84 (1.03)
14	Miguel	Polly	2.74 (1.11)	3.30 (1.09)	3.88 (1.06)	3.05 (0.40)	1.57 (0.89)	2.45 (0.98)	2.86 (0.99)
15	Isabela	Nuance Vocalizer	3.08 (1.17)	3.45 (1.07)	3.89 (1.07)	3.08 (0.34)	1.94 (0.98)	3.08 (1.25)	3.26 (1.02)
16	Diego	Nuance Vocalizer	2.60 (1.27)	2.84 (1.21)	3.46 (1.21)	3.19 (0.54)	2.46 (1.34)	2.51 (1.17)	2.69 (1.16)
17	Sofía	Watson	2.73 (1.14)	3.15 (1.17)	3.91 (1.00)	2.81 (0.42)	1.57 (0.97)	2.19 (0.92)	2.86 (0.93)
18	Rosa	AT&T Natural Voices	1.84 (0.87)	2.32 (0.98)	3.17 (1.07)	3.12 (0.67)	2.44 (1.25)	1.76 (0.70)	1.96 (0.85)
19	Alberto	AT&T Natural Voices	2.14 (1.22)	2.12 (1.05)	2.86 (1.09)	3.02 (0.53)	2.43 (1.34)	1.77 (0.84)	1.88 (0.87)
24	Elena	MS Azure	4.56 (0.69)	4.44 (0.82)	4.68 (0.59)	3.10 (0.44)	1.35 (0.76)	4.20 (1.07)	4.35 (0.79)
25	Tomás	MS Azure	4.18 (1.03)	4.04 (0.91)	4.57 (0.69)	3.14 (0.40)	1.39 (0.84)	3.85 (1.14)	4.09 (0.89)
26	Wavenet-A	Google Cloud TTS	3.23 (1.33)	3.76 (1.16)	4.36 (0.87)	3.48 (0.62)	1.58 (0.89)	2.91 (1.03)	3.46 (1.07)
27	Wavenet-B	Google Cloud TTS	3.85 (1.12)	4.15 (0.99)	4.47 (0.72)	3.16 (0.43)	1.52 (0.90)	3.23 (1.12)	3.99 (0.79)

Cuadro 4.2: Evaluaciones promedio recibidas por cada voz y sistema para cada dimensión, con su desvío estándar

En líneas generales, todas las voces de la prueba tuvieron buenas evaluaciones en todas las dimensiones. Como podría esperarse, se destacaron las dos voces naturales, aunque el desempeño de las voces 24 y 25 estuvo muy cerca, se estima, porque, pese a que había varios sistemas para el español de Buenos Aires, en este caso las voces generadas tenían alta calidad, con buena entonación y sin defectos de algún tipo.

#### 4.6 CONCLUSIONES

En este capítulo se resumieron los distintos tipos de pruebas de calidad que pueden aplicarse a un sistema de habla artificial. Adicionalmente, se mostraron los resultados de una prueba de evaluación de tres sistemas de conversión de texto a habla para español latinoamericano, entre los que está el sistema Aromo, en el que se trabajó como parte del doctorado. Así, se encontró que es importante que un sistema esté entrenado con oraciones de distinta longitud, y de varias frases melódicas, de forma de poder responder adecuadamente a los requerimientos de las estructuras de las oraciones utilizadas en los sistemas de información por voz.

Se encontró que las pruebas disponibles actualmente tienen dimensiones que correlacionan entre sí, además de que no prueban adecuadamente los sistemas de conversión de texto a habla de uso general, donde no se tiene preestablecido un dominio de aplicación. Así, se propuso una prueba perceptual completa, que incluye varias partes en las cuales se evalúan distintos aspectos del habla artificial, considerando nuevas dimensiones, como las relacionadas a atributos paralingüísticos —e.g., confiabilidad de la voz—. La realización de la prueba final, y comparación con otras pruebas de uso extensivo como las ITU, quedará como trabajo futuro.

Se presentó el diseño del *corpus* EVALPERCEP2023, creado especialmente para la evaluación de habla artificial del español de Buenos Aires. Para esto se realizó una prueba perceptual inicial, que consistió de varias dimensiones perceptuales simples: naturalidad, entonación, claridad, velocidad, defectos, familiaridad y evaluación general. Este corpus y los juicios obtenidos en la prueba básica serán utilizados posteriormente para la creación de un modelo para la predicción automática de la calidad del habla artificial, como se describe en el [Capítulo 5](#).

## EVALUACIÓN AUTOMÁTICA DE LA CALIDAD DEL HABLA

---

En el [Capítulo 4](#) se presentó el problema de la evaluación de la calidad de habla de un sistema TTS, y se mencionó que los sistemas suelen evaluarse por medio de pruebas perceptuales donde se le presentan elocuciones generadas por uno o más sistemas a diferentes sujetos; estos deben indicar su grado de preferencia y, en algunos casos, como en el diseño propuesto por la prueba recomendación ITU P.85 [\[211\]](#), se incluyen varios tipos de preguntas por cada una de las unidades de audio que se presentan. Adicionalmente, se mencionaron otras propuestas para evaluación objetiva del habla —también llamada *instrumental*—, entre las que se destacan Möller y colaboradores [\[53, 83, 86, 140, 149\]](#).

En este capítulo se presentarán atributos acústicos del habla y predicciones obtenidas por medio de otros sistemas que permitirán obtener una estimación de distintas dimensiones de la evaluación de calidad de habla de un sistema, así como la definición de una métrica general.

### 5.1 CONTEXTO ACTUAL DE LA EVALUACIÓN AUTOMÁTICA

Muy recientemente el área de la evaluación automática de la calidad del habla artificial tuvo un resurgimiento, materializado en el desafío VoiceMOS [\[91\]](#), que se realizó como parte de uno de los congresos más importantes del área de procesamiento automático del habla. Como se resume en las motivaciones del desafío, aun con recursos para hacer *crowdsourcing* para que mucha gente evalúe las elocuciones generadas por los sistemas, la tarea sigue siendo costosa y requiere mucho tiempo. Adicionalmente, los datos obtenidos por medio de estas plataformas no son totalmente confiables, además de que no hay un control preciso de cómo se realizan las pruebas, por lo que deben tomarse medidas especiales para poder usar esos datos, como se describe en [\[100, 143, 233, 234\]](#), entre otros trabajos.

Como parte de este desafío se presentó un *corpus* de oraciones generadas por sistemas TTS y de conversión de voz, descrito en [\[33\]](#) y analizado en [\[28\]](#), con sus respectivas evaluaciones perceptuales, las cuales se originaron en otros desafíos, como las distintas ediciones del *Blizzard Challenge (BC)* y del *Voice Conversion Challenge (VCC)*. En una de las partes del desafío se presentan elocuciones sólo en idioma inglés, mientras en la otra parte se debía evaluar elocuciones en otras lenguas, para lo cual, finalmente, se eligió el chino. Así, los modelos

buscados estaban orientados a ser multilingües o que pudieran ser adaptados a otras lenguas sin problemas.

Todos los sistemas que participaron existosamente del desafío utilizaron redes neuronales profundas, en algunos casos usando aprendizaje supervisado [147, 177, 195, 210] y autosupervisado [76, 92, 127, 203, 227], entre otros enfoques.

Adicionalmente a la predicción de la calidad general del habla (o MOS), también hay trabajos que utilizan las redes neuronales profundas para la evaluación de la inteligibilidad [235], naturalidad [134], ambas a la vez [153], así como otras dimensiones de la señal de sonido como el ruido, coloración, discontinuidades y sonoridad [133, 136].

## 5.2 CARACTERÍSTICAS PARA LA EVALUACIÓN AUTOMÁTICA DE LA CALIDAD DEL HABLA

Como parte de esta tesis se propone un conjunto de características para la descripción del habla artificial para su posterior utilización para la predicción de distintas dimensiones asociadas a la calidad.

Debido a que la evaluación que se quiere hacer no está orientada a un tipo o contexto de uso en particular, las características deberán ser lo suficientemente generales para poder analizarse para diferentes tipos de elocuciones (por ejemplo, oraciones largas o cortas). Como parte de esta selección se optó por crear caracterizaciones nuevas, extender otras existentes y aprovechar medidas generadas en el estado del arte.

### *Naturalidad*

Para la naturalidad se optó por utilizar la evaluación propuesta como parte del modelo NISQA-TTS desarrollado por Mittag y Möller [134], que utiliza redes neuronales profundas y, como en el *VoiceMOS Challenge* mencionado anteriormente, está entrenado usando los datos del *Blizzard Challenge* y del *Voice Conversion Challenge*.

Así, se aplicará el modelo de NISQA-TTS para obtener un valor de naturalidad para la elocución que está siendo evaluada.

### *Inteligibilidad*

Continuando la idea presentada por Peiró-Lilja y col. [153], se utilizará un sistema de Reconocimiento Automático de Habla (RAH) para la evaluación de la inteligibilidad del habla. Los autores proponen realizar el reconocimiento por medio de sistemas que no utilizan modelos del lenguaje, en tanto, aducen, eso podría ser un mejor indicador de la inteligibilidad de la elocución.

Por el contrario, creemos que el modelo del lenguaje no puede escindirse del oyente, ya que el modelo opera en el reconocimiento

de qué es lo que dice la elocución. Por consiguiente, la utilización de reconocedores que estén orientados a reconocer *tokens* más pequeños, sin intervención de un modelo de lenguaje, creemos que no es concordante con el reconocimiento humano de las palabras contenidas en la elocución.

Así, para poder evaluar la inteligibilidad se realiza el reconocimiento de habla por medio de la herramienta *OpenAI Whisper*<sup>1</sup>[168], usando los dos modelos más pequeños multilingüaje, llamados *tiny* y *base*. De esta forma, al contar con los textos objetivos de cada elocución, se puede calcular la distancia de edición de Levenshtein [120]. Una vez obtenida esta distancia, el valor se normaliza por la cantidad de caracteres del texto original. Previamente al cálculo de la distancia se realiza una normalización de la oración original y el texto reconocido por el sistema de RAH, donde ambos textos se convierten a mayúscula y se eliminan signos de puntuación. Se eligió este sistema RAH por sobre otros porque es de código abierto y, entre otras ventajas más, presenta diferentes modelos preentrenados y de diferente tamaños, además de permitir el trabajo multilingüe (de hecho, según sus propios datos [168], el desempeño del reconocedor es el más alto para el español).

#### *Manejo de pausas*

Tomando como punto de partida la salida del sistema de RAH, se utiliza esta información para identificar en qué lugares se marcaron pausas por medio de signos de puntuación. Así, luego de normalizar la oración original y el texto reconocido, reemplazando cada palabra entera por un carácter, lo que permite identificar la posición del signo de puntuación dentro de la oración, se realizará el cálculo de la distancia de Levenshtein, y luego este número se normalizará por la cantidad total de signos de puntuación en la oración. Todas las oraciones originales terminan con punto, y se puede agregar uno para quienes no lo tienen, por lo que no hay riesgo de tener una división por cero.

#### *Degradación de la señal de habla*

Siguiendo lo propuesto por Mittag y col. [133, 136], se utilizarán las evaluaciones de la calidad de la señal incluidos en el modelo NISQA, ya que, más allá de que las elocuciones a evaluar no están en el contexto del canal telefónico u otro con ancho de banda acotado o con compresión, estos atributos nos servirán como criterios generales para evaluar la calidad de la señal de habla.

---

<sup>1</sup> <https://github.com/openai/whisper>

- **Ruido:** degradación de la señal originada en ruido de fondo, o generado por circuitos o codificación de la señal.
- **Coloracion:** degradación de la señal causada por distorsiones en la respuesta de frecuencias, por ejemplo, introducidas por limitaciones de ancho de banda, codificación de baja cantidad de bits, o por el manejo de la pérdida de paquetes.
- **Discontinuidades:** distorsiones aisladas o no estacionarias que se introducen por la pérdida de paquetes por recorte de la señal.
- **Sonoridad:** defectos en el habla originados en variaciones en el nivel del sonido, tanto más bajo como más alto de lo esperado.

#### *Calidad general del habla*

Se incluirá entre los atributos a las evaluaciones generales de calidad del habla (MOS) obtenidas con los tres modelos disponibles en NISQA:

- **MOS del modelo NISQA:** obtenido a partir de habla natural con degradación.
- **MOS del modelo NISQA-MOSonly:** obtenido a partir de habla natural con degradación, de un modelo que sólo hace la estimación de MOS, y genera las mismas estimaciones que el modelo anterior.
- **MOS del modelo NISQA-TTS:** obtenido a partir de habla sintética y evaluaciones perceptuales.

#### *Estado del hablante*

Como se comentó en los capítulos 3 y 4, la agradabilidad de una voz artificial se ve afectada por atributos paralingüísticos, emociones, atributos de personalidad y estado interno aparente del hablante (e. g., ver [111]). Así, se utilizó la herramienta openSMILE<sup>2</sup> [47] para obtener varios grupos de atributos que se utilizaron como parte de competencias de detección de emociones, y atributos paralingüísticos, entre otras cosas, que fueron parte de Interspeech:

- 2009 Emotion Challenge (IS09\_emotion) [182]
- 2010 Paralinguistic Challenge (IS10\_paraling) [184]
- 2011 Speaker State Challenge (IS11\_speaker\_state) [185]
- 2012 Speaker Trait Challenge (IS12\_speaker\_trait) [186]
- 2013 ComParE (IS13\_ComParE) [187]

<sup>2</sup> <https://audeering.github.io/opensmile>



En todos los casos anteriores se tiene la misma cantidad de características sin importar la duración de cada audio, debido a que se trata de valores estadísticos (e. g., máximo, mínimo, posición de máximo, etc.). Esto es algo relevante, en tanto no se sabe, a priori, los tipos de oraciones que podrían utilizarse en diferentes contextos de prueba, por lo que se decidió usar una cantidad de datos variable, sin que haya que hacer recortes en segmentos de audio de tamaños acotados, ya que se pierde parte de la visión general de cada elocución.

### 5.3 CREACIÓN DE LOS MODELOS Y PRUEBA

Dados los atributos definidos anteriormente, se tomaron los datos recolectados en la prueba perceptual básica (descrita en la [Subsección 4.5.4](#)) y se les aparearon los atributos para cada uno de los estímulos, de acuerdo a como se describe más abajo. De esos datos, se quitaron los correspondientes a las elocuciones de dos de las voces, las cuales fueron reservadas para la validación, como se describe más abajo.

Para la prueba de los modelos se realizó validación cruzada de cinco partes (*5-fold CV*), utilizando los siguientes modelos de regresión:

- Random Forest (RF): regresión con 100 estimadores
- SVM con kernel radial (SVR-RBF)
- SVM con kernel polinomial (SVR-Poly)

Así, se definieron los siguientes conjuntos de datos para realizar validación. Donde se indica 'all\_features' se incluyen las nueve dimensiones definidas más arriba por cada elocución, y lo que varía es qué valores de las siete dimensiones que surgen de los juicios realizados por los evaluadores (i. e., naturalidad, entonación, claridad, velocidad, defectos, familiaridad y evaluación general).

- allfeatures\_allevs: Un registro por cada evaluación, con todas las características por elocución (i. e., se repiten por cada juicio realizado).
- allfeatures\_meanestimulo: Un registro por elocución, con todas sus características, con el promedio de evaluación por cada una de las siete dimensiones por elocución.
- allfeatures\_meanestimulo\_allevs : Un registro por evaluación, con todas sus características y evaluaciones recibidas, más un registro por elocución, todas las características y las evaluaciones promedio por elocución.
- allfeatures\_meanvoice: Un registro por elocución, con las características de cada una y el promedio de las evaluaciones por voz.

Dimensión	$R^2$ prom.	$R^2$ desv.	NMSE prom	NMSE desv.
General	0.67	0.17	-0.54	0.28
Naturalidad	0.66	0.17	-0.7	0.35
Familiaridad	0.62	0.19	-0.71	0.36
Entonacion	0.62	0.19	-0.62	0.32
Claridad	0.54	0.23	-0.57	0.3
Defectos	0.44	0.27	-0.67	0.34
Velocidad	0.42	0.29	-0.13	0.07

Cuadro 5.1:  $R^2$  y NMSE por cada dimensión usando regresión R.Forest con atributos básicos

- `allfeatures_meanvoice_meanestimulo_allevs`: Un registro por cada evaluación, más dos registros por cada elocución, uno el promedio de las evaluaciones por voz y otro con el promedio de evaluaciones por elocución, en todos los casos con las nueve dimensiones de evaluación.
- `is09_emotion_allevs`: Un registro por cada evaluación, con 384 características acústicas de cada elocución más las siete dimensiones de evaluación.
- `is10_paraling_allevs`: Un registro por cada evaluación, con 1581 características acústicas de cada elocución más las siete dimensiones de evaluación.
- `is11_speakerstate_allevs`: Un registro por cada evaluación, con 4368 características acústicas de cada elocución más las siete dimensiones de evaluación.
- `is12_speakertrait_allevs`: Un registro por cada evaluación, con 5757 características acústicas de cada elocución más las siete dimensiones de evaluación.
- `is13_ComParE_allevs`: Un registro por cada evaluación, con 6373 características acústicas de cada elocución más las siete dimensiones de evaluación.

## 5.4 RESULTADOS Y DISCUSIÓN

### 5.4.1 *Desempeño por dimensión de evaluación*

En esta sección se hará un resumen de los resultados por cada una de las dimensiones de las evaluaciones definidas anteriormente, así como por cada técnica de regresión y juego de datos. A partir de estos datos se elegirá un *set* de características y predictores que expliquen

mejor la variación en todas las dimensiones, para que luego se pueda integrar estos modelos para la predicción de la evaluación general percibida de cada sistema.

Se utilizaron dos formas de evaluar cada set-modelo, con  $R^2$  y el promedio del error cuadrado negativo (NMSE, de inglés *Negative Mean Squared Error*). En la segunda medida se usa la versión negativa en tanto que un valor más alto indica un mejor resultado.

En el Cuadro 5.1 se puede ver el desempeño de un regresión con *Random Forests (RF)* sólo utilizando las dimensiones definidas en este trabajo de tesis. Así, se calculó el promedio del  $R^2$  y NMSE por cada dimensión, con sus respectivos desvíos. Allí se puede ver que existe una buena correlación y bajo error para la mayor parte de las dimensiones a predecir, por lo que se aprecia que con las pocas dimensiones definidas se puede obtener buenos resultados aún para voces que nunca fueron vistas. Posteriormente, será necesario realizar otros análisis incluyendo más información de acuerdo a los juegos de datos definidos previamente, como se realiza a continuación.

Así, se encuentran algunos cuadros con la evaluación de los diferentes grupos de atributos para la predicción de cada una de las dimensiones de evaluación del habla, los cuales se generaron por medio de una validación cruzada de cinco partes, y el valor presentado en cada fila es el promedio, con su respectivo desvío estándar.

Así, se tienen a disposición el desempeño de cada set-modelo, y en negrita se marca el mejor resultado según cada una de las dos medidas, tanto para el modelo con las nueve dimensiones de evaluación como para los set de datos de información de emociones y paralingüística. En este último caso sólo se armaron modelos con los atributos de las competencias, para analizar el aporte *per se* de cada uno.

Como se puede ver en el Cuadro 5.2, en la dimensión 'Claridad' puede verse que el mejor desempeño se lo obtuvo con el modelo que utilizó la información de todas las evaluaciones y el promedio de cada una de las elocuciones con una regresión usando *Random Forests (RF)*. Por otro lado, el menor error se alcanzó con el conjunto de datos que sólo tiene los atributos de cada elocución y el promedio obtenido por cada cada voz en las dimensiones de evaluación, usando SVM con kernel polinomial de grado tres.

Para la dimensión 'Defectos' el mejor resultado también se obtuvo con el juego de datos `allfeatures_meanestimulo_allevals` y la regresión con RF, mientras que el error mínimo se obtuvo con con el promedio de evaluaciones por cada voz. Más allá de que el valor de  $R^2$  no es alto, es importante destacar que se eligió una cantidad acotada de características del habla artificial, presentadas más arriba, por lo que es importante ver qué parte de los valores se pueden explicar con este modelo simple, como se aprecia en el Cuadro 5.3.

Para el caso de la 'Entonación' (ver Cuadro 5.4), nuevamente el modelo ganador `allfeatures_meanestimulo_allevals-RF`, y pese a

su simpleza, tiene un buen valor de  $R^2$ , teniendo en cuenta que las evaluaciones de la entonación y la prosodia en general son difíciles de resolver de manera completa. Por otro lado, el menor error se obtuvo también con los datos que tienen el promedio de evaluaciones por cada elocución y usan el regresor SVM polinomial.

En lo respectivo a la 'Familiaridad' se observa el mejor resultado con el juego de datos `allfeatures_meanvoice_meanestimulo_alleva1s` usando *Random Forests*, como se aprecia en el Cuadro 5.5. En este caso, el mejor desempeño con los errores también se alcanzan con este modelo. Además, en este tipo de dimensiones se aprecia un mejor desempeño de los juegos de datos relacionados con paralingüística y características del hablante, lo que tiene sentido, ya que la familiaridad está asociada a este tipo de información.

Viendo el Cuadro 5.6 se aprecia que el modelo `allfeatures_meanvoice_meanestimulo_alleva1s-RF` tiene el mejor resultado para la dimensión 'Naturalidad', en ambas métricas de evaluación, por lo que este modelo se presenta como el más informativo por cada dimensión.

En lo respectivo a la dimensión 'Velocidad', en general se obtuvieron no muy buenos resultados en la medida  $R^2$ , aunque el error es bajo. Esto último puede deberse a que, para este caso, el mejor valor posible está en la mitad del puntaje (esto es, en 3), y la mayor parte de los sistemas manejan adecuadamente la velocidad, por lo que la distribución de valores para todos los sistemas está alrededor del promedio. Ver Cuadro 5.6 para más detalles de los resultados.

Por último, en lo respectivo a la evaluación general, el modelo que mejor desempeño tiene es `allfeatures_meanestimulo_alleva1s-RF`, seguido de cerca por el modelo que incorpora la información promedio por cada voz. En lo respectivo a los errores, este último modelo obtiene los mejores resultados.

De acuerdo a los resultados mencionados anteriormente, se armó un resumen (ver sección Cuadro 5.10) donde se indica cuál es el modelo-método que obtiene los mejores resultados. Así, si consta un  $R^2$  o una E, esto indica que esa fila es del modelo que tienen mejor resultado en esa dimensión de evaluación.

#### 5.4.2 Modelos elegidos para la predicción de la calidad general

De acuerdo a los resultados obtenidos anteriormente, resumidos en el Cuadro 5.9, se seleccionaron aquellos modelos que registraron mejores resultados tanto en la explicación de las evaluaciones como en tener el menor error por cada una de las seis dimensiones. Además, se seleccionó uno de los modelos que utilizaron los sets de datos de emociones, características del hablante y paralingüística, que fue el que tuvo la mayor cantidad de mejores resultados entre ellos.

Set	Método	R <sub>2</sub> (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	0.104 (0.034)	-1.068 (0.157)
allfeatures_allevs	SVR_poly	0.067 (0.062)	-1.116 (0.197)
allfeatures_allevs	SVR_rbf	0.015 (0.081)	-1.182 (0.232)
allfeatures_meanestimulo	RF	0.252 (0.088)	-0.415 (0.102)
allfeatures_meanestimulo	SVR_poly	0.293 (0.067)	-0.392 (0.089)
allfeatures_meanestimulo	SVR_rbf	0.242 (0.092)	-0.422 (0.108)
allfeatures_meanestimulo_allevs	RF	<b>0.565 (0.219)</b>	-0.502 (0.28)
allfeatures_meanestimulo_allevs	SVR_poly	0.145 (0.051)	-0.887 (0.29)
allfeatures_meanestimulo_allevs	SVR_rbf	0.12 (0.047)	-0.913 (0.298)
allfeatures_meanvoice	RF	-0.983 (1.598)	-0.333 (0.229)
allfeatures_meanvoice	SVR_poly	-0.445 (0.991)	<b>-0.250 (0.159)</b>
allfeatures_meanvoice	SVR_rbf	-0.374 (0.84)	-0.259 (0.178)
allfeatures_meanvoice_meanestimulo_allevs	RF	0.522 (0.207)	-0.458 (0.292)
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.207 (0.066)	-0.724 (0.359)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.192 (0.076)	-0.742 (0.372)
is09_emotion_allevs	RF	0.212 (0.031)	-0.938 (0.133)
is09_emotion_allevs	SVR_poly	-0.089 (0.061)	-1.294 (0.163)
is09_emotion_allevs	SVR_rbf	-0.043 (0.062)	-1.246 (0.205)
is10_paraling_allevs	RF	0.232 (0.045)	-0.914 (0.124)
is10_paraling_allevs	SVR_poly	-0.08 (0.026)	-1.281 (0.137)
is10_paraling_allevs	SVR_rbf	-0.016 (0.041)	-1.209 (0.165)
is11_speakerstate_allevs	RF	<b>0.238 (0.038)</b>	<b>-0.907 (0.122)</b>
is11_speakerstate_allevs	SVR_poly	-0.084 (0.066)	-1.287 (0.162)
is11_speakerstate_allevs	SVR_rbf	-0.007 (0.028)	-1.197 (0.145)
is12_speakertrait_allevs	RF	0.237 (0.031)	-0.908 (0.124)
is12_speakertrait_allevs	SVR_poly	-0.079 (0.048)	-1.284 (0.167)
is12_speakertrait_allevs	SVR_rbf	-0.008 (0.027)	-1.198 (0.142)
is13_ComParE_allevs	RF	0.233 (0.034)	-0.912 (0.119)
is13_ComParE_allevs	SVR_poly	-0.077 (0.046)	-1.28 (0.162)
is13_ComParE_allevs	SVR_rbf	-0.002 (0.028)	-1.19 (0.141)

Cuadro 5.2: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Claridad'

Set	Método	Rz (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	0.064 (0.043)	-0.687 (0.076)
allfeatures_allevs	SVR_poly	0.049 (0.044)	-0.698 (0.078)
allfeatures_allevs	SVR_rbf	-0.030 (0.066)	-0.758 (0.1)
allfeatures_meanestimulo	RF	0.221 (0.072)	-0.246 (0.061)
allfeatures_meanestimulo	SVR_poly	0.259 (0.068)	-0.232 (0.052)
allfeatures_meanestimulo	SVR_rbf	0.209 (0.043)	-0.248 (0.056)
allfeatures_meanestimulo_allevs	RF	<b>0.519 (0.242)</b>	-0.335 (0.174)
allfeatures_meanestimulo_allevs	SVR_poly	0.122 (0.051)	-0.555 (0.171)
allfeatures_meanestimulo_allevs	SVR_rbf	0.073 (0.024)	-0.581 (0.17)
allfeatures_meanvoice	RF	-3.692 (5.559)	-0.127 (0.035)
allfeatures_meanvoice	SVR_poly	-2.97 (4.869)	<b>-0.100 (0.019)</b>
allfeatures_meanvoice	SVR_rbf	-3.325 (5.346)	-0.109 (0.021)
allfeatures_meanvoice_meanestimulo_allevs	RF	0.408 (0.301)	-0.312 (0.186)
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.212 (0.114)	-0.443 (0.232)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.177 (0.102)	-0.458 (0.234)
is10_paraling_allevs	RF	0.16 (0.064)	-0.614 (0.056)
is10_paraling_allevs	SVR_poly	-0.223 (0.059)	-0.895 (0.076)
is10_paraling_allevs	SVR_rbf	-0.043 (0.047)	-0.767 (0.091)
is11_speakerstate_allevs	RF	0.157 (0.059)	-0.617 (0.06)
is11_speakerstate_allevs	SVR_poly	-0.341 (0.092)	-0.989 (0.146)
is11_speakerstate_allevs	SVR_rbf	0.012 (0.028)	-0.725 (0.079)
is12_speakertrait_allevs	RF	0.157 (0.062)	-0.617 (0.059)
is12_speakertrait_allevs	SVR_poly	-0.344 (0.091)	-0.991 (0.145)
is12_speakertrait_allevs	SVR_rbf	0.01 (0.029)	-0.727 (0.08)
is13_ComParE_allevs	RF	0.162 (0.052)	-0.614 (0.058)
is13_ComParE_allevs	SVR_poly	-0.342 (0.091)	-0.989 (0.145)
is13_ComParE_allevs	SVR_rbf	0.017 (0.032)	-0.722 (0.082)

Cuadro 5.3: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Defectos'

Set	Método	R2 (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	0.132 (0.076)	-1.238 (0.149)
allfeatures_allevs	SVR_poly	0.187 (0.062)	-1.159 (0.127)
allfeatures_allevs	SVR_rbf	0.158 (0.05)	-1.202 (0.126)
allfeatures_meanestimulo	RF	0.269 (0.095)	-0.646 (0.123)
allfeatures_meanestimulo	SVR_poly	0.319 (0.061)	-0.603 (0.103)
allfeatures_meanestimulo	SVR_rbf	0.263 (0.051)	-0.655 (0.12)
allfeatures_meanestimulo_allevs	RF	<b>0.667 (0.172)</b>	-0.464 (0.243)
allfeatures_meanestimulo_allevs	SVR_poly	0.234 (0.055)	-0.997 (0.222)
allfeatures_meanestimulo_allevs	SVR_rbf	0.201 (0.05)	-1.039 (0.228)
allfeatures_meanvoice	RF	-0.506 (1.093)	-0.521 (0.316)
allfeatures_meanvoice	SVR_poly	-0.179 (0.608)	<b>-0.419 (0.236)</b>
allfeatures_meanvoice	SVR_rbf	-0.136 (0.455)	-0.454 (0.302)
allfeatures_meanvoice_meanestimulo_allevs	RF	0.627 (0.166)	-0.446 (0.245)
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.289 (0.095)	-0.847 (0.337)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.261 (0.098)	-0.88 (0.349)
is10_paraling_allevs	RF	0.354 (0.041)	-0.918 (0.051)
is10_paraling_allevs	SVR_poly	-0.056 (0.144)	-1.497 (0.155)
is10_paraling_allevs	SVR_rbf	0.087 (0.061)	-1.299 (0.092)
is11_speakerstate_allevs	RF	0.345 (0.056)	-0.929 (0.04)
is11_speakerstate_allevs	SVR_poly	-0.089 (0.057)	-1.554 (0.158)
is11_speakerstate_allevs	SVR_rbf	0.132 (0.045)	-1.235 (0.073)
is12_speakertrait_allevs	RF	0.348 (0.048)	-0.925 (0.031)
is12_speakertrait_allevs	SVR_poly	-0.091 (0.057)	-1.557 (0.158)
is12_speakertrait_allevs	SVR_rbf	0.126 (0.042)	-1.243 (0.07)
is13_ComParE_allevs	RF	0.34 (0.037)	-0.938 (0.032)
is13_ComParE_allevs	SVR_poly	-0.093 (0.045)	-1.559 (0.137)
is13_ComParE_allevs	SVR_rbf	0.132 (0.043)	-1.235 (0.071)

Cuadro 5.4: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Entonación'

Set	Método	R <sub>2</sub> (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	0.222 (0.063)	-1.325 (0.123)
allfeatures_allevs	SVR_poly	0.192 (0.087)	-1.373 (0.142)
allfeatures_allevs	SVR_rbf	0.163 (0.044)	-1.424 (0.097)
allfeatures_meanestimulo	RF	0.325 (0.069)	-0.795 (0.086)
allfeatures_meanestimulo	SVR_poly	0.273 (0.082)	-0.856 (0.102)
allfeatures_meanestimulo	SVR_rbf	0.227 (0.042)	-0.909 (0.058)
allfeatures_meanestimulo_allevs	RF	0.722 (0.142)	-0.462 (0.245)
allfeatures_meanestimulo_allevs	SVR_poly	0.229 (0.041)	-1.209 (0.191)
allfeatures_meanestimulo_allevs	SVR_rbf	0.2 (0.038)	-1.256 (0.204)
allfeatures_meanvoice	RF	-0.366 (0.613)	-1.08 (0.584)
allfeatures_meanvoice	SVR_poly	-0.39 (0.623)	-1.055 (0.479)
allfeatures_meanvoice	SVR_rbf	-0.216 (0.372)	-1.031 (0.569)
allfeatures_meanvoice_meanestimulo_allevs	RF	<b>0.738 (0.124)</b>	<b>-0.406 (0.233)</b>
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.26 (0.062)	-1.082 (0.308)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.236 (0.069)	-1.118 (0.325)
is10_paraling_allevs	RF	0.497 (0.033)	-0.858 (0.081)
is10_paraling_allevs	SVR_poly	0.011 (0.081)	-1.685 (0.176)
is10_paraling_allevs	SVR_rbf	0.129 (0.042)	-1.482 (0.107)
is11_speakerstate_allevs	RF	0.492 (0.039)	-0.863 (0.06)
is11_speakerstate_allevs	SVR_poly	-0.014 (0.014)	-1.726 (0.093)
is11_speakerstate_allevs	SVR_rbf	0.094 (0.03)	-1.541 (0.084)
is12_speakertrait_allevs	RF	0.499 (0.039)	-0.851 (0.057)
is12_speakertrait_allevs	SVR_poly	-0.016 (0.013)	-1.729 (0.089)
is12_speakertrait_allevs	SVR_rbf	0.094 (0.03)	-1.542 (0.083)
is13_ComParE_allevs	RF	0.496 (0.04)	-0.856 (0.059)
is13_ComParE_allevs	SVR_poly	-0.027 (0.049)	-1.748 (0.131)
is13_ComParE_allevs	SVR_rbf	0.093 (0.034)	-1.544 (0.089)

Cuadro 5.5: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Familiaridad'



Set	Método	R <sub>2</sub> (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	0.222 (0.07)	-1.509 (0.12)
allfeatures_allevs	SVR_poly	0.199 (0.08)	-1.553 (0.135)
allfeatures_allevs	SVR_rbf	0.186 (0.062)	-1.583 (0.148)
allfeatures_meanestimulo	RF	0.387 (0.044)	-0.753 (0.027)
allfeatures_meanestimulo	SVR_poly	0.363 (0.052)	-0.782 (0.028)
allfeatures_meanestimulo	SVR_rbf	0.316 (0.04)	-0.843 (0.059)
allfeatures_meanestimulo_allevs	RF	0.683 (0.161)	-0.6 (0.324)
allfeatures_meanestimulo_allevs	SVR_poly	0.256 (0.071)	-1.323 (0.276)
allfeatures_meanestimulo_allevs	SVR_rbf	0.244 (0.064)	-1.346 (0.29)
allfeatures_meanvoice	RF	-0.208 (0.663)	-0.838 (0.422)
allfeatures_meanvoice	SVR_poly	-0.198 (0.551)	-0.842 (0.367)
allfeatures_meanvoice	SVR_rbf	-0.1 (0.333)	-0.873 (0.427)
allfeatures_meanvoice_meanestimulo_allevs	RF	<b>0.689 (0.147)</b>	<b>-0.538 (0.319)</b>
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.313 (0.089)	-1.132 (0.405)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.3 (0.099)	-1.154 (0.421)
is10_paraling_allevs	RF	0.442 (0.058)	-1.089 (0.157)
is10_paraling_allevs	SVR_poly	-0.018 (0.092)	-1.989 (0.28)
is10_paraling_allevs	SVR_rbf	0.11 (0.093)	-1.731 (0.188)
is11_speakerstate_allevs	RF	0.434 (0.06)	-1.099 (0.124)
is11_speakerstate_allevs	SVR_poly	-0.091 (0.044)	-2.119 (0.089)
is11_speakerstate_allevs	SVR_rbf	0.077 (0.048)	-1.793 (0.083)
is12_speakertrait_allevs	RF	0.447 (0.061)	-1.075 (0.123)
is12_speakertrait_allevs	SVR_poly	-0.087 (0.052)	-2.112 (0.09)
is12_speakertrait_allevs	SVR_rbf	0.075 (0.047)	-1.797 (0.083)
is13_ComParE_allevs	RF	0.445 (0.063)	-1.078 (0.129)
is13_ComParE_allevs	SVR_poly	-0.086 (0.048)	-2.11 (0.103)
is13_ComParE_allevs	SVR_rbf	0.073 (0.053)	-1.801 (0.104)

Cuadro 5.6: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión ‘Naturalidad’

Set	Método	R <sub>2</sub> (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	-0.08 (0.054)	-0.232 (0.025)
allfeatures_allevs	SVR_poly	-0.014 (0.017)	-0.219 (0.031)
allfeatures_allevs	SVR_rbf	-0.017 (0.016)	-0.22 (0.032)
allfeatures_meanestimulo	RF	-0.095 (0.131)	-0.097 (0.009)
allfeatures_meanestimulo	SVR_poly	-0.054 (0.042)	-0.095 (0.014)
allfeatures_meanestimulo	SVR_rbf	-0.044 (0.061)	-0.094 (0.016)
allfeatures_meanestimulo_allevs	RF	<b>0.502 (0.246)</b>	-0.103 (0.053)
allfeatures_meanestimulo_allevs	SVR_poly	0.008 (0.021)	-0.182 (0.052)
allfeatures_meanestimulo_allevs	SVR_rbf	-0.002 (0.015)	-0.184 (0.052)
allfeatures_meanvoice	RF	-2.662 (2.173)	-0.032 (0.023)
allfeatures_meanvoice	SVR_poly	-2.482 (2.675)	-0.034 (0.03)
allfeatures_meanvoice	SVR_rbf	-2.216 (2.617)	<b>-0.031 (0.026)</b>
allfeatures_meanvoice_meanestimulo_allevs	RF	-0.049 (1.049)	-0.102 (0.057)
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.037 (0.03)	-0.146 (0.084)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.024 (0.016)	-0.147 (0.084)
is10_paraling_allevs	RF	0.046 (0.053)	-0.206 (0.027)
is10_paraling_allevs	SVR_poly	-0.086 (0.081)	-0.234 (0.033)
is10_paraling_allevs	SVR_rbf	0.01 (0.014)	-0.215 (0.034)
is11_speakerstate_allevs	RF	0.043 (0.044)	-0.207 (0.034)
is11_speakerstate_allevs	SVR_poly	-0.005 (0.02)	-0.217 (0.033)
is11_speakerstate_allevs	SVR_rbf	-0.013 (0.019)	-0.219 (0.034)
is12_speakertrait_allevs	RF	0.048 (0.033)	-0.206 (0.034)
is12_speakertrait_allevs	SVR_poly	-0.005 (0.02)	-0.217 (0.033)
is12_speakertrait_allevs	SVR_rbf	-0.013 (0.016)	-0.219 (0.034)
is13_ComParE_allevs	RF	0.045 (0.036)	-0.207 (0.034)
is13_ComParE_allevs	SVR_poly	-0.005 (0.021)	-0.217 (0.032)
is13_ComParE_allevs	SVR_rbf	-0.01 (0.015)	-0.219 (0.035)

Cuadro 5.7: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Velocidad del habla'

Set	Método	R <sub>2</sub> (DesvEst)	NMSE (DesvEst)
allfeatures_allevs	RF	0.181 (0.119)	-1.367 (0.209)
allfeatures_allevs	SVR_poly	0.212 (0.071)	-1.314 (0.106)
allfeatures_allevs	SVR_rbf	0.186 (0.045)	-1.359 (0.098)
allfeatures_meanestimulo	RF	0.3 (0.122)	-0.801 (0.117)
allfeatures_meanestimulo	SVR_poly	0.307 (0.081)	-0.797 (0.086)
allfeatures_meanestimulo	SVR_rbf	0.271 (0.061)	-0.841 (0.081)
allfeatures_meanestimulo_allevs	RF	<b>0.718 (0.147)</b>	-0.465 (0.263)
allfeatures_meanestimulo_allevs	SVR_poly	0.252 (0.059)	-1.16 (0.218)
allfeatures_meanestimulo_allevs	SVR_rbf	0.233 (0.054)	-1.188 (0.219)
allfeatures_meanvoice	RF	-0.448 (0.898)	-0.853 (0.475)
allfeatures_meanvoice	SVR_poly	-0.27 (0.65)	-0.734 (0.335)
allfeatures_meanvoice	SVR_rbf	-0.167 (0.475)	-0.756 (0.396)
allfeatures_meanvoice_meanestimulo_allevs	RF	0.708 (0.128)	<b>-0.434 (0.236)</b>
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly	0.296 (0.082)	-1.014 (0.341)
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf	0.28 (0.086)	-1.037 (0.349)
is10_paraling_allevs	RF	0.435 (0.061)	-0.948 (0.151)
is10_paraling_allevs	SVR_poly	-0.039 (0.095)	-1.734 (0.172)
is10_paraling_allevs	SVR_rbf	0.118 (0.061)	-1.473 (0.129)
is11_speakerstate_allevs	RF	0.46 (0.049)	-0.901 (0.077)
is11_speakerstate_allevs	SVR_poly	-0.019 (0.015)	-1.704 (0.121)
is11_speakerstate_allevs	SVR_rbf	0.13 (0.042)	-1.456 (0.142)
is12_speakertrait_allevs	RF	0.448 (0.045)	-0.921 (0.082)
is12_speakertrait_allevs	SVR_poly	-0.02 (0.015)	-1.705 (0.119)
is12_speakertrait_allevs	SVR_rbf	0.128 (0.041)	-1.459 (0.14)
is13_ComParE_allevs	RF	0.453 (0.044)	-0.912 (0.079)
is13_ComParE_allevs	SVR_poly	-0.022 (0.034)	-1.707 (0.123)
is13_ComParE_allevs	SVR_rbf	0.129 (0.041)	-1.458 (0.138)

Cuadro 5.8: Desempeño de diferentes modelos y sets de datos para la predicción de la dimensión 'Evaluación General'

Set	Método	Cla	Def	Ent	Fam	Nat	Vel	Gral
allfeatures_allevs	RF							
allfeatures_allevs	SVR_poly							
allfeatures_allevs	SVR_rbf							
allfeatures_meanestimulo	SVR_poly							
allfeatures_meanestimulo	SVR_rbf							
<b>allfeatures_meanestimulo_allevs</b>	<b>RF</b>	R2	R2	R2			R2	R2
allfeatures_meanestimulo_allevs	SVR_poly							
allfeatures_meanestimulo_allevs	SVR_rbf							
allfeatures_meanvoice	RF							
<b>allfeatures_meanvoice</b>	<b>SVR_poly</b>	E	E	E				
<b>allfeatures_meanvoice</b>	<b>SVR_rbf</b>						E	
<b>allfeatures_meanvoice_meanestimulo_allevs</b>	<b>RF</b>				R2, E	R2, E		E
allfeatures_meanvoice_meanestimulo_allevs	SVR_poly							
allfeatures_meanvoice_meanestimulo_allevs	SVR_rbf							
is09_emotion_allevs	RF							
is09_emotion_allevs	SVR_poly							
is09_emotion_allevs	SVR_rbf							
is10_paraling_allevs	RF			R2, E				
is10_paraling_allevs	SVR_poly							
is10_paraling_allevs	SVR_rbf							
is11_speakerstate_allevs	RF	R2, E						R2, E
is11_speakerstate_allevs	SVR_poly							
is11_speakerstate_allevs	SVR_rbf							
<b>is12_speakertrait_allevs</b>	<b>RF</b>				R2, E	R2, E	R2, E	
is12_speakertrait_allevs	SVR_poly							
is12_speakertrait_allevs	SVR_rbf							
is13_ComParE_allevs	RF		R2, E					
is13_ComParE_allevs	SVR_poly							
is13_ComParE_allevs	SVR_rbf							

Cuadro 5.9: Resumen de mejores modelos por cada dimensión de evaluación de acuerdo a R2 y NMSE

1. **allfeatures\_meanestimulo\_allevs-RF:** El modelo de Regresión de RandomForest entrenado con el set de datos que tiene un registro por evaluación, con todas sus características y evaluaciones recibidas, más un registro por elocución, con todas las características y las evaluaciones promedio por elocución.
2. **allfeatures\_meanvoice-SVR\_poly:** El modelo de Regresión con SVM con kernel polinómico de grado tres, entrenado con el set de datos con un registro por elocución, con las características de cada una y el promedio de las evaluaciones por voz.
3. **allfeatures\_meanvoice-SVR\_rbf:** El modelo de Regresión con SVM con kernel radial, entrenado con el set de datos con un registro por elocución, con las características de cada una y el promedio de las evaluaciones por voz.
4. **allfeatures\_meanvoice\_meanestimulo\_allevs-RF:** El modelo de Regresión de RandomForest entrenado con el set de datos con un registro por cada evaluación, más dos registros por cada elocución, uno el promedio de las evaluaciones por voz y otro con el promedio de evaluaciones por elocución, en todos los casos con las nueve dimensiones de evaluación.
5. **is12\_speakertrait\_allevs-RF:** El modelo de Regresión de RandomForest entrenado con el set de datos de la competencia de reconocimiento de características del hablante.

A partir de estos modelos se realizarán tres modelos que integrarán distintos tipos de información:

- **A - Predicción directa de la evaluación general:** Con los cinco modelos seleccionados previamente se realizarán modelos nuevos que integren esa información de distinta manera. Así, se utilizará la salida de los cinco modelos elegidos para estimar la calida general del habla artificial por cada elocución.
- **B - Predicción de dimensiones con integración por regresión:** se realizará la predicción de las siete dimensiones evaluadas por cada elocución, utilizando los mejores modelos según  $R^2$  y NMSE para cada una, y, con esa información, y sumada a la predicción por el modelo de características del hablante elegido, se realizará varios modelos nuevos de regresión para estimar la calidad general del habla.
- **C - Predicción de dimensiones con integración *naive*:** similar a B en que se crearán dos estimadores por cada dimensión de la evaluación de la calidad, sumado a un estimador general por el set de datos de características del hablante, pero en vez de hacer una regresión para predecir la calidad general se le dará esa información como entrada a un modelo entrenado para

relacionar las seis dimensiones de evaluación con la evaluación general de la elocución.

#### 5.4.3 Validación de los modelos

Para la validación de los modelos propuestos se decidió separar todas las elocuciones de dos de las voces (26 y 27), siendo ambas del mismo sistema, una masculina y otra femenina, ya que ambos recibieron evaluaciones en el promedio o ligeramente más arriba que las otras voces, y no se trata de casos extremos, de forma de evitar que se obtuvieran. Así, el objetivo es poder evaluar el poder de generalización del modelo creado, ya que este se pondrá a prueba con una voz que nunca se escuchó, ni tampoco se tuvo acceso a otra voz del mismo sistema. Más allá de que esta decisión pone a prueba los modelos que se fueran a entrenar, se trata de algo necesario, por cuanto deben poder utilizarse en el contexto de otros sistemas nunca vistos.

Para los modelos se realizó primero un entrenamiento de los distintos estimadores con un conjunto de voces, de acuerdo a su definición, luego con otras voces nunca vistas se realizó el ajuste de las regresiones necesarias para la integración de la información de cada predictor y, finalmente, los modelos entrenados fueron evaluados con las voces elegidas para la validación.

En el Cuadro 5.10 se pueden ver los resultados, evaluados con las medidas  $R^2$  y error cuadrado promedio negativo (NMSE, por sus siglas en inglés). Como se puede apreciar, los resultados no son buenos en forma global, en parte por lo estricto de la prueba, pero sí pueden servir para analizar comparativamente cada una de las propuestas.

De acuerdo a las pruebas realizadas el mejor desempeño lo tuvo el modelo de regresión con perceptrón multicapa, donde se usaron las predicciones generadas para cada una de las siete dimensiones de calidad definidas anteriormente.

### 5.5 CONCLUSIONES

En este capítulo se presentaron algunas dimensiones para la caracterización de habla artificial para su evaluación de calidad en forma automática. Las dimensiones elegidas permiten explicar gran parte de la variación de los valores de siete dimensiones de la evaluación de cada elocución, por lo que, pese a ser pocas, se presentan como informativas.

Se probaron con diferentes juegos de datos, de forma de poder predecir las dimensiones de evaluación y la calidad de forma integral. En este último caso no se obtuvieron buenos resultados, lo que puede deberse a que con los predictores que se crearon se estaba perdiendo información de cada elocución. Así, se deberán explorar nuevas opcio-

Tipo de modelo	Método	R <sub>2</sub>	NMSE
A-Predicción directa	Media	-2.454	-1.146
	Reg. Lineal	-2.941	-1.308
	Random Forest	-2.877	-1.286
	SVR_poly	-2.748	-1.244
	Reg. MLP	-0.615	-0.536
B-Predicción de dimensiones con integración por regresión	Reg. Lineal	-3.445	-1.475
	Random Forest	-2.836	-1.273
	SVR_poly	-2.773	-1.252
	Reg. MLP	<b>-0.085</b>	<b>-0.360</b>
C-Predicción de dimensiones con integración naive	Reg. Lineal	-5.504	-2.158
	Random Forest	-7.968	-2.976
	SVR_poly	-10.640	-3.863
	Reg. MLP	-7.395	-2.786

Cuadro 5.10: Evaluación de modelos con los datos de validación

nes para armar modelos que aprovechen las características definidas en esta tesis junto con otros atributos acústicos de cada onda.

Se presentaron dos medidas generadas a partir de un sistema de RAH, uno para la evaluación objetiva de la inteligibilidad y el otro para evaluar la ubicación de las pausas por medio de la identificación de los signos de puntuación en el texto reconocido. Esto último es especialmente útil, debido a que las pausas y otros atributos prosódicos no suelen ser de fácil evaluación, por ejemplo, debido a que existen muchas realizaciones que puede tener un texto, donde todas son válidas, y no hay una forma precisa de determinar cuál es la mejor de forma automática.

Actualmente el problema de la evaluación de la calidad del habla tiene un fuerte resurgimiento, en especial, se cree, debido a la disponibilidad total de tecnología para el uso habitual de técnicas de aprendizaje profundo, y es un área que presentará muchos desafíos, en particular, ahora que gran parte de los sistemas disponibles tiene una buena calidad, y es necesario hilar más fino para poder diferenciar el desempeño de cada uno, ahora que la inteligibilidad y naturalidad ya dejan de ser un problema.

## 5.6 TRABAJO FUTURO

Debido a que la evaluación de la calidad del habla artificial, pensándola en forma general, requiere de trabajar con muchos datos para cubrir la variedad de realizaciones de los distintos sistemas, se planea ampliar la recolección de datos de la prueba básica realizada, además de incorporar datos de otras fuentes, como la base utilizada para el

*VoiceMOS Challenge*, descrita en [33]. Para ello se profundizará en el trabajo con redes neuronales profundas, que actualmente domina el estado del arte, y en la generación no supervisada de etiquetas y el trabajo con *embeddings* para el habla.

Se espera poder profundizar en las diferencias y cómo aprovechar mejor las predicciones realizadas a nivel de voz, sistema y, hasta de evaluador, tomando como base el trabajo de Chinen y col. [28]. Se puede tomar como punto de partida la caracterización de los evaluadores realizada en [92].

Relacionado con esto último, queda pendiente cómo contemplar la distribución de puntajes recibidos por cada elocución de cada sistema, más allá de tomar el promedio, debido a las limitaciones que tiene, o la mediana. Así, se plantea como trabajo futuro cómo integrar la información obtenida de cada elocución evaluada, de forma de usar la información de las distribuciones para comparar el desempeño de los sistemas, usando para ellos estimadores robustos.



## PROMINENCIA EN EL HABLA

---

Se puede definir a la prominencia como la propiedad por la cual se percibe a ciertas unidades lingüísticas como destacadas con respecto a su entorno [202]. En particular, hablamos de *prominencia prosódica* como una característica que se puede evaluar en el habla, en tanto es una medida perceptual acerca de qué partes de una elocución se destacan por sobre otras a su alrededor.

La evaluación de los atributos prosódicos del habla, y en particular, la prominencia, son determinantes para la calidad percibida de un sistema. La información de qué partes de una frase son más prominentes guarda relación con el sentido que se le otorga a una frase, además de que es relevante al momento de que evaluar la calidad de una elocución. Así, aún cuando se usan las mismas unidades segmentales, se percibe como mejor a una elocución que tiene una curva de entonación natural [216]. En general, se suelen estudiar los atributos prosódicos en forma general, por medio de atributos acústicos del habla [150], o utilizando modelos de entonación como el de Fujisaki [84].

En este capítulo describiremos en forma general el foco y la prominencia en el habla, y mostraremos el diseño experimental desarrollado para obtener información particular de la prominencia para nuestra variante del español, con el fin de poder construir clasificadores que permitan detectar la presencia de sílabas prominentes, y que esta información pueda ser luego utilizada para analizar segmentos de habla artificial. Esta parte del trabajo de tesis está descripto en varias publicaciones [44, 66, 72, 74, 137].

### 6.1 FOCO EN EL HABLA

Durante la producción de habla, la información de la estructura entonativa de una oración se refleja en el énfasis relativo dado a sus palabras de contenido y en las características tonales de los segmentos finales de cada frase. El foco que tiene una oración puede explicitarse por medio del orden elegido para las palabras, debido que en el español el orden de los constituyentes es relativamente libre [113] y permite esta flexibilidad. Dado que las sílabas con acento léxico pueden presentar un aumento en la frecuencia fundamental ( $F_0$ ), energía y duración y, además, una mayor estabilidad espectral, si estas sílabas son enfatizadas en el discurso, algunos o todos estos parámetros pueden ser modificados.

Por ejemplo, los acentos tonales altos se integrarán a los rasgos acústicos presentes en la sílaba, aumentando o disminuyendo el tono fundamental. El correlato perceptual de este proceso de énfasis e integración es la prominencia silábica [60]. Junto con los aspectos acústicos de los acentos lexicales y tonales coexisten otros factores a nivel segmental y suprasegmental que afectan la percepción de prominencia, tales como la información lingüística y paralingüística. Los aspectos pragmáticos asociados a la información nueva que se transmite, también pueden influenciar los juicios de prominencia.

Debido a que la función focal se define como la selección de un segmento de la oración como el de mayor jerarquía pragmática, por su relevancia en el procesamiento informativo del destinatario, su realización necesita de recursos léxico-gramaticales y fónicos. Dado que en español las estructuras sintácticas son flexibles, resulta importante establecer el papel del factor acústico para generar la percepción del foco. A pesar de que la estructura de la información se pueda jerarquizar principalmente a partir del orden de palabras, es la prosodia la que determina la interpretación final frente a estructuras sintácticamente similares.

Además, el foco sintáctico se manifiesta en la superficie con una mayor prominencia prosódica y así, permite distinguir entre estructuras marcadas de las no marcadas [238]. La diferencia entre los parámetros prosódicos (entonación, duración, sonoridad, pausa), los sintácticos y los informativos a la hora de marcar el foco radica en que los prosódicos siempre estarán presentes y el foco se da por la naturaleza de su manifestación. El valor del foco prosódico difiere en el tipo de discurso. Como el habla espontánea no está muy organizada ni focalizada, la prosodia cumple un rol fundamental para estructurar el plano sintáctico, es decir, organizar el discurso en unidades menores. A nivel pragmático, la prosodia comunica estados de ánimo e intereses elocutivos. En el nivel de la cláusula se parte de la idea de que cualquier oración es respuesta de una pregunta supuesta, y el foco marcaría el constituyente que motiva la respuesta a esa pregunta implícita. Dentro de este dominio, Ladd [114] define el foco como una tendencia a acentuar o desacentuar porciones de una oración relacionadas con el significado; así, los autores indican que las oraciones tendrían ítems marcados con los rasgos [-foco] o [+foco] según la intención del hablante. El alcance del foco puede variar: puede abarcar toda la oración (foco neutral o normal) o estar sobre un constituyente como el predicado (foco amplio) o sobre un solo ítem léxico (foco estrecho). El foco puede transmitir varios tipos de intereses comunicativos, agregar información nueva, corregir información precedente, seleccionar un elemento entre otros ofrecidos por el contexto o contrastar información supuesta por el oyente.

## 6.2 PERSPECTIVAS EN EL ESTUDIO DE LA PROMINENCIA PROSÓDICA

A efectos de comprender el enfoque seguido en este trabajo, es relevante describir el resumen de las distintas perspectivas en el trabajo con la *prominencia prosódica*, como hacen Wagner y col. [222], quienes ubican en tres grupos a las diferentes visiones que hay sobre el tema, como se describe más abajo. En este trabajo se tienen en cuenta aspectos que pertenecen a las tres perspectivas, ya que una de las visiones que tiene es la de la integración de información con el fin de utilizarlo para la evaluación de la prominencias en el habla.

### *Perspectiva funcional*

Aquí se tiene en cuenta la función de la prominencia prosódica para la comunicación, en tanto es indicativo de la estructura de la información, indicación de contexto, acento de frase, orden de las palabras y categoría gramatical.

Dentro de esta categoría se incluyen aspectos paralingüísticos, que tienen que ver con emociones, atributos de la personalidad y otras expresiones que se comunican a través de la voz [64], y que se manifiestan por variaciones prosódicas entre las que está la prominencia.

### *Perspectiva física*

En la perspectiva física, la mira está en los atributos que surgen del análisis de la señal de audio, como frecuencia fundamental, duración, e intensidad, entre otros. En los casos puramente físicos, se observa la voz como algo continuo, más que como algo categórico, que está puesto dentro de un contexto de comunicación.

### *Perspectiva cognitiva*

Para el caso de la perspectiva cognitiva, la mirada está puesta en el procesamiento perceptual, y, en especial, en el procesamiento neuronal de bajo nivel, y las conexiones entre redes neuronales en el cerebro. Sin embargo, esta perspectiva está relacionada con el enfoque funcional, en tanto se puede estudiar la prominencia como parte de procesos concretos, como la atención y la memoria, y también con el enfoque físico, donde buscan los correlatos con las señales cerebrales registradas.

### 6.3 EVALUACIÓN DEL FOCO Y PROMINENCIAS PARA EL ESPAÑOL DE BUENOS AIRES

En el estudio del foco en español, Toledo [204] caracteriza cuáles son los correlatos acústicos de las señales prosódicas. Para ello, utiliza un *corpus* de oraciones con estructura profunda similar, pero con distintas marcaciones focales según el contexto discursivo que surgía de una interrogación. Posteriormente, analiza contrastivamente la *F<sub>0</sub>*, la intensidad y duración de estas emisiones, con el fin de describir acústicamente la marcación focal. Los resultados muestran una variabilidad en el uso de estos patrones acústicos, lo que hace concluir que, en español, la marcación prosódica es asistemática y dependería más de una opción del hablante que del contexto. Debido a la falta de consenso en la bibliografía acerca de los factores acústicos que inciden en la marcación focal, Face [50] realiza un estudio experimental de la producción de foco en el español madrileño, con el fin de determinar el valor de la altura tonal para este fenómeno. Para ello, investiga la altura tonal de ciertas palabras, manipulando la posición oracional— inicial, media y final— en diferentes contextos focales. Los resultados muestran que, de acuerdo a la posición del foco, la altura tonal podría indicar varias cosas: dentro de una misma frase podría marcar que el hablante se está aproximando al foco de la misma o indicar la palabra focal por elevarse más que en una frase sin foco; después de la palabra focal, el hablante reduce el tono para marcar el final del foco. Si bien lo califica como complejo, el comportamiento de la altura tonal identifica que las variaciones en un mismo rasgo pueden comunicar un solo aspecto, el foco.

Por otra parte, Labastía [113] realiza un análisis cualitativo del foco en el discurso espontáneo de un hablante del español de Buenos Aires, en el que tuvo en cuenta la *F<sub>0</sub>*, la duración y la intensidad. Los resultados concluyen que el español de Buenos Aires tiende a mantener el acento focal en el constituyente final, lo que genera que unidades enunciativas menores como las palabras sean resaltadas, en lugar de unidades más largas como las frases. Además, muestra que la ubicación del acento final puede tener una interpretación del foco contrastiva o no contrastiva.

En [39], Dorta Luis realiza un estudio multidimensional del foco en español en el que considera factores sintácticos, semánticos, pragmáticos, discursivos y prosódicos para describirlo. El análisis de un discurso oral espontáneo revela que si bien los recursos prosódicos son los más relevantes a la hora de señalar el foco, la marcación focal no siempre tiene como objetivo indicar información nueva, sino que depende de los intereses del hablante.

Si bien existen estudios en español acerca de la interacción entre la percepción de los datos acústicos y la expectativa lingüística —i. e., reconocimiento de la modalidad enunciativa, identificación del foco,

prominencia silábica—, en ellos no se analiza a los distintos aspectos involucrados de manera conjunta. En un estudio sobre el español madrileño, Face [51] investiga experimentalmente el rol de las pistas acústicas en la percepción de las frases interrogativas y entonativas. Mediante la manipulación de la  $F_0$ , se evaluó si un grupo de sujetos podía distinguir si un enunciado era interrogativo o declarativo a partir de oraciones completas y de otras incompletas. Los resultados muestran que los sujetos pueden reconocer acertadamente curvas entonacionales desde las primeras frases del enunciado y que pueden identificar de forma temprana cualquier alteración en la  $F_0$ . Eso sugiere que esta pista acústica es la más relevante para el reconocimiento de la modalidad enunciativa.

Por otra parte, en el estudio de la percepción del foco en el español, en [116] Lang-Rigal se propone identificar cuáles son los correlatos acústicos para la percepción del foco amplio y el foco estrecho en el español de Buenos Aires. Para ello, graba un solo enunciado (“Manolo viene mañana”) con foco neutro —en toda la oración— y con foco estrecho —en el sujeto. Posteriormente, realizó una tarea de percepción en la que los sujetos debían indicar la palabra de mayor prominencia de la frase. Los resultados muestran que la percepción del foco estrecho en el sujeto (“Manolo”) se debe a tres factores: un aumento tonal en la sílaba acentuada de esa palabra, el aumento de la duración de la misma sílaba y una pausa después de la palabra focalizada. Además, este estudio aporta evidencia de que la percepción del foco se ve afectada por las características acústicas de los elementos no focalizados.

En otras lenguas existen estudios que identifican los factores acústicos que intervienen en la percepción de la prominencia silábica. Eriksson, Grabe y Traunmüller [42, 43] diseñaron una serie de experimentos en sueco con el objetivo de determinar hasta qué punto el esfuerzo vocal puede indicar la identificación subjetiva de la prominencia silábica. Los experimentos involucraban diferentes ensayos de percepción de enunciados en sueco con distintos grupos experimentales: un grupo de hablantes del sueco, otro de hablantes del inglés sin conocimiento alguno del sueco y, por último, un grupo de sujetos que debían evaluar la prominencia a partir de los enunciados escritos. Así, en un caso se evaluaba la percepción de los datos con interferencia lingüística, otro sin interferencia lingüística y la expectativa lingüística sin datos acústicos, respectivamente. Los resultados muestran que los hablantes del sueco y del inglés evalúan la prominencia de manera similar, aunque otorgando distinto peso a cada una de las pistas acústicas, mientras que las evaluaciones de las frases escritas (expectativa lingüística) fueron muy diferentes de las evaluaciones de las frases escuchadas. Los autores identifican así tres factores relevantes en la percepción acústica de la prominencia silábica: el esfuerzo vocal, el tono y la duración. Sin embargo, no logran establecer cuál de estos

tres factores es el más relevante para la asignación de la prominencia. Las diferencias entre los resultados de las evaluaciones de los oyentes ingleses y de los suecos y las diferencias entre los que tuvieron acceso a los datos acústicos frente a los que no, indican que la información y las expectativas lingüísticas juegan un rol fundamental en la percepción de la prominencia y a veces esta percepción no tiene en cuenta los factores acústicos reales, como las influencias del procesamiento descendente (conocido comunmente como procesamiento *top-down* [35]).

Como parte de este mismo trabajo de tesis, por otra parte, en [137] se describe el estudio integral realizado sobre estos fenómenos —reconocimiento de la modalidad, identificación del foco y percepción de la prominencia silábica—, pero para el alemán. Así, como parte del trabajo se exploró hasta qué punto la información lingüística como la modalidad, el foco y la prominencia silábica pueden ser extraídas de los rasgos prosódicos de un enunciado. Primeramente, se analizaron los rasgos prosódicos de diferentes enunciados que incluían dos modalidades (interrogativa y declarativa) y tres focos (neutro, sujeto y predicado). Se observaron la *F<sub>0</sub>*, la intensidad, la duración y calidad de voz y se los relacionó con las evaluaciones subjetivas de prominencia hechas por los sujetos. Se encontró que las partes focalizadas tienen rangos de *F<sub>0</sub>* expandidos y un aumento de la intensidad y duración, y que las estructuras no focalizadas presentan el efecto contrario. Los resultados perceptuales parecen no coincidir enteramente con los rasgos acústicos. El foco amplio no fue correctamente identificado. Los sujetos elegían un foco por defecto —el último, lo que podría explicarse nuevamente con información lingüística. No se contaba con estudios para el español que se propusieran analizar estos fenómenos (modalidad enunciativa, foco y prominencia silábica) sistemáticamente, de manera conjunta y tanto en percepción como en producción, por lo que se propuso seguir las mismas líneas de investigación que se iniciaron para el alemán, en español rioplatense. En este contexto, se consideró al foco como el fragmento del enunciado producido por el hablante con mayor relevancia prosódica y prominencia como la percepción subjetiva de esta producción.

De acuerdo a lo anterior, se definieron los siguientes objetivos, que se desarrollan con el diseño experimental detallado más abajo:

1. Determinar hasta qué punto los oyentes pueden reconocer la modalidad de un enunciado únicamente a partir de pistas prosódicas
2. Investigar la identificación correcta del foco por parte de los oyentes de acuerdo a pistas prosódicas
3. Observar la percepción subjetiva de la prominencia silábica en cada una de las condiciones focales

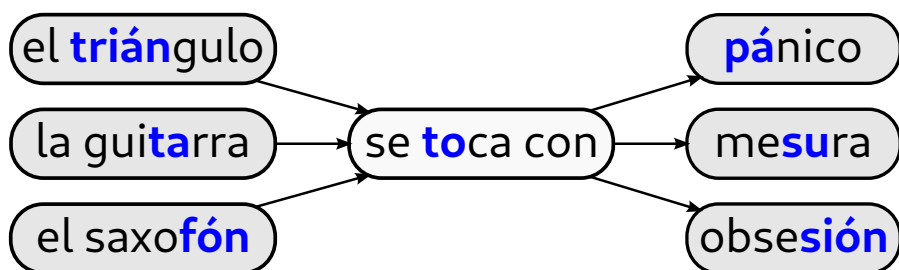


Figura 6.1: Estructura de las oraciones y sílabas acentuadas

4. Correlacionar todas estas medidas subjetivas con los datos acústicos de las producciones orales (Fo, duración, intensidad)
5. Entrenar clasificadores para evaluar la detección de sílabas prominentes en oraciones, a partir de los atributos acústicos definidos anteriormente

### 6.3.1 *Diseño experimental*

A continuación se resume el diseño experimental diseñado y ejecutado para la recolección de datos de foco y prominencia en habla. En el [Apéndice B](#) pueden verse más detalles sobre el diseño de la prueba.


#### *Corpus*

Se utilizaron oraciones con la estructura: sujeto (SN) y predicado (V+SP). El núcleo del SN y del SP presentaban variaciones en el acento léxico, que podía ser oxítono, paroxítono o proparoxítono. La combinación de cada uno de estos sintagmas generó un total de nueve oraciones básicas (ver [Figura 6.1](#)). Las oraciones pertenecen al *corpus* AMPER (Atlas Multimedia de la Prosodia del Espacio Románico) en sus modalidades declarativa e interrogativa absolutas[73]. Se les pidió a diez hablantes nativos de español, cinco mujeres y cinco hombres, sin entrenamiento y hablantes de la variedad rioplatense, sin problemas en su aparato fonatorio, que reprodujeran las oraciones anteriores con distintas variantes, lo que quedó grabado en una sala no tratada acústicamente con un micrófono dinámico. Los hablantes emitieron tres repeticiones para cada una de las nueve oraciones a una velocidad de habla semi-rápida. Los hablantes fueron instruidos sobre el objetivo de la prueba con la indicación del tipo de foco que debía generar. Los tipos solicitados fueron: sin foco o foco neutro, foco estrecho en la palabra de contenido del sujeto (triángulo, guitarra y saxofón) y foco estrecho en la palabra de contenido del predicado (pánico, medida, obsesión). El *corpus* final a utilizar para el estudio perceptual, acústico y de reconocimiento automático contuvo 540 elocuciones: 9 oraciones  $\times$  3 focos  $\times$  2 modalidades  $\times$  10 hablantes.



**Evaluación de prominencia en oraciones** [ oración 1 de 540 ]

1. Escuche el siguiente audio las veces que lo necesite:



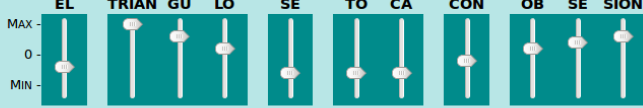
2. ¿Qué tipo de oración cree que es? **[Ayuda]**

3. Seleccione la pregunta que mejor se responda con esta frase

- ¿QUÉ PASA?
- ¿CÓMO se toca el triángulo?
- ¿QUÉ se toca con obsesión?
- ¿el PIANO se toca con obsesión?
- ¿el triángulo se toca con ALEGRÍA?

4. Por cada sílaba seleccione el nivel de prominencia que tiene de acuerdo a lo escuchado. Las sílabas que se destacan tienen un valor positivo, y aquellas débiles un valor negativo. Todos los valores deben estar por sobre el mínimo. **[Ayuda]**

**EL TRIÁN GU LO SE TO CA CON OB SE SIÓN**



[ oración 1 de 540 ]

Siguiente oración >> Guardar y seguir en otro momento

Figura 6.2: Pantalla del sitio web para la obtención de datos de foco y prominencia

### *Diseño de la prueba*

A partir de los audios obtenidos, se diseñó una prueba experimental en una página web (ver diseño en [Figura 6.2](#)) en la que los usuarios debían, en primer lugar, escuchar un estímulo, y luego clasificar la modalidad oracional: declarativa, interrogativa o ambigua en caso que no tenga seguridad sobre la modalidad. Seguidamente, se les pedía que seleccionaran una de cinco preguntas que mejor se respondiera con esa frase (esto tenía que ver con la identificación del foco: foco neutro, en el sujeto, en el predicado o contrastivo) y, por último, se solicitó que para cada sílaba de la oración se señalara el nivel de prominencia en una escala negativa y positiva (que luego quedó registrada entre el  $-5$  y el  $+5$ ) para lo cual se usó una matriz de cursores deslizables siguiendo la idea desarrollada por Eriksson, Thunberg y Traunmüller [43] para el sueco. De esta forma, la marcación de la prominencia subjetiva terminó describiendo una curva para toda la oración, como se aprecia en la sección 4 de la [Figura 6.2](#).

Participaron de este experimento 7 sujetos (cuatro hombres y tres mujeres), todos hablantes nativos de Buenos Aires y de alta escolaridad, que evaluaron los 540 estímulos a lo largo de cinco sesiones, para un total de más de 2 horas de evaluación, con aproximadamente 110 presentaciones por sesión, en forma aleatoria.

En total los sujetos evaluaron 547 oraciones, en varias sesiones, siendo las primeras 7 a modo de práctica del funcionamiento de la herramienta de registro de respuestas.



#### 6.4 DETECCIÓN AUTOMÁTICA DE LA PROMINENCIA

Como está descrito en [74], se implementó un detector automático de prominencia a nivel de sílabas empleando rasgos prosódicos obtenidos de las emisiones acústicas. Así partir de las puntuaciones de prominencia obtenidas mediante la evaluación perceptual, se definió para cada oyente las sílabas prominentes como aquellas en las que se observó un pico en las puntuaciones informadas. Posteriormente, se aplicó un proceso de votación simple para obtener las etiquetas de prominencia de consenso, que conformaron las salidas deseadas del detector de prominencias propuesto. Como observaciones o parámetros de entrada para este detector se utilizaron para describir cada sílaba los parámetros prosódicos propuestos por Rosenberg en [174]: duración en segundos, valores mínimos, máximos y desvío estándar de  $F_0$ , intensidad y énfasis espectral [80]. Para el caso de la frecuencia fundamental ( $F_0$ ), se obtuvo tanto en Hertz como en escala Erb y de semitonos.

Para todos los casos se trabajó con los valores normalizados de todos estos parámetros empleando *z-score* por locutor, con el fin de descartar las diferencias en los rangos de valores usados por cada uno. Una vez obtenidos los vectores de observaciones con sus correspondientes salidas deseadas, se construyeron los conjuntos de entrenamiento y evaluación siguiendo la metodología de validación cruzada de diez particiones, dejando en cada partición todos los datos de un hablante como conjunto de evaluación.

Finalmente, se crearon tres clasificadores usando el programa Weka [226] para construir y evaluar los resultados obtenidos, de forma de cubrir tres de los métodos más comunes y que presentan distintas características: árboles de clasificación (J48) [166], máquinas de vectores soporte con kernel lineal entrenado con el Algoritmo de optimización mínima secuencial, del inglés *Sequential Minimum Optimization (SMO)* [161] y regresión logística [23].

Así, a partir de los mismos datos recolectados anteriormente, en [44] se crearon clasificadores usando otros conjuntos de atributos. A los mencionados anteriormente se agregó el énfasis espectral, que es una medida que refleja la contribución relativa de las altas frecuencias del espectro a la intensidad general, ya que es relevante para la clasificación de categorías prosódicas. [80, 208]

Estos atributos se obtuvieron a través del software Praat [17]. Para el cálculo de la  $F_0$  se siguió el algoritmo de dos pasadas propuesto por Boersma [16]. Adicionalmente, para contemplar las posibles diferencias entre las medidas lineales y perceptuales, se utilizaron las medidas en Hertz, ERBs y semitonos, normalizados por hablante usando *z-score*. La estimación de las intensidades se realizó con el método estándar de Praat, usando un paso de 5ms, y también fueron normalizadas usando *z-score*, de forma de tomar estos valores relativos

a la distribución de cada hablante. Por otro lado, el énfasis espectral se calculó como la diferencia, en dB, entre la intensidad general y la intensidad de la señal luego de aplicar un filtro pasabajos, separando la frecuencia fundamental del resto de los armónicos y, así, obtener una medida normalizada de la energía en las bandas de alta frecuencia, como está descrito en [80].

Finalmente, se usaron como atributos el promedio, el máximo y el mínimo de  $F_0$ , intensidad y énfasis espectral, luego de haber sido normalizadas por *z-score*, por el mismo motivo mencionado anteriormente. A esto se sumó la duración silábica, que también fue normalizada por cada hablante.

Nuevamente se implementaron tres clasificadores citados anteriormente (J48, SMO y Regresión Logística), usando el software Weka, y realizó validación cruzada de 10 partes, por cada hablante del cual se tenían elocuciones. A partir de todos los modelos generados se realizó un promedio del desempeño de cada clasificador, con su desvío estándar.

## 6.5 RESULTADOS Y DISCUSIÓN

### 6.5.1 Modalidad y Foco

Los resultados sobre la percepción de modalidad muestran porcentajes de identificación altos para las oraciones afirmativas (99.59 %) e interrogativas (97.04 %) indicando que las pistas acústicas emitidas son bien reconocidas, como se muestra en el Cuadro 6.1. La modalidad de interrogación muestra más errores, probablemente, debido a la producción interrogativa con un acento bitonal de frase+juntura de tipo retroflejo, típico del español de Buenos Aires. Este efecto puede dar la sensación de afirmación (1.24 %) o ambigüedad (1.72 %) en unas pocas emisiones.

En lo que respecta a la percepción del foco, como se aprecia en el Cuadro 6.2, se ve una tendencia a ubicar a este en el objeto. En particular, las oraciones de foco neutro fueron percibidas como con foco en el objeto en un alto porcentaje (76.40 %). Aquí podemos especular que los oyentes al recibir las oraciones declarativas de foco neutro, aun sin indicadores acústicos específicos, presuponen que la información de carácter pragmático se completa habitualmente en el predicado. En particular, las oraciones interrogativas de foco neutro reciben, además, rasgos acústicos que favorecen la percepción del foco en el predicado. En las estructuras sujeto-verbo-objeto, prototípicas en lenguas como el español, se asume que el predicado es más informativo que el sujeto, es decir, se reconoce como información no presupuesta o no conocida. Si la información prosódica no indica lo contrario, esa es la estructura no marcada y por ello, el oyente reconocerá al predicado como la información nueva que está focalizada temáticamente [238]. En los

Modalidad pretendida	Modalidad percibida [%]		
	Declarativa	Interrogativa	Ambigua
Declarativa	99.59	0.29	0.12
Interrogativa	1.24	97.04	1.72

Cuadro 6.1: Prominencias. Identificación de modalidad de la oración

Foco pretendido	% Foco Percibido por la mayoría (Afirmativas)				% Foco Percibido por la mayoría (Interrogativas)			
	Neut.	Suj.	Obj.	Amb.	Neut.	Suj.	Obj.	Amb.
Neutro	<b>13.48</b>	1.12	76.40	8.99	<b>1.14</b>	40.91	45.45	12.50
Sujeto	2.25	<b>92.13</b>	3.37	2.25	0.00	<b>94.38</b>	3.37	2.25
Objeto	0.00	0.00	<b>97.65</b>	2.35	0.00	22.47	<b>75.28</b>	2.25

Cuadro 6.2: Prominencias. Identificación de las condiciones de foco

contornos de Fo se verifica una característica adicional, no contemplada en el análisis del modelo, dada por los alineamientos tonales. En declarativas e interrogativas de foco neutro, los acentos tonales se desplazan sistemáticamente de las sílabas acentuadas lexicalmente en el sujeto y objeto a la siguiente sílaba.

Como se puede ver en la [Figura 6.3](#), en elocuciones declarativas con foco en el sujeto y en el objeto, el acento tonal muestra una clara alineación con la sílaba acentuada correspondiente, aun cuando el comando no posea la máxima energía. La información acústica de foco estrecho en interrogativas no sigue el mismo patrón. En este caso los acentos tonales se desplazan en un número importante a la sílaba siguiente, como ocurre para el foco neutro. Estos resultados permiten afirmar que la detección de foco basada únicamente en Fo o en los comandos del modelo de entonación no son suficientes como indicadores acústicos. La alineación de los máximos de Fo son medidas complementarias para la determinación del foco en el sujeto y el objeto en oraciones declarativas, pero no para las interrogativas, que presentan la misma desalineación que las oraciones de foco neutro. La percepción de las prominencias indica una marcación de niveles decrecientes para las sílabas con foco en el objeto para declarativas, en el sujeto para interrogativas, en el objeto para interrogativas y para el

	J48	SMO	REG. LOG.
Promedio	88.24 %	90.60 %	90.78 %
Desvío Estándar	3.57 %	3.11 %	3.43 %

Cuadro 6.3: Tasas de detección de prominencias silábicas promedio para tres clasificadores

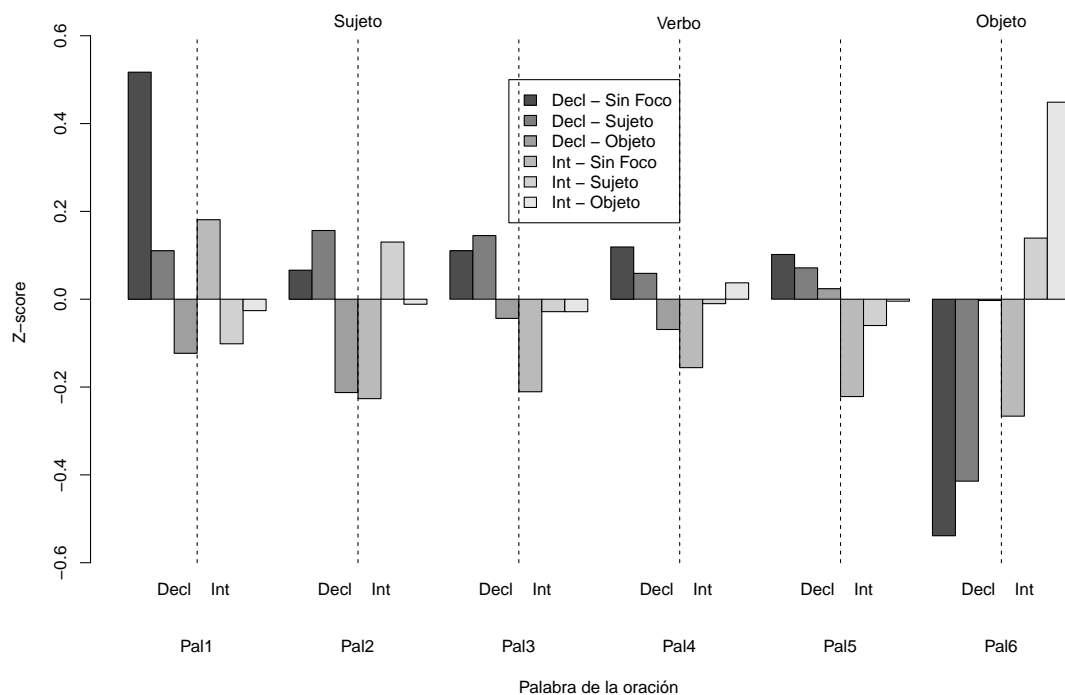


Figura 6.3: Prominencia normalizada con z-score por evaluador, de acuerdo a la posición en la oración, por modalidad y foco percibido

sujeto en declarativas respectivamente. Las prominencias percibidas en cada sílaba no están asociadas directamente con la palabra que se indicó con foco. Sin embargo, el valor de prominencia máxima observada en el objeto resulta ser un buen discriminador de la marcación de foco en todas las condiciones.

Puede especularse como hipótesis futura que la información de foco se encuentra indicada por la prominencia al final de la frase, al igual que la modalidad. Cuando se emplea la información acústica completa de las vocales acentuadas lexical y tonalmente, la detección de las prominencias es satisfactoria como se indica en el Cuadro 6.3.

### 6.5.2 Detección de prominencias con y sin contexto

Para poder evaluar adecuadamente a cada clasificador binario que se entrenó, y tener en cuenta no sólo la tasa de aciertos (en este caso, la detección correcta de un segmento prominente) o de errores, sino también tener la información de cuántos de los segmentos con prominencia son detectados, se optó por usar la medida F [125]. Esta es la media armónica entre la precisión (en inglés, *precision*) y la exhaustividad (en inglés, *recall*), como se indica en las ecuaciones 6.1, 6.2 y 6.3. Allí los positivos verdaderos representan aquellos segmentos de habla clasificados como prominentes, y que efectivamente lo eran; los falsos negativos son aquellos casos en que no se detectó prominencia, pero sí la había (también llamados “errores de Tipo I”); los falsos positivos

Atributos	Algoritmo		
	J48	SMO Medida F	Reg. Log
Fo_ST	0.71 (0.09)	0.71 (0.02)	0.71 (0.02)
Fo_Erb	0.74 (0.02)	0.71 (0.02)	0.75 (0.02)
Fo_Hz	0.74 (0.03)	0.71 (0.02)	0.76 (0.02)
Fo_completo	0.76 (0.03)	0.71 (0.02)	0.76 (0.02)
Energía	0.77 (0.04)	0.72 (0.04)	0.78 (0.05)
ÉnfasisEspec	0.77 (0.03)	0.71 (0.02)	0.78 (0.04)
Duración	0.84 (0.03)	0.84 (0.03)	0.85 (0.03)
Todos	<b>0.88</b> (0.03)	<b>0.90</b> (0.03)	<b>0.91</b> (0.03)

Cuadro 6.4: Resultados de la detección de prominencia de acuerdo los conjuntos de atributos utilizados, en promedio y desv. est. de la medida F

Contexto		Algoritmo		
		J48	SMO Medida F	Reg. Log.
Prev.	Post.			
0	0	0.88 (0.03)	0.90 (0.03)	0.91 (0.03)
0	1	0.89 (0.03)	0.92 (0.03)	0.92 (0.03)
1	0	0.90 (0.03)	0.93 (0.03)	0.93 (0.03)
1	1	0.91 (0.03)	0.94 (0.03)	0.93 (0.03)
1	2	<b>0.92</b> (0.02)	0.94 (0.02)	<b>0.94</b> (0.03)
2	1	0.91 (0.02)	0.94 (0.03)	0.93 (0.03)
2	2	0.91 (0.03)	<b>0.95</b> (0.03)	<b>0.94</b> (0.03)

Cuadro 6.5: Resultados de la detección de prominencia de acuerdo a la información de contexto silábica, en promedio y desv. est. de la medida F

son aquellos casos donde se predijo prominencia, pero en la realidad no la había (también “errores de tipo II”); el total de predicciones son todas las clasificaciones hechas; y el total de positivos es la cantidad de muestras que efectivamente habían sido etiquetadas como prominentes. De esta forma, valores de la medida F cercanos a 1 indican que las prominencias detectadas son efectivamente ciertas, y que se puede detectar todas las prominencias identificadas en las oraciones de prueba.

$$\text{medida-F} = 2 \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (6.1)$$

$$\text{precision} = \frac{\text{positivos verdaderos}}{\text{total predicciones}} = \frac{\text{positivos verdaderos}}{\text{positivos verdaderos} + \text{falsos positivos}} \quad (6.2)$$

$$\text{recall} = \frac{\text{positivos verdaderos}}{\text{total de positivos}} = \frac{\text{positivos verdaderos}}{\text{positivos verdaderos} + \text{falsos negativos}} \quad (6.3)$$

Como se aprecia en el [Cuadro 6.4](#), cuando se toma sólo la información de la sílaba a ser analizada, la duración es la que obtiene los mejores resultados, seguido de la energía, énfasis espectral y, por último, los distintos valores de Fo. Con respecto a esto último, tanto la Fo en escala lineal como perceptual dan los mismos resultados, y todos los atributos de Fo permiten obtener un mejor resultado.

Teniendo en cuenta que el foco y la prominencia son fenómenos de contraste [39], también se analizó el aporte de agregar la información de las sílabas a la derecha y la izquierda de la que está en evaluación. Así, para detectar la prominencia de una sílaba, se agregaron todos los atributos de cero, una o dos sílabas adyacentes, a la derecha o a la izquierda: (0,0) para denotar que no se toma ninguna sílaba adyacente, (1,0) para indicar que se toma la información de una sílaba a la derecha, (1,1) para incluir la información de una sílaba adyacente a la derecha y otra a la izquierda, y de la misma forma con el resto de los casos.

Estos resultados se muestran en el [Cuadro 6.5](#). Allí se aprecia que los mejores resultados se obtienen con SVM de kernel lineal (SMO), cuando se incluye la información de las dos sílabas anteriores y las dos posteriores. Esto era de esperarse, ya que es habitual que se obtenga un mejor desempeño con más información de contexto, y más aún en casos como la prominencia, donde hay información que se comunica a partir del contraste de una unidad con las unidades alrededor de la misma.

Complementariamente a lo anterior, se buscó ver el impacto concreto de variar el contexto que se toma para detectar la prominencia de una sílaba para los distintos atributos (todas las medidas de  $F_0$ , duración y energía), como se resume en la [Figura 6.4](#). Allí se aprecian los valores promedio y desvío estándar de la Medida-F, por medio de una validación cruzada de 10 partes, para los distintos atributos.

En lo que respecta a la **duración**, cuyos resultados se ven en la parte a) de la [Figura 6.4](#), al comparar el desempeño sin contexto y con un contexto de una sílaba — i. e., (1 – 0) y (0 – 1), se encontraron diferencias estadísticas significativas ( $p < 0,05$ ) en una prueba T de a pares entre las variantes (0 – 0) y (0 – 1) para dos de los tres clasificadores (J48 y SMO), mientras que no encontraron diferencias significativas entre (0 – 0) y (1 – 0). Asimismo, se comparó el desempeño de (1 – 1) con (1 – 2) y (2 – 1), donde se encontraron diferencias significativas entre (1 – 1) y (1 – 2) para los tres clasificadores, mientras que no hubo diferencias para (1 – 1) con (2 – 1).

Para la **energía**, cuyos resultados se ven en la parte b) de la [Figura 6.4](#), se aprecian notables diferencias al incorporar información de contexto. Así, existen diferencias significativas entre (0 – 0) y (0 – 1) para todos los clasificadores, mientras que al comparar (0 – 0) con (1 – 0) se prueba la significancia para SMO y Regresión Logística, pero no para J48.

En lo que tiene que ver con otros atributos de la **Frecuencia Fundamental (Fo)**, cuyos resultados se ven en la parte c) de la [Figura 6.4](#), nuevamente se aprecian notables diferencias al incorporar información de una sílaba anterior. Así, se prueban diferencias significativas entre (0 – 0) y (1 – 0) para todos los clasificadores, mientras que al comparar (0 – 0) con (0 – 1) se prueba la significancia sólo para Regresión Logística y J48.

Más arriba se pudo ver que, considerados individualmente, la duración y la energía son claves para la detección de prominencia en una sílaba si se consideran sólo sus atributos propios, como ya se había visto para el inglés [107, 174, 191]. Sin embargo, la importancia de la Frecuencia Fundamental ( $F_0$ ) crece a medida que se incorpora información contextual de otras sílabas, como ya se reportó para el español de España [122]. En este caso se detectó un comportamiento distinto con respecto a la energía y la duración, ya que en estos casos se obtuvieron mejores resultados si se incorporaba más información de sílabas subsiguientes, mientras que para la  $F_0$  era más importante la información de al menos una sílaba anterior. Sin embargo, estos resultados deberán ser puestos a prueba, con otros *corpora*, fuera del contexto de laboratorio, con más datos y con textos más variados.

Por otro lado, también se encontró que el énfasis espectral sirvió por sí sólo para la la detección de prominencias, aunque no con el mismo desempeño que duración, energía y  $F_0$ .

En cuanto al desempeño general, se obtuvo el mejor resultado para la Medida-F con 0.95 (con una precisión de 94.75 %), para el caso Máquinas de Vectores de Soporte (SVM) con kernel lineal, tomando la información de dos sílabas antes y después de la que está siendo evaluada, y usando todos los atributos disponibles. Más allá del alto desempeño alcanzado, que es superior a lo que se indica en otros reportes de la bibliografía, es importante tener en cuenta las características limitadas y poco variadas de las estructuras de las oraciones utilizadas en este estudio. Sin embargo, es muy promisorio con vistas a trabajos futuros.

## 6.6 CONCLUSIONES

A partir de los análisis realizados anteriormente, se vio que la indicación de foco no está asociada directamente con el nivel de prominencia percibida en el segmento focalizado. El valor de prominencia máxima en el objeto resulta ser el indicador que discrimina con significancia estadística la marcación de foco en todas las condiciones. Los máximos de Fo y los comandos del modelo de entonación se asocian mejor con los niveles de prominencia. Los parámetros acústicos directos como los máximos de Fo, la energía en dB, la duración y el espectro vocálico se complementan para dar la información acústica de prominencia, como se verifica al emplear estos datos en el sistema de clasificación automático. Se concluye en este trabajo que el foco estaría marcado por el nivel de prominencia en el objeto y que los niveles de prominencia se asocian con los componentes acústicos con una precisión cercana al 90 %.

Puede especularse como hipótesis futura que la información de foco se encuentra —al igual que la modalidad— indicada por la prominencia al final de la frase. Cuando se emplea la información acústica completa de las vocales acentuadas lexical y tonalmente, la detección de las prominencias es satisfactoria como se indica en el [Cuadro 6.3](#).

En lo que respecta al análisis detallado en la construcción de clasificadores para la detección de prominencias, se encontró que primero la duración y luego la energía permitían obtener mejores resultados si sólo se miraba la información de la sílaba evaluada. Para el caso en que se toma información de una o más sílabas precedentes, la Fo toma los primeros lugares como pista para la detección de prominencia, lo que puede explicarse en que habitualmente los cambios de Fo están antes o después de la sílaba prominente, como se reportó en [72, 74, 137], lo que no queda registrado sólo con la información de Fo máxima y mínima si se mira una sola sílaba.

Adicionalmente, se encontró que todos los atributos en conjunto permiten llegar a un mejor desempeño que cada uno visto en forma



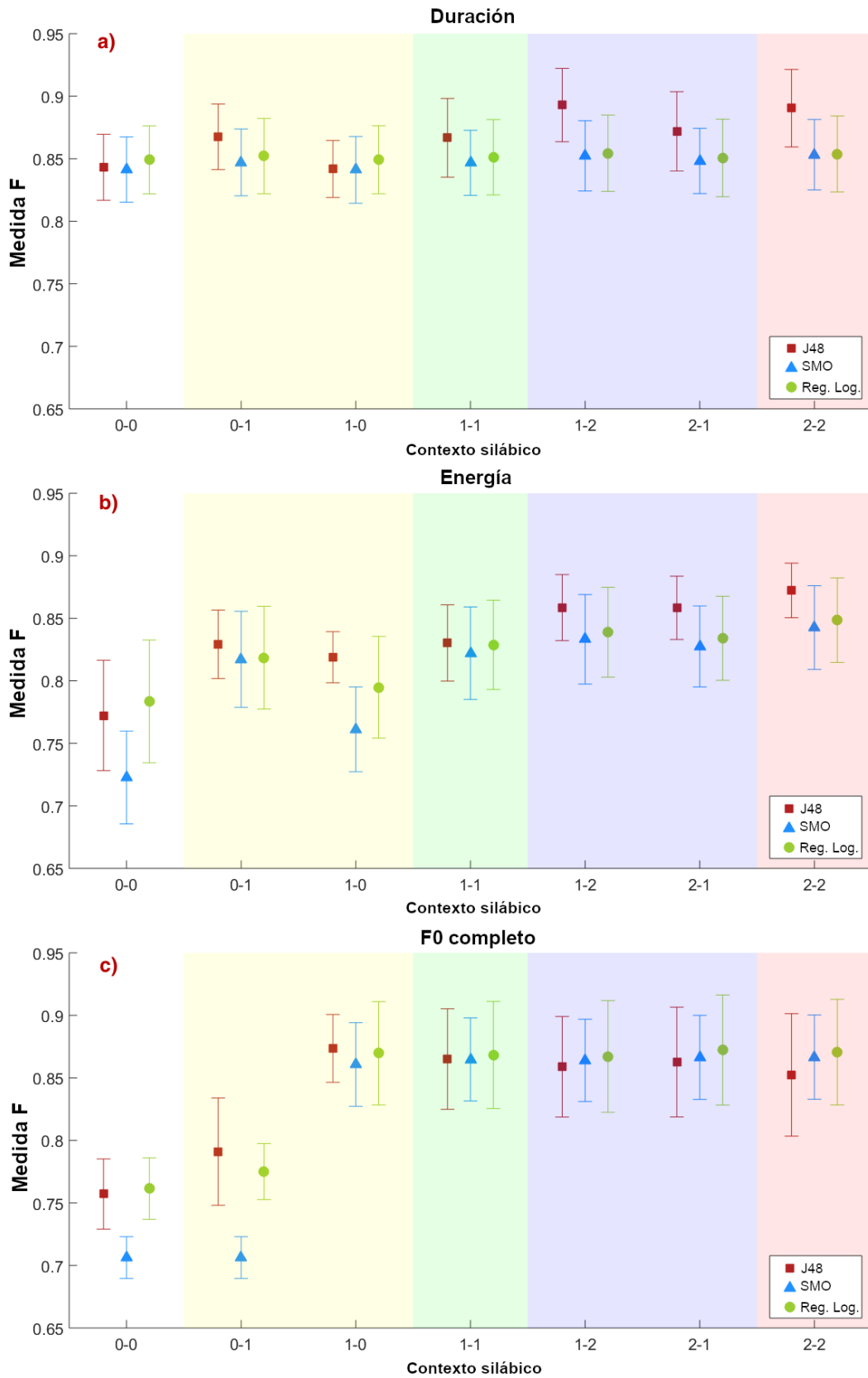


Figura 6.4: Medida F (Prod y Desv. Est) para cada clasificador, atributos y sílabas de contexto. a) Duración b) Energía c) Atributos de Fo

individual, lo que denota que hay información complementaria entre los atributos utilizados.

Finalmente, se estudió el efecto de agregar información contextual silábica para la tarea de detección, lo que permitió que mejorara el desempeño de los clasificadores. Se apreció un mayor aporte informativo de las sílabas posteriores, aunque, en cualquier caso, mejora el funcionamiento al agregar información de ambos lados de la sílaba que está siendo evaluada.

El mejor desempeño en la tarea de detección de las sílabas prominentes se obtuvo utilizando Máquinas de Vectores de Soporte (SVM), con kernel lineal (SMO), para una Medida-F de 94.75% utilizando dos sílabas a cada lado de la que está siendo evaluada. Sin embargo, estos resultados deberán ser puestos a prueba en contextos sintácticos más variados, y con oraciones de mayor longitud, de forma de que puedan utilizarse como referencia para la detección de prominencias en oraciones de distintas variantes del español. Con respecto a esto último, es necesario realizar este mismo estudio utilizando variantes de otras regiones del país, de forma de que puedan obtenerse aquellos atributos que mejor permiten detectar la prominencia en este contexto más general.

Finalmente, el contar con clasificadores confiables y simples para la detección de prominencias puede ser de importancia para su utilización en la evaluación de habla artificial, aunque también puede ser de utilidad para el análisis del habla producida por personas que están aprendiendo una otra lengua [56, 103]. De esta forma, la adecuada marcación de la prominencia y otros atributos prosódicos relacionados con la entonación puede ser utilizada como indicador de la calidad de una elocución, en tanto son factores determinantes para el éxito de la implementación de estas voces (e. g., en el uso para publicidad [173], o en robots asistentes en el ámbito médico [96]), y está asociado a una mayor naturalidad [130].

## CARACTERIZACIÓN DE HABLA ALEGRE Y ENOJADA

---

En este capítulo se resume el trabajo para el diseño, recopilación y estudio preliminar de un *corpus* de habla natural expresiva para el español de Buenos Aires, en el cual se buscó generar emociones de alegría y enojo en los participante por medio de la evocación de recuerdos y pensamientos personales, con el fin de poder caracterizar y evaluar estos aspectos en el habla artificial, y, posteriormente, para su utilización en reconocimiento de emociones y comprensión de habla, entre otras posibles aplicaciones.

### 7.1 CORPUS DE HABLA ALEGRE Y ENOJADA

Por medio de este trabajo se buscó obtener un conjunto de rasgos de la voz a nivel segmental —i. e., atributos de consonantes, vocales y su articulación— que permita caracterizar las variaciones del habla en estado neutral, de enojo y alegría de un mismo hablante del idioma español de Buenos Aires.

De esta manera se podría utilizar esta información para la síntesis de habla expresiva y para la detección de expresividad en el habla y clasificación subsiguiente.

En la actualidad no existen antecedentes del análisis de expresividad en habla espontánea del idioma español, y, en general, se registran pocos antecedentes de habla expresiva para el español rioplatense, por lo que este *corpus* y el análisis preliminar sobre el mismo poseen un valor *per se*.

Según Nordström [148] las características del habla alegre y enojada pueden resumirse de la siguiente manera:

- Alegría intensa: Está relacionada a un aumento del promedio del Pitch (medido a través de la  $F_0$ ) y de la energía de las partes vocalizadas del habla alrededor de los 1000Hz.
- Enojo intenso: Está asociado a un aumento del promedio y desviación estándar del Pitch (medido a través de la  $F_0$ ) y un aumento en la energía en las frecuencias más altas, del formante 3 en adelante, lo que evidencia con un decremento en el índice de Hammarberg [75].

Entre otros atributos, el habla enojada y la alegre pueden diferenciarse porque se codifican usando distintas frecuencias del espectro. Para tener en cuenta a esa diferencias, se usaron los coeficientes de mel-cepstrum. Sin embargo, se sabe que que no es posible hallar una

diferencia notable entre enojo y alegría moderados [148], como ocurre con la expresión de otras emociones en el habla. Adicionalmente, se desea incorporar nuevos atributos para la detección y generación de habla a los ya registrados [13, 108, 148, 163, 170, 172].

### *Objetivo*

Obtener un conjunto de rasgos de la voz a nivel segmental, que permitan describir y detectar habla expresiva de enojo y alegría, de forma de que puedan ser utilizados para la síntesis y el reconocimiento de habla.

### *Hipótesis*

Así como se puede caracterizar el habla de una persona a través de atributos suprasegmentales — e.g., formantes,  $F_0$ , velocidad del habla, intensidad, pausas, léxico, entre otros—, también existen variaciones a nivel segmental —i.e., características de vocales y consonantes, y su articulación—. De esta manera, los atributos segmentales varían de acuerdo al estado de hablante, las características del mensaje que se desea comunicar y la relación con el oyente.

#### *7.1.1 Diseño del Experimento*

Se armó un corpus de habla espontánea a partir del cual se buscará obtener patrones a nivel segmental que permitan caracterizar habla enojada, alegre y neutral. En el [Apéndice D](#) se incluyen detalles del diseño del experimento.

Se tomaron grabaciones a 10 participantes, 5 hombres y 5 mujeres, de entre 22 y 30 años y con formación universitaria, hablantes nativos del español de Buenos Aires, y sin problemas de audición ni enfermedades neurológicas.

#### *7.1.2 Recolección de datos*

Se realizó una entrevista de una hora aproximada de duración. Para la misma se utilizaron preguntas abiertas, que podrían modificarse ligeramente, o incorporar otras nuevas, a partir de los resultados obtenidos en cada entrevista. Las charlas se grabaron en estéreo en 16khz y 16bits, utilizando un grabador portátil TASCAM DR-08.

## 7.2 DISEÑO DE LA ENTREVISTA

### *Tareas Preliminares*

Tiempo estimado: 10 minutos.

Se le pedirá al individuo que para el día de la sesión lleve consigo una o más fotografías que representen momentos y/o personas importantes de su vida. Al comenzar el experimento se le pedirá al individuo que complete un formulario con los siguientes datos:

- Datos generales (edad, sexo, altura, peso)
- Formación (nivel alcanzado, conocimiento y nivel de otros idiomas)
- Contexto (tiempo y lugar de exposición a distintas lenguas, por ej. casa, trabajo, etc.)
- Salud de su voz (fuma, tiene que forzar la voz habitualmente, patologías previas relacionadas con la voz)
- Preferencias (club de fútbol/basket/rugby favorito, 4 personalidades públicas que más le agradan y 4 que más le desagradan, día favorito de la semana)

La información provista por el usuario será utilizada para enriquecer la entrevista, de forma de tomar esa información para evocar los recuerdos positivos y negativos del usuario, además de para tomar como referencia de habla neutra.

### *Introducción*

Tiempo estimado: 10 minutos.

En la primera parte de la entrevista se realizarán los siguientes pasos:

1. Se le pide al individuo que diga su nombre y edad, cuál es estado físico y de ánimo al momento, qué hizo antes de venir a la entrevista y qué hará después.
2. Luego se le pide describir cómo está compuesta su familia y que actividades realiza cada uno.
3. Se le pregunta sobre qué actividades realiza actualmente y cuáles le gustaría realizar.

### *Núcleo*

Tiempo estimado: 70 minutos.

El núcleo de la entrevista presenta cuatro etapas. Las dos primeras, de inducción de enojo y alegría, que se alternarán de forma equitativa para evitar el *efecto de orden*<sup>1</sup>.

<sup>1</sup> Donde el orden en el que se presentan los estímulos induce cambios en los resultados obtenidos

1. Inducción de enojo: Se le pide al individuo que desarrolle los siguientes puntos.

- Indicar cuáles son las cosas que más le disgustan y lo enojan de su trabajo, ciudad, familia y de personalidades públicas y por qué.
- Describir alguna situación/actividad que lo moleste más y que crea que no se está haciendo nada para su resolución.
- Comentar si votará y si sabe a quién en las próximas elecciones presidenciales, y que comente por qué no elegiría las otras opciones.
- Describir el último hecho que le generó enojo y lo recree con detalles.
- Describir el hecho más exasperante de su vida que recuerde y lo recree con detalles. En caso de disponer de alguna fotografía relacionada, se le pide utilizala para describir la situación.
- Leer dos veces las frases con connotación negativa (grupo 3).

2. Inducción de alegría:

- Se le pide a la persona que indique cuáles son las cosas que más le gustan y lo entusiasman de su trabajo, ciudad, familia y de personalidades públicas y por qué.
- Describir alguna situación/actividad que lo entusiasme más y le genere más alegría.
- Se le pide al individuo que indique con detalles qué haría si ganara 10 millones de dólares en la lotería. Se le pide al individuo que indique cuál es el sueño de su vida y qué haría para alcanzarlo.
- Se le pide al individuo que describa el último hecho alegre de su vida y lo recree con detalles.
- Se le pide al individuo que describa el hecho más alegre de su vida y lo recree con detalles. En caso de disponer de alguna fotografía relacionada, se le pide utilizala para describir la situación.
- Leer dos veces las frases con connotación positiva (grupo 2).

3. Lectura de textos:

- Lectura de las frases con connotación ambigua (grupo 1), en forma neutra.
- Lectura de frases ambiguas (grupo 1), en forma alegre (dos veces).

- Lectura de frases ambiguas (grupo 1), en forma neutra.
- Lectura de frases ambiguas (grupo 1), en forma enojada (dos veces).
- Lectura de frases ambiguas (grupo 1), en forma neutra.

### Cierre

Tiempo estimado: 10 minutos.

De manera de tener más información de tipo neutral, y para terminar la charla conociendo cuál es la percepción del sujeto experimental acerca de cómo se expresan las emociones en el habla, se siguen los siguientes pasos:

1. Se le solicita al individuo que comente si detecta los cambios de voz de la gente relacionadas con su estado de ánimo.
2. Luego se le pide que describa el habla enojada y el habla alegre/entusiasta y brinde ejemplos.
3. Se le pide que evalúe la calidad de las representaciones que hizo de habla enojada y alegre e indique qué cosas le faltaron o sobraron.

### 7.3 ANÁLISIS PRELIMINAR DE DIFERENCIAS ENTRE HABLA ALEGRE Y ENOJADA

El objetivo de este trabajo es hacer una caracterización preliminar del habla alegre y enojada de parte del corpus descripto en la [Sección 7.1](#). Así, se evaluarán los siguientes atributos del habla:

- **Frecuencia Fundamental:** También denominada  $F_0$ , la Frecuencia Fundamental es la frecuencia más baja presente en una señal. En este contexto particular esta característica representa el tono de la voz (o “pitch”). Es proporcional al número de veces que vibran las cuerdas vocales por segundo. Se puede estimar para el caso de la señal de habla, pero para tener la medida más precisa se deben utilizar otros instrumentos de medición como, por ejemplo, un electroglotógrafo, que permite obtener con precisión la frecuencia de la vibración de las cuerdas.
- **Formantes:** Comúnmente denominados  $F_1, F_2, \dots$ , los formantes son bandas de frecuencia donde se concentra la mayor parte de la energía de un segmento temporal de la señal. En términos del habla, estos reflejan la resonancia natural y las modulaciones de los articuladores en el tracto vocal. Así, los formantes proveen información muy valiosa respecto de cómo los sonidos fueron producidos, y en particular sobre la geometría del tracto vocal, por lo que son indicadores de la calidad de la voz.

- **Energía:** La energía total está relacionada con el intensidad con la que habla una persona, y se sabe que para dar énfasis en el habla se utiliza el aumento en la energía, además de otros rasgos. Además, la energía es afectada por el grado de activación general de una persona — e.g., empieza a hablar gritando porque está exaltada—.
- **Jitter:** Es una medida de perturbación, calculada como el promedio de variación de frecuencias de un ciclo a otro, entendiendo un “ciclo” como una ventana de tiempo de longitud definida según el caso. En particular, la presencia del Jitter suele ser percibida como un “temblor” en la voz [54].
- **Shimmer:** Al igual que el Jitter, es una medida de perturbación. La diferencia es que el Shimmer mide la variación de amplitud, en contraposición con la frecuencia. De esta manera, es otra forma de caracterizar la “temblor” en la voz [54].

### 7.3.1 Caracterización del habla con MFCC

Los Coeficientes Cepstrales en las Frecuencias de Mel (o, inglés, *MFCC*) [36] son una representación del espectro de energía de corto plazo de un sonido, obtenido al transformar una señal de audio a través de una serie de pasos cuyo objetivo es el de imitar a la cóclea humana [218]. Estos atributos son utilizados en la caracterización del habla, así como también en la clasificación de música, entre otros sonidos. Es el resultado de hacer la siguiente serie de procesamientos da como producto los coeficientes: Onda > DFT > Log-Amplitude Spectrum > Mel-Scaling > Transforma de coseno discreta > MFCCs

En particular interesan los MFCCs de índices más bajos, pues codifican información acerca de los formantes de la señal, relacionadas a la respuesta frecuencial del tracto vocal.

En el proceso, se tomó la log-amplitud al espectro pues el humano no percibe la amplitud del sonido linealmente, sino logarítmicamente —aunque no estrictamente, sino mas bien como una tendencia—. Luego se lo pasa a Escala Mel [197], que consiste en pasar de una representación lineal de las frecuencias del espectro a una escala perceptual, que es una escala logarítmica de la frecuencia en Hz multiplicada por una constante, como se aprecia en la [Figura 7.1](#). Esta escala, al igual que la log-amplitud, es usada por captar mejor la percepción humana de las diferentes frecuencias del espectro.

En el presente trabajo usaremos los coeficientes mencionados para caracterizar las señales del habla, tratando de encontrar patrones a partir de las diferencias de las distintas emociones.



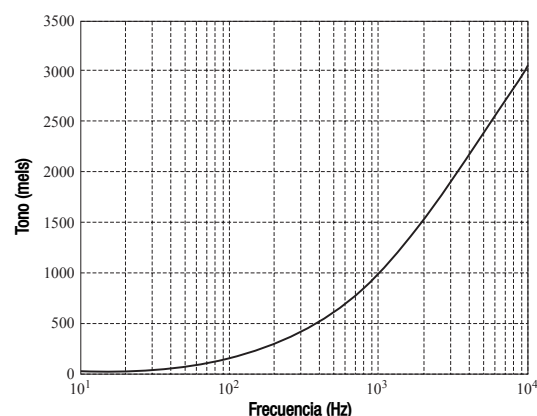


Figura 7.1: Relación de Escala Mel con respecto a las frecuencias

### 7.3.2 Materiales y métodos

El objetivo de este trabajo es caracterizar las emociones del habla enojada y alegre con el objetivo de ver en qué se diferencian —o qué similitudes tienen—. Para eso, teniendo un *corpus* de frases asociadas a sujetos con distintas emociones, se usó la aplicación popular de procesamiento de voz “Praat” para extraer características del habla. Así, se obtuvieron las siguientes características: promedio, desviación estandar, mínimo, máximo, y mediana de  $F_0$ ,  $F_1$ ,  $F_2$  y  $F_3$ ; promedio de los coeficientes de Mel cepstrum 1 al 4; y duración del audio. Esto fue como resultado de considerar el trabajo de GeMAPS [48]. A partir de ahí se hizo un análisis de datos, evaluando si existía la presencia de patrones que pudiesen ayudar a discriminar entre las dos emociones.

### 7.3.3 Corpus utilizado

Se utilizó una parte del *corpus* descrito en la [Sección 7.1](#), en particular, la parte de habla leída con connotación de alegría, enojo y ambigüa, como se describe el [Apéndice D](#).

Para este análisis preliminar sólo se utilizó la información de cuatro sujetos del *corpus*, dos hombres y dos mujeres.

### 7.3.4 Análisis realizados

Se realizó un análisis comparativo del habla alegre y enojada usando las caracterizaciones descriptas anteriormente:

- Frecuencia Fundamental y formantes
- Jitter y Shimmer
- MFCC
- Duración

Para ello, se crearon gráficos comparativos de las grabaciones de las mismas frases (ambiguas), contrastando las métricas obtenidas para las connotaciones alegre, enojada y neutral.

#### 7.4 RESULTADOS

Podemos observar en las Figuras 7.2 y 7.3 la media de  $F_0$  y una correlación con la emoción subyacente de la frase.

Para los atributos de Jitter y Shimmer se encontró una tendencia particular, como se aprecia en la Figura 7.4.

Por otra parte, en los hablantes hombres se encontró que el primer coeficiente de mel-cepstrum presenta una separación entre alegría y enojo notable, lo que no ocurre en las mujeres. Estos resultados pueden verse en Figura 7.5.

Al igual que en los resultados vistos para  $F_0$ , el orden de las emociones se mantiene —i. e., cuál presenta un mayor valor de la métrica analizada—.

Por último, se realizó un análisis de las diferencias de duración promedio para cada sujeto a lo largo de todas sus frases, como se puede ver en la Figura 7.6.

#### 7.5 DISCUSIÓN

Haciendo la extracción de características y el análisis de datos, como se esperaba, pudo verse que la clasificación de emociones del habla es sumamente compleja y se dificulta la discriminación clara entre las emociones estudiadas [117, 218].

##### *Contraste por $F_0$*

Lo primero que se notó fue que, observando la Figura 7.2, resulta destacable que la variación del promedio de  $F_0$  según la emoción cobra una amplitud mucho mayor para los dos sujetos hombres en comparación con las métricas obtenidas para las mujeres. Esto indicaría, en términos de habla, una mayor percepción en cambio de tono entre un habla neutra a una alegre o enojada. Para los hombres es mucho más fácil de percibir, mientras que las mujeres presentan cambios más sutiles en el tono de voz, haciendo un poco más complejo diferenciar el habla alegre de la enojada.

Más aún, esta diferencia es particularmente notable en la Figura 7.3, donde una vez más resulta muy notoria la diferencia promedio entre los promedios de  $F_0$ . Se destaca, además, que no existe necesariamente una tendencia muy notoria entre ambas emociones y el habla neutra.

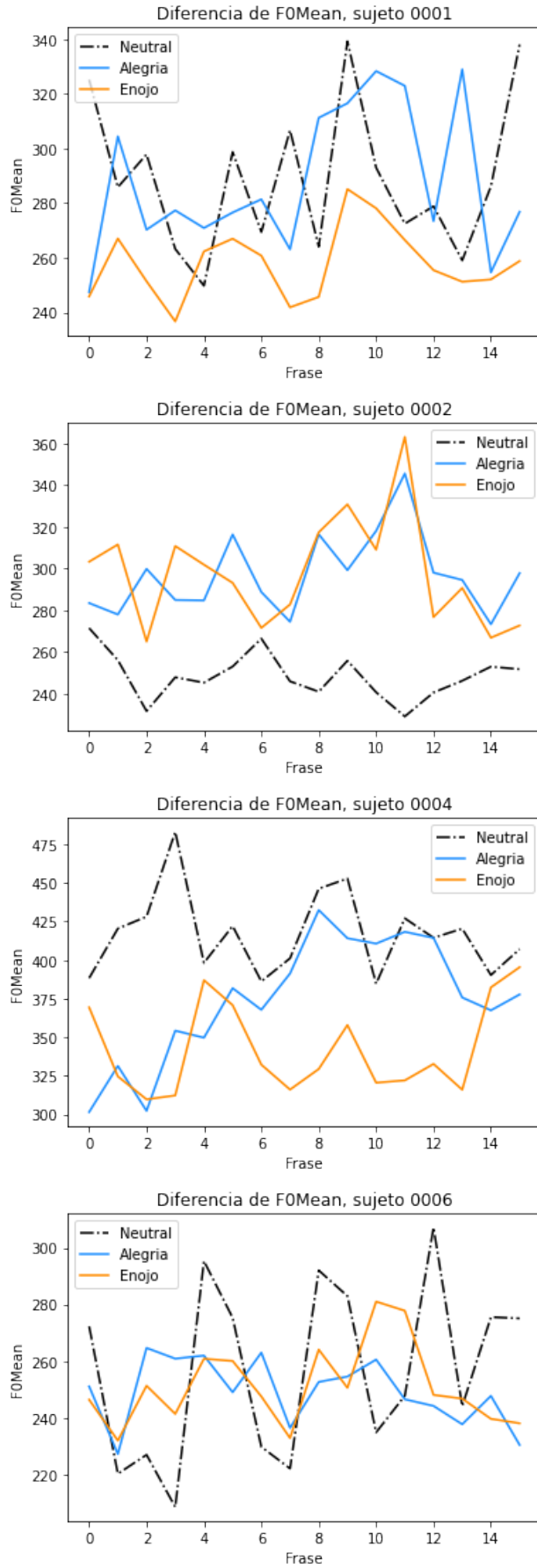


Figura 7.2: Diferencia de Fo promedio para los 4 sujetos

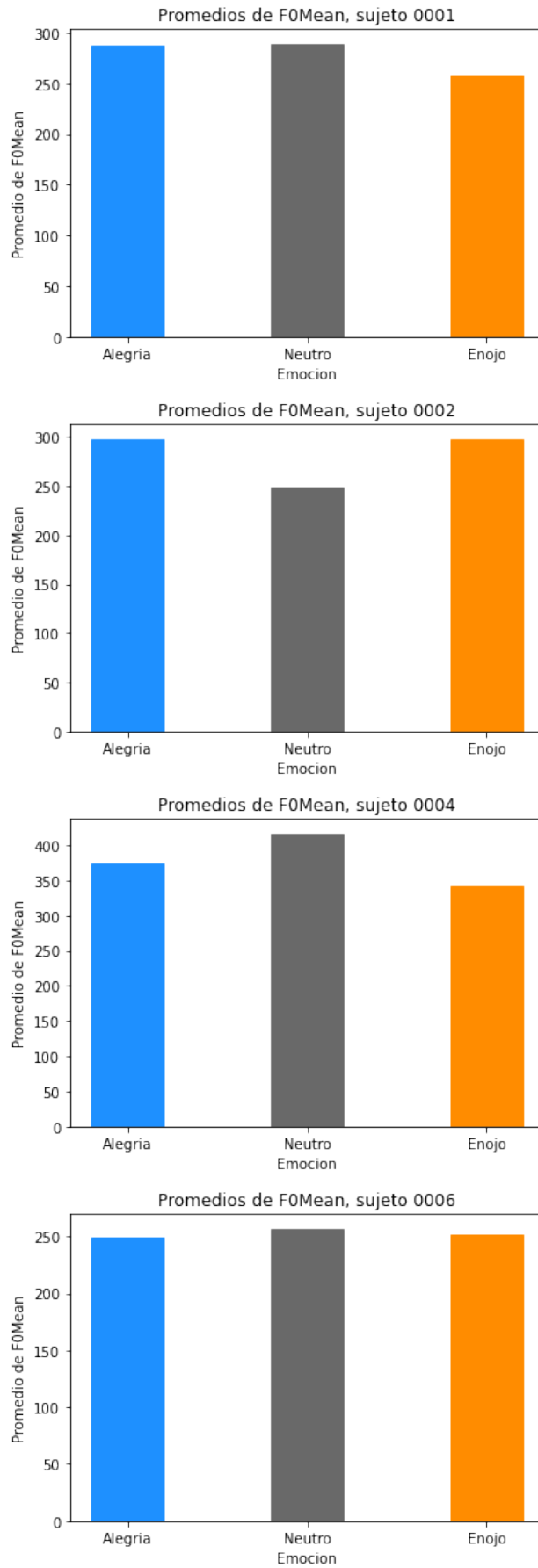


Figura 7.3: Promedio de Fo sobre todas las frases para los 4 sujetos

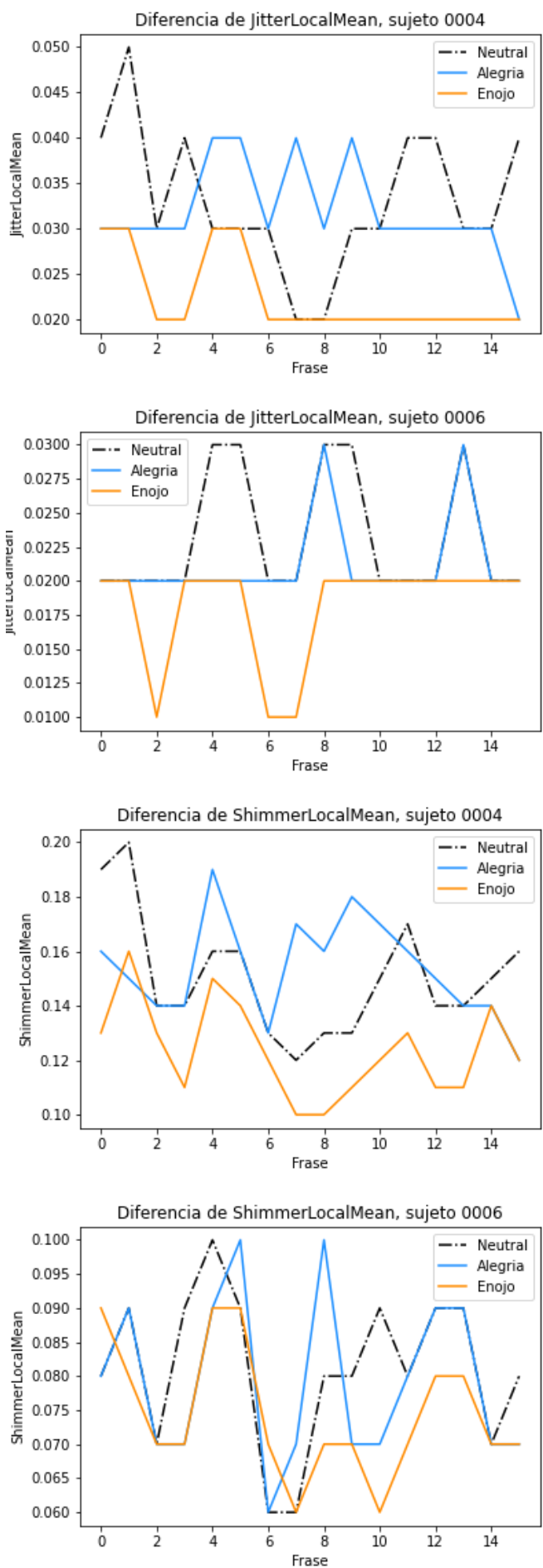


Figura 7.4: Jitter y Shimmer para los sujetos 4 y 6

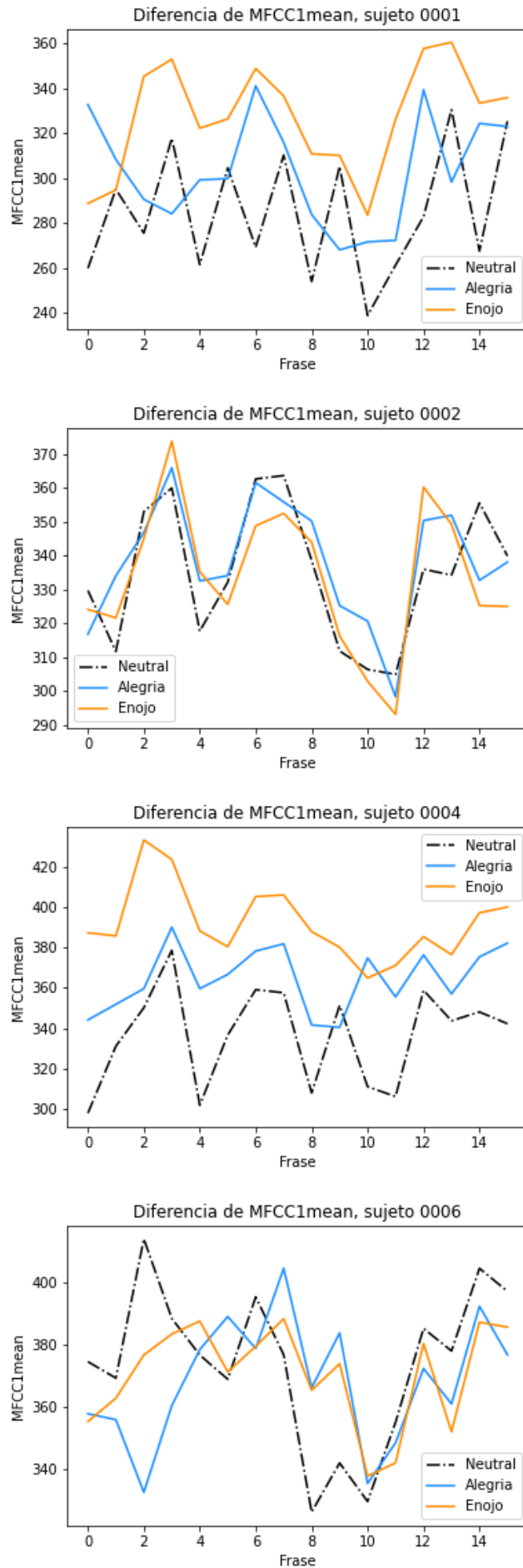


Figura 7.5: Coeficientes de Mel cepstrum para todos los sujetos

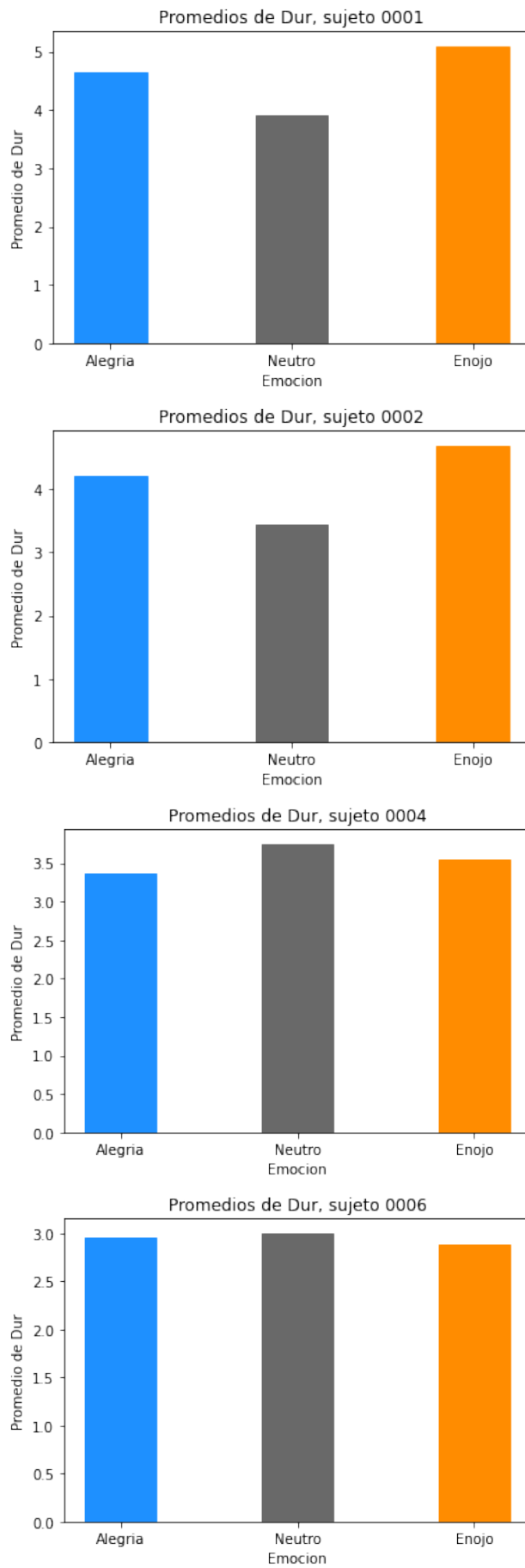


Figura 7.6: Promedio de duración sobre todas las frases para los 4 sujetos

### *Contraste por Jitter y Shimmer*

En cuanto a las métricas de Jitter y Shimmer, se encontró que en ambas los valores para habla alegre suelen estar por encima de las del habla enojada, aunque nuevamente no hay una tendencia marcada para el habla neutra. En particular, se puede ver que el habla enojada suele tener valores de Jitter y Shimmer más bajos en comparación con la neutra —y, por supuesto, con la alegre—. Se puede notar que los sujetos exhibidos en la figura 7.5 son un hombre y una mujer, y los resultados vistos en ellos son representativos del resto de los resultados obtenidos para los otros sujetos.

Es razonable que ambas métricas presenten un comportamiento similar, debido a la manera en la que son obtenidas.

Recordando las implicancias que tienen estos parámetros en la voz de una persona, esto también es esperable. El Jitter y Shimmer están asociados al “temblor” de una voz. Se podría entender, luego, que cuando las personas están enojadas suelen encontrarse más resolutas, decididas. Esto se reflejaría en el habla.

### *Contraste por Mel cepstrum*

Se obtuvieron resultados relevantes a partir de los coeficientes de Mel, análogos a los encontrados para la Fo. Así, se volvió a observar que para los hombres existe una amplitud apreciable entre las métricas para alegría y enojo, aunque esta vez es el enojo la emoción que predomina por sobre el habla neutral y la alegría. Por otro lado, en las mujeres nuevamente se notó que la diferencia se hace más difícil de percibir, al igual que lo que se vio para Fo, lo que puede apreciarse en la [Figura 7.5](#).

### *Contraste por duración*

Finalmente, se analizaron los resultados para las duraciones de los audios, en particular para el promedio por emoción, y como se aprecia en la [Figura 7.6](#), la duración de una frase puede verse afectada por la emoción subyacente en el habla de una persona. Esto puede percibirse, por ejemplo, como mayor velocidad en el habla de la persona que está bajo la emoción.

Cabe destacar que en este análisis no se pudieron obtener resultados concluyentes ya que, como se ve en la figura anterior, no se encuentra un patrón. Esto podría explicarse debido a que no todas las personas hablan de la misma manera, y es muy posible que una misma persona presente formas diferentes de expresarse según la emoción. Por lo tanto, es posible que se encuentre variabilidad entre personas para una característica tan fácil de manipular como la velocidad del habla.



## 7.6 CONCLUSIONES

Luego de un extensivo análisis por distintas características del habla en una importante cantidad de frases, queda claro que hay algunos atributos más relevantes que otros para la caracterización de habla alegre y enojada. Así, como se esperaba, la  $F_0$  resulta un indicador para detectar habla alegre y enojada. Para el caso particular de esta selección del *corpus*, los hombres parecen presentar mayor amplitud entre los valores de  $F_0$  para habla alegre y enojada. Esto implica que su tono suele encontrarse más sujeto a variaciones según la emoción que para el caso de las mujeres analizadas. Resta analizar a futuro si esto es así debido a cuestiones físicas del aparato fonatorio y del cuerpo masculino o tiene otro origen.

Por otra parte, se vio que la voz alegre suele “temblar” más que la enojada, que generalmente parecería ser más estable, lo que podría explicarse en la firmeza que se tiene en la mayoría de las circunstancias cuando se habla de manera enojada.

En cuanto a los coeficientes de mel-cepstrum, se observó que en su mayoría no presentan resultados concluyentes, destacando únicamente el coeficiente MFCC<sub>1</sub>, para el que se notó un comportamiento muy similar al de la  $F_0$ , exceptuando la inversión de la emoción superior en valores, en este caso el enojo en contraposición de la alegría como ocurría con la  $F_0$ .

Finalmente, en lo que respecta a las duraciones, se aprecia que, si bien es un atributo que se ve afectado por las emociones, además de que está directamente relacionada a la velocidad del habla, no resulta un parámetro de referencia claro para la posible clasificación del habla alegre o enojada. Sin embargo, sería interesante en un futuro investigarlo en mayor detalle, por ejemplo, haciendo un análisis segmental de las duraciones, y haciendo normalización por hablante.

Como se mencionó en el [Capítulo 4](#), la generación de habla expresiva y natural es un problema aún no resuelto, y más aún en variantes fuera del español ibérico. Así, la caracterización del habla alegre y enojada, entre otros tipos de habla expresiva, y el entrenamiento y uso de clasificadores simples y rápidos para su reconocimiento pueden ser de gran utilidad para el análisis de voces artificiales, además de su aplicación en evaluación de segunda lengua (L2) —que guarda algunas similitudes con la evaluación de habla artificial— teniendo en cuenta que el habla natural se tiene como referencia de expresividad, sin por esto querer ‘copiar’ sus características<sup>2</sup>. Además, esta caracterización de habla expresiva tienen un valor de por sí, que se espera profundizar en próximos trabajos.

<sup>2</sup> Esto es de especial interés, debido a que se puede llegar a un escenario de “valle inquietante”, donde el habla artificial se torna poco agradable y extraña al oyente humano, debido a su cercanía con el habla natural [126]



## CIERRE

---

Este capítulo se resumirán los principales Resultados, Conclusiones y Trabajos Futuros del trabajo de tesis de doctorado.

### 8.1 RESUMEN DE RESULTADOS

1. Elaboración de materiales para la evaluación de habla artificial para la variante del español de Buenos Aires, en tanto que los contenidos actualmente disponibles están mayormente para el inglés y el alemán, y, en menor medida, para el español de España.
2. Análisis y descripción detallada del problema de la evaluación y los escenarios de uso para sistemas de habla artificial.
3. Caracterización y detección automática de la prominencia para el español de Buenos Aires.
4. Diseño de una prueba para la evaluación perceptual del habla, que incluye información útil para la posterior evaluación paralingüística (e. g., confianza en la voz). Además, se diseñó otra prueba básica más corta, que es especialmente útiles para evaluaciones rápidas de sistemas, ya que contienen un conjunto mínimo de dimensiones en las que evaluar a los sistemas de voces artificiales.
5. Diseño de un conjunto de características para la evaluación automática de la calidad del habla artificial.
6. Creación de un *corpus* de habla expresiva espontánea y leída.
7. Caracterización preliminar de habla alegre y enojada natural.

### 8.2 CONCLUSIONES

A continuación se presentan las principales conclusiones de este trabajo de tesis:

1. Es necesario crear más recursos para el trabajo con habla para las distintas variantes del español de Argentina y otras zonas de la región, tanto por el impacto en el desarrollo de tecnologías asociadas, y, más importante, como parte de protección del acervo cultural de esta región del mundo.

2. El contar con sistemas de generación de habla artificial con variantes locales permitiría contar con sistemas más agradable y usables.
3. Es necesario contar con evaluaciones de calidad que permitan acortar y abaratar los tiempos de desarrollo y ajuste de los sistemas, es especial para contextos como el de nuestra región.
4. El crear pruebas orientadas a la calidad de experiencia de los usuarios, contemplando varios aspectos de la evaluación del habla y no sólo aspectos de la señal, permite obtener evaluaciones más representativas de la idiosincracia de los usuarios, pero sin perder generalidad.
5. Es esencial contar con escenarios de uso claros para los sistemas donde está acotado su contexto de utilización, de forma de orientar el diseño de pruebas y la evaluación de los resultados de esas pruebas.
6. El aporte en recursos para el desarrollo de sistemas de procesamiento de lenguaje natural, entre los que están los sistemas TTS, representan un aporte significativo al desarrollo del área en nuestra región, lo cual puede tener un impacto concreto en la industria de multimedia entretenimiento, en el contexto de proyectos de I+D.
7. La comprensión de cómo funciona la percepción humana del habla es imprescindible para el desarrollo de pruebas de calidad y, finalmente, para el desarrollo de tecnologías que puedan adaptarse a las necesidades de los usuarios.
8. La calidad del habla artificial debe evaluarse desde varios aspectos, tomando en cuenta los escenarios de uso definidos anteriormente.
9. Si se diseñan pruebas más orientadas a la experiencia del usuario, y que representen de manera más variada los posibles escenarios de uso, además de contemplar particularidades propias de las variantes de cada región, se pueden obtener sistemas de conversión de texto a habla de mejor calidad y, finalmente, que sean más utilizados en todo tipo de ámbitos.
10. Es importante agregar atributos paralingüísticos, como la confiabilidad del hablante, además de las clásicas pruebas de inteligibilidad y naturalidad, entre otros, ya que están asociados a mejores resultados en las pruebas con los usuarios.
11. Se pueden definir características del habla artificial para poder realizar la evaluación automática de la calidad. Estas características permiten explicar gran parte de la variación de ciertas

dimensiones de la calidad, y son promisorias para integrarse con otras características del habla para mejorar la predicción de la calidad general de los sistemas TTS.

12. El análisis de la producción de la prominencia y el foco son importantes al momento de evaluar la calidad de una elocución. Luego, su caracterización para nuestra variante del español, así como la construcción de clasificadores para poder reconocerla adecuadamente, constituyen aportes relevantes al área.
13. Se puede detectar la prominencia en habla con buenos resultados, utilizando pocos atributos y contexto silábico (Medida-F  $\approx$  95 %).
14. Es necesario contar con recursos para el estudio de la expresividad en el habla y, en particular, para poder trabajar con la alegría y el enojo, que son útiles para su uso en sistemas de diálogo, entre otros, además de para comprender mejor sus características generales.
15. Es posible caracterizar en forma general al habla alegre y enojada, por medio de la utilidad de pocos atributos acústicos.

### 8.3 POSIBLE TRABAJO FUTURO

1. Implementación en conjunto de los métodos presentados en este trabajo para la evaluación integral de sistemas de conversión de texto a habla.
2. Utilización de las técnicas exploradas en esta tesis para la evaluación del habla de personas que están aprendiendo español u otras lenguas (L2) [102, 115, 219].
3. Caracterización y evaluación de vocalizaciones no verbales en el habla artificial del español (e.g., risas [124], interjecciones [209] y otras expresiones [171]).
4. Evaluación de atributos paralingüísticos de las voces (e.g., actitudes [199]).
5. Profundización en el desarrollo de técnicas para la evaluación automática de la calidad del habla, en particular, aprovechando las grandes bases de datos disponibles y los últimos desarrollos de las técnicas de aprendizaje profundo.
6. Trabajo con estimadores robustos para la predicción de las distintas dimensiones de la calidad del habla artificial, como alternativa para utilización de los promedios, y para contemplar la distribución de predicciones que tiene un mismo sistema, de forma de poder armar un ranking de sistemas para su comparación.

7. Caracterización detallada del habla alegre y enojada natural a partir del *corpus* realizado, continuando la evaluación preliminar.
8. Desarrollo de un clasificador y evaluador de emociones en el habla artificial y natural.
9. Utilización de las técnicas desarrolladas en este trabajo para la evaluación de habla natural, en especial para el estudio de problemas en la producción del lenguaje.
10. Considerar aspectos específicos para ciertas técnicas de TTS del estado del arte, como, por ejemplo, las que utilizan distintos tipos de redes neuronales, que obtienen buenos resultados en las evaluaciones perceptuales [32].



## DISEÑO DE LA PRUEBA DE EVALUACIÓN DE CALIDAD DEL HABLA

---

### A.1 ORACIONES UTILIZADAS

#### A.1.1 Prueba ITU

##### *Tarea 1*

- Señor Di Fiori: el televisor con pantalla 3D, de 52 cm., código: tres uno seis, de 42000\$, se le enviará en una semana.
- Señorita Pérez: la cafetera exprés de acero inoxidable, triple pocillo, con código 5734, de 1235\$, se le enviará en 3 días hábiles.
- Señora Embe: la máquina de coser Sínger, de puntada invisible, código 4197, de 1399\$, se le enviará en 3 horas.

##### *Tarea 2*

- Atención. Aerolíneas Argentinas anuncia que el número de vuelo, seis cuatro tres, con destino a Italia, saldrá a las 12:10hs, de la terminal B. Puerta de embarque 5.
- Atención. American Eralyns anuncia que el número de vuelo, ocho seis tres, con destino a Méjico, saldrá a las 8:30hs, de la terminal C. Puerta de embarque 14.
- Atención. Yapán Eralyns anuncia que el número de vuelo, 1594, con destino a Tokio, saldrá a las 18hs, de la terminal A. Puerta de embarque 12.

##### *Tarea 3*

- Estimado Jorge Pérez. La factura de Telatel del mes de enero, por 214\$, está pendiente de pago. Le pedimos que se acerque a la sucursal Flores para regularizar la situación. Gracias.
- Estimado Joaquín González. La factura de Tututel del mes de febrero, por 453\$, está pendiente de pago. Le pedimos que se acerque a la sucursal Palermo para regularizar la situación. Gracias.
- Estimada Susana Fachineli. La factura de Tatatel del mes de marzo, por 784\$, está pendiente de pago. Le pedimos que se

acerque a la sucursal Belgrano, para regularizar la situación.  
Gracias.

#### A.1.2 *Oraciones prueba SUS*

- El viento dulce armó un libro de panqueques
- Los salames escribían melodías sabrosas
- El insecto francés conduce el elevador
- Saben de su amor por el almíbar con patas
- El avión pintaba los días de melón cocido
- El amor de la fruta de zapatos y la maceta
- El pincel construyó algunos océanos exitosos
- Se vacunan con semillas de llave inglesa
- A las dos se suben los meses colorados
- La receta es con fiambre y bulones de plástico
- Los meses cocinan zapatos de bambú
- Las hojas del perchero querían contar
- Podemos atrasar el camión con chocolate de acero
- Pintaban pieles con acero quirúrgico
- La bicicleta contiene cinco elefantes voladores
- El humo de la estación cantaba muy feliz
- El cuadro es una creación con harina de caqui
- El micrófono de la papa quedaba en Brasil
- Sin barco nos dirigimos con presión a la peluca
- La sal endulzaba la puerta de madera
- El chanco no escribe las pinturas de flan con agua
- Todos sabían que la migraña bailaba con Violeta
- Estamos por comer la galaxia sin honor
- El melón ya sabía la verdad del hielo
- Los militares extranjeros beben un plato de cemento
- Cantó con la botella del tubo fluorescente



- Las flores bebieron un ascensor divertido
- El oso panda estudiaba el rollo de cocina
- El viento amargo armó un libro de maní
- La cama corría maratón con el balero
- El caballo de detergente conduce la heladera
- Tuvieron sueños de miel con madera plastificada
- La piraña cantó con el mate de la biblioteca
- La bolsa mostraba los ojos con alquitrán
- La perra quería dominar al café del mar
- No querían condimentar las uvas de cemento
- Estudiaba el tribunal con pelos del río
- Vamos a comer empanadas de neumáticos
- El cantante toma el corcho con cera paraguaya
- La nieve se enamoró de un enano de polenta
- Las lluvias llegan a carcajadas de un mantel
- Ganaron una casa de humo celeste
- El camión danzaba con luces de miel rugosa
- Exprimieron tubos de acero con azúcar
- El libro chillón cantaba crema de zapatos
- Mezclaron arroz crudo y alambres de papel
- Ellos corrían con signos y choripanes
- Le gustan las películas con pecas y alfajor
- Milanesas con aire otoñal de anillo
- La piedra tocaba la guitarra con las ramas

*Oraciones prueba MOS*

- De cada seis pacientes que se van a hacer un estudio, sólo atienden a dos
- El plantel volvió a entrenar ayer en el Parque General San Martín
- La autora convirtió el material en un éxito de taquilla global
- Si hubo un responsable dentro del gobierno, será sancionado
- En las próximas horas, habría más cambios en el Gabinete
- Y la principal preocupación de los jugadores es zafar de la Promoción
- Este fin de semana, quedó como único puntero del torneo local
- Hay gustos, que se pagan carísimo
- Este no es el momento adecuado para discutir
- Si alguien tiene pruebas, que las presente ante la Justicia
- Los dirigentes del gremio, confían en que la Presidente los recibirá
- Parece que sabían los movimientos de la familia
- El sector de informática, es el nuevo generador de empleo del país
- La propuesta es refinanciar, y así salir de la depresión económica
- La segunda semana fue totalmente exitosa
- Este es un partido clave en la batalla por evitar la Promoción
- Lo que ocurrió aquí, es algo muy terrible
- Esperamos que resulte según lo previsto
- El resto de la escena, se completa en forma virtual
- En los próximos años, se estima que el clima recrudecerá lentamente

## DISEÑO DE LA PRUEBA DE EVALUACIÓN DE PROMINENCIA

---

### B.1 ORACIONES UTILIZADAS

Se utilizaron nueve textos distintos

Se emplearon nueve oraciones declarativas y nueve interrogativas, mediante la combinación en el sujeto y predicado de tres palabras de contenido trisilábicas con acento léxico.

1. El saxofón se toca con obsesión
2. La guitarra se toca con obsesión
3. El triángulo se toca con obsesión
4. El saxofón se toca con medida
5. La guitarra se toca con medida
6. El triángulo se toca con medida
7. El saxofón se toca con pánico
8. La guitarra se toca con pánico
9. El triángulo se toca con pánico

### B.2 INSTRUCCIONES PARA LOS SUJETOS

El objetivo de la evaluación perceptual fue determinar si los oyentes rescatan tanto la información de modalidad como la del foco y obtener los niveles de prominencia prosódica percibida de cada sílaba en las dos modalidades: declarativa e interrogativa. Para la prueba perceptual la instrucción fue *“Usted puede oprimir la tecla de reproducción las veces que lo desee. Luego debe decidir el tipo de modalidad de la oración entre las opciones declarativa, interrogativa o ambigua. Luego debe indicar la pregunta o respuesta que mejor se corresponde con la afirmación o la interrogación escuchada: Por último debe posicionar los cursores para indicar el grado de prominencia de cada sílaba.”* La tarea se desarrolló en 5 sesiones, con aproximadamente 110 estímulos presentados en forma aleatoria. Siete oyentes participaron en la prueba, todos hablantes nativos de Buenos Aires. La tarea demanda 135 minutos en total (15 seg por oración). El uso de una matriz de cursores deslizables para evaluar la prominencia tomó la idea desarrollada por Eriksson, Grabe y Traunmüller [42] para el sueco.

Seguidamente, se les pedía que seleccionaran una de cinco preguntas que mejor se respondiera con esa frase (identificación del foco: foco neutro, en el sujeto, en el predicado o contrastivo) y, por último, se solicitó que para cada sílaba de la oración se señalara el nivel de prominencia en una escala del -5 al +5. La marcación de la prominencia subjetiva generó así una curva para toda la oración. En total evaluaron 547 oraciones, las primeras 7 funcionaron a modo de práctica. Participaron de este experimento 7 sujetos (cuatro hombres y tres mujeres) de alta escolaridad.

## DISEÑO DE LA PRUEBA DE EVALUACIÓN DE CALIDAD DEL HABLA

---

### C.1 ORACIONES UTILIZADAS

#### *Sistemas de Información*

- Señor Di Fiori: el televisor con pantalla 3D, de 52 centímetros, código 3-1-6, de 42000 pesos, se le enviará en una semana.
- Señorita Pérez: la cafetera exprés de acero inoxidable, triple pocillo, con código 5734, de 1735 pesos, se le enviará en 3 días hábiles.
- Señora Embe: la máquina de coser Sínger, de puntada invisible, código 4197, de 1399 pesos, se le enviará en 3 horas.
- Estimado Jorge Pérez. La factura de Telatel del mes de enero, por 214 pesos, está pendiente de pago. Le pedimos que se acerque a la sucursal Flores para regularizar la situación. Gracias.
- Estimado Joaquín González. La factura de Tututel del mes de febrero, por 453 pesos, está pendiente de pago. Le pedimos que se acerque a la sucursal Palermo para regularizar la situación. Gracias.
- Estimada Susana Fachineli. La factura de Tatatel del mes de marzo, por 784 pesos, esta pendiente de pago. Le pedimos que se acerque a la sucursal Belgrano, para regularizar la situación. Gracias.
- Atención. Aerolíneas Argentinas anuncia que el número de vuelo, 6-4-3, con destino a Italia, saldrá a las 12:10 horas, de la terminal B. Puerta de embarque 5. (Entrenamiento)

#### *Noticias*

- La segunda semana fue totalmente exitosa.
- La autora convirtió el material en un éxito de taquilla global.
- El plantel volvió a entrenar ayer en el Parque General San Martín.
- Los trabajadores bancarios de todo el país harán una huelga mañana que paralizará la actividad financiera, en reclamo de un

aumento salarial del 30 por ciento y en protesta por la demora en el inicio de las negociaciones.

- En la ópera su nombre no necesita presentación. Marcelo Álvarez, uno de los grandes tenores del mundo, dueño de una de las voces más bellas y reconocibles de su generación, y protagonista aclamado en las mejores salas líricas.
- Mari es voluntaria en Casa de la Bondad, otra de las obras de la fundación. Allí, ayuda a pacientes que están solos o que no tienen recursos, a atravesar la última etapa de una enfermedad.
- El sector de informática es el nuevo generador de empleo del país. (Entrenamiento)

### *Expresividad*

- Nunca pensé que esto podía estar pasando. Es increíble. (Ambiguo)
- Este premio va dedicado para ustedes, la gente del comité de selección, que siempre pensaron lo mismo de mí. (Ambiguo)
- Este es el momento más alegre de mi vida. Lo soñé y ahora estoy por alcanzarlo. (Alegría)
- Acabo de llegar de una cena con mis ex compañeros de la escuela. Fue sensacional el reencuentro. (Alegría)
- Me cansaron. Les aseguro que no vengo nunca más a este lugar y menos con ustedes. (Enojo)
- Es el sonido más horrible que escuché en mi vida. Como mínimo me van a tener que devolver la entrada. (Enojo)
- Ya estoy preparando las valijas para las vacaciones. En algunas horas más estoy en la playa, con sol y sin preocupaciones. (Entrenamiento)

### *Oraciones Semánticamente Impredecibles (SUS)*

- Los salames escribían melodías sabrosas.
- El insecto francés conduce el elevador.
- El pincel construyó algunos océanos exitosos.
- El viento dulce armó un libro de panqueques.
- El avión pintaba los días de melón cocido.
- El chancho no escribe las pinturas de flan con agua.
- El oso panda estudiaba el rollo de cocina. (Entrenamiento)

## C.2 PRUEBA BÁSICA: INSTRUCCIONES PARA LOS PARTICIPANTES

## BIENVENIDOX A LA PRUEBA INICIAL DE EVALPERCEP2023

¡Gracias por participar de la prueba!"

Ahora va a escuchar distintos audios con voces en español generadas por computadora, y luego de cada una deberá responder algunas preguntas.

Deberá tener en cuenta el audio escuchado, así como el mensaje transmitido y cómo este está pronunciado.

Sólo podrá escuchar una vez cada audio, por lo que le pedimos que preste mucha atención.

**IMPORTANTE:** Regule el volumen al un nivel en el que pueda escuchar todo y se sienta con comodidad.

\* Escriba su nombre de pila:

Primero vamos a comenzar con algunos ejemplos.

Va a escuchar un audio, y luego deberá responder algunas preguntas. Aproveche para regular el volumen a una intensidad que le permita escuchar todo, pero que le sea agradable.

Ahora comenzamos con la prueba.

Va a escuchar un audio, y luego deberá responder algunas preguntas.

Escuche atentamente el siguiente audio porque se reproducirá sólo una vez (!)

### ¿CÓMO EVALÚA LA VOZ ESCUCHADA?

#### \* NATURALIDAD

Escriba un valor entre 1 y 5 para indicar que tan natural le parece la voz escuchada, en relación a la voz humana, donde (1) indica 'Artificial' y (5) 'Natural'

#### \* ENTONACIÓN

Escriba un valor entre 1 y 5 para indicar si la voz escuchada tiene una buena entonación y usa bien las pausas al hablar, comparada con el habla humana común y de acuerdo texto de la oración, donde (1) indica 'Mala entonación' y (5) 'Buena entonación'

#### \* CLARIDAD

Escriba un valor entre 1 y 5 para indicar qué tan bien pronunciado

estaba el mensaje, donde (1) significa 'Poco claro' y (5) 'Muy claro'

**\* VELOCIDAD**

Escriba un valor entre 1 y 5 para evaluar la velocidad del mensaje, donde (1) indica 'Demasiado lento' y (5) que es 'Demasiado rápido'

**\* DEFECTOS**

Escriba un valor entre 1 y 5 para evaluar si encontró algún defecto en el sonido, como ruidos extraños, chirridos, silbidos o interrupciones, donde (1) indica 'Ningún defecto' y (5) que es 'Muchos defectos'.

**\* FAMILIARIDAD**

Escriba un valor entre 1 y 5 para evaluar qué tan familiar le resultó la forma de hablar de la voz, donde (1) indica 'Nada familiar' y (5) que es 'Muy familiar'.

*¿CÓMO ES SU EVALUACIÓN GENERAL DE LA VOZ QUE ESCUCHÓ?*

Escriba un valor entre 1 y 5 para indicar qué le pareció la voz, donde (1) indica 'No me gustó' y (5) 'Me gustó'



## DISEÑO DE CORPUS PARA LA EVALUACIÓN DE HABLA ALEGRE Y ENOJADA

---

### D.1 FORMULARIO A COMPLETAR PREVIO A LA GRABACIÓN

Día y Hora:

Lugar:

Estado de ánimo:

Estado del tiempo:

#### D.1.1 *Datos del Entrevistado*

##### *Generales*

Nombre:

Edad:

Altura y peso:

Educación:

Idiomas (nivel desde 1-inicial a 5-experto):

##### *Contexto*

¿Sus parientes directos hablaban otro idioma/dialecto o tenían algún acento distinto del porteño?

Indique grado y lugar de exposición a otras lenguas (por ej. casa, trabajo, etc.):

##### *De la salud de la voz*

¿Fuma?:

¿Tiene que forzar la voz cotidianamente?:

¿Tuve/tiene problemas neurológicos?:

¿Tuvo/tiene patologías relacionadas con el tracto vocal (boca, cuerdas vocales, lengua, nariz, etc.):

##### *Gustos*

Equipo de fútbol/basquet/rugby/otro deporte favorito:

Nombre las cuatro (4) personas (públicas o no) que más le agradan/entusiasmen:

Nombre las cuatro (4) personas (públicas o no) que más le desagradan/odie:

*Experiencias*

¿Cuándo y cuál fue la última vez en la que se sintió enojado por algo/alguien?

¿Cuándo y cuál fue la última vez en la que se sintió alegre/entusiasmado por algo/alguien?

¿Cuándo y cuál fue la ocasión en la que se sintió más enojado por algo/alguien en toda su vida?

¿Cuándo y cuál fue la ocasión en la que se sintió más alegre/entusiasmado por algo/alguien en toda su vida?

*Consentimiento de utilización de información*

La información de este formulario así como las grabaciones realizadas como parte de este experimento serán utilizados sólo con fines académicos. Las grabaciones completas no serán publicadas sin mi consentimiento, aunque algunos segmentos de las mismas pueden ser utilizadas con fines de demostración y, siempre, resguardando mi identidad.

Fecha y Lugar:

Firma y Aclaración:

**D.2 FRASES DE CONNOTACIÓN ALEGRE, ENOJADA Y AMBIGUA****D.2.1 Grupo 1: Frases con connotación ambigua**

- Nunca pensé que esto podía estar pasando. Es increíble.
- Y ahora va a venir el experto a mostrarnos cómo hacer las cosas.
- Me tomó por sorpresa ¿Pensás que voy a creer que no lo hicieron a propósito?
- Yo en la ciudad, y no hay nadie en las calles hasta el 31 de enero.
- Una hora de trabajo y llevo dos páginas escritas.
- ¿A que no sabés? El arreglo me costó doscientos ventidós pesos.
- Viste. Nadie podía creer que Carlos era capaz de hacerlo.
- Este premio va dedicado para ustedes, la gente del comité de selección, que siempre pensaron lo mismo de mí.

**D.2.2 Grupo 2: Frases con connotación alegre**

- Este es el momento más alegre de mi vida. Lo sñé y ahora estoy por alcanzarlo.
- ¡Felicitaciones! Te ganaste el primer premio del concurso.

- Ya estoy preparando las valijas para las vacaciones. En algunas horas más estoy en la playa, con sol y sin preocupaciones.
- Tengo buenas noticias. Voy a ser papá ¡Y parece que va a ser una nena!
- Acabo de llegar de una cena con mis ex compañeros de la escuela. Fue sensacional el reencuentro.
- No me vas a creer. Ayer salí con la chica que conocimos en el casamiento de Marcelo y quedé flechado.
- Es un genio espectacular ¡Parece que tuviera la pelota atada!
- ¡Por fin! ¡Vamos! ¡Esta vez se nos tenía que dar!

### D.2.3 Grupo 3: Frases con connotación de enojo

- El planteo no tiene ni pies ni cabeza ¿Te parece que puedo estar contento?
- Me cansaron... Les aseguro que no vengo nunca más a este lugar y menos con ustedes.
- Es imposible creer que sigan vendiendo este producto ¡Me siento estafado!
- No, la verdad que no acepto la propuesta ni hoy ni mañana. Espero que esto sea una broma de mal gusto.
- Si hoy no aparecés por el entrenamiento, olvidate que te incluya en la lista para el partido del domingo.
- ¿No te parece que es algo caro dos mil pesos por el arreglo? Me llevo el auto y no te pago nada.
- Buscate otras excusas porque de esta no vas a salir así de fácil. No te creo nada.
- Es el sonido más horrible que escuché en mi vida. Cómo mínimo me van a tener que devolver la entrada.



## BIBLIOGRAFÍA

---

- [1] E. M. Albornoz. «Modelado de estructuras prosódicas para el reconocimiento automático del habla». Tesis de doctorado. Universidad Nacional del Litoral, 2011. URL: <http://hdl.handle.net/11185/442> (vid. pág. 17).
- [2] Amazon. 'Alexa, ¡Hola!': Alexa can now speak Spanish in the U.S. 2019. URL: <https://blog.aboutamazon.com/devices/alexa-hola> (visitado 29-03-2020) (vid. pág. 1).
- [3] J.-N. Antons. *Neural Correlates of Quality Perception for Complex Speech Signals*. T-Labs Series in Telecommunication Services. Cham: Springer International Publishing, 2015, pág. 97. ISBN: 978-3-319-15520-3. DOI: [10.1007/978-3-319-15521-0](https://doi.org/10.1007/978-3-319-15521-0). URL: <http://link.springer.com/10.1007/978-3-319-15521-0> (vid. pág. 35).
- [4] J.-N. Antons, K. U. R. Laghari, S. Arndt, R. Schleicher, S. Moller, D. O'Shaughnessy y T. H. Falk. «Cognitive, affective, and experience correlates of speech quality perception in complex listening conditions». En: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, mayo de 2013, págs. 3672-3676. ISBN: 978-1-4799-0356-6. DOI: [10.1109/ICASSP.2013.6638343](https://doi.org/10.1109/ICASSP.2013.6638343). URL: <http://ieeexplore.ieee.org/document/6638343/> (vid. pág. 35).
- [5] J.-n. Antons, R. Schleicher, S. Arndt, S. Möller, S. Member, A. K. Porbadnigk y G. Curio. «Analyzing Speech Quality Perception Using Electroencephalography». En: 6.6 (2012), págs. 721-731 (vid. págs. 33, 35).
- [6] S. Arndt, J.-N. Antons, R. Gupta, K. ur Rehman Laghari, R. Schleicher, S. Moller y T. H. Falk. «Subjective quality ratings and physiological correlates of synthesized speech». En: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, jul. de 2013, págs. 152-157. ISBN: 978-1-4799-0738-0. DOI: [10.1109/QoMEX.2013.6603229](https://doi.org/10.1109/QoMEX.2013.6603229). URL: <http://ieeexplore.ieee.org/document/6603229/> (vid. pág. 35).
- [7] S. Arndt, J.-n. Antons, R. Schleicher, S. Moller y G. Curio. «Using Electroencephalography to Measure Perceived Video Quality». En: *IEEE Journal of Selected Topics in Signal Processing* 8.3 (jun. de 2014), págs. 366-376. ISSN: 1932-4553. DOI: [10.1109/JSTSP.2014.2313026](https://doi.org/10.1109/JSTSP.2014.2313026). URL: <http://ieeexplore.ieee.org/document/6777327/> (vid. pág. 35).

- [8] M. F. Assaneo, D. Ramirez Butavand, M. A. Trevisan y G. B. Mindlin. «Discrete Anatomical Coordinates for Speech Production and Synthesis». En: *Frontiers in Communication* 4. April (abr. de 2019), págs. 1-13. ISSN: 2297-900X. DOI: [10.3389/fcomm.2019.00013](https://doi.org/10.3389/fcomm.2019.00013). URL: <https://www.frontiersin.org/article/10.3389/fcomm.2019.00013/full> (vid. pág. 16).
- [9] L. Baghai-Ravary y S. W. Beet. *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. Ed. por A. Neustein. SpringerBriefs in Electrical and Computer Engineering. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-4573-9. DOI: [10.1007/978-1-4614-4574-6](https://doi.org/10.1007/978-1-4614-4574-6). URL: <http://link.springer.com/10.1007/978-1-4614-4574-6> (vid. pág. 35).
- [10] A. Baird, E. Parada-Cabaleiro, C. Fraser, S. Hantke y B. Schuller. «The Perceived Emotion of Isolated Synthetic Audio». En: *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion - AM'18*. New York, New York, USA: ACM Press, 2018, págs. 1-8. ISBN: 9781450366090. DOI: [10.1145/3243274.3243277](https://doi.org/10.1145/3243274.3243277). URL: <http://dl.acm.org/citation.cfm?doid=3243274.3243277> (vid. pág. 28).
- [11] R. K. Balasubramanium, J. S. Bhat, M. Srivastava y A. Eldose. «Cepstral analysis of sexually appealing voice.» En: *Journal of voice : official journal of the Voice Foundation* 26.4 (jul. de 2012), págs. 412-5. ISSN: 1557-8658. DOI: [10.1016/j.jvoice.2011.03.011](https://doi.org/10.1016/j.jvoice.2011.03.011). URL: <http://www.ncbi.nlm.nih.gov/pubmed/21724369> (vid. pág. 28).
- [12] N. Barbot, O. Boëffard, J. Chevelu y A. Delhay. «Large Linguistic Corpus Reduction with SCP Algorithms». En: *Computational Linguistics* 41.3 (sep. de 2015), págs. 355-383. ISSN: 0891-2017. DOI: [10.1162/COLI\\_a\\_00225](https://doi.org/10.1162/COLI_a_00225). URL: [https://www.mitpressjournals.org/doi/abs/10.1162/COLI%7B%5C\\_%7Da%7B%5C\\_%7D00225](https://www.mitpressjournals.org/doi/abs/10.1162/COLI%7B%5C_%7Da%7B%5C_%7D00225) (vid. pág. 3).
- [13] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson y N. Amir. «The Automatic Recognition of Emotions in Speech». En: *Emotion-Oriented Systems*. 2011, págs. 71-99. ISBN: 978-3-642-15183-5. DOI: [10.1007/978-3-642-15184-2\\_6](https://doi.org/10.1007/978-3-642-15184-2_6). URL: [http://centaur.reading.ac.uk/30305/%20http://link.springer.com/10.1007/978-3-642-15184-2%7B%5C\\_%7D6](http://centaur.reading.ac.uk/30305/%20http://link.springer.com/10.1007/978-3-642-15184-2%7B%5C_%7D6) (vid. pág. 94).
- [14] C. Benoît, M. Grice y V. Hazan. «The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences». En: *Speech Communication* 18.4 (jun. de 1996), págs. 381-392. ISSN: 01676393. DOI: [10.1016/0167-6393\(96\)00026-X](https://doi.org/10.1016/0167-6393(96)00026-X). URL: <https://linkinghub.elsevier.com/retrieve/pii/016763939600026X> (vid. pág. 33).

- [15] A. W. Black, H. Zen y K. Tokuda. «Statistical Parametric Speech Synthesis». En: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. IEEE, 2007, págs. IV-1229-IV-1232. ISBN: 1-4244-0727-3. DOI: [10.1109/ICASSP.2007.367298](https://doi.org/10.1109/ICASSP.2007.367298). URL: <http://ieeexplore.ieee.org/document/4218329/> (vid. pág. 16).
- [16] P. Boersma. «Acurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound». En: *IEA Proceedings 17 17* (1993), págs. 97-110. URL: [http://www.fon.hum.uva.nl/paul/papers/Proceedings%7B%5C\\_%7D1993.pdf](http://www.fon.hum.uva.nl/paul/papers/Proceedings%7B%5C_%7D1993.pdf) (vid. pág. 83).
- [17] P. Boersma y D. Weenink. *Praat: doing phonetics by computer*. 2021. URL: <http://www.praat.org/> (vid. pág. 83).
- [18] A. Bonafonte, S. Pascual y G. Dorca. «Spanish Statistical Parametric Speech Synthesis Using a Neural Vocoder». En: *Interspeech 2018*. Vol. 2018-Septe. September. ISCA: ISCA, sep. de 2018, págs. 1998-2001. DOI: [10.21437/Interspeech.2018-2417](https://doi.org/10.21437/Interspeech.2018-2417). URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2018/abstracts/2417.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2018/abstracts/2417.html) (vid. pág. 17).
- [19] S. Callesano y P. M. Carter. «Latinx perceptions of Spanish in Miami: Dialect variation, personality attributes and language use». En: *Language & Communication* 67 (jul. de 2019), págs. 84-98. ISSN: 02715309. DOI: [10.1016/j.langcom.2019.03.003](https://doi.org/10.1016/j.langcom.2019.03.003). URL: <https://linkinghub.elsevier.com/retrieve/pii/S027153091830377X> (vid. pág. 3).
- [20] N. Campbell. «Evaluation of Speech Synthesis». En: *Evaluation of Text and Speech Systems*. Dordrecht: Springer Netherlands, 2007. Cap. 2, págs. 29-64. DOI: [10.1007/978-1-4020-5817-2\\_2](https://doi.org/10.1007/978-1-4020-5817-2_2). URL: [http://link.springer.com/10.1007/978-1-4020-5817-2%7B%5C\\_%7D2](http://link.springer.com/10.1007/978-1-4020-5817-2%7B%5C_%7D2) (vid. págs. 31-33).
- [21] F. Campillo Díaz, F. Méndez Pazó, M. Arza Rodríguez, E. Rodríguez Banga, F. M. Pazó, M. Arza Rodríguez y E. Rodríguez Banga. «The GTM-UVigo Systems for Albayzín 2010 Text-to-Speech Evaluation». En: *FALA 2010 (VI Jornadas en Tecnología del Habla)*. 2010, págs. 349-352 (vid. págs. 1, 45).
- [22] E. D. Casserly y D. B. Pisoni. «Speech perception and production». En: *WIREs Cognitive Science* 1.5 (sep. de 2010), págs. 629-647. ISSN: 1939-5078. DOI: [10.1002/wcs.63](https://doi.org/10.1002/wcs.63). URL: <https://onlinelibrary.wiley.com/doi/10.1002/wcs.63> (vid. pág. 23).
- [23] S. L. Cessie y J. C. V. Houwelingen. «Ridge Estimators in Logistic Regression». En: *Applied Statistics* 41.1 (1992), pág. 191. ISSN: 00359254. DOI: [10.2307/2347628](https://doi.org/10.2307/2347628). URL: <https://onlinelibrary.wiley.com/doi/10.2307/2347628> (vid. pág. 83).

- [24] J.-D. Chen y N. Campbell. «Objective Distance Measures for Assessing Concatenative Speech Synthesis». En: *6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*. 1999. URL: [https://www.isca-speech.org/archive/eurospeech%7B%5C\\_%7D1999/e99%7B%5C\\_%7D0611.html](https://www.isca-speech.org/archive/eurospeech%7B%5C_%7D1999/e99%7B%5C_%7D0611.html) (vid. págs. 32, 35).
- [25] Y.-J. Chen, T. Tu, C.-c. Yeh y H.-Y. Lee. «End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning». En: *INTERSPEECH 2019*. Vol. 2019-Sept. ISCA: ISCA, sep. de 2019, págs. 2075-2079. DOI: 10.21437/Interspeech.2019-2730. URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2019/abstracts/2730.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2019/abstracts/2730.html) (vid. págs. 17, 46).
- [26] Y. Chen y col. «Sample Efficient Adaptive Text-to-Speech». En: *ICLR 2019*. Sep. de 2019, págs. 1-15. arXiv: 1809.10460. URL: <http://arxiv.org/abs/1809.10460> (vid. pág. 45).
- [27] T. Chi, P. Ru y S. A. Shamma. «Multiresolution spectrotemporal analysis of complex sounds». En: *The Journal of the Acoustical Society of America* 118.2 (ago. de 2005), págs. 887-906. ISSN: 0001-4966. DOI: 10.1121/1.1945807. URL: <http://asa.scitation.org/doi/10.1121/1.1945807> (vid. pág. 36).
- [28] M. Chinen, J. Skoglund, C. K. A. Reddy, A. Ragano y A. Hines. «Using Rater and System Metadata to Explain Variance in the VoiceMOS Challenge 2022 Dataset». En: *Interspeech 2022*. Vol. 2022-Sept. September. ISCA: ISCA, sep. de 2022, págs. 4531-4535. DOI: 10.21437/Interspeech.2022-799. arXiv: 2209.06358. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/chinen22%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/chinen22%7B%5C_%7Dinterspeech.html) (vid. págs. 55, 74).
- [29] E. K. Chiou, N. L. Schroeder y S. D. Craig. «How we trust, perceive, and learn from virtual humans: The influence of voice quality». En: *Computers & Education* 146.October 2019 (mar. de 2020), pág. 103756. ISSN: 03601315. DOI: 10.1016/j.compedu.2019.103756. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0360131519303094> (vid. págs. 3, 28, 46).
- [30] R. Clark, H. Silen, T. Kenter y R. Leith. «Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs». En: *10th ISCA Speech Synthesis Workshop*. September. ISCA: ISCA, sep. de 2019, págs. 99-104. DOI: 10.21437/SSW.2019-18. arXiv: 1909.03965. URL: [http://www.isca-speech.org/archive/SSW%7B%5C\\_%7D2019/abstracts/SSW10%7B%5C\\_%7D0%7B%5C\\_%7D3-1.html](http://www.isca-speech.org/archive/SSW%7B%5C_%7D2019/abstracts/SSW10%7B%5C_%7D0%7B%5C_%7D3-1.html) (vid. pág. 31).



- [31] R. A. J. Clark y K. E. Dusterhoff. «Objective Methods for Evaluating Synthetic Intonation». En: *6th European Conference on Speech Communication and Technology*. 1999. URL: [https://www.isca-speech.org/archive/eurospeech%7B%5C\\_%7D1999/e99%7B%5C\\_%7D1623.html](https://www.isca-speech.org/archive/eurospeech%7B%5C_%7D1999/e99%7B%5C_%7D1623.html) (vid. págs. 32, 35).
- [32] M. Cohn y G. Zellou. «Perception of Concatenative vs. Neural Text-To-Speech (TTS): Differences in Intelligibility in Noise and Language Attitudes». En: *Interspeech 2020*. ISCA: ISCA, oct. de 2020, págs. 1733-1737. DOI: 10.21437/Interspeech.2020-1336. URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2020/abstracts/1336.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2020/abstracts/1336.html) (vid. pág. 112).
- [33] E. Cooper y J. Yamagishi. «How do Voices from Past Speech Synthesis Challenges Compare Today?» En: *11th ISCA Speech Synthesis Workshop (SSW 11)*. August. ISCA: ISCA, ago. de 2021, págs. 183-188. DOI: 10.21437/SSW.2021-32. URL: [https://www.isca-speech.org/archive/ssw%7B%5C\\_%7D2021/cooper21%7B%5C\\_%7Dssw.html](https://www.isca-speech.org/archive/ssw%7B%5C_%7D2021/cooper21%7B%5C_%7Dssw.html) (vid. págs. 55, 74).
- [34] R. Dall, J. Yamagishi y S. King. «Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation». En: *7th International Conference on Speech Prosody 2014*. ISCA: ISCA, mayo de 2014, págs. 1012-1016. DOI: 10.21437/SpeechProsody.2014-191. URL: [http://www.isca-speech.org/archive/SpeechProsody%7B%5C\\_%7D2014/abstracts/196.html](http://www.isca-speech.org/archive/SpeechProsody%7B%5C_%7D2014/abstracts/196.html) (vid. págs. 5, 46).
- [35] M. H. Davis e I. S. Johnsrude. «Hearing speech sounds: Top-down influences on the interface between audition and speech perception». En: *Hearing Research* 229.1-2 (jul. de 2007), págs. 132-147. ISSN: 03785955. DOI: 10.1016/j.heares.2007.01.014. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0378595507000263%20http://www.ncbi.nlm.nih.gov/pubmed/17317056> (vid. págs. 21, 80).
- [36] S. B. Davis y P. Mermelstein. «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences». En: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (ago. de 1980), págs. 357-366. ISSN: 0096-3518. DOI: 10.1109/TASSP.1980.1163420. URL: <http://ieeexplore.ieee.org/document/1163420/> (vid. págs. 21, 36, 98).
- [37] Doh-Suk Kim. «ANIQUE: an auditory model for single-ended speech quality estimation». En: *IEEE Transactions on Speech and Audio Processing* 13.5 (sep. de 2005), págs. 821-831. ISSN: 1063-6676. DOI: 10.1109/TSA.2005.851924. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1495466%20http://ieeexplore.ieee.org/document/1495466/> (vid. pág. 35).

- [38] Doh-Suk Kim y A. Tarraf. «Enhanced Perceptual Model For Non-Intrusive Speech Quality Assessment». En: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 1. IEEE, 2006, págs. I-829-I-832. ISBN: 1-4244-0469-X. DOI: [10.1109/ICASSP.2006.1660149](https://doi.org/10.1109/ICASSP.2006.1660149). URL: <http://ieeexplore.ieee.org/document/1660149/> (vid. pág. 35).
- [39] J. Dorta Luis. «La focalización prosódica: funcionalidad en los niveles lingüístico y pragmático». En: *Estudios de fonética experimental* 17.17 (2008), págs. 105-138. ISSN: 1575-5533 (vid. págs. 78, 88).
- [40] M. Dubiel, M. Halvey, P. O. Gallegos y S. King. «Persuasive Synthetic Speech». En: *Proceedings of the 2nd Conference on Conversational User Interfaces*. New York, NY, USA: ACM, jul. de 2020, págs. 1-9. ISBN: 9781450375443. DOI: [10.1145/3405755.3406120](https://doi.org/10.1145/3405755.3406120). URL: <https://dl.acm.org/doi/10.1145/3405755.3406120> (vid. pág. 28).
- [41] D. M. Eberhard, G. F. Simons y C. D. Fennig. *Ethnologue: Languages of the World. Twenty-third edition*. 2020. URL: <http://www.ethnologue.com> (visitado 29-03-2020) (vid. pág. 1).
- [42] A. Eriksson, E. Grabe y H. Traunmüller. «Perception of Syllable Prominence by Listeners with and without Competence in the Tested Language». En: *Speech Prosody* (2002), págs. 1-4. URL: [https://www.isca-speech.org/archive%7B%5C\\_%7Dopen/sp2002/sp02%7B%5C\\_%7D275.html](https://www.isca-speech.org/archive%7B%5C_%7Dopen/sp2002/sp02%7B%5C_%7D275.html) (vid. págs. 79, 117).
- [43] A. Eriksson, G. C. Thunberg y H. Traunmüller. «Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing». En: *EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology*. 2001, págs. 399-402. ISBN: 8790834100. URL: [https://www.isca-speech.org/archive/eurospeech%7B%5C\\_%7D2001/e01%7B%5C\\_%7D0399.html](https://www.isca-speech.org/archive/eurospeech%7B%5C_%7D2001/e01%7B%5C_%7D0399.html) (vid. págs. 79, 82).
- [44] D. Evin, C. Cossio-Mercado, H. Torres, J. Gurlekian y H. Mixdorff. «Automatic prominence detection in argentinian Spanish». En: *Proceedings of the International Conference on Speech Prosody*. 2018, págs. 680-684. DOI: [10.21437/SpeechProsody.2018-138](https://doi.org/10.21437/SpeechProsody.2018-138) (vid. págs. 75, 83).
- [45] D. A. Evin. «Incorporación de información suprasegmental en el proceso de reconocimiento automático del habla». Tesis Doctoral. 2011. URL: [http://digital.bl.fcen.uba.ar/Download/Tesis/Tesis%7B%5C\\_%7D4920%7B%5C\\_%7DEvin.pdf](http://digital.bl.fcen.uba.ar/Download/Tesis/Tesis%7B%5C_%7D4920%7B%5C_%7DEvin.pdf) (vid. pág. 22).
- [46] F. Eyben, S. Buchholz y N. Braunschweiler. «Unsupervised clustering of emotion and voice styles for expressive TTS». En: *2012 IEEE International Conference on Acoustics, Speech and Signal*

- Processing (ICASSP)*. i. IEEE, mar. de 2012, págs. 4009-4012. ISBN: 978-1-4673-0046-9. DOI: [10.1109/ICASSP.2012.6288797](https://doi.org/10.1109/ICASSP.2012.6288797). URL: <http://ieeexplore.ieee.org/document/6288797/> (vid. pág. 16).
- [47] F. Eyben, M. Wöllmer y B. Schuller. «openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor». En: *Proceedings of the international conference on Multimedia - MM '10*. New York, New York, USA: ACM Press, 2010, pág. 1459. ISBN: 9781605589336. DOI: [10.1145/1873951.1874246](https://doi.org/10.1145/1873951.1874246). URL: <http://dl.acm.org/citation.cfm?doid=1873951.1874246> (vid. pág. 58).
- [48] F. Eyben y col. «The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing». En: *IEEE Transactions on Affective Computing* 7.2 (abr. de 2016), págs. 190-202. ISSN: 1949-3045. DOI: [10.1109/TAFFC.2015.2457417](https://doi.org/10.1109/TAFFC.2015.2457417). URL: <http://ieeexplore.ieee.org/document/7160715/> (vid. pág. 99).
- [49] T. Ezzat, J. Bouvrie y T. Poggio. «Spectro-Temporal Analysis of Speech Using 2-D Gabor Filters». En: *Interspeech 2007*. 2007, págs. 506-509 (vid. pág. 36).
- [50] T. Face. «El foco y la altura tonal en español». En: *Boletín de Lingüística* 17 (2002), págs. 30-52. ISSN: 0798-9709. URL: <https://www.redalyc.org/articulo.oa?id=34701703> (vid. pág. 78).
- [51] T. L. Face. «The role of intonational cues in the perception of declaratives and absolute interrogatives in Castilian Spanish». En: *Estudios de fonética experimental* 16 (2007), págs. 186-225. ISSN: 2385-3573 (vid. pág. 79).
- [52] T. Falk, Qingfeng Xu y Wai-Yip Chan. «Non-Intrusive GMM-Based Speech Quality Measurement». En: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*. Vol. 1. IEEE, 2005, págs. 125-128. ISBN: 0-7803-8874-7. DOI: [10.1109/ICASSP.2005.1415066](https://doi.org/10.1109/ICASSP.2005.1415066). URL: <http://ieeexplore.ieee.org/document/1415066/> (vid. pág. 35).
- [53] T. H. Falk y S. Moller. «Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems». En: *IEEE Signal Processing Letters* 15 (2008), págs. 781-784. ISSN: 1070-9908. DOI: [10.1109/LSP.2008.2006709](https://doi.org/10.1109/LSP.2008.2006709). URL: <http://ieeexplore.ieee.org/document/4682563/> (vid. págs. 32, 35, 55).
- [54] M. Farrús, J. Hernando y P. Ejarque. «Jitter and Shimmer Measurements for Speaker Recognition Mireia». En: *Interspeech 2007* (2007), págs. 778-781. ISSN: 17519675. URL: [https://repositori.upf.edu/bitstream/handle/10230/28250/Farrus%7B%5C\\_%7DInterspeech2007%7B%5C\\_%7Djitt.pdf?sequence=1%7B%5C%7D&DisAllowed=y%7B%5C%7D0Ahttp://digital-library](https://repositori.upf.edu/bitstream/handle/10230/28250/Farrus%7B%5C_%7DInterspeech2007%7B%5C_%7Djitt.pdf?sequence=1%7B%5C%7D&DisAllowed=y%7B%5C%7D0Ahttp://digital-library)

- [theiet.org/content/journals/10.1049/iet-spr.2008.0147](http://theiet.org/content/journals/10.1049/iet-spr.2008.0147) (vid. pág. 98).
- [55] R. Fernandez y R. Picard. «Recognizing affect from speech prosody using hierarchical graphical models». En: *Speech Communication* 53.9-10 (nov. de 2011), págs. 1088-1103. ISSN: 01676393. DOI: [10.1016/j.specom.2011.05.003](https://doi.org/10.1016/j.specom.2011.05.003). URL: <http://linkinghub.elsevier.com/retrieve/pii/S0167639311000689> (vid. pág. 28).
- [56] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash y K. Precoda. «Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems». En: *Speech Communication* 69 (2015), págs. 31-45. ISSN: 01676393. DOI: [10.1016/j.specom.2015.02.002](https://doi.org/10.1016/j.specom.2015.02.002). URL: <http://dx.doi.org/10.1016/j.specom.2015.02.002> (vid. pág. 92).
- [57] J. L. Flanagan. *Speech Analysis Synthesis and Perception*. 2.<sup>a</sup> ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1972. ISBN: 978-3-662-01564-3. DOI: [10.1007/978-3-662-01562-9](https://doi.org/10.1007/978-3-662-01562-9). URL: <http://link.springer.com/10.1007/978-3-662-01562-9> (vid. pág. 14).
- [58] L. A. Forero Mendoza, E. Cataldo, M. M. Vellasco, M. A. Silva y J. A. Apolinário. «Classification of Vocal Aging Using Parameters Extracted From the Glottal Signal». En: *Journal of Voice* 28.5 (sep. de 2014), págs. 532-537. ISSN: 08921997. DOI: [10.1016/j.jvoice.2014.02.001](https://doi.org/10.1016/j.jvoice.2014.02.001). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0892199714000216> (vid. pág. 35).
- [59] L. Formiga, A. Trilla, F. Alías, I. Iriondo y J. C. Socoró. «Adaptation of the URL-TTS system to the 2010 Albayzin Evaluation Campaign». En: *FALA 2010 (VI Jornadas en Tecnología del Habla)*. 2010, págs. 363-370 (vid. págs. 1, 45).
- [60] D. B. Fry. «Experiments in the Perception of Stress». En: *Language and Speech* 1.2 (abr. de 1958), págs. 126-152. ISSN: 0023-8309. DOI: [10.1177/002383095800100207](https://doi.org/10.1177/002383095800100207). URL: <http://journals.sagepub.com/doi/10.1177/002383095800100207> (vid. pág. 76).
- [61] T. Ganchev. *Contemporary Methods for Speech Parameterization*. New York, NY: Springer New York, 2011. ISBN: 978-1-4419-8446-3. DOI: [10.1007/978-1-4419-8447-0](https://doi.org/10.1007/978-1-4419-8447-0). URL: <http://link.springer.com/10.1007/978-1-4419-8447-0> (vid. pág. 35).
- [62] L. Gauder, A. Gravano, L. Ferrer, P. Riera y S. Brussino. «A protocol for collecting speech data with varying degrees of trust». En: *SMM19, Workshop on Speech, Music and Mind 2019*. September. ISCA: ISCA, sep. de 2019, págs. 6-10. DOI: [10.21437/SMM.2019-2](https://doi.org/10.21437/SMM.2019-2). URL: [http://www.isca-speech.org/archive/SMM%7B%5C\\_%7D2019/abstracts/SMM19%7B%5C\\_%7Dpaper%7B%5C\\_%7D4.html](http://www.isca-speech.org/archive/SMM%7B%5C_%7D2019/abstracts/SMM19%7B%5C_%7Dpaper%7B%5C_%7D4.html) (vid. pág. 28).

- [63] H. Giles y A. C. Billings. «Assessing Language Attitudes: Speaker Evaluation Studies». En: *The Handbook of Applied Linguistics*. Oxford, UK: Blackwell Publishing Ltd, 2004. Cap. 7, págs. 187-209. ISBN: 0631228993. DOI: [10.1002/9780470757000.ch7](https://doi.org/10.1002/9780470757000.ch7). URL: <https://onlinelibrary.wiley.com/doi/10.1002/9780470757000.ch7> (vid. págs. 3, 28, 46).
- [64] C. Gobl. «The role of voice quality in communicating emotion, mood and attitude». En: *Speech Communication* 40.1-2 (abr. de 2003), págs. 189-212. ISSN: 01676393. DOI: [10.1016/S0167-6393\(02\)00082-1](https://doi.org/10.1016/S0167-6393(02)00082-1). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167639302000821> (vid. pág. 77).
- [65] V. Grancharov y W. B. Kleijn. «Speech Quality Assessment». En: *Springer Handbook of Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Cap. 5, págs. 83-99. DOI: [10.1007/978-3-540-49127-9\\_5](https://doi.org/10.1007/978-3-540-49127-9_5). URL: [http://link.springer.com/10.1007/978-3-540-49127-9\\_5](http://link.springer.com/10.1007/978-3-540-49127-9_5) (vid. págs. 32, 33, 35).
- [66] M. Güemes, B. Sampetro, C. Cossio-Mercado y J. Gurlekian. «La relación entre foco y prosodia: análisis de la percepción de las prominencias acentuales en un corpus del español de Buenos Aires». En: *ELUA. Estudios de Lingüística Universidad de Alicante* 30 (2016), págs. 129-139. ISSN: 0212-7636. DOI: [10.14198/ELUA2016.30.06](https://doi.org/10.14198/ELUA2016.30.06). URL: <http://hdl.handle.net/10045/60771> (vid. pág. 75).
- [67] A. Guevara-Rukoz, I. Demirsahin, F. He, S. H. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna y O. Kjartansson. «Crowdsourcing latin american Spanish for low-resource text-to-speech». En: *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*. May. 2020, págs. 6504-6513. ISBN: 9791095546344. URL: <https://aclanthology.org/2020.lrec-1.801> (vid. pág. 1).
- [68] Guo Chen y V. Parsa. «Bayesian Model Based Non-Intrusive Speech Quality Evaluation». En: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*. Vol. 1. IEEE, 2005, págs. 385-388. ISBN: 0-7803-8874-7. DOI: [10.1109/ICASSP.2005.1415131](https://doi.org/10.1109/ICASSP.2005.1415131). URL: <http://ieeexplore.ieee.org/document/1415131/> (vid. pág. 35).
- [69] J. Gurlekian, L. Colantoni y H. Torres. «El alfabeto fonético SAMPA y el diseño de corpora fonéticamente balanceados». En: *Fonoaudiológica* 47.3 (2001), págs. 58-69 (vid. págs. 9, 47).
- [70] J. Gurlekian, C. Cossio-Mercado, H. Torres y M. E. Vaccari. «Subjective evaluation of a high quality text-to-speech system for Argentine Spanish». En: *IberSPEECH 2012 - VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*. 2012, págs. 241-250. URL: <https://www.researchgate.net/profile/>

- Christian%7B%5C\_%7DCossio-Mercado/publication/265955190%7B%5C\_%7DSubjective%7B%5C\_%7DEvaluation%7B%5C\_%7Dof%7B%5C\_%7Da%7B%5C\_%7DHigh%7B%5C\_%7DQuality%7B%5C\_%7DText-to-Speech%7B%5C\_%7DSystem%7B%5C\_%7Dfor%7B%5C\_%7DArgentine%7B%5C\_%7DSpanish/ (vid. págs. 34, 36, 46).
- [71] J. Gurlekian, H. Mixdorff, D. Evin, H. Torres y H. Pfitzinger. «Alignment of Fo model parameters with final and non-final accents in Argentinean Spanish». En: *Speech Prosody 2010*. January. 2010, págs. 3-6. URL: [https://www.isca-speech.org/archive/sp2010/sp10%7B%5C\\_%7D131.html](https://www.isca-speech.org/archive/sp2010/sp10%7B%5C_%7D131.html) (vid. págs. 9).
- [72] J. Gurlekian, H. Mixdorff, H. Torres, C. Cossio-Mercado y D. Evin. «Acoustic correlates of perceived syllable prominence in Spanish». En: *Proceedings of the International Conference on Speech Prosody*. Mayo de 2016, págs. 673-677. DOI: 10.21437/SpeechProsody.2016-138. URL: [http://www.isca-speech.org/archive/SpeechProsody%7B%5C\\_%7D2016/abstracts/275.html](http://www.isca-speech.org/archive/SpeechProsody%7B%5C_%7D2016/abstracts/275.html) (vid. págs. 75, 90).
- [73] J. Gurlekian y G. Toledo. «Datos preliminares del Amper-Argentina : las oraciones declarativas e interrogativas absolutas sin expansión». En: *Language Design. Journal of Theoretical and Experimental Linguistics* (2008), págs. 213-220. ISSN: 1139-4218. URL: <https://ddd.uab.cat/record/148547> (vid. págs. 81).
- [74] J. Gurlekian, H. Torres, D. Evin, H. Mixdorff, C. Cossio-Mercado y M. Güemes. «Estudio del foco : las prominencias acentuales , el modelado acústico y la detección automática». En: *Fernández Planas, A. Ma. (ed.) 53 reflexiones sobre aspectos de la fonética y otros temas de lingüística*. 2016, págs. 209-219. ISBN: 9788460898306 (vid. págs. 75, 83, 90).
- [75] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg y L. Wedin. «Perceptual and Acoustic Correlates of Abnormal Voice Qualities». En: *Acta Oto-Laryngologica* 90.1-6 (1980), págs. 441-451. ISSN: 0001-6489. DOI: 10.3109/00016488009131746. URL: <http://www.tandfonline.com/doi/full/10.3109/00016488009131746> (vid. págs. 93).
- [76] J. Hao, S. Ye, C. Lu, F. Dong, J. Liu y D. Pi. «Soft-label Learn for No-Intrusive Speech Quality Assessment». En: *Interspeech 2022*. Vol. 2022-Sept. September. ISCA: ISCA, sep. de 2022, págs. 3303-3307. DOI: 10.21437/Interspeech.2022-10400. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/hao22%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/hao22%7B%5C_%7Dinterspeech.html) (vid. págs. 56).
- [77] M. Hariharan, M. P. Paulraj y Sazali Yaacob. «Identification of vocal fold pathology based on Mel Frequency Band Energy Coefficients and singular value decomposition». En: *2009 IEEE*



- International Conference on Signal and Image Processing Applications*. IEEE, 2009, págs. 514-517. ISBN: 978-1-4244-5560-7. DOI: [10.1109/ICSIPA.2009.5478710](https://doi.org/10.1109/ICSIPA.2009.5478710). URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5478710%20http://ieeexplore.ieee.org/document/5478710/> (vid. pág. 35).
- [78] T. Hastie, R. Tibshirani y J. Friedman. *The Elements of Statistical Learning*. 2.<sup>a</sup> ed. Springer Series in Statistics. New York, NY: Springer New York, 2009. ISBN: 978-0-387-84857-0. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7). URL: <http://link.springer.com/10.1007/978-0-387-84858-7> (vid. pág. 32).
- [79] M. Heckmann, X. Domont, F. Joubin y C. Goerick. «A hierarchical framework for spectro-temporal feature extraction». En: *Speech Communication* 53.5 (mayo de 2011), págs. 736-752. ISSN: 01676393. DOI: [10.1016/j.specom.2010.08.006](https://doi.org/10.1016/j.specom.2010.08.006). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167639310001408> (vid. pág. 36).
- [80] M. Heldner. «Spectral emphasis as an additional source of information in accent detection». En: *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. 2001 (vid. págs. 83, 84).
- [81] H. Hermansky. «Perceptual linear predictive (PLP) analysis of speech». En: *The Journal of the Acoustical Society of America* 87.4 (abr. de 1990), págs. 1738-1752. ISSN: 0001-4966. DOI: [10.1121/1.399423](https://doi.org/10.1121/1.399423). URL: <http://asa.scitation.org/doi/10.1121/1.399423> (vid. pág. 36).
- [82] F. Hinterleitner. *Quality of Synthetic Speech*. T-Labs Series in Telecommunication Services. Singapore: Springer Singapore, 2017. ISBN: 978-981-10-3733-7. DOI: [10.1007/978-981-10-3734-4](https://doi.org/10.1007/978-981-10-3734-4). URL: <http://link.springer.com/10.1007/978-981-10-3734-4> (vid. págs. 15, 31).
- [83] F. Hinterleitner, S. Möller, C. Norrenbrock y U. Heute. «Perceptual quality dimensions of text-to-speech systems». En: *INTERSPEECH 2011*. August. 2011, págs. 2177-2180 (vid. págs. 35, 55).
- [84] F. Hinterleitner, C. Norrenbrock y S. Möller. «On the Use of Fujisaki Parameters for the Quality Prediction of Synthetic Speech». En: *Proc. of the 23th Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*. 2012, págs. 112-119 (vid. págs. 35, 75).
- [85] F. Hinterleitner, C. R. Norrenbrock y S. Möller. «Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech». En: *8th ISCA Speech Synthesis Workshop*. 2013, págs. 147-151. URL: <https://www.isca->

- [speech.org/archive/ssw8/ssw8%7B%5C\\_%7D147.html](http://speech.org/archive/ssw8/ssw8%7B%5C_%7D147.html) (vid. pág. 31).
- [86] F. Hinterleitner, M. Sebastian, T. H. Falk y T. Polzehl. «Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009». En: *Blizzard Challenge Workshop 2010*. 2010 (vid. págs. 35, 55).
- [87] F. Hinterleitner, B. Weiss y S. Moller. «Influence of corpus size and content on the perceptual quality of a unit selection MaryTTS voice». En: *2016 IEEE Workshop on Spoken Language Technology, SLT 2016 - Proceedings*. 2017, págs. 680-685. ISBN: 9781509049035. DOI: [10.1109/SLT.2016.7846336](https://doi.org/10.1109/SLT.2016.7846336) (vid. pág. 49).
- [88] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku y K. Tokuda. «Singing Voice Synthesis Based on Generative Adversarial Networks». En: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mayo de 2019, págs. 6955-6959. ISBN: 978-1-4799-8131-1. DOI: [10.1109/ICASSP.2019.8683154](https://doi.org/10.1109/ICASSP.2019.8683154). URL: <https://ieeexplore.ieee.org/document/8683154/> (vid. pág. 17).
- [89] T. Hofffeld, P. E. Heegaard, M. Varela y S. Möller. «QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS». En: *Quality and User Experience 1.1* (dic. de 2016), pág. 2. ISSN: 2366-0139. DOI: [10.1007/s41233-016-0002-1](https://doi.org/10.1007/s41233-016-0002-1). URL: <http://link.springer.com/10.1007/s41233-016-0002-1> (vid. págs. 5, 36).
- [90] J. I. J. I. Hualde. *Los sonidos del español*. Cambridge: Cambridge University Press, 2013. ISBN: 9780511719943. DOI: [10.1017/CB09780511719943](https://doi.org/10.1017/CB09780511719943). URL: <http://ebooks.cambridge.org/ref/id/CB09780511719943> (vid. pág. 9).
- [91] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda y J. Yamagishi. «The VoiceMOS Challenge 2022». En: *Interspeech 2022*. Vol. 2022-Sept. September. ISCA: ISCA, sep. de 2022, págs. 4536-4540. DOI: [10.21437/Interspeech.2022-970](https://doi.org/10.21437/Interspeech.2022-970). arXiv: [2203.11389](https://arxiv.org/abs/2203.11389). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/huang22f%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/huang22f%7B%5C_%7Dinterspeech.html) (vid. pág. 55).
- [92] W.-C. Huang, E. Cooper, J. Yamagishi y T. Toda. «LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech». En: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2022-May. IEEE, mayo de 2022, págs. 896-900. ISBN: 978-1-6654-0540-9. DOI: [10.1109/ICASSP43922.2022.9747222](https://doi.org/10.1109/ICASSP43922.2022.9747222). arXiv: [2110.09103](https://arxiv.org/abs/2110.09103). URL: <https://ieeexplore.ieee.org/document/9747222/> (vid. págs. 56, 74).



- [93] *Introducción a las etiquetas EAGLES*. URL: <https://www.cs.upc.edu/%7B-%7Dnlp/tools/parole-sp.html> (visitado 01-06-2021) (vid. pág. 18).
- [94] ITU-T. *P.85 : A method for subjective performance assessment of the quality of speech voice output devices*. 1994. URL: <https://www.itu.int/rec/T-REC-P.85-199406-I> (vid. págs. 47, 50).
- [95] ITU-T. *P.85 Amendment 1: Evaluation of speech output for audiobook reading tasks*. 2013 (vid. pág. 45).
- [96] J. James, B. T. Balamurali, C. I. Watson y B. MacDonald. «Empathetic Speech Synthesis and Testing for Healthcare Robots». En: *International Journal of Social Robotics* (sep. de 2020). ISSN: 1875-4791. DOI: 10.1007/s12369-020-00691-4. URL: <https://doi.org/10.1007/s12369-020-00691-4> <http://link.springer.com/10.1007/s12369-020-00691-4> (vid. pág. 92).
- [97] A. C. Janska. «Further Investigation of MDS as a Tool for Evaluation of Speech Quality of Synthesized Speech». Master of Science Thesis. University of Edinburgh, 2009 (vid. pág. 34).
- [98] U. Jekosch. *Voice and Speech Quality Perception*. Signals and Communication Technology. Berlin/Heidelberg: Springer-Verlag, 2005. ISBN: 3-540-24095-0. DOI: 10.1007/3-540-28860-0. URL: <http://link.springer.com/10.1007/3-540-28860-0> (vid. pág. 31).
- [99] M. Jessen. «Speaker classification in forensic phonetics and acoustics». En: *Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4343 LNAI (2007), págs. 180-204. ISSN: 16113349. DOI: 10.1007/978-3-540-74200-5\_10 (vid. pág. 28).
- [100] R. Z. Jimenez, B. Naderi y S. Moller. «Effect of Environmental Noise in Speech Quality Assessment Studies using Crowdsourcing». En: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, mayo de 2020, págs. 1-6. ISBN: 978-1-7281-5965-2. DOI: 10.1109/QoMEX48832.2020.9123144. URL: <https://ieeexplore.ieee.org/document/9123144/> (vid. pág. 55).
- [101] D. Jurafsky y J. H. Martin. *Speech and Language Processing*. 2.<sup>a</sup> ed. Prentice Hall, 2008. ISBN: 978-0131873216. URL: <http://www.cs.colorado.edu/%7B-%7Dmartin/slp.html> (vid. pág. 32).
- [102] S. Kabashima, Y. Inoue, D. Saito y N. Minematsu. «DNN-Based Scoring of Language Learners' Proficiency Using Learners' Shadowings and Native Listeners' Responsive Shadowings». En: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, dic. de 2018, págs. 971-978. ISBN: 978-1-5386-4334-1. DOI: 10.1109/SLT.2018.8639645. URL: <https://ieeexplore.ieee.org/document/8639645/> (vid. pág. 111).

- [103] H. Kallio, A. Suni, P. Virkkunen y J. Simko. «Prominence-based evaluation of L2 prosody». En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2018-Sept. September. ISCA: ISCA, sep. de 2018, págs. 1838-1842. DOI: [10.21437/Interspeech.2018-1873](https://doi.org/10.21437/Interspeech.2018-1873). URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2018/abstracts/1873.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2018/abstracts/1873.html) (vid. pág. 92).
- [104] W. von Kempelen. *Mechanismus der menschlichen Sprache*. Ed. por F. Brackhane, R. Sproat y J. Trouvain. 1791 (vid. pág. 14).
- [105] S. M. Ketrow. «Attributes of a telemarketer's voice and persuasiveness. A review and synthesis of the literature». En: *Journal of Direct Marketing* 4.3 (1990), págs. 7-21. ISSN: 08920591. DOI: [10.1002/dir.4000040304](https://doi.org/10.1002/dir.4000040304). URL: <https://onlinelibrary.wiley.com/doi/10.1002/dir.4000040304> (vid. pág. 28).
- [106] D. H. Klatt. «Software for a cascade/parallel formant synthesizer». En: *The Journal of the Acoustical Society of America* 67.3 (mar. de 1980), págs. 971-995. ISSN: 0001-4966. DOI: [10.1121/1.383940](https://doi.org/10.1121/1.383940). URL: <http://asa.scitation.org/doi/10.1121/1.383940> (vid. págs. 14, 15).
- [107] G. Kochanski, E. Grabe, J. Coleman y B. Rosner. «Loudness predicts prominence: Fundamental frequency lends little». En: *The Journal of the Acoustical Society of America* 118.2 (ago. de 2005), págs. 1038-1054. ISSN: 0001-4966. DOI: [10.1121/1.1923349](https://doi.org/10.1121/1.1923349). URL: <http://asa.scitation.org/doi/10.1121/1.1923349> (vid. pág. 89).
- [108] M. Kockmann, L. Burget y J. Černocký. «Application of speaker- and language identification state-of-the-art techniques for emotion recognition». En: *Speech Communication* 53.9-10 (nov. de 2011), págs. 1172-1185. ISSN: 01676393. DOI: [10.1016/j.specom.2011.01.007](https://doi.org/10.1016/j.specom.2011.01.007). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167639311000082> (vid. pág. 94).
- [109] F. Köster, D. Guse, M. Wältermann y S. Möller. «Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech». En: *Proc. 41st German Annual Conference on Acoustics (DAGA)*. 2015, págs. 150-153 (vid. pág. 49).
- [110] B. Krenn, S. Schreitter y F. Neubarth. «Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application». En: *AI and Society* 32.1 (2017), págs. 65-77. ISSN: 14355655. DOI: [10.1007/s00146-014-0569-0](https://doi.org/10.1007/s00146-014-0569-0). URL: <http://dx.doi.org/10.1007/s00146-014-0569-0> (vid. págs. 3, 46).
- [111] K. Kühne, M. H. Fischer e Y. Zhou. «The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study». En:

- Frontiers in Neurobotics* 14.December (dic. de 2020), págs. 1-15. ISSN: 1662-5218. DOI: [10.3389/fnbot.2020.593732](https://doi.org/10.3389/fnbot.2020.593732). URL: <https://www.frontiersin.org/articles/10.3389/fnbot.2020.593732/full> (vid. págs. 32, 58).
- [112] K. Kyriakopoulos, M. Gales y K. Knill. «Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features». En: August (2017), págs. 59-64. DOI: [10.21437/slate.2017-11](https://doi.org/10.21437/slate.2017-11) (vid. pág. 28).
- [113] L. O. Labastía. «Prosodic prominence in Argentinian Spanish». En: *Journal of Pragmatics* 38.10 (oct. de 2006), págs. 1677-1705. ISSN: 03782166. DOI: [10.1016/j.pragma.2005.03.019](https://doi.org/10.1016/j.pragma.2005.03.019). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0378216606000452> (vid. págs. 75, 78).
- [114] D. R. Ladd. «Even, Focus, and Normal Stress». En: *Journal of Semantics* 2.2 (ene. de 1983), págs. 157-170. ISSN: 0167-5133. DOI: [10.1093/semant/2.2.157](https://doi.org/10.1093/semant/2.2.157). URL: <https://academic.oup.com/jos/article-lookup/doi/10.1093/semant/2.2.157> (vid. pág. 76).
- [115] F. Landini, L. Ferrer y H. Franco. «Adaptation Approaches for Pronunciation Scoring with Sparse Training Data». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10458 LNAI. 2017, págs. 87-97. ISBN: 9783319664286. DOI: [10.1007/978-3-319-66429-3\\_8](https://doi.org/10.1007/978-3-319-66429-3_8). URL: [http://link.springer.com/10.1007/978-3-319-66429-3\\_8](http://link.springer.com/10.1007/978-3-319-66429-3_8) (vid. pág. 111).
- [116] J. Lang-Rigal. «Perception of Narrow Focus Prosody in Buenos Aires Spanish». En: *5th Conference on Laboratory Approaches to Romance Phonology*. Ed. por S. M. Alvord. Somerville, MA, 2011, págs. 118-126. URL: <http://www.lingref.com/cpp/larp/5/index.html> (vid. pág. 79).
- [117] S. Latif, A. Qayyum, M. Usman y J. Qadir. «Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages». En: *2018 International Conference on Frontiers of Information Technology (FIT)*. IEEE, dic. de 2018, págs. 88-93. ISBN: 978-1-5386-9355-1. DOI: [10.1109/FIT.2018.00023](https://doi.org/10.1109/FIT.2018.00023). arXiv: [1812.10411](https://arxiv.org/abs/1812.10411). URL: <https://ieeexplore.ieee.org/document/8616972/> (vid. pág. 100).
- [118] C.-C. Lee, E. Mower, C. Busso, S. Lee y S. Narayanan. «Emotion recognition using a hierarchical binary decision tree approach». En: *Speech Communication* 53.9-10 (nov. de 2011), págs. 1162-1171. ISSN: 01676393. DOI: [10.1016/j.specom.2011.06.004](https://doi.org/10.1016/j.specom.2011.06.004). URL: <http://linkinghub.elsevier.com/retrieve/pii/S0167639311000884> (vid. pág. 35).

- [119] T. E. Lehnert, S. Krolak-Schwerdt y T. Hörstermann. «Language and nationality attitudes as distinct factors that influence speaker evaluations: Explicit versus implicit attitudes in Luxembourg». En: *Language & Communication* 61 (jul. de 2018), págs. 58-70. ISSN: 02715309. DOI: [10.1016/j.langcom.2018.01.005](https://doi.org/10.1016/j.langcom.2018.01.005). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0271530917301295> (vid. pág. 28).
- [120] V. I. Levenshtein. «Binary codes capable of correcting deletions, insertions and reversals». En: *Soviet Physics - Doklady* 10.8 (1966), págs. 707-710 (vid. pág. 57).
- [121] J. Llisterri, C. Carbó, M. J. Machuca, C. De la Mota, M. Riera y A. Ríos. «La conversión de texto en habla: aspectos lingüísticos». En: (2004), págs. 145-186 (vid. pág. 8).
- [122] J. Llisterri, M. Machuca, C. de la Mota, M. Riera y A. Ríos. «The perception of lexical stress in Spanish». En: *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, 2003, págs. 2023-2026. ISBN: 1876346485. URL: [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15%7B%5C\\_%7D2023.html](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15%7B%5C_%7D2023.html) (vid. pág. 89).
- [123] J. Lorenzo-Trueba, O. Watts, R. Barra-Chicote, J. Yamagishi, S. King y J. M. Montero. «Simple4All proposals for the Albayzin Evaluations in Speech Synthesis». En: *IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*. 2012, págs. 653-662 (vid. págs. 1, 45).
- [124] B. Ludusan y P. Wagner. «No laughing matter. An investigation into the acoustic cues marking the use of laughter». En: *International Congress of Phonetic Sciences*. 2019, págs. 2179-2182. URL: <https://pub.uni-bielefeld.de/record/2933731> (vid. pág. 111).
- [125] C. Manning y H. Schütze. *Foundations of Statistical Natural Language Processing*. 1999. ISBN: 9780262133609 (vid. pág. 86).
- [126] T. Männistö-Funk y T. Sihvonen. «Voices from the Uncanny Valley». En: *Digital Culture & Society* 4.1 (mar. de 2018), págs. 45-64. ISSN: 2364-2122. DOI: [10.14361/dcs-2018-0105](https://doi.org/10.14361/dcs-2018-0105). URL: <http://www.degruyter.com/view/j/dcs.2018.4.issue-1/dcs-2018-0105/dcs-2018-0105.xml> (vid. págs. 11, 107).
- [127] P. Manocha y A. Kumar. «Speech Quality Assessment through MOS using Non-Matching References». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 654-658. DOI: [10.21437/Interspeech.2022-407](https://doi.org/10.21437/Interspeech.2022-407). arXiv: 2206.12285. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/manocha22c%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/manocha22c%7B%5C_%7Dinterspeech.html) (vid. pág. 56).

- [128] A. Mariniak. «A Global Framework for the Assessment of Synthetic Speech without Subjects». En: *3rd European Conference on Speech Communication and Technology EUROSPEECH'93*. September. 1993, págs. 1683-1686 (vid. pág. 35).
- [129] M. Martínez Soler y C. Cossio Mercado. «Generating and Evaluating Coarse-Grained Instructions in a Virtual Environment». En: *IBERAMIA 2010 - Workshop on NLP and Web-based technologies*. 2010. URL: <https://cs.famaf.unc.edu.ar/~7B~%7Dlaura/nlpw/nlpw/program.html> (vid. pág. 10).
- [130] C. Mayo, R. a.J. Clark y S. King. «Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis». En: *Speech Communication* 53.3 (mar. de 2011), págs. 311-326. ISSN: 01676393. DOI: 10.1016/j.specom.2010.10.003. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167639310001627> (vid. págs. 32, 34, 35, 92).
- [131] T. Merritt, R. A. J. Clark, Z. Wu, J. Yamagishi y S. King. «Deep neural network-guided unit selection synthesis». En: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2016-May. IEEE, mar. de 2016, págs. 5145-5149. ISBN: 978-1-4799-9988-0. DOI: 10.1109/ICASSP.2016.7472658. URL: <http://ieeexplore.ieee.org/document/7472658/> (vid. pág. 7).
- [132] G. Mittag, R. Cutler, Y. Hosseinkashi, M. Revow, S. Srinivasan, N. Chande y R. Aichner. «DNN No-Reference PSTN Speech Quality Prediction». En: *Interspeech 2020*. ISCA, oct. de 2020, págs. 2867-2871. DOI: 10.21437/Interspeech.2020-2760. URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2020/abstracts/2760.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2020/abstracts/2760.html) (vid. pág. 35).
- [133] G. Mittag y S. Möller. «Quality Degradation Diagnosis for Voice Networks — Estimating the Perceived Noisiness, Coloration, and Discontinuity of Transmitted Speech». En: *Interspeech 2019*. ISCA: ISCA, sep. de 2019, págs. 3426-3430. DOI: 10.21437/Interspeech.2019-2636. URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2019/abstracts/2636.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2019/abstracts/2636.html) (vid. págs. 35, 56, 57).
- [134] G. Mittag y S. Möller. «Deep Learning Based Assessment of Synthetic Speech Naturalness». En: *Interspeech 2020*. ISCA: ISCA, oct. de 2020, págs. 1748-1752. DOI: 10.21437/Interspeech.2020-2382. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2020/mittag20%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2020/mittag20%7B%5C_%7Dinterspeech.html) (vid. págs. 35, 56).
- [135] G. Mittag y S. Möller. «Full-Reference Speech Quality Estimation with Attentional Siamese Neural Networks». En: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mayo de 2020, págs. 346-350.

- ISBN: 978-1-5090-6631-5. DOI: [10.1109/ICASSP40776.2020.9053951](https://doi.org/10.1109/ICASSP40776.2020.9053951). URL: <https://ieeexplore.ieee.org/document/9053951/> (vid. pág. 35).
- [136] G. Mittag, B. Naderi, A. Chehadi y S. Möller. «NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets». En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 4* (2021), págs. 2818-2822. ISSN: 19909772. DOI: [10.21437/Interspeech.2021-299](https://doi.org/10.21437/Interspeech.2021-299). arXiv: [2104.09494](https://arxiv.org/abs/2104.09494) (vid. págs. 35, 56, 57).
- [137] H. Mixdorff, C. Cossio-Mercado, A. Hönemann, J. Gurlekian, D. Evin y H. Torres. «Acoustic correlates of perceived syllable prominence in German». En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2015, págs. 51-55 (vid. págs. 75, 80, 90).
- [138] G. Mohammadi y A. Vinciarelli. «Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features». En: *IEEE Transactions on Affective Computing* (2012), págs. 1-14. ISSN: 1949-3045. DOI: [10.1109/T-AFFC.2012.5](https://doi.org/10.1109/T-AFFC.2012.5). URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6175005> (vid. pág. 28).
- [139] S. H. Mohammadi y A. Kain. «An overview of voice conversion systems». En: *Speech Communication* 88 (abr. de 2017), págs. 65-82. ISSN: 01676393. DOI: [10.1016/j.specom.2017.01.008](https://doi.org/10.1016/j.specom.2017.01.008). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167639315300698> (vid. pág. 16).
- [140] S. Möller, F. Hinterleitner, T. H. Falk y T. Polzehl. «Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems». En: *INTERSPEECH 2010*. 2010, págs. 1325-1328. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2010/i10%7B%5C\\_%7D1325.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2010/i10%7B%5C_%7D1325.html) (vid. págs. 32, 35, 55).
- [141] R. Näätänen, T. Kujala e I. Winkler. «Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses». En: *Psychophysiology* 48.1 (ene. de 2011), págs. 4-22. ISSN: 00485772. DOI: [10.1111/j.1469-8986.2010.01114.x](https://doi.org/10.1111/j.1469-8986.2010.01114.x). URL: <http://doi.wiley.com/10.1111/j.1469-8986.2010.01114.x> (vid. pág. 35).
- [142] E. Nachmani y L. Wolf. «Unsupervised Polyglot Text-to-speech». En: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mayo de 2019, págs. 7055-7059. ISBN: 978-1-4799-8131-1. DOI: [10.1109/ICASSP.2019.8683519](https://doi.org/10.1109/ICASSP.2019.8683519). URL: <https://ieeexplore.ieee.org/document/8683519/> (vid. págs. 17, 45).



- [143] B. Naderi, T. Hosfeld, M. Hirth, F. Metzger, S. Moller y R. Z. Jimenez. «Impact of the Number of Votes on the Reliability and Validity of Subjective Speech Quality Assessment in the Crowdsourcing Approach». En: *2020 12th International Conference on Quality of Multimedia Experience, QoMEX 2020* (2020). DOI: [10.1109/QoMEX48832.2020.9123115](https://doi.org/10.1109/QoMEX48832.2020.9123115). arXiv: [2003.11300](https://arxiv.org/abs/2003.11300) (vid. pág. 55).
- [144] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku y K. Tokuda. «Singing voice synthesis based on convolutional neural networks». En: (abr. de 2019). arXiv: [1904.06868](https://arxiv.org/abs/1904.06868). URL: <http://arxiv.org/abs/1904.06868> (vid. pág. 11).
- [145] C. Nass y K. M. Lee. «Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction.» En: *Journal of Experimental Psychology: Applied* 7.3 (2001), págs. 171-181. ISSN: 1939-2192. DOI: [10.1037/1076-898X.7.3.171](https://doi.org/10.1037/1076-898X.7.3.171). URL: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.7.3.171> (vid. pág. 28).
- [146] E. Navas, I. Hernaez, D. Erro, J. Salaberria, B. Oyharçabal y M. Padilla. «Developing a Basque TTS for the Navarrese-Lapurdean Dialect». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8854. 2014, págs. 11-20. ISBN: 9783319136226. DOI: [10.1007/978-3-319-13623-3\\_2](https://doi.org/10.1007/978-3-319-13623-3_2). URL: [http://link.springer.com/10.1007/978-3-319-13623-3\\_2](http://link.springer.com/10.1007/978-3-319-13623-3_2) (vid. pág. 1).
- [147] H. Nguyen, K. Li y M. Unoki. «Automatic Mean Opinion Score Estimation with Temporal Modulation Features on Gammatone Filterbank for Speech Assessment». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 4526-4530. DOI: [10.21437/Interspeech.2022-528](https://doi.org/10.21437/Interspeech.2022-528). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/nguyen22b%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/nguyen22b%7B%5C_%7Dinterspeech.html) (vid. pág. 56).
- [148] H. Nordström. «Emotional Communication in the Human Voice». PhD Thesis. Stockholm University, 2019. ISBN: 9789177977353. URL: <http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-167973> (vid. págs. 93, 94).
- [149] C. Norrenbrock, U. Heute, F. Hinterleitner y S. Möller. «Aperiodicity Analysis for Quality Estimation of Text-to-Speech Signals». En: *Interspeech 2011*. August. 2011, págs. 2193-2196. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2011/i11%7B%5C\\_%7D2193.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2011/i11%7B%5C_%7D2193.html) (vid. págs. 35, 55).

- [150] C. R. Norrenbrock, U. Heute y F. Hinterleitner. «On the Use of Vocal-Tract Approximations for Instrumental Quality Assessment». En: *ITG-Fachbericht 236: Sprachkommunikation*. VDE VERLAG GMBH, 2012 (vid. pág. 75).
- [151] C. R. Norrenbrock, F. Hinterleitner, U. Heute y S. Möller. «Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals». En: 19.5 (2012), págs. 255-258 (vid. pág. 35).
- [152] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior y K. Kavukcuoglu. «WaveNet: A Generative Model for Raw Audio». En: (sep. de 2016), págs. 1-15. arXiv: 1609.03499. URL: <http://arxiv.org/abs/1609.03499> (vid. pág. 17).
- [153] A. Peiró-Lilja, G. Cámbara, M. Farrús y J. Luque. «Naturalness and Intelligibility Monitoring for Text-to-Speech Evaluation». En: *Speech Prosody 2022*. May. Mayo de 2022, págs. 445-449. DOI: 10.21437/SpeechProsody.2022-91. URL: [https://www.isca-speech.org/archive/speechprosody%7B%5C\\_%7D2022/peirililja22%7B%5C\\_%7Dspeechprosody.html](https://www.isca-speech.org/archive/speechprosody%7B%5C_%7D2022/peirililja22%7B%5C_%7Dspeechprosody.html) (vid. pág. 56).
- [154] E. a. Peterson, N. Roy, S. N. Awan, R. M. Merrill, R. Banks y K. Tanner. «Toward Validation of the Cepstral Spectral Index of Dysphonia (CSID) as an Objective Treatment Outcomes Measure». En: *Journal of Voice* 27.4 (jul. de 2013), págs. 401-410. ISSN: 08921997. DOI: 10.1016/j.jvoice.2013.04.002. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0892199713000787> (vid. pág. 35).
- [155] D. Picovici y A. Mahdi. «Output-based objective speech quality measure using self-organizing map». En: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. Vol. 1. IEEE, 2003, págs. I-476-I-479. ISBN: 0-7803-7663-3. DOI: 10.1109/ICASSP.2003.1198821. URL: <http://ieeexplore.ieee.org/document/1198821/> (vid. pág. 35).
- [156] S. Pinto, P. Tremblay, A. Basirat y M. Sato. «The impact of when, what and how predictions on auditory speech perception». En: *Experimental Brain Research* 237.12 (2019), págs. 3143-3153. ISSN: 14321106. DOI: 10.1007/s00221-019-05661-5. URL: <https://doi.org/10.1007/s00221-019-05661-5> (vid. pág. 21).
- [157] K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. O'Connor, S. Röder, P. W. Andrews, B. Fink, L. M. DeBruine, B. C. Jones y D. R. Feinberg. «Vocal indicators of body size in men and women: a meta-analysis». En: *Animal Behaviour* 95 (sep. de 2014), págs. 89-99. ISSN: 00033472. DOI: 10.1016/j.anbehav.2014.06.011. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0003347214002632> (vid. pág. 28).



- [158] D. Pisoni, H. Nusbaum y B. Greene. «Perception of synthetic speech generated by rule». En: *Proceedings of the IEEE* 73.11 (1985), págs. 1665-1676. ISSN: 0018-9219. DOI: [10.1109/PROC.1985.13346](https://doi.org/10.1109/PROC.1985.13346). URL: <http://ieeexplore.ieee.org/document/1457614/> (vid. pág. 24).
- [159] D. B. Pisoni. «Perception of Synthetic Speech». En: *Progress in Speech Synthesis*. New York, NY: Springer New York, 1997, págs. 541-560. DOI: [10.1007/978-1-4612-1894-4\\_43](https://doi.org/10.1007/978-1-4612-1894-4_43). URL: [http://link.springer.com/10.1007/978-1-4612-1894-4%7B%5C\\_%7D43](http://link.springer.com/10.1007/978-1-4612-1894-4%7B%5C_%7D43) (vid. pág. 24).
- [160] D. B. Pisoni y R. E. Remez. «The Handbook of Speech Perception». En: (ene. de 2005). Ed. por D. B. Pisoni y R. E. Remez. DOI: [10.1002/9780470757024](https://doi.org/10.1002/9780470757024). URL: <http://doi.wiley.com/10.1002/9780470757024> (vid. pág. 22).
- [161] J. C. Platt. «Fast training of support vector machines using sequential minimal optimization». En: *Advances in kernel methods*. Ed. por S. B., B. C.J.C. y S. A.J. 1.<sup>a</sup> ed. MIT Press, 1999. Cap. 12, págs. 185-208. ISBN: 9780262194167. URL: <https://mitpress.mit.edu/books/advances-kernel-methods> (vid. pág. 83).
- [162] T. Polzehl, B. Naderi, F. Föster y S. Möller. «Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments». En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2015-Janua. 2015, págs. 2794-2798 (vid. pág. 28).
- [163] T. Polzehl, A. Schmitt, F. Metze y M. Wagner. «Anger recognition in speech using acoustic and linguistic cues». En: *Speech Communication* 53.9-10 (nov. de 2011), págs. 1198-1209. ISSN: 01676393. DOI: [10.1016/j.specom.2011.05.002](https://doi.org/10.1016/j.specom.2011.05.002). URL: <http://linkinghub.elsevier.com/retrieve/pii/S0167639311000677> (vid. pág. 94).
- [164] T. Polzehl, K. Schoenenberg, M. Sebastian, F. Metze, G. Mohammadi y A. Vinciarelli. «On Speaker-Independent Personality Perception and Prediction from Speech». En: (2012), págs. 258-261 (vid. pág. 28).
- [165] A. K. Porbadnigk, M. S. Treder, B. Blankertz, J.-N. Antons, R. Schleicher, S. Möller, G. Curio y K.-R. Müller. «Single-trial analysis of the neural correlates of speech quality perception». En: *Journal of Neural Engineering* 10.5 (oct. de 2013), pág. 056003. ISSN: 1741-2560. DOI: [10.1088/1741-2560/10/5/056003](https://doi.org/10.1088/1741-2560/10/5/056003). URL: <http://www.ncbi.nlm.nih.gov/pubmed/23902853> <https://iopscience.iop.org/article/10.1088/1741-2560/10/5/056003> (vid. pág. 35).

- [166] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Elsevier, 1993. ISBN: 9780080500584. DOI: [10.1016/C2009-0-27846-9](https://doi.org/10.1016/C2009-0-27846-9). URL: <https://linkinghub.elsevier.com/retrieve/pii/S20090278469> (vid. pág. 83).
- [167] L. R. Rabiner y R. W. Schafer. *Theory and application of digital speech processing*. Prentice Hall, 2010, pág. 1060. ISBN: 9780136034292 (vid. pág. 21).
- [168] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey e I. Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision*. Dic. de 2022. DOI: [10.48550/arXiv.2212.04356](https://doi.org/10.48550/arXiv.2212.04356). arXiv: [2212.04356](https://arxiv.org/abs/2212.04356). URL: <http://arxiv.org/abs/2212.04356> (vid. pág. 57).
- [169] J. Raine, K. Pisanski, A. Oleszkiewicz, J. Simner y D. Reby. «Human Listeners Can Accurately Judge Strength and Height Relative to Self from Aggressive Roars and Speech». En: *iScience* 4 (jun. de 2018), págs. 273-280. ISSN: 25890042. DOI: [10.1016/j.isci.2018.05.002](https://doi.org/10.1016/j.isci.2018.05.002). URL: <https://linkinghub.elsevier.com/retrieve/pii/S2589004218300579> (vid. pág. 28).
- [170] K. S. Rao y S. G. Koolagudi. *Emotion Recognition using Speech Features*. SpringerBriefs in Electrical and Computer Engineering. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-5142-6. DOI: [10.1007/978-1-4614-5143-3](https://doi.org/10.1007/978-1-4614-5143-3). URL: <http://link.springer.com/10.1007/978-1-4614-5143-3> (vid. pág. 94).
- [171] G. Renunathan Naidu, S. Lebai Lutfi, A. Azazi, J. Lorenzo-Trueba y J. Martinez. «Cross-Cultural Perception of Spanish Synthetic Expressive Voices Among Asians». En: *Applied Sciences* 8.3 (mar. de 2018), pág. 426. ISSN: 2076-3417. DOI: [10.3390/app8030426](https://doi.org/10.3390/app8030426). URL: <http://www.mdpi.com/2076-3417/8/3/426> (vid. pág. 111).
- [172] P. Riera, L. Ferrer, A. Gravano y L. Gauder. «No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems». En: *SMM19, Workshop on Speech, Music and Mind 2019*. September. ISCA: ISCA, sep. de 2019, págs. 11-15. DOI: [10.21437/SMM.2019-3](https://doi.org/10.21437/SMM.2019-3). URL: [http://www.isca-speech.org/archive/SMM%7B%5C\\_%7D2019/abstracts/SMM19%7B%5C\\_%7Dpaper%7B%5C\\_%7D9.html](http://www.isca-speech.org/archive/SMM%7B%5C_%7D2019/abstracts/SMM19%7B%5C_%7Dpaper%7B%5C_%7D9.html) (vid. págs. 28, 94).
- [173] E. Rodero. «Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices». En: *Computers in Human Behavior* 77 (dic. de 2017), págs. 336-346. ISSN: 07475632. DOI: [10.1016/j.chb.2017.08.044](https://doi.org/10.1016/j.chb.2017.08.044). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0747563217305058> (vid. pág. 92).

- [174] A. Rosenberg. «Automatic detection and classification of prosodic events». PhD Thesis. Columbia University, 2009. URL: [http://www.cs.columbia.edu/nlp/theses/andrew%7B%5C\\_%7Drosenberg.pdf](http://www.cs.columbia.edu/nlp/theses/andrew%7B%5C_%7Drosenberg.pdf) (vid. págs. 83, 89).
- [175] H. F. Rueda Ch., C. V. Correa P. y H. Arguello Fuentes. «Design and Development of a Speech Synthesis Software for Colombian Spanish Applied To Communication Through Mobile Devices». En: *DYNA* 79.173 (2012), págs. 71-80. URL: [http://www.scielo.org.co/scielo.php?script=sci%7B%5C\\_%7Darttext%7B%5C%&%7Dpid=S0012-73532012000300023](http://www.scielo.org.co/scielo.php?script=sci%7B%5C_%7Darttext%7B%5C%&%7Dpid=S0012-73532012000300023) (vid. págs. 1, 45).
- [176] S. A. Sabab y M. H. Ashmafee. «Blind Reader: An intelligent assistant for blind». En: *2016 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, dic. de 2016, págs. 229-234. ISBN: 978-1-5090-4090-2. DOI: 10.1109/ICCITECHN.2016.7860200. URL: <http://ieeexplore.ieee.org/document/7860200/> (vid. pág. 10).
- [177] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi y H. Saruwatari. «UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 4521-4525. DOI: 10.21437/Interspeech.2022-439. arXiv: 2204.02152. URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/saeki22c%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/saeki22c%7B%5C_%7Dinterspeech.html) (vid. pág. 56).
- [178] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martínez-Liconá, H. L. Rufiner y J. Goddard. «Deep Learning for Emotional Speech Recognition». En: 2014, págs. 311-320. ISBN: 9780735440425. DOI: 10.1007/978-3-319-07491-7\_32. URL: [http://link.springer.com/10.1007/978-3-319-07491-7%7B%5C\\_%7D32](http://link.springer.com/10.1007/978-3-319-07491-7%7B%5C_%7D32) (vid. pág. 28).
- [179] N. Schilling y A. Marsters. «Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes». En: *Annual Review of Applied Linguistics* 35 (mar. de 2015), págs. 195-214. ISSN: 0267-1905. DOI: 10.1017/S0267190514000282. URL: [https://www.cambridge.org/core/product/identifier/S0267190514000282/type/journal%7B%5C\\_%7Darticle](https://www.cambridge.org/core/product/identifier/S0267190514000282/type/journal%7B%5C_%7Darticle) (vid. pág. 28).
- [180] M. Schröder. «The SEMAINE API: towards a standards-based framework for building emotion-oriented systems». En: *Advances in human-computer interaction* (2010), págs. 1-29. URL: <http://dl.acm.org/citation.cfm?id=1809194> (vid. pág. 31).
- [181] J. Schroeter. «Basic Principles of Speech Synthesis». En: *Springer Handbook of Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Cap. 19, págs. 413-428. DOI: 10.1007/978-3-540-49127-9\_19. URL: [http://link.springer.com/10.1007/978-3-540-49127-9%7B%5C\\_%7D19](http://link.springer.com/10.1007/978-3-540-49127-9%7B%5C_%7D19) (vid. pág. 7).

- [182] B. Schuller, S. Steidl y A. Batliner. «The INTERSPEECH 2009 emotion challenge». En: *Interspeech 2009*. ISCA: ISCA, sep. de 2009, págs. 312-315. DOI: [10.21437/Interspeech.2009-103](https://doi.org/10.21437/Interspeech.2009-103). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2009/schuller09%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2009/schuller09%7B%5C_%7Dinterspeech.html) (vid. pág. 58).
- [183] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller y S. Narayanan. «Paralinguistics in speech and language—State-of-the-art and the challenge». En: *Computer Speech & Language* 27.1 (ene. de 2013), págs. 4-39. ISSN: 08852308. DOI: [10.1016/j.csl.2012.02.005](https://doi.org/10.1016/j.csl.2012.02.005). URL: <http://linkinghub.elsevier.com/retrieve/pii/S0885230812000162> (vid. pág. 28).
- [184] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller y S. S. Narayanan. «The INTERSPEECH 2010 paralinguistic challenge». En: *Interspeech 2010*. September. ISCA: ISCA, sep. de 2010, págs. 2794-2797. DOI: [10.21437/Interspeech.2010-739](https://doi.org/10.21437/Interspeech.2010-739). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2010/schuller10b%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2010/schuller10b%7B%5C_%7Dinterspeech.html) (vid. pág. 58).
- [185] B. Schuller, S. Steidl, A. Batliner, F. Schiel y J. Krajewski. «The INTERSPEECH 2011 speaker state challenge». En: *Interspeech 2011*. August. ISCA: ISCA, ago. de 2011, págs. 3201-3204. DOI: [10.21437/Interspeech.2011-801](https://doi.org/10.21437/Interspeech.2011-801). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2011/schuller11b%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2011/schuller11b%7B%5C_%7Dinterspeech.html) (vid. pág. 58).
- [186] B. Schuller y col. «The INTERSPEECH 2012 speaker trait challenge». En: *Interspeech 2012*. Vol. 1. ISCA: ISCA, sep. de 2012, págs. 254-257. DOI: [10.21437/Interspeech.2012-86](https://doi.org/10.21437/Interspeech.2012-86). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2012/schuller12%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2012/schuller12%7B%5C_%7Dinterspeech.html) (vid. pág. 58).
- [187] B. Schuller y col. «The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism». En: *Interspeech 2013*. August. ISCA: ISCA, ago. de 2013, págs. 148-152. DOI: [10.21437/Interspeech.2013-56](https://doi.org/10.21437/Interspeech.2013-56). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2013/schuller13%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2013/schuller13%7B%5C_%7Dinterspeech.html) (vid. pág. 58).
- [188] S. K. Scott e I. S. Johnsrude. «The neuroanatomical and functional organization of speech perception.» En: *Trends in neurosciences* 26.2 (feb. de 2003), págs. 100-7. ISSN: 0166-2236. DOI: [10.1016/S0166-2236\(02\)00037-1](https://doi.org/10.1016/S0166-2236(02)00037-1). URL: <http://www.ncbi.nlm.nih.gov/pubmed/12536133> (vid. pág. 36).

- [189] M. Shamsi, J. Chevelu, N. Barbot y D. Lolive. «Corpus design for expressive speech: impact of the utterance length». En: *10th International Conference on Speech Prosody 2020*. May. ISCA: ISCA, mayo de 2020, págs. 955-959. DOI: [10.21437/speechprosody.2020-195](https://doi.org/10.21437/speechprosody.2020-195). URL: [http://www.isca-speech.org/archive/SpeechProsody%7B%5C\\_%7D2020/abstracts/199.html](http://www.isca-speech.org/archive/SpeechProsody%7B%5C_%7D2020/abstracts/199.html) (vid. pág. 15).
- [190] S. Shirali-Shahreza y G. Penn. «MOS Naturalness and the Quest for Human-Like Speech». En: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, dic. de 2018, págs. 346-352. ISBN: 978-1-5386-4334-1. DOI: [10.1109/SLT.2018.8639599](https://doi.org/10.1109/SLT.2018.8639599). URL: <https://ieeexplore.ieee.org/document/8639599/> (vid. pág. 50).
- [191] R. Silipo y S. Greenberg. «Prosodic Stress Revisited : Reassessing the Role of Fundamental Frequency». En: *Proceedings of the NIST Speech Transcription Workshop*. 2000 (vid. pág. 89).
- [192] D. Sityaev, K. Knill y T. Burrows. «Comparison of the ITU-T P . 85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems». En: *INTERSPEECH 2006 - ICSLP Ninth International Conference on Spoken Language Processing*. 2006, págs. 1077-1080 (vid. pág. 34).
- [193] A. Sluijter, E. Bosgoed, J. Kerkhoff, E. Meier, T. Rietveld, A. Sanderman, M. Swerts y J. Terken. «Evaluation of speech synthesis systems for Dutch in telecommunication applications». En: *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*. 1998. URL: [http://www.isca-speech.org/archive/ssw3/ssw3%7B%5C\\_%7D213.html](http://www.isca-speech.org/archive/ssw3/ssw3%7B%5C_%7D213.html) (vid. pág. 34).
- [194] E. Sohoglu, J. E. Peelle, R. P. Carlyon y M. H. Davis. «Predictive Top-Down Integration of Prior Knowledge during Speech Perception». En: *Journal of Neuroscience* 32.25 (jun. de 2012), págs. 8443-8453. ISSN: 0270-6474. DOI: [10.1523/JNEUROSCI.5069-11.2012](https://doi.org/10.1523/JNEUROSCI.5069-11.2012). URL: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5069-11.2012> (vid. pág. 21).
- [195] A. Stan. «The ZevomOS entry to VoiceMOS Challenge 2022». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 4516-4520. DOI: [10.21437/Interspeech.2022-105](https://doi.org/10.21437/Interspeech.2022-105). arXiv: [2206.07448](https://arxiv.org/abs/2206.07448). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/stan22%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/stan22%7B%5C_%7Dinterspeech.html) (vid. pág. 56).
- [196] D. Stanton, Y. Wang y R. Skerry-Ryan. «Predicting Expressive Speaking Style from Text in End-To-End Speech Synthesis». En: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, dic. de 2018, págs. 595-602. ISBN: 978-1-5386-4334-1. DOI: [10.1109/SLT.2018.8639682](https://doi.org/10.1109/SLT.2018.8639682). arXiv: [1808.01410](https://arxiv.org/abs/1808.01410). URL: <https://ieeexplore.ieee.org/document/8639682/> (vid. pág. 17).

- [197] S. S. Stevens y J. Volkman. «The Relation of Pitch to Frequency: A Revised Scale». En: *The American Journal of Psychology* 53.3 (jul. de 1940), pág. 329. ISSN: 00029556. DOI: [10.2307/1417526](https://doi.org/10.2307/1417526). URL: <https://www.jstor.org/stable/1417526?origin=crossref> (vid. pág. 98).
- [198] R. S. Sutton y A. G. Barto. *Reinforcement Learning*. 2.<sup>a</sup> ed. The MIT Press, 2018. ISBN: 9780262193986 (vid. pág. 32).
- [199] Z. S. Syed, J. Schroeter, K. Sidorov y D. Marshall. «Computational Paralinguistics: Automatic Assessment of Emotions, Mood and Behavioural State from Acoustics of Speech». En: *Interspeech 2018*. Vol. 2018-Septe. September 2018. ISCA: ISCA, sep. de 2018, págs. 511-515. DOI: [10.21437/Interspeech.2018-2019](https://doi.org/10.21437/Interspeech.2018-2019). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2018/syed18%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2018/syed18%7B%5C_%7Dinterspeech.html) (vid. pág. 111).
- [200] P. Taylor. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009. ISBN: 9780511816338. DOI: [10.1017/CB09780511816338](https://doi.org/10.1017/CB09780511816338). URL: <http://ebooks.cambridge.org/ref/id/CB09780511816338> (vid. págs. 14, 32).
- [201] P. Taylor y A. W. Black. «Assigning phrase breaks from part-of-speech sequences». En: *Computer Speech & Language* 12.2 (abr. de 1998), págs. 99-117. ISSN: 08852308. DOI: [10.1006/csla.1998.0041](https://doi.org/10.1006/csla.1998.0041). URL: <https://linkinghub.elsevier.com/retrieve/pii/S088523089800419> (vid. pág. 8).
- [202] J. Terken. «Fundamental frequency and perceived prominence of accented syllables». En: *The Journal of the Acoustical Society of America* 89.4 (abr. de 1991), págs. 1768-1776. ISSN: 0001-4966. DOI: [10.1121/1.401019](https://doi.org/10.1121/1.401019). URL: <http://asa.scitation.org/doi/10.1121/1.401019> (vid. pág. 75).
- [203] X. Tian, K. Fu, S. Gao, Y. Gu, K. Wang, W. Li y Z. Ma. «A Transfer and Multi-Task Learning based Approach for MOS Prediction». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 5438-5442. DOI: [10.21437/Interspeech.2022-10022](https://doi.org/10.21437/Interspeech.2022-10022). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/tian22d%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/tian22d%7B%5C_%7Dinterspeech.html) (vid. pág. 56).
- [204] G. Toledo. «Señales prosódicas del foco». En: *Revista Argentina de Lingüística* 5.1-2 (1998), págs. 205-230 (vid. pág. 78).
- [205] H. M. Torres, J. A. Gurlekian y C. Cossio-Mercado. «Aromo: Argentine Spanish TTS System». En: *IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*. 2012, págs. 416-421 (vid. págs. 1, 18, 20, 45).



- [206] H. M. Torres, J. A. Gurlekian, D. A. Evin y C. G. Cossio Mercado. «Emilia: a speech corpus for Argentine Spanish text to speech synthesis». En: *Language Resources and Evaluation* 53.3 (sep. de 2019), págs. 419-447. ISSN: 1574-020X. DOI: [10.1007/s10579-019-09447-7](https://doi.org/10.1007/s10579-019-09447-7). URL: <http://link.springer.com/10.1007/s10579-019-09447-7> (vid. págs. 16, 19, 20).
- [207] H. M. Torres. «Generación automática de la prosodia para un sistema de conversión de texto a habla». Tesis Doctoral. Universidad de Buenos Aires, 2008 (vid. págs. 7-9, 18).
- [208] H. Traunmüller y A. Eriksson. «Acoustic effects of variation in vocal effort by men, women, and children». En: *The Journal of the Acoustical Society of America* 107.6 (jun. de 2000), págs. 3438-3451. ISSN: 0001-4966. DOI: [10.1121/1.429414](https://doi.org/10.1121/1.429414). URL: <http://asa.scitation.org/doi/10.1121/1.429414> (vid. pág. 83).
- [209] J. Trouvain y K. P. Truong. «Comparing non-verbal vocalisations in conversational speech corpora». En: *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)*. January. European Language Resources Association (ELRA), 2012, págs. 36-39. URL: <http://doc.utwente.nl/80906/> (vid. pág. 111).
- [210] W.-C. Tseng, W.-T. Kao y H.-y. Lee. «DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 4541-4545. DOI: [10.21437/Interspeech.2022-11247](https://doi.org/10.21437/Interspeech.2022-11247). arXiv: [2204.03219](https://arxiv.org/abs/2204.03219). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/tseng22b%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/tseng22b%7B%5C_%7Dinterspeech.html) (vid. pág. 56).
- [211] UIT-T. *P.85: Método para la Evaluación Subjetiva de la Calidad Vocal de los Dispositivos Generadores de Voz*. 1994 (vid. págs. 2, 34, 45, 55).
- [212] UIT-T. *P.800: Métodos de determinación subjetiva de la calidad de transmisión*. 1996 (vid. págs. 33, 45).
- [213] UIT-T. *P.50: Voces artificiales*. 1999 (vid. pág. 7).
- [214] UIT-T. *P.563: Método basado en un solo extremo para la evaluación objetiva de la calidad vocal en aplicaciones de telefonía de banda estrecha*. 2004 (vid. pág. 35).
- [215] UIT-T. *P.800.1: Terminología de las notas medias de opinión*. 2006 (vid. pág. 45).
- [216] M. Vainio, J. Jarvikivi, S. Werner, N. Volk y J. Valikangas. «Effect of prosodic naturalness on segmental acceptability in synthetic speech». En: *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002*. IEEE, 2002, págs. 143-146. ISBN: 0-7803-7395-2. DOI:

- 10.1109/WSS.2002.1224394. URL: <http://ieeexplore.ieee.org/document/1224394/> (vid. pág. 75).
- [217] Y. Vazquez Alvarez y M. Huckvale. «The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems». En: *7th International Conference on Spoken Language Processing, ICSLP 2002*. 2002, págs. 329-332 (vid. pág. 34).
- [218] K. Venkataramanan y H. R. Rajamohan. «Emotion Recognition from Speech». En: *Cognitive Technologies 9783319436647* (dic. de 2019), págs. 409-428. arXiv: 1912.10458. URL: <http://arxiv.org/abs/1912.10458> (vid. págs. 28, 98, 100).
- [219] J. Vidal, L. Ferrer y L. Brambilla. «EpaDB: A Database for Development of Pronunciation Assessment Systems». En: *Interspeech 2019*. Vol. 2019-Sept. ISCA: ISCA, sep. de 2019, págs. 589-593. DOI: 10.21437/Interspeech.2019-1839. URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2019/abstracts/1839.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2019/abstracts/1839.html) (vid. pág. 111).
- [220] S. T. Völkel, R. Schödel, D. Buschek, C. Stachl, V. Winterhalter, M. Bühner y H. Hussmann. «Developing a Personality Model for Speech-based Conversational Agents Using the Psychological Approach». En: (2020). DOI: 10.1145/3313831.3376210. arXiv: 2003.06186. URL: [http://arxiv.org/abs/2003.06186%7B%5C\\_%7D0Ahttp://dx.doi.org/10.1145/3313831.3376210](http://arxiv.org/abs/2003.06186%7B%5C_%7D0Ahttp://dx.doi.org/10.1145/3313831.3376210) (vid. pág. 28).
- [221] P. Wagner y S. Betz. «Speech Synthesis Evaluation: Realizing a Social Turn». En: 28. *Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2017*. 2017, págs. 167-173. URL: <http://www.essv.de/paper.php?id=234> (vid. págs. 2, 5).
- [222] P. Wagner y col. «Different Parts of the Same Elephant: a Roadmap To Disentangle and Connect Different Perspectives on Prosodic Prominence». En: *18th International Congress of Phonetic Sciences (ICPhS 2015)*. 2015. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0202.pdf> (vid. pág. 77).
- [223] P. Wagner y col. «Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program». En: *10th ISCA Speech Synthesis Workshop*. September. ISCA: ISCA, sep. de 2019, págs. 105-110. DOI: 10.21437/SSW.2019-19. URL: [http://www.isca-speech.org/archive/SSW%7B%5C\\_%7D2019/abstracts/SSW10%7B%5C\\_%7D0%7B%5C\\_%7D3-2.html](http://www.isca-speech.org/archive/SSW%7B%5C_%7D2019/abstracts/SSW10%7B%5C_%7D0%7B%5C_%7D3-2.html) (vid. págs. 5, 10, 46).
- [224] Y. Wang y col. «Tacotron: Towards End-to-End Speech Synthesis». En: *Interspeech 2017*. ISCA: ISCA, ago. de 2017, págs. 4006-4010. DOI: 10.21437/Interspeech.2017-1452. URL: <http://www>.



- [isca-speech.org/archive/Interspeech%7B%5C\\_%7D2017/abstracts/1452.html](http://isca-speech.org/archive/Interspeech%7B%5C_%7D2017/abstracts/1452.html) (vid. pág. 17).
- [225] S. J. Winters y D. B. Pisoni. *Perception and Comprehension of Synthetic Speech*. Inf. téc. 2004 (vid. pág. 24).
- [226] I. H. Witten, E. Frank, M. A. Hall y C. J. Pal. *Data Mining*. Elsevier, 2017. ISBN: 9780128042915. DOI: [10.1016/C2015-0-02071-8](https://doi.org/10.1016/C2015-0-02071-8). URL: <https://linkinghub.elsevier.com/retrieve/pii/C20150020718> (vid. pág. 83).
- [227] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino e Y. Zhao. «Fusion of Self-supervised Learned Models for MOS Prediction». En: *Interspeech 2022*. Vol. 2022-Septe. September. ISCA: ISCA, sep. de 2022, págs. 5443-5447. DOI: [10.21437/Interspeech.2022-10262](https://doi.org/10.21437/Interspeech.2022-10262). arXiv: [2204.04855](https://arxiv.org/abs/2204.04855). URL: [https://www.isca-speech.org/archive/interspeech%7B%5C\\_%7D2022/yang220%7B%5C\\_%7Dinterspeech.html](https://www.isca-speech.org/archive/interspeech%7B%5C_%7D2022/yang220%7B%5C_%7Dinterspeech.html) (vid. pág. 56).
- [228] G. N. Yannakakis, R. Cowie y C. Busso. «The Ordinal Nature of Emotions: An Emerging Approach». En: *IEEE Transactions on Affective Computing* (2018). ISSN: 1949-3045. DOI: [10.1109/TAFFC.2018.2879512](https://doi.org/10.1109/TAFFC.2018.2879512). URL: <https://ieeexplore.ieee.org/document/8521685/> (vid. pág. 4).
- [229] G. N. Yannakakis y J. Hallam. «Ranking vs. Preference: A Comparative Study of Self-reporting». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 6974 LNCS. PART 1. 2011, págs. 437-446. ISBN: 9783642245992. DOI: [10.1007/978-3-642-24600-5\\_47](https://doi.org/10.1007/978-3-642-24600-5_47). URL: [http://link.springer.com/10.1007/978-3-642-24600-5%7B%5C\\_%7D47](http://link.springer.com/10.1007/978-3-642-24600-5%7B%5C_%7D47) (vid. pág. 4).
- [230] G. N. Yannakakis y H. P. Martínez. «Ratings are Overrated!» En: *Frontiers in ICT* 2.JUL (jul. de 2015), pág. 1. ISSN: 2297-198X. DOI: [10.3389/fict.2015.00013](https://doi.org/10.3389/fict.2015.00013). URL: <http://journal.frontiersin.org/Article/10.3389/fict.2015.00013/abstract> (vid. pág. 4).
- [231] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi y T. Kitamura. «Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis». En: *6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*. 1999. URL: [https://www.isca-speech.org/archive/eurospeech%7B%5C\\_%7D1999/e99%7B%5C\\_%7D2347.html](https://www.isca-speech.org/archive/eurospeech%7B%5C_%7D1999/e99%7B%5C_%7D2347.html) (vid. pág. 16).
- [232] H. Zen, K. Tokuda y A. W. Black. «Statistical parametric speech synthesis». En: *Speech Communication* 51.11 (nov. de 2009), págs. 1039-1064. ISSN: 01676393. DOI: [10.1016/j.specom.2009.04.004](https://doi.org/10.1016/j.specom.2009.04.004). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167639309000648>

- 20<http://www.sciencedirect.com/science/article/pii/S0167639309000648> (vid. pág. 16).
- [233] R. Zequeira Jimenez, L. F. Gallardo y S. Möller. «Scoring voice likability using pair-comparison: Laboratory vs. crowdsourcing approach». En: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, mayo de 2017, págs. 1-3. ISBN: 978-1-5386-4024-1. DOI: [10.1109/QoMEX.2017.7965678](https://doi.org/10.1109/QoMEX.2017.7965678). URL: <http://ieeexplore.ieee.org/document/7965678/> (vid. pág. 55).
- [234] R. Zequeira Jiménez, L. Fernández Gallardo y S. Möller. «Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing.» En: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. New York, New York, USA: ACM Press, 2018, págs. 1127-1130. ISBN: 9781450356404. DOI: [10.1145/3184558.3191545](https://doi.org/10.1145/3184558.3191545). URL: <http://dl.acm.org/citation.cfm?doid=3184558.3191545> (vid. pág. 55).
- [235] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang e Y. Tsao. «Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features». En: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), págs. 54-70. ISSN: 2329-9290. DOI: [10.1109/TASLP.2022.3205757](https://doi.org/10.1109/TASLP.2022.3205757). arXiv: [2111.02363](https://arxiv.org/abs/2111.02363). URL: <https://ieeexplore.ieee.org/document/9905733/> (vid. pág. 56).
- [236] X. Zhang, Y. C. Wu y L. L. Holt. «The Learning Signal in Perceptual Tuning of Speech: Bottom Up Versus Top-Down Information». En: *Cognitive Science* 45.3 (mar. de 2021), págs. 1-24. ISSN: 0364-0213. DOI: [10.1111/cogs.12947](https://doi.org/10.1111/cogs.12947). URL: <https://onlinelibrary.wiley.com/doi/10.1111/cogs.12947> (vid. pág. 21).
- [237] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg y B. Ramabhadran. «Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning». En: *INTERSPEECH 2019*. ISCA: ISCA, sep. de 2019, págs. 2080-2084. DOI: [10.21437/Interspeech.2019-2668](https://doi.org/10.21437/Interspeech.2019-2668). URL: [http://www.isca-speech.org/archive/Interspeech%7B%5C\\_%7D2019/abstracts/2668.html](http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2019/abstracts/2668.html) (vid. págs. 17, 45).
- [238] M. L. Zubizarreta. «Las funciones informativas: tema y foco». En: *Gramática descriptiva de la lengua española*. Ed. por I. Bosque y V. Demonte. Espasa Calpe, 1999, págs. 4215-4244. ISBN: 84-239-7917-2 (vid. págs. 76, 84).