

Reconocimiento Automático del Habla

2023-2024

Conversión de texto en habla



DEPARTAMENT DE SISTEMES
INFORMÀTICS I COMPUTACIÓ



MIARFID-RAH mcastro@dsic.upv.es

How Speech Synthesizers Work

Cómo funcionan los sintetizadores de voz



Open the pod bay doors, HAL. I'm sorry Dave,
I'm afraid I can't do that.

Índice

- Introducción
- Estructura de un sistema TTS
- Análisis del texto
- Modelado prosódico
- Técnicas de síntesis
- Evaluación
- *Trending topics*

Este material está elaborado a partir del material de Eva Navas, Inma Hernández, Ibon Saratxaga y Daniel Erro para el curso de “Speech Technologies” impartido en el Master Erasmus Mundus Language & Communication Technologies
<http://aholab.ehu.eus>

Introducción

- La voz es la forma más simple de comunicación entre humanos:
 - ¿Por qué no entre humanos y máquinas?
- Síntesis de habla
 - Es la producción artificial del habla
- Conversión de texto en habla (“Text-To-Speech” TTS)
 - Producir automáticamente una señal de voz con el mensaje contenido en el texto



Lectura artificial de texto

Introducción

- Conversión TTS → multidisciplinar
 - Procesado de señal
 - Fonología
 - Morfología
 - Sintaxis
 - Programación
 - Electrónica
 - *Machine Learning!*
 -

Introducción

<http://speech.zone/courses/speech-synthesis/module-1-introduction/historical-examples/>



VODER (1939) - Early Speech Synthesizer.mp4

Dudley's "Voder"

1939

1978



[Speak N' Spell commercial 1980](#)

<http://www.speaknspell.co.uk/>

infovox

1994

1997

ivona™
Text-To-Speech



The Festival Speech
Synthesis System
[Demo](#)

Google
Assistant
[Demo](#)

[loquendo](#)

Introducción

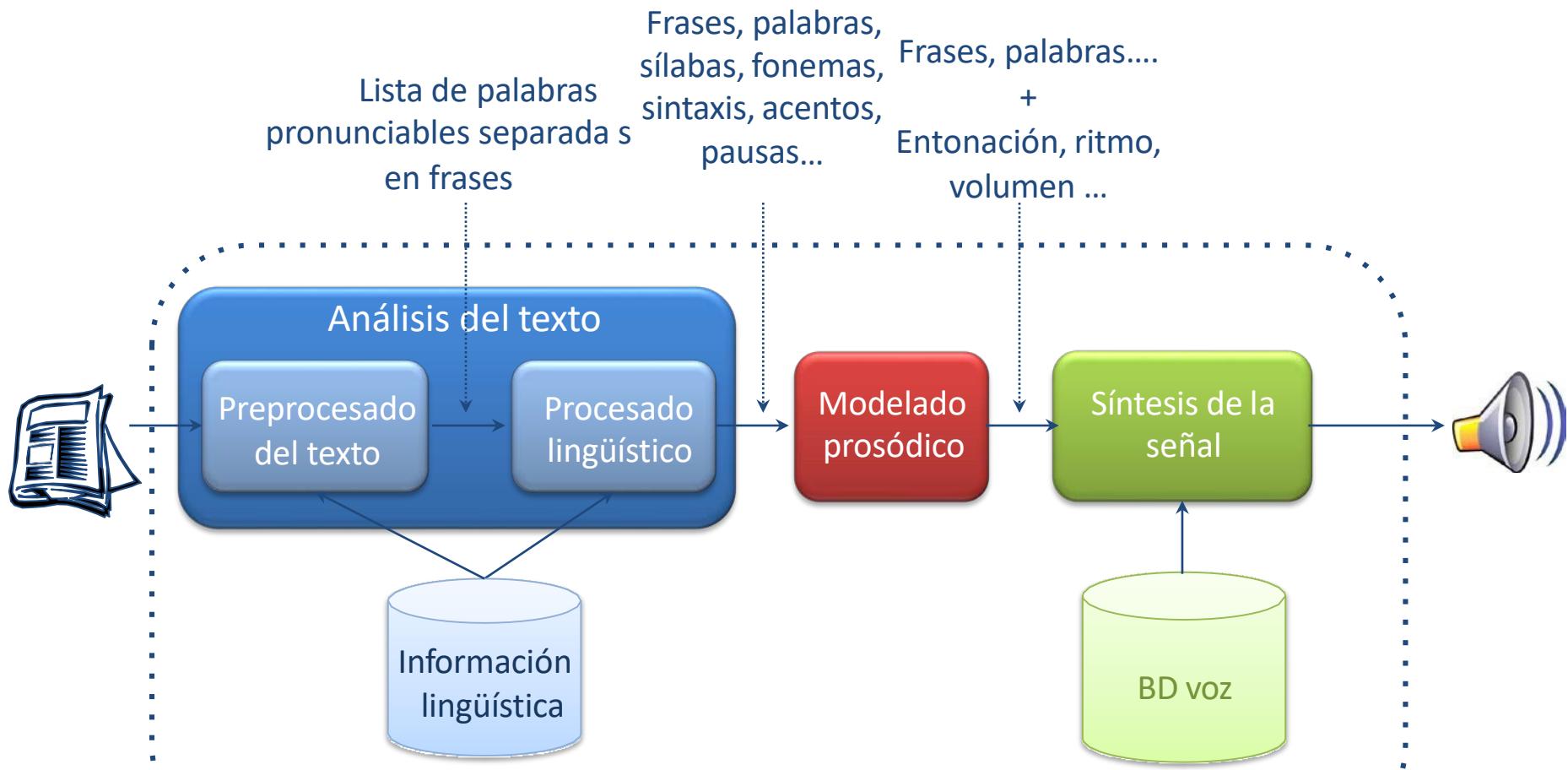
- Aplicaciones:
 - Asistentes virtuales
 - Búsqueda por voz
 - Call centers
 - Videojuegos
 - Traducción voz-voz
 - Lectores de documentos/web para ciegos
 - Herramientas de comunicación alternativa
 - Interacción con robots
 - Enseñanza de idiomas
 - Información en aeropuertos, estaciones, trenes...
 -



Índice

- Introducción
- Estructura de un sistema TTS
- Análisis del texto
- Modelado prosódico
- Técnicas de síntesis
- Evaluación
- Trending topics

Estructura de un sistema TTS



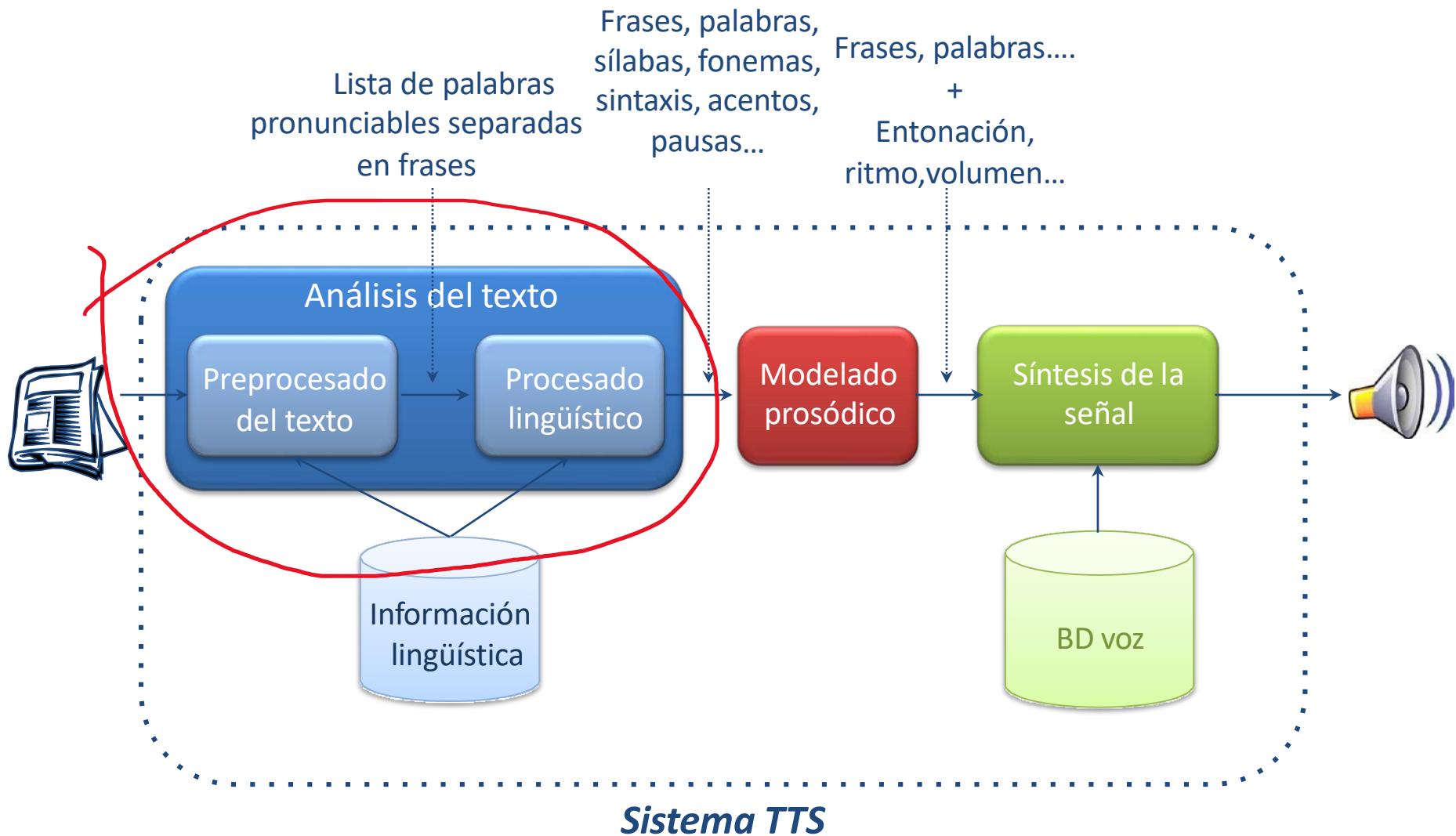
Sistema TTS

*En los sistemas actuales, las fronteras entre módulos no están tan claras.

Índice

- Introducción
- Estructura de un sistema TTS
- Análisis del texto
 - Preprocesado del texto
 - Procesado lingüístico
- Modelado prosódico
- Técnicas de síntesis
- Evaluación
- Trending topics

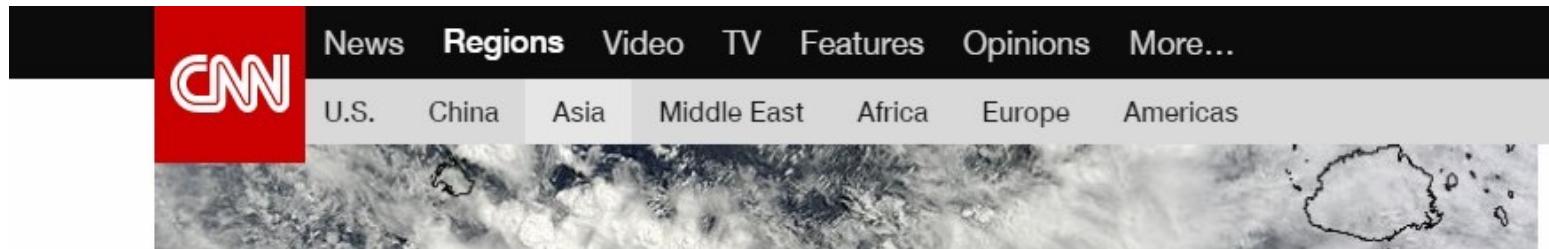
Estructura de un sistema TTS



Sistema TTS

*En los sistemas actuales, las fronteras entre módulos no están tan claras.

Análisis del texto



Cyclone Pam heads towards the South Pacific islands of Vanuatu (left) on March 11.



Story highlights

Tropical Cyclone Pam is a Category 5 storm that is heading towards Vanuatu

Storm is projected to hit country's capital Friday night local time

Meteorological Services warned of "very destructive winds and very rough to phenomenal seas with heavy swells."



Cyclone Pam's projected path

(CNN)—Tropical Cyclone Pam, one of the strongest storms seen in the South Pacific in years, is churning towards the country of Vanuatu and triggering storm warnings in other island nations.

The Category 5 storm is whipping up winds of 160 miles per hour (260 kph), triggering concerns of torrential rainfall, flooding and landslides. The Vanuatu

Evacuation alerts have been issued for several parts of the country.

The storm is expected to intensify in the next 12 to 24 hours as it heads southwards, according to the Joint Typhoon Warning Center. The storm is projected to pass Vanuatu's capital, Port Vila on Friday evening. The capital sits on the coastline, which is vulnerable to storm surges during powerful cyclones.

Análisis del texto

- Dependiente del idioma
- Dos partes:
 - *Preprocesado del texto*: transforma la entrada en una secuencia de **frases**, cada una formada por una secuencia de **palabras legibles**.
 - *Análisis lingüístico*: transforma texto legible en una secuencia de etiquetas fonéticas y lingüísticas.

Análisis del texto: Preprocesado

- Convierte el texto en pronunciable
 - Identifica patrones y los convierte en palabras (normalización o expansión)
- Dependiente del idioma
- Usa diccionarios o lexicones
- Reduce la variabilidad tipográfica
 - Mayúsculas, caracteres especiales, números...
- Tareas:
 - Tarea 1: de texto a frases
 - Tarea 2: de frases a patrones
 - Tarea 3: de patrones a palabras

Análisis de texto: Preprocesado, Tarea 1 (de texto a frases)

- Tratamiento de los signos de puntuación
 - Básico para detectar el fin de frase
 - Algunos signos de puntuación son difíciles de interpretar
 - Punto
 - Abreviaturas
 - Números
 - Signo de multiplicación
 - Signo de puntuación
 - Direcciones WEB
 - Guiones
 - Fechas
 - Como una coma
 - Palabras compuestas
 - Continuador de línea
 - Resultados deportivos



Clarkson was suspended after a "fracas" with a BBC producer

Top Gear host Jeremy Clarkson initiated the BBC investigation which prompted his suspension, after he informed BBC bosses of the alleged "fracas".

BBC News understands that the star phoned BBC head of television, Danny Cohen, to report the incident.

Producer Oisin Tymon, with whom the altercation took place, is not believed to have filed his own complaint.

Interviews are expected to be held with the star and other parties next week, and the show has been taken off-air.

Clarkson has expressed regret over the incident, which his co-presenter James May labelled "a bit of a dust-up".

An online petition calling for the star's reinstatement - set up by political blogger Guido Fawkes - has accrued more than 800,000 signatures since Tuesday.

Análisis de texto: Preprocesado , Tarea 2 (de frases a patrones)

- Identificación de tokens
 - Patrones: palabras potenciales
 - Normalmente delimitados por espacios en blanco o signos de puntuación

Esto es un ejemplo sencillo.

A veces, faltan espacios por errores de escritura...

Otras veces los patrones son más complejos como en 10cm, por ejemplo.

Análisis de texto: Preprocesado, Tarea 3 (de patrones a palabras)

- Expansión o normalización
 - Acrónimos ccoo  comisiones obreras
 - Abreviaturas dcha.  derecha
 - Números 12  doce
 - Fechas 1/11/15  uno de noviembre de dos mil quince
 - Horas 10:30  las diez y media
 - Unidades 1 kg  un kilo
 - URL, email ...
- Problemas
 - Ambigüedad en acrónimos y siglas
 - Concordancia de género y número
 - Incorrekiones ortográficas

Análisis de texto: Procesado lingüístico

- Estructura



Análisis de texto: Procesado lingüístico

- Categorización
 - Asignar categoría a las palabras (*Part of Speech*, POS): nombre, verbo, preposición... además de relaciones sintáticas
 - Por regla, con métodos estadísticos o ML (DNN)
 - La categoría influye en la colocación de las pausas, acentuación y en la predicción de la prosodia (entonación)
 - Debe resolver problemas de disambigüación

I watch my watch. I read the book.

Análisis de texto: Procesado lingüístico

- Pausas
 - Colocar pausas donde sean necesarias, además de las indicadas por los signos de puntuación
 - Interdependencias entre prosodia y pausas, por eso a veces forma parte de la prosodia
 - Afectan a la transcripción fonética (coarticulación)
 - Importantes para obtener naturalidad (y para entender el significado en algunos casos)

Análisis de texto: Procesado lingüístico

- Transcripción fonética
 - Convertir la secuencia de caracteres ortográficos en secuencia de fonemas
 - Por regla o usando diccionarios, estadísticos¹, DNN²
 - Uso de alfabetos específicos para representar los fonemas
 - IPA
 - SAMPA
 - Uso de la transcripción estándar
 - Dificultad dependiente de la lengua
 - Problemas en nombres propios y palabras extranjeras
 - Problemas de mezcla de idiomas (catalán-castellano euskera-castellano, euskera-francés) (especialmente difícil si la lengua es morfológicamente compleja):

Banoa Leroy-Merlinera (“Me voy al Leroy-Merlin”)

- (1) Maximilian Bisani, Hermann Ney. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication* Elsevier : North-Holland, 2008, 50 (5), pp.434. <10.1016/j.specom.2008.01.002>.
- (2) Kaisheng Yao, Geoffrey Zweig. Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion, INTERSPEECH 2015, <https://arxiv.org/abs/1506.00196v3>

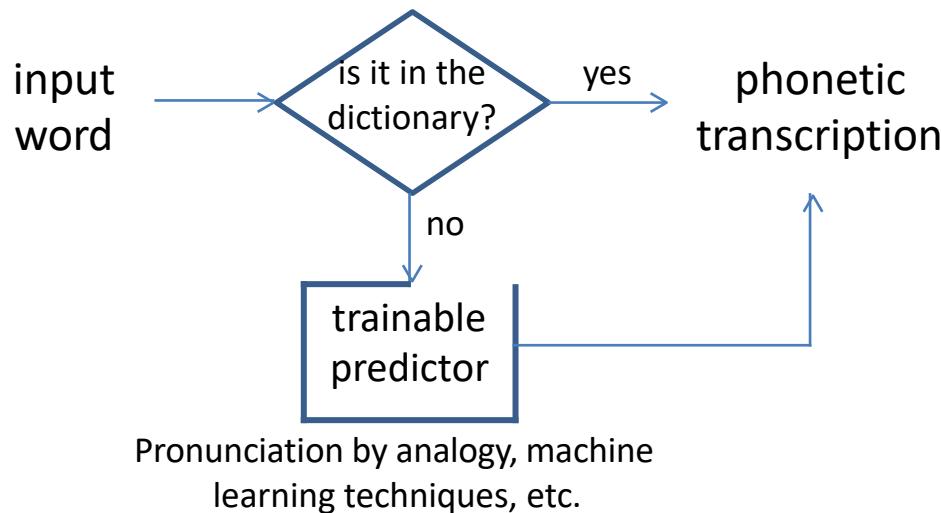
Análisis de texto: Procesado lingüístico

- De grafemas a fonemas en español

```
a: /a/  
b: if between vowels, /B/; else, /b/  
c: if ce/ci, /T/; else if ch, tS; else, /k/  
d: if between vowels, /D/; else, /d/  
e: /e/  
f: /f/  
g: if ge/gi, /x/; else if between vowels, /G/; else, /g/  
h: if ch, /tS/; else, //  
i: if before vowel, /j/; else, /i/  
j: /x/  
k: /k/  
l: if ll, /L/; else, /l/  
m: /m/  
n: /n/  
ñ: /J/  
o: /o/  
p: /p/  
q: /k/  
r: if rr or beginning of word, /rr/; else, /r/  
s: /s/  
t: /t/  
u: if gue/gui/que/qui, //; else if before vowel, /w/; else /u/  
v: if between vowels, /B/; else, /b/  
w: /w/  
x: /k/+/s/  
y: if before consonant, /i/; else, /j/  
z: /T/
```

Análisis de texto: Procesado lingüístico

- De grafemas a fonemas en inglés



Análisis de texto: Procesado lingüístico

- Silabificación
 - Dividir las palabras en sílabas
 - Excepciones → su-bli-me sub-lin-gual
 - Palabras problemáticas
- Acentuación
 - Asignar el acento a las sílabas correspondientes
 - Palabras según el acento
 - Tónicas → Necesaria la categoría gramatical para
 - Átonas → determinarlo

Análisis del texto: Todas las etapas

- Preprocesado:
 - Tarea 1: texto → frases
 - Tarea 2: frases → patrones
 - Tarea 3: patrones → palabra(s)
- Procesado lingüístico:
 - Tarea 1: categorías (part-of-speech tagging)
 - Tarea 2: inserción de pausas
 - Tarea 3: transcripción fonética
 - Tarea 4: silabificación
 - Tarea 5: acentuación

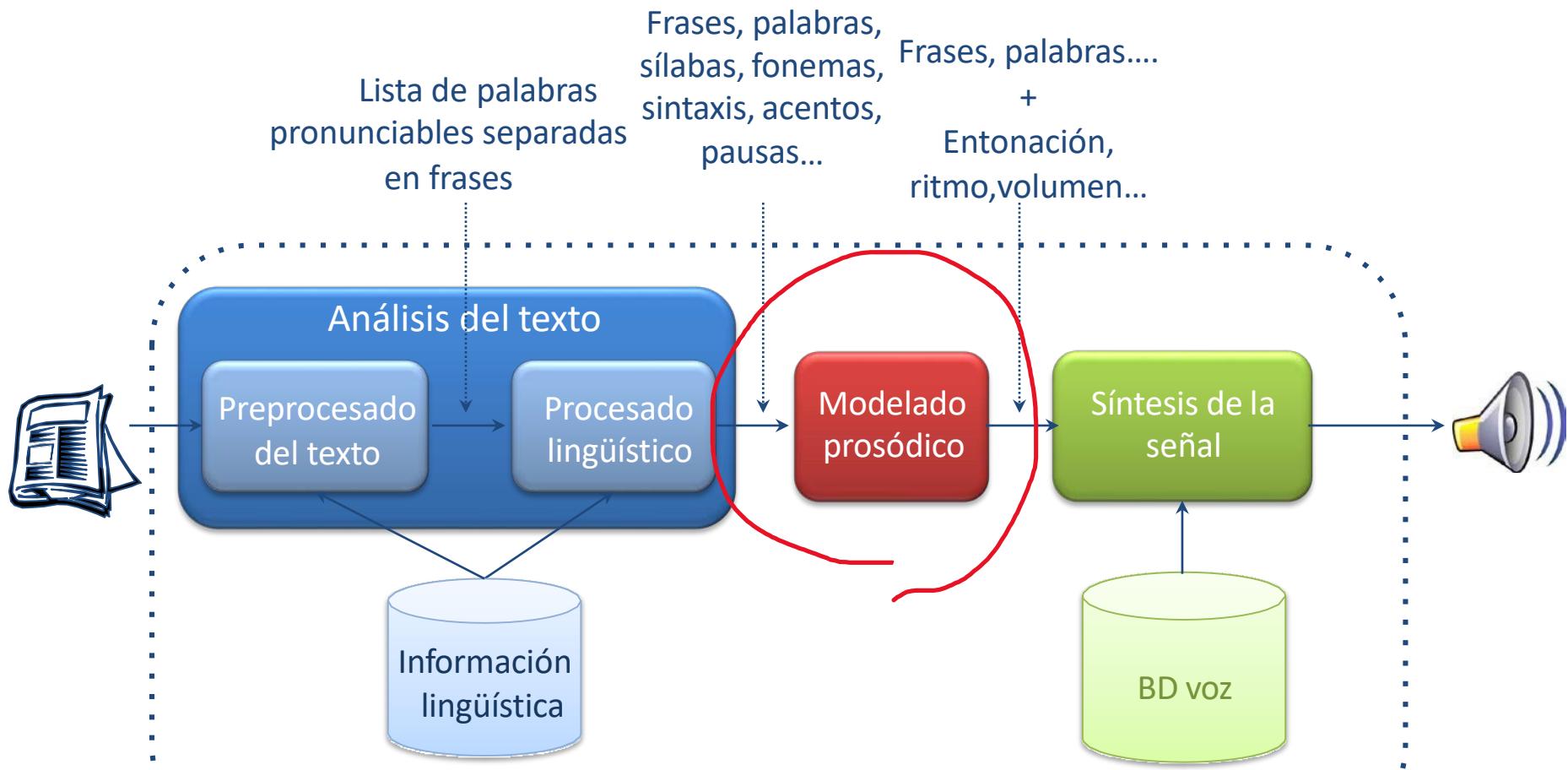
Análisis del texto

- Algunos analizadores de texto existentes:
 - Festival (Inglés y otros)
 - AhoTTS (Euskara, Castellano)
 - Festcat (Catalán)
 - Cotovia (Gallego)
- Herramientas de procesado de lenguaje natural para etiquetado POS y análisis sintáctico:
 - Freeling (análisis sintáctico en diferentes lenguas)
 - Wordnet
 - MORFEUS (Euskara)

Índice

- Introducción
- Estructura de un sistema TTS
- Análisis del texto
- Modelado prosódico
- Técnicas de síntesis
- Evaluación
- Trending topics

Estructura de un sistema TTS



Sistema TTS

*En los sistemas actuales, las fronteras entre módulos no están tan claras.

Modelado prosódico

- El texto de entrada se ha transformado en etiquetas: etiquetas fonéticas (fonemas), etiquetas lingüísticas (palabras, sílabas, acentos ...) y etiquetas prosódicas básicas (pausas)
- Objetivo del modelado prosódico: establecer una correspondencia entre las etiquetas y las **características prosódicas**

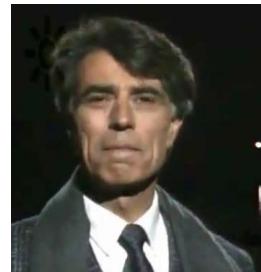
Modelado prosódico

- ¿Cuáles son las *dimensiones de la prosodia*?
 - Pausas
 - Entonación: Se relaciona con las variaciones tonales en el habla, como subir o bajar la voz al final de una oración para indicar una pregunta o mantenerla estable para una afirmación.
 - Ritmo: Incluye la cadencia y el patrón de acentuación en el habla. La velocidad a la que se habla, la duración de las sílabas y las pausas entre palabras también forman parte de esta dimensión.
 - Intensidad: Se refiere al volumen o la fuerza con la que se emite el habla.
 - Calidad de voz (normalmente no se considera en TTS)
 - Grado de articulación (tampoco considerado en TTS)
 - ...

Modelado prosódico

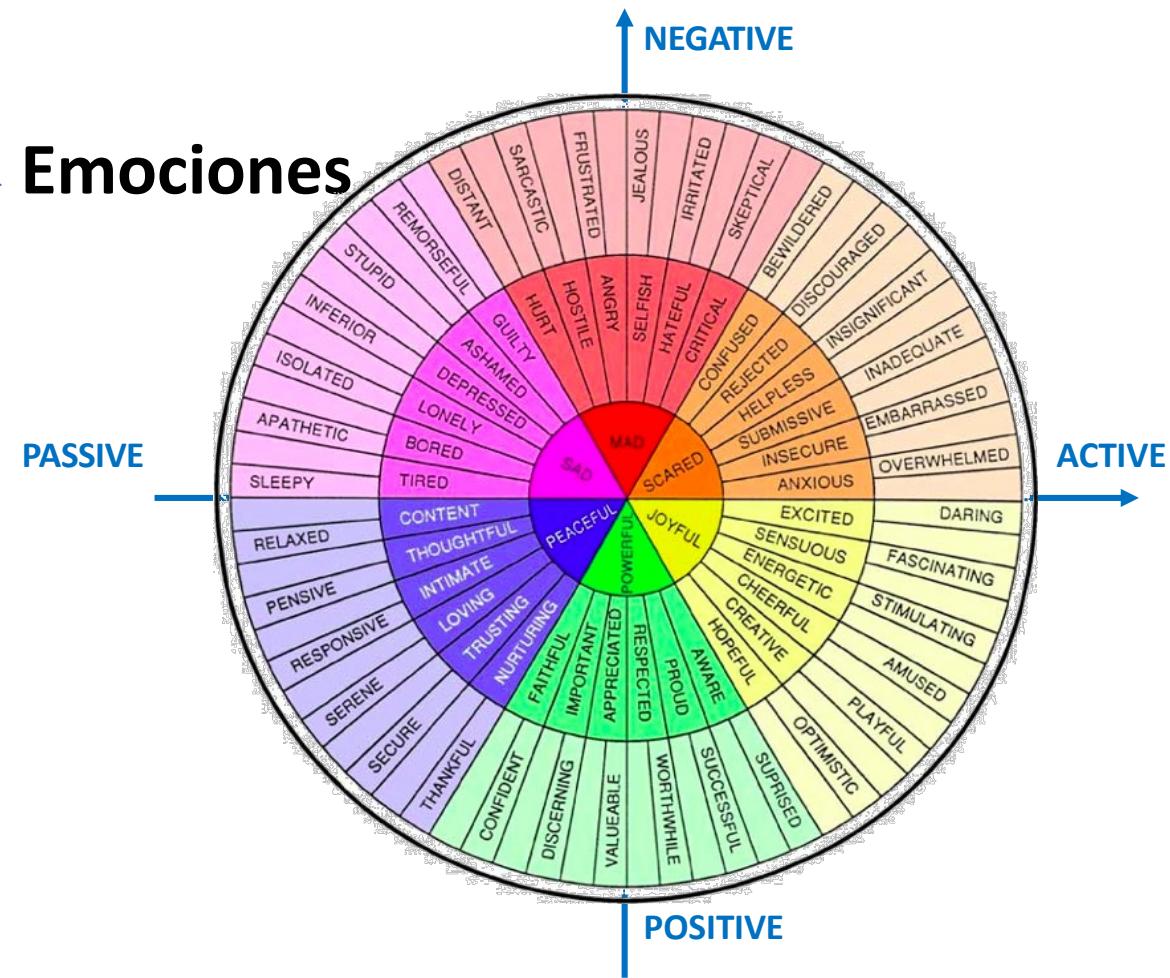
- ¿Cuáles son las *dimensiones de la prosodia*?
 - Pausas
 - Entonación
 - Ritmo
 - Intensidad

Estilos de habla



Modelado prosódico

- ¿Cuáles son las *dimensiones de la prosodia*?
 - Pausas
 - Entonación
 - Ritmo
 - Intensidad



Modelado prosódico

- ¿Cuáles son las *propiedades acústicas* relacionadas con las dimensiones de la prosodia?
 - Pausas → ...
 - Entonación → frecuencia fundamental
 - Ritmo → duración de los fonemas (y duración de las pausas!)
 - Intensidad → energía de la señal vocal

Modelado prosódico

- El texto de entrada se ha transformado en etiquetas (módulo 1): etiquetas fonéticas (fonemas), etiquetas lingüísticas (palabras, sílabas, acentos...) y etiquetas prosódicas básicas (pausas)
- Objetivo del módulo 2: establecer una correspondencia entre las etiquetas y la frecuencia fundamental, la duración de los fonemas y la energía de la señal local
- ¡No existe un mapeo correcto único!

Modelado prosódico

- La ubicación de las *pausas* depende de...
Información morfosintáctica en un contexto de palabras lo suficientemente largo, número de sílabas desde la última pausa, número de sílabas hasta el siguiente signo de puntuación ...
- La *entonación* depende de...
Tipo de frase, longitud de la frase, número de grupos de acentos en la frase, longitud de cada grupo de acentos, ubicación de la palabra acentuada en el grupo de acentos, número de sílabas del grupo de acentos...
- Las *duraciones* dependen de...
Tipo de fonema, contexto fonético, ubicación dentro de la sílaba y la palabra, acento...
- La *intensidad* depende de...
Duración, valor de pitch, contexto fonético, ubicación en la frase...

Modelado prosódico

- ¿Es todo igual de importante? Importancia relativa
 - *Pausas*: las pausas incorrectas hacen que el habla sintética sea muy antinatural, hay que ser moderado
 - *Entonación*: crucial para la naturalidad
 - *Duración*: importante pero más fácil que la entonación
 - *Intensidad*: relativamente menos importante

Modelado prosódico: Técnicas

- Aprender (automáticamente) la correspondencia entre etiquetas y características prosódicas a partir de un corpus (etiquetado)
- Modelos de entonación:
 - ToBI (Tone and Break Index): extraer etiquetas de entonación significativas predecibles a partir del texto -> Métodos de etiquetado automático. Uso limitado
 - Modelo de "Fujisaki superpositional": descripción matemática del contorno de entonación, incluyendo la localización de los acentos; durante la síntesis, los parámetros de la entonación se predicen a partir del texto con métodos estadísticos y se genera la curva de entonación
- Duración: reglas “a mano” (D. Klatt), árboles de decisión...
-

¹ E. Navas, I. Hernández, I. Luengo (2006) An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS, IEEE Trans. On SAP, 14(4)

Modelado prosódico

- ¿Cómo podemos diseñar un *TTS emocional*?
 - Entrenar usando etiquetas de emociones
 - Durante la síntesis, dos posibilidades:
 - La emoción es seleccionada por el usuario
 - La emoción es predicha a partir del texto (¡difícil!)



Text-to-Speech oddcast®

Language: English Voice: Alan

Emotional sound Click on a sound below to add it to the text

Ah Ah_01 Ah_02

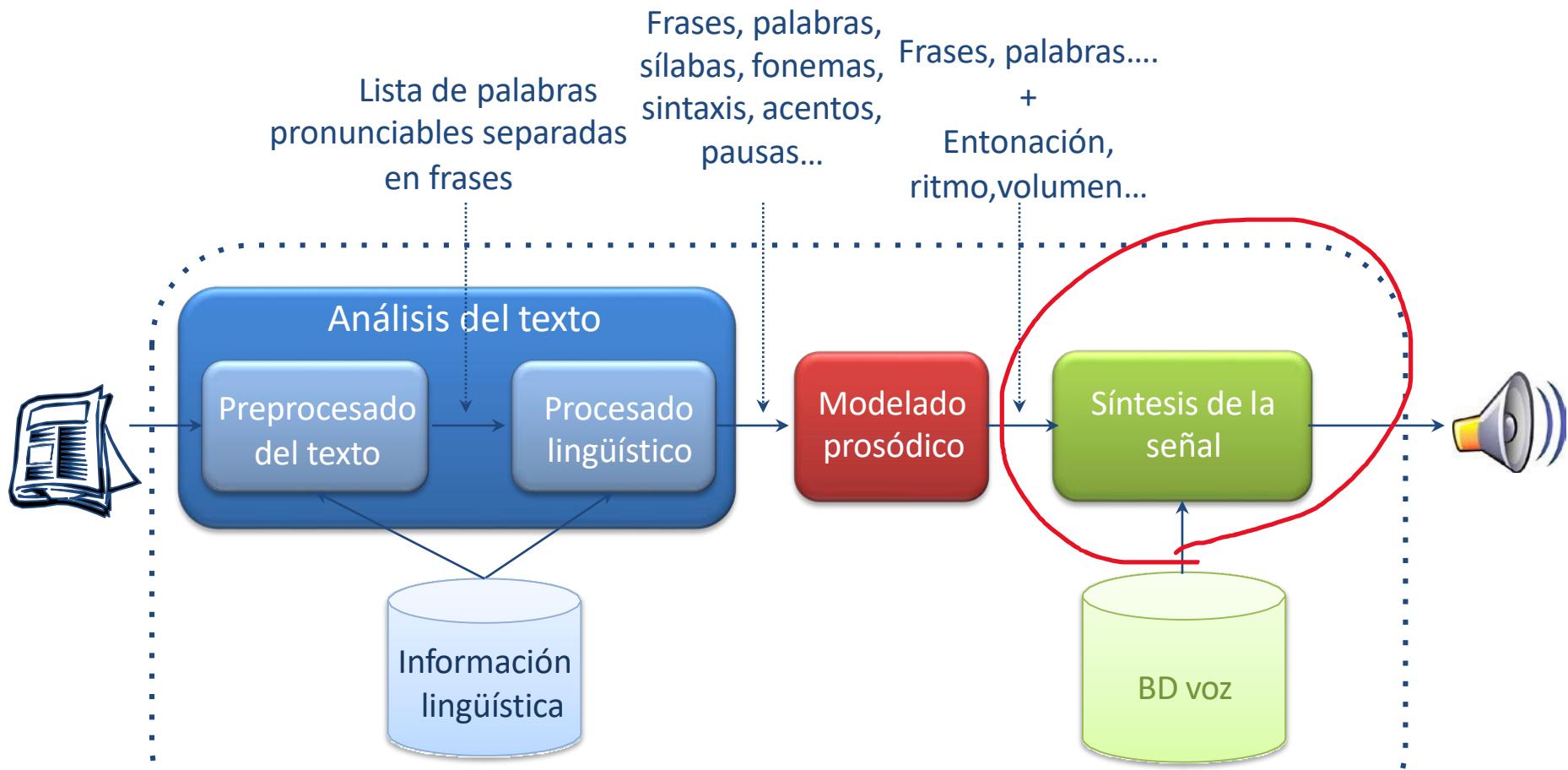
Enter Text:

Say It

Índice

- Introducción
- Estructura de un sistema TTS
- Análisis del texto
- Modelado prosódico
- Técnicas de síntesis
- Evaluación
- Trending topics

Estructura de un sistema TTS

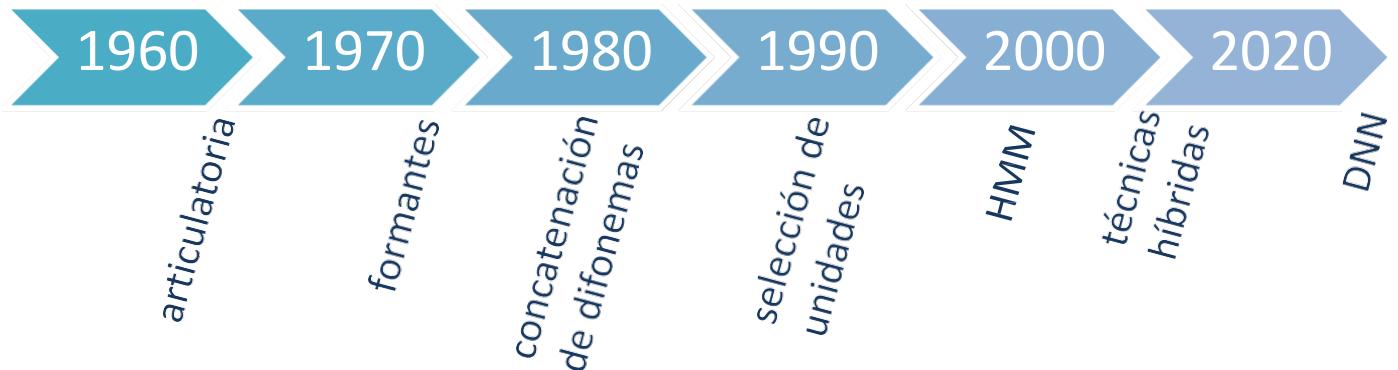


Sistema TTS

*En los sistemas actuales, las fronteras entre módulos no están tan claras.

Técnicas de síntesis

CONOCIMIENTO



Tipos de técnicas de síntesis

Paramétricas

Articulatoria

Formantes

HMM

DNN

Concatenación

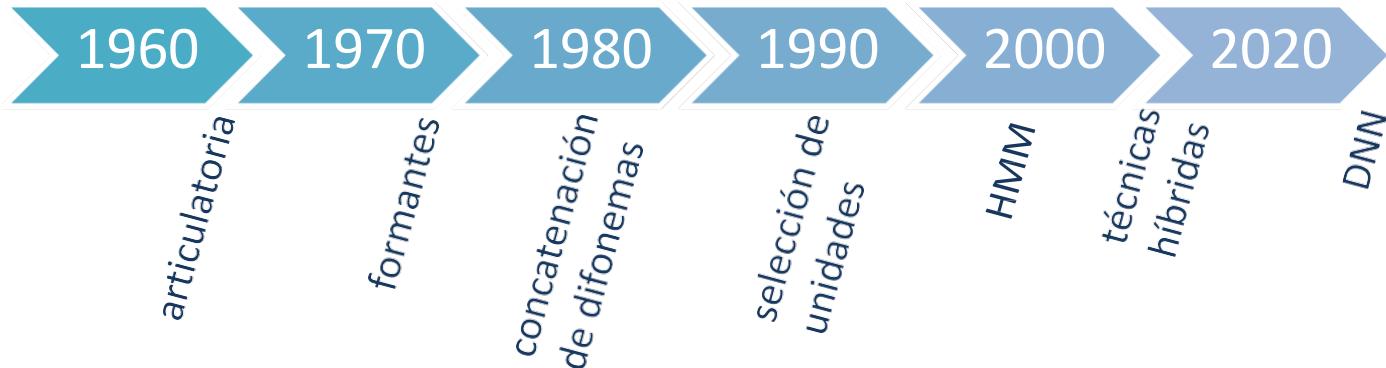
Inventario fijo

Selección de unidades

Técnicas de síntesis

CONOCIMIENTO

DATOS



Tipos de técnicas de síntesis

Paramétricas

Articulatoria

Formantes

HMM

DNN

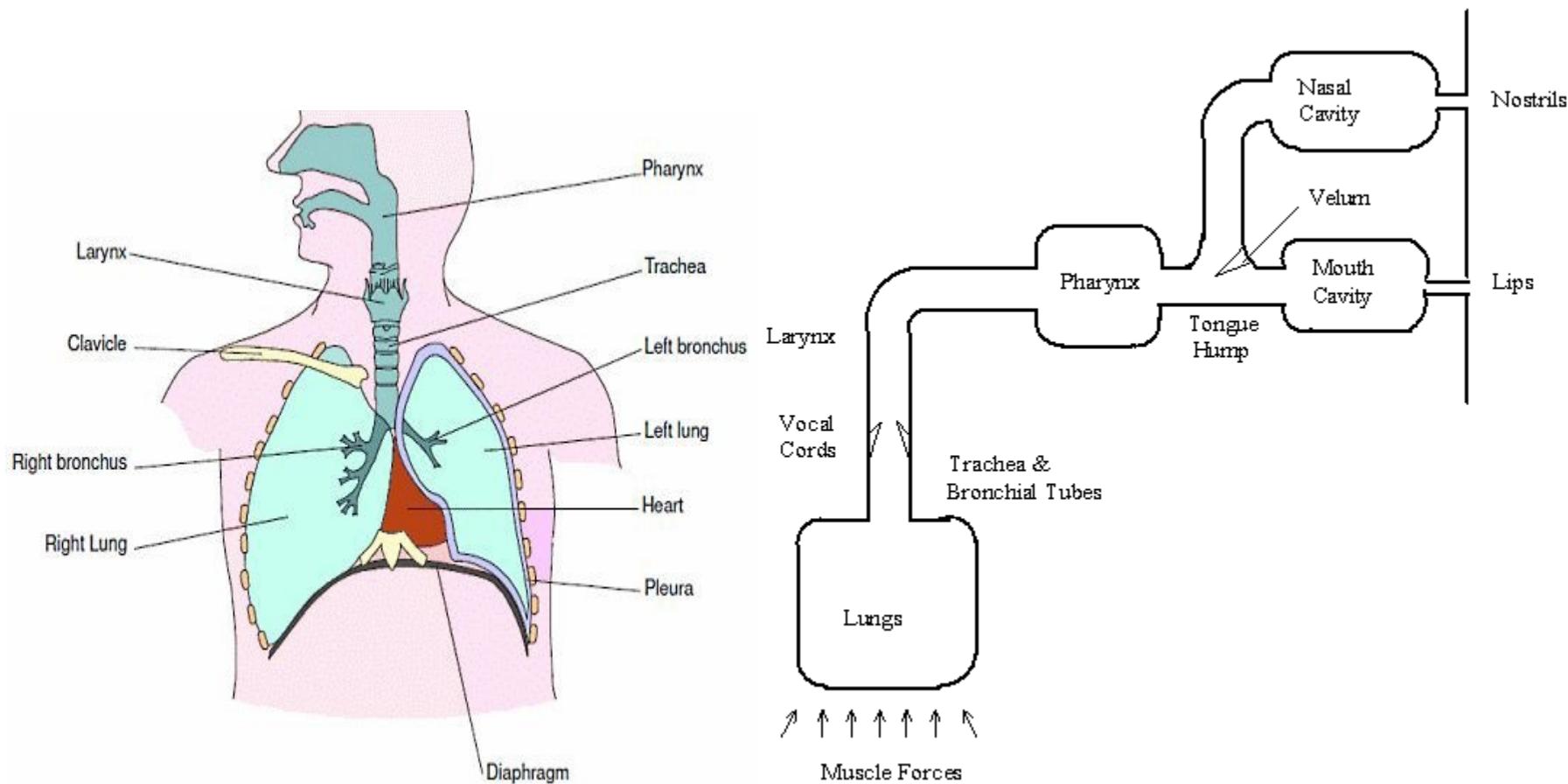
Concatenación

Inventario fijo

Selección de unidades

Técnicas de síntesis: Producción de la voz

- Los primeros intentos tratan de reproducir el aparato fonador humano

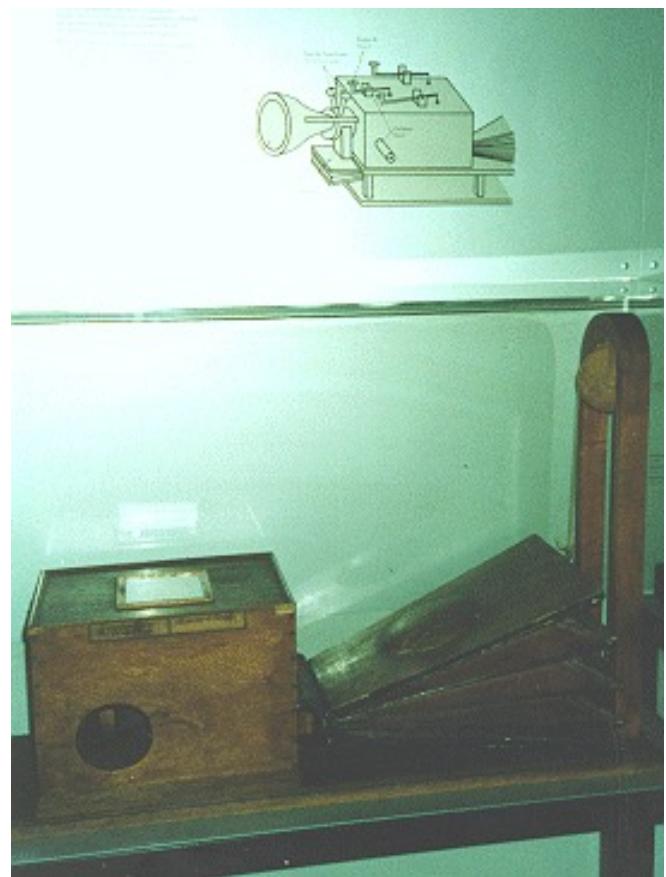


Técnicas de síntesis: Articulatoria

- Los primeros intentos para producir voz artificial fueron articulatorios

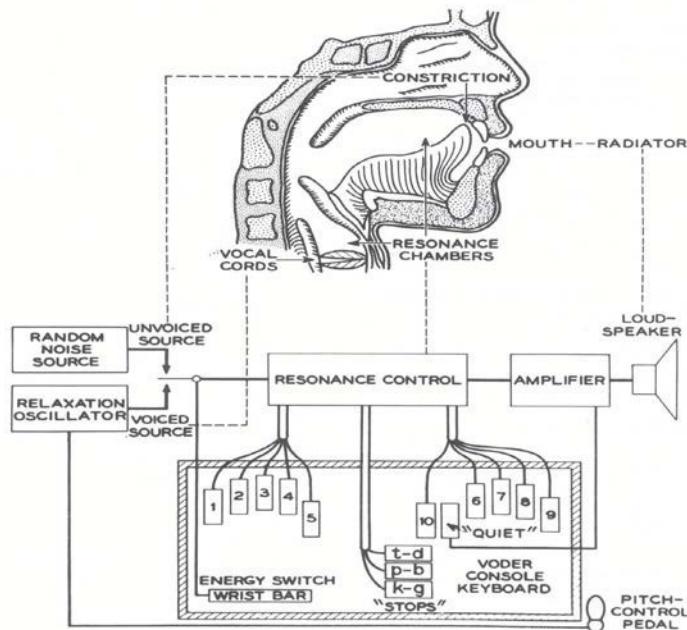
Machine for speaking by Von Kempelen (1791)

[Demo](#)



Técnicas de síntesis: Modelos eléctricos

1939: Modelos eléctricos: VODER de Dudley



Técnicas de síntesis: Modelos eléctricos

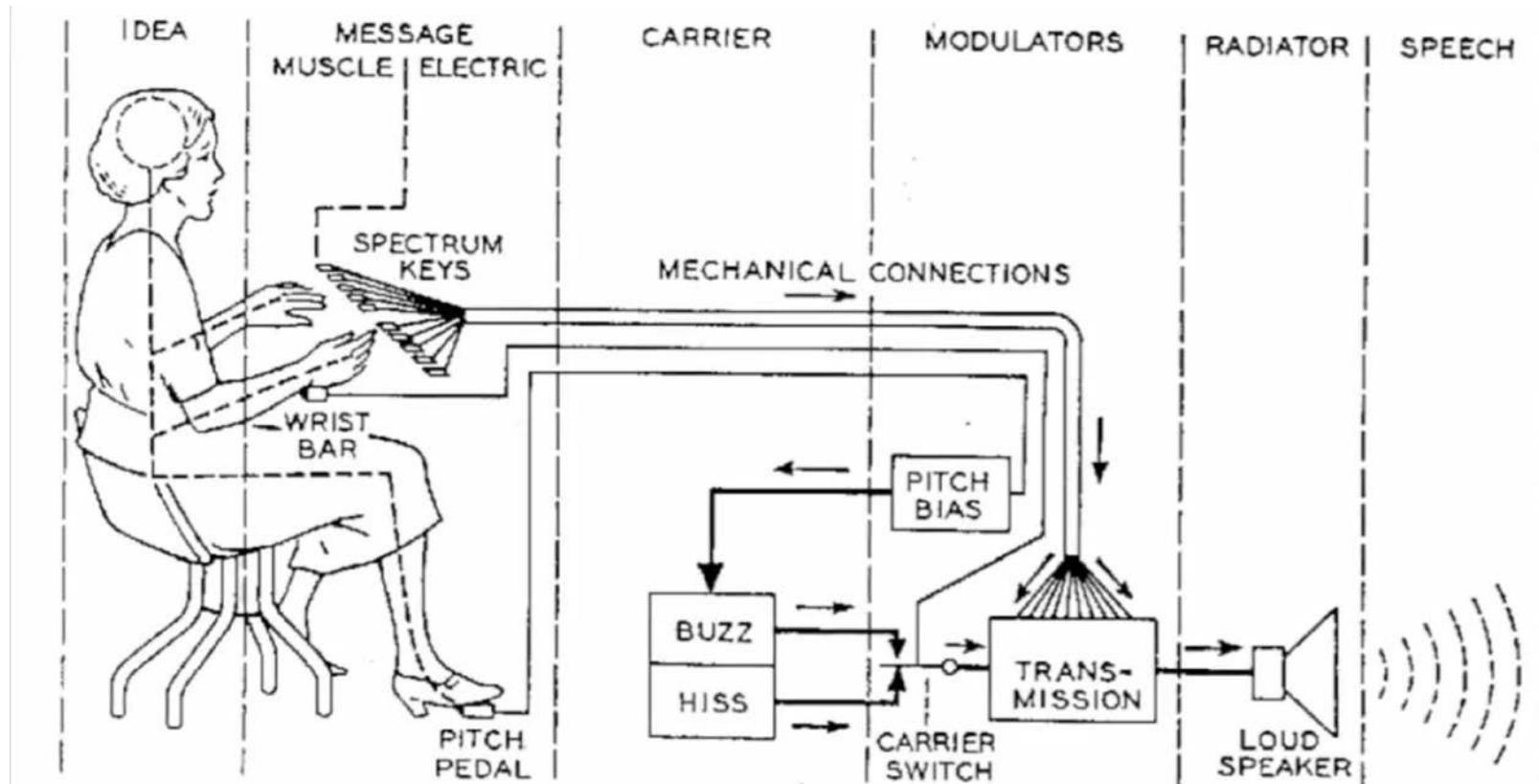
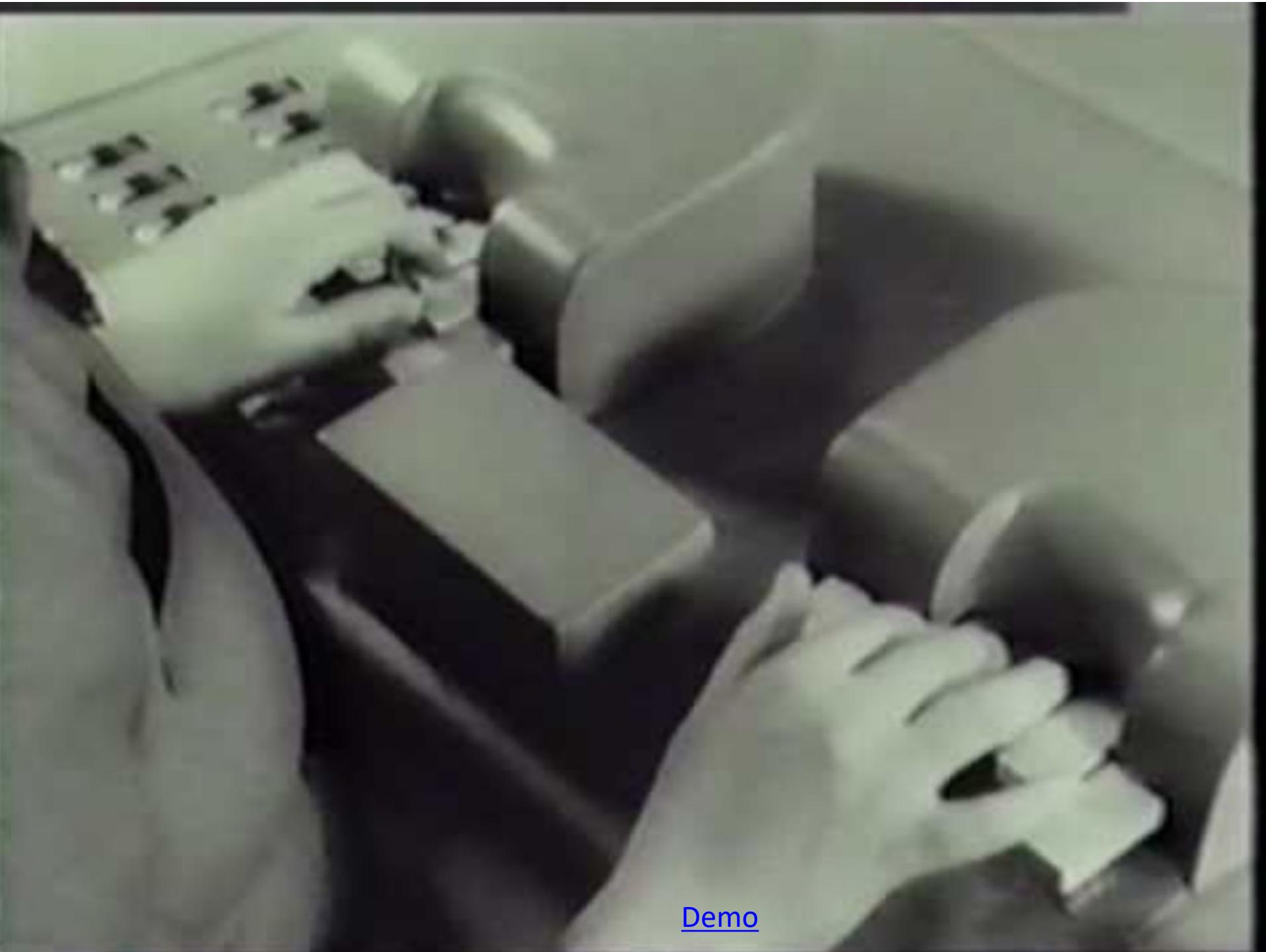
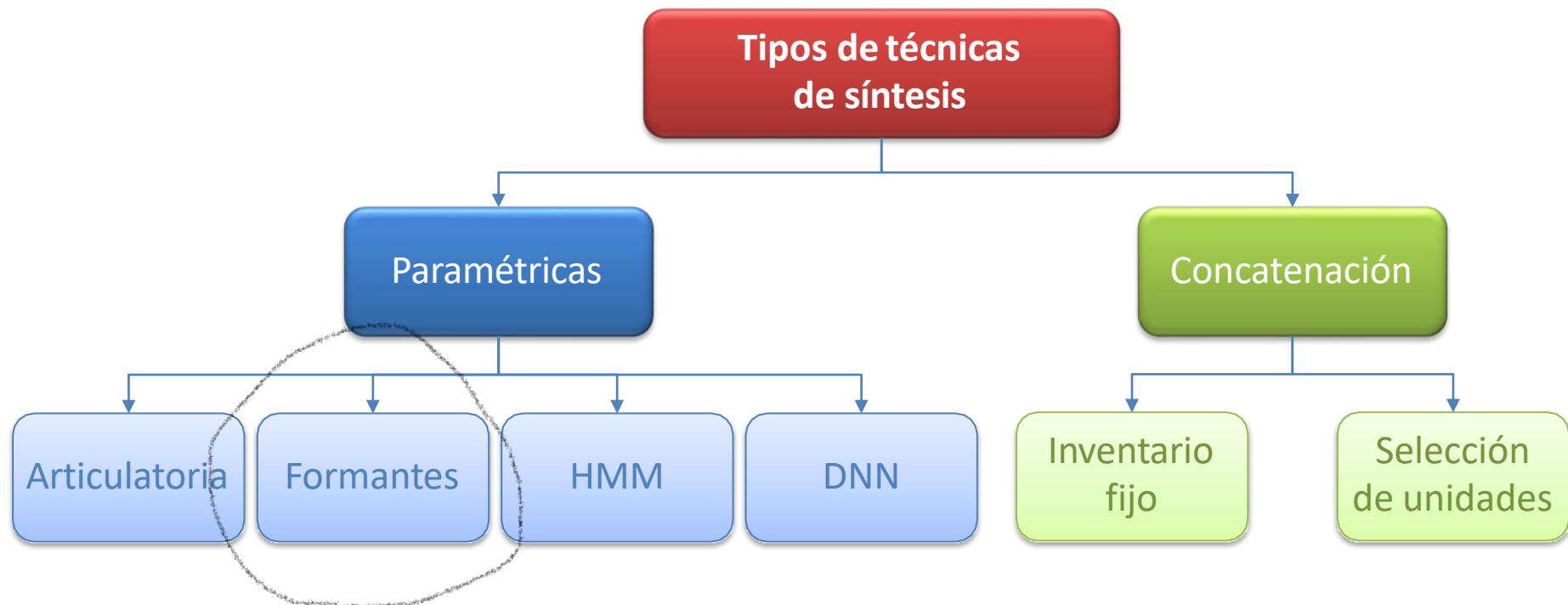
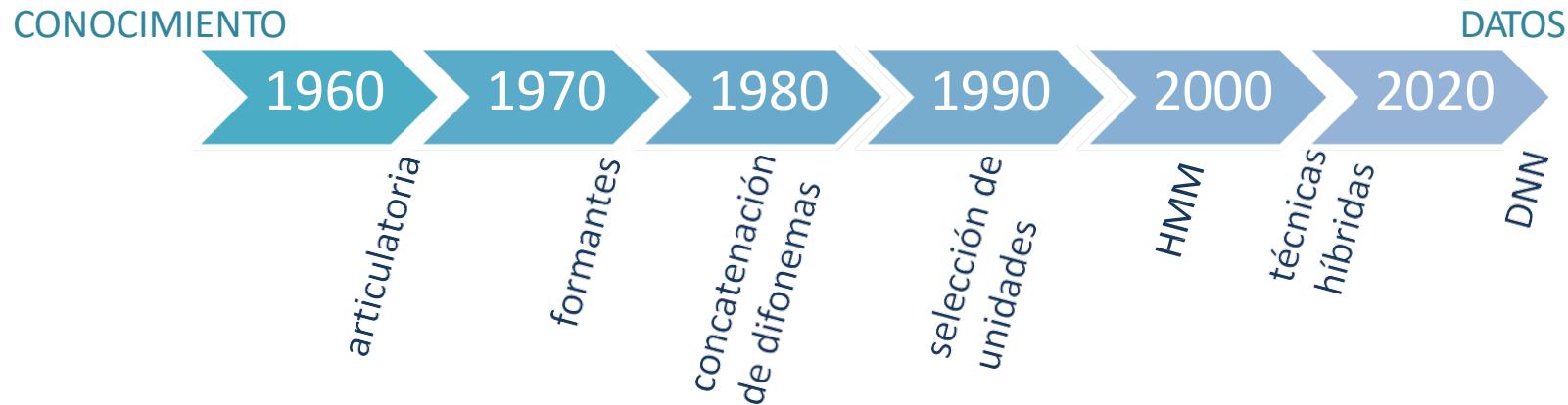


Fig. 8—Schematic circuit of the voder.



[Demo](#)

Técnicas de síntesis



Técnicas de síntesis: Por formantes

- Aproximación al proceso de producción basada en el conocimiento. También llamado *síntesis basada en reglas* (aunque los sistemas por concatenación también tienen componentes basados en reglas)
- La síntesis de formantes no utiliza muestras de voz humana, la voz sintetizada se crea con un modelo matemático:
 - 1) El audio de salida es creado a partir de la síntesis aditiva y un modelo acústico (síntesis mediante modelado físico).
 - 2) Parámetros como la frecuencia fundamental, fonación y niveles de ruido se modulan a través del tiempo para crear una forma de onda de una voz artificial.

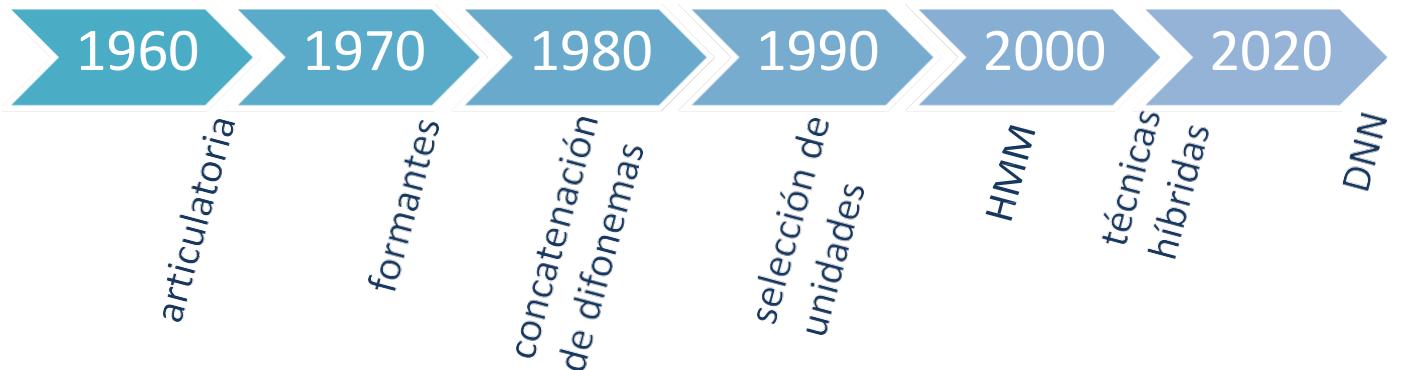
Técnicas de síntesis: Por formantes

- La voz generada es suave por el diseño del sistema
- Es posible crear nuevas voces y efectos
- Difícil estimar el valor adecuado para todos los parámetros
- Largo tiempo de desarrollo: difícil localizar y corregir errores
- Baja calidad y zumbidos



Técnicas de síntesis

CONOCIMIENTO



Tipos de técnicas de síntesis

Paramétricas

Articulatoria

Formantes

HMM

DNN

Concatenación

Inventario fijo

Selección de unidades

Técnicas de síntesis: Por concatenación

- Generan la señal sintética concatenando señales naturales (segmentos de voz pregrabados)
- Los segmentos pueden ser oraciones completas, palabras, sílabas, fonemas , difonemas...
- Las unidades de voz a concatenar se almacenan en una base de datos de voz (en forma de formas de onda o espectrogramas), etiquetadas en función de sus propiedades acústicas (por ejemplo, su frecuencia fundamental)
- En tiempo de ejecución, la secuencia deseada se crea determinando la mejor cadena de unidades candidatas de la base de datos (selección de unidades)
 - Si el número de mensajes y contextos prosódicos es reducido →concatenación directa
 - Para generar cualquier mensaje →uso de técnicas de procesado de señal: PSOLA, MBROLA, HNS

Técnicas de síntesis: Por concatenación

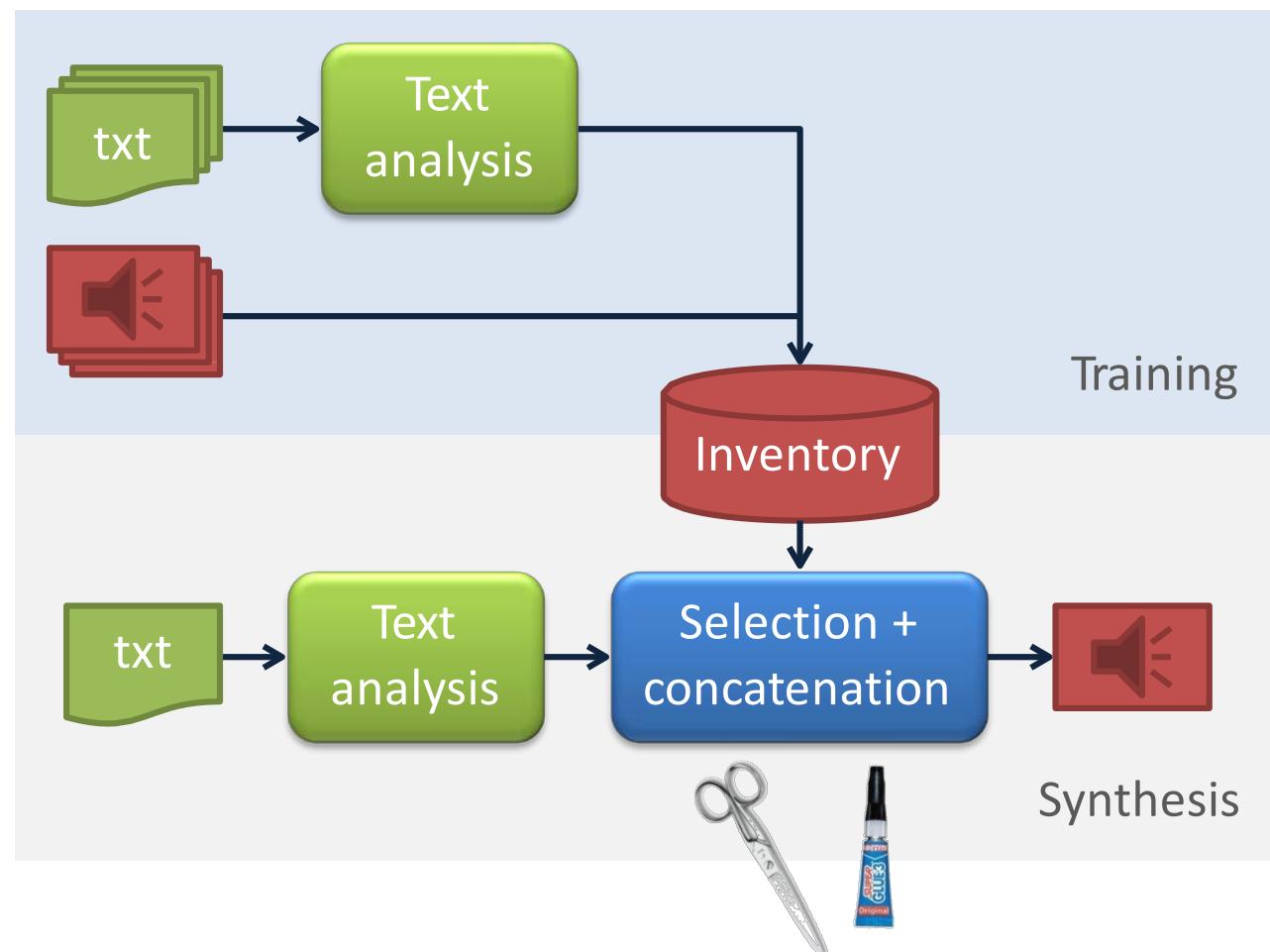
- Base de datos de voz: Inventario fijo
 - Se graba un ejemplo de cada combinación de fonemas (difonemas, trifonemas en algunos casos) posible en una lengua en un contexto fonético estable
 - No es una solución óptima:
 - Modificaciones prosódicas degradan la calidad
 - Las dependencias fonéticas implican más que el fonema anterior/siguiente
 - Grabar todos los trifonemas en todos los contextos prosódicos posibles no es factible

Técnicas de síntesis: Por concatenación

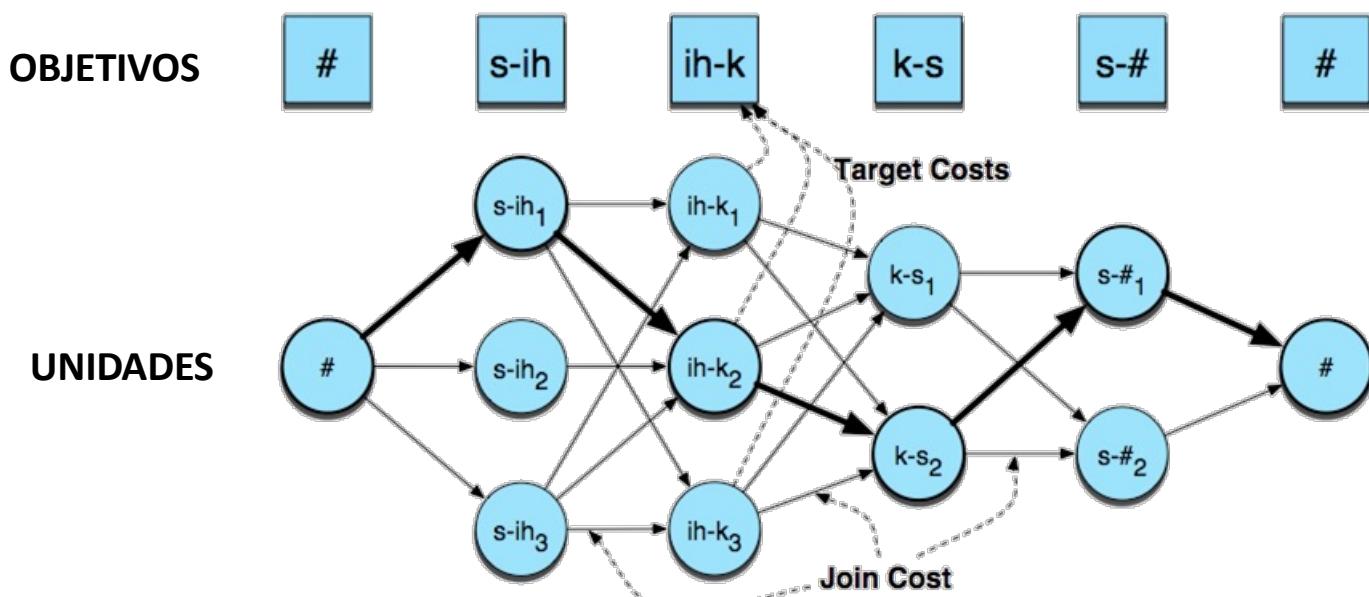
- Selección de unidades



En la base de datos de voz natural hay miles de ejemplos de cada fonema en distintos contextos fonéticos y prosódicos (pocas modificaciones necesarias)



Técnicas de síntesis: Por concatenación



- Se aplica el algoritmo de Viterbi para seleccionar la secuencia más adecuada
- Para concatenar las unidades seleccionadas finalmente se aplican técnicas PSOLA
- PSOLA: modifica el tono y la duración de una señal de voz

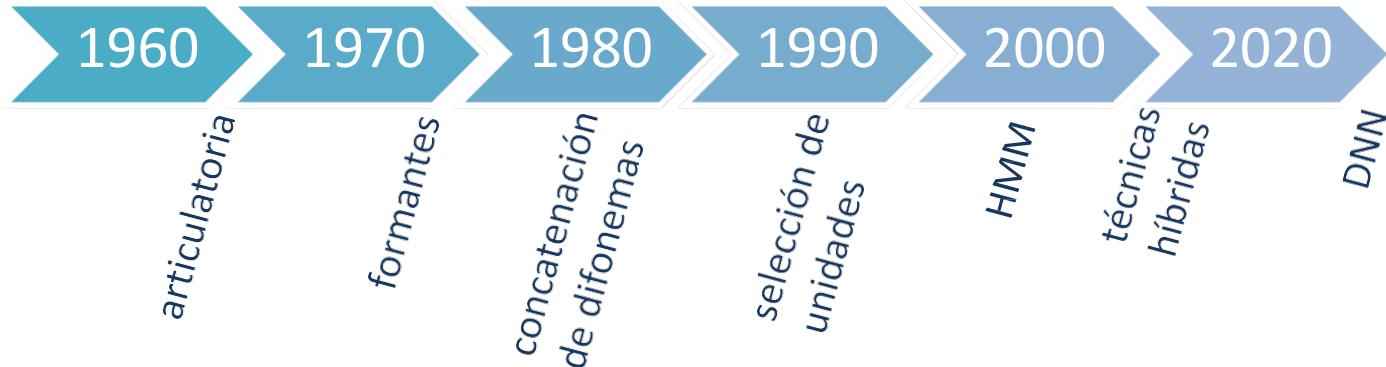
Técnicas de síntesis: Por concatenación

- Buena calidad en dominios restringidos, pero...
 - Calidad de la voz sintética muy variable
 - Los errores en la detección de las fronteras entre fonemas generan discontinuidades muy molestas
 - La selección requiere carga computacional elevada si el inventario es muy grande
 - Para añadir voces nuevas se debe grabar una nueva base de datos o aplicar técnicas de conversión de voz

Técnicas de síntesis

CONOCIMIENTO

DATOS



Tipos de técnicas de síntesis

Paramétricas

Articulatoria

Formantes

HMM

DNN

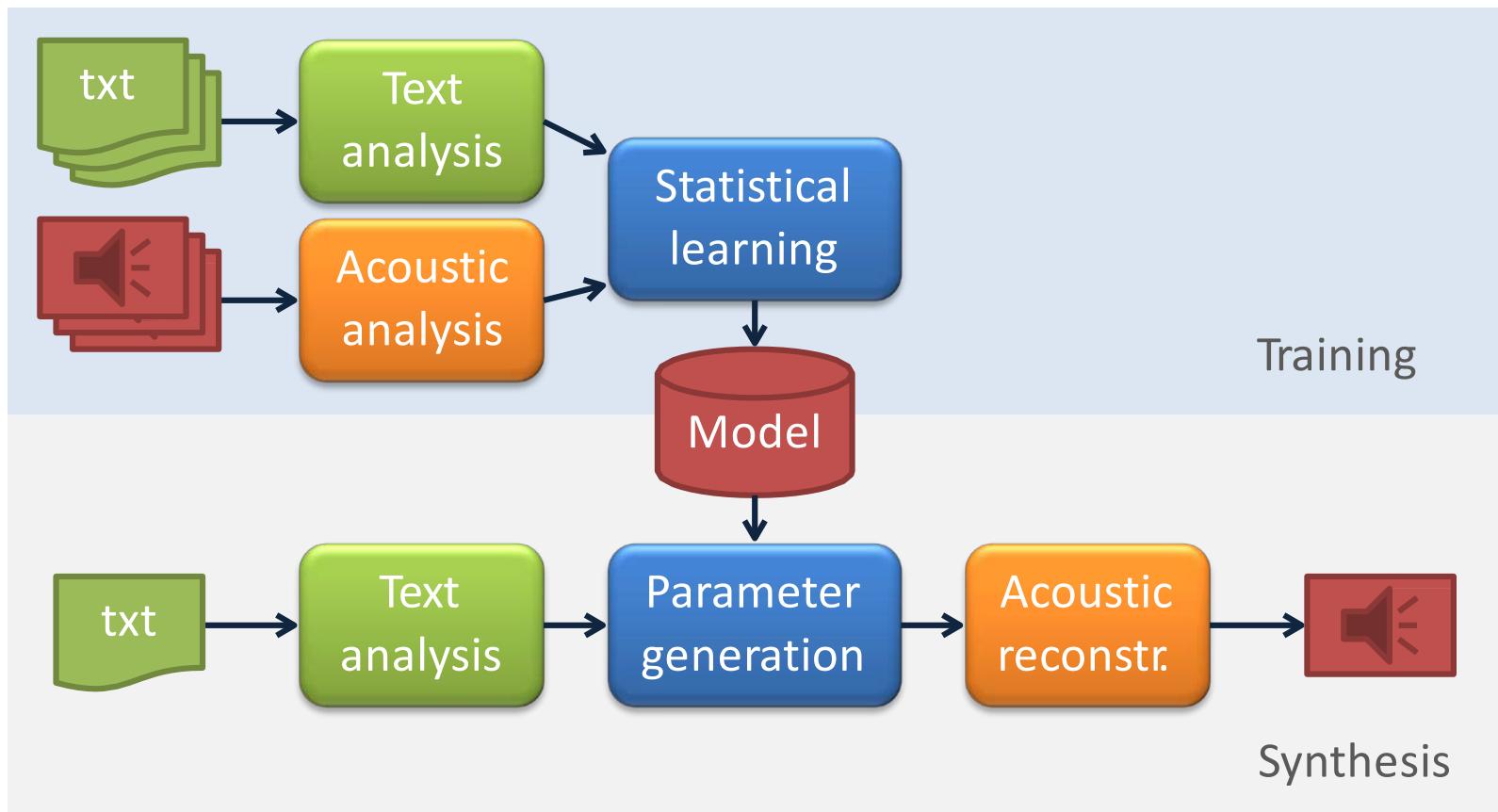
Concatenación

Inventario fijo

Selección de unidades

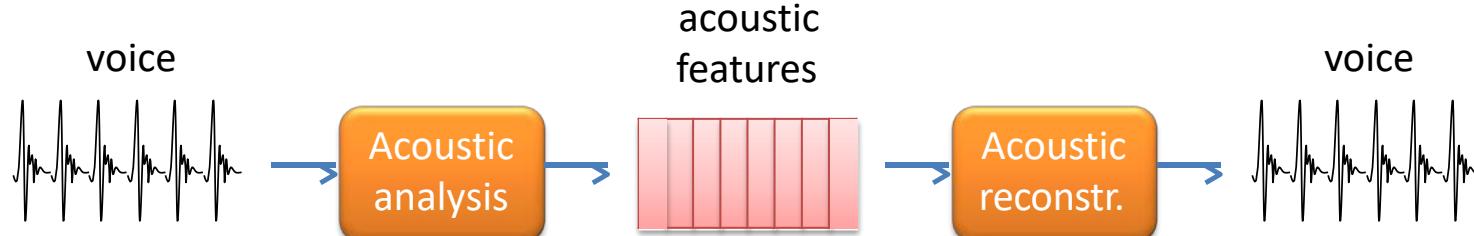
Técnicas de síntesis: Basada en HMM

- Síntesis estadístico paramétrica
 - Dos fases: entrenamiento de los modelos y síntesis



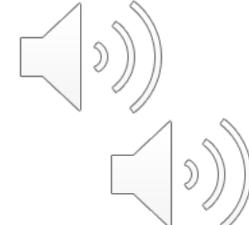
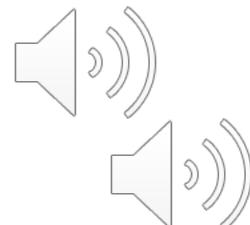
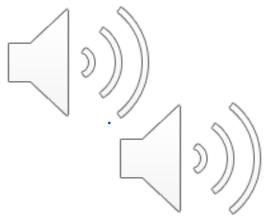
Técnicas de síntesis: Basada en HMM

- Los parámetros de la voz se generan mediante un modelo HMM
- HMM Speech Synthesis System (HTS) disponible en <http://hts.sp.nitech.ac.jp/> 
- Uso de vocoders (*Voice coders*) para parametrizar la voz
 - MLSA, STRAIGHT, WORLD, AHOCODER



Técnicas de síntesis: Basada en HMM

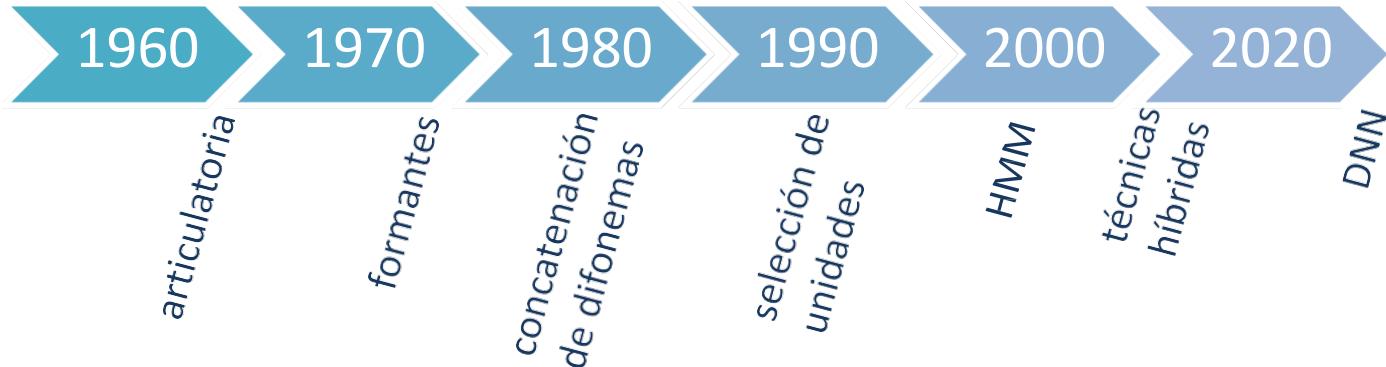
- La duración se modela explícitamente: HSMM
- Voz estable y suave
- Calidad de voz moderada debido al uso de vocoder
- Las voces se pueden cambiar cambiando los parámetros
- Facilidad para adaptar los modelos a nuevas voces
- Posibilidad de generar nuevas voces y estilos por interpolación



Técnicas de síntesis

CONOCIMIENTO

DATOS



Tipos de técnicas de síntesis

Paramétricas

Articulatoria

Formantes

HMM

DNN

Concatenación

Inventario fijo

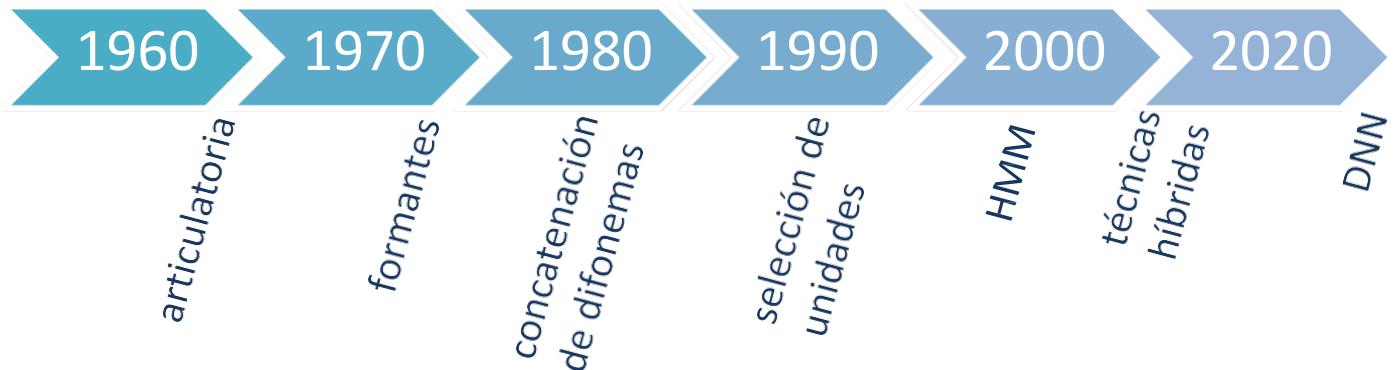
Selección de unidades

Técnicas de síntesis: Síntesis híbrida

- Utilizar los parámetros obtenidos con los modelos para encontrar unidades óptimas en la base de datos o utilizar unidades naturales para mejorar los parámetros generados por los modelos

Técnicas de síntesis

CONOCIMIENTO



Tipos de técnicas de síntesis

Paramétricas

Articulatoria

Formantes

HMM

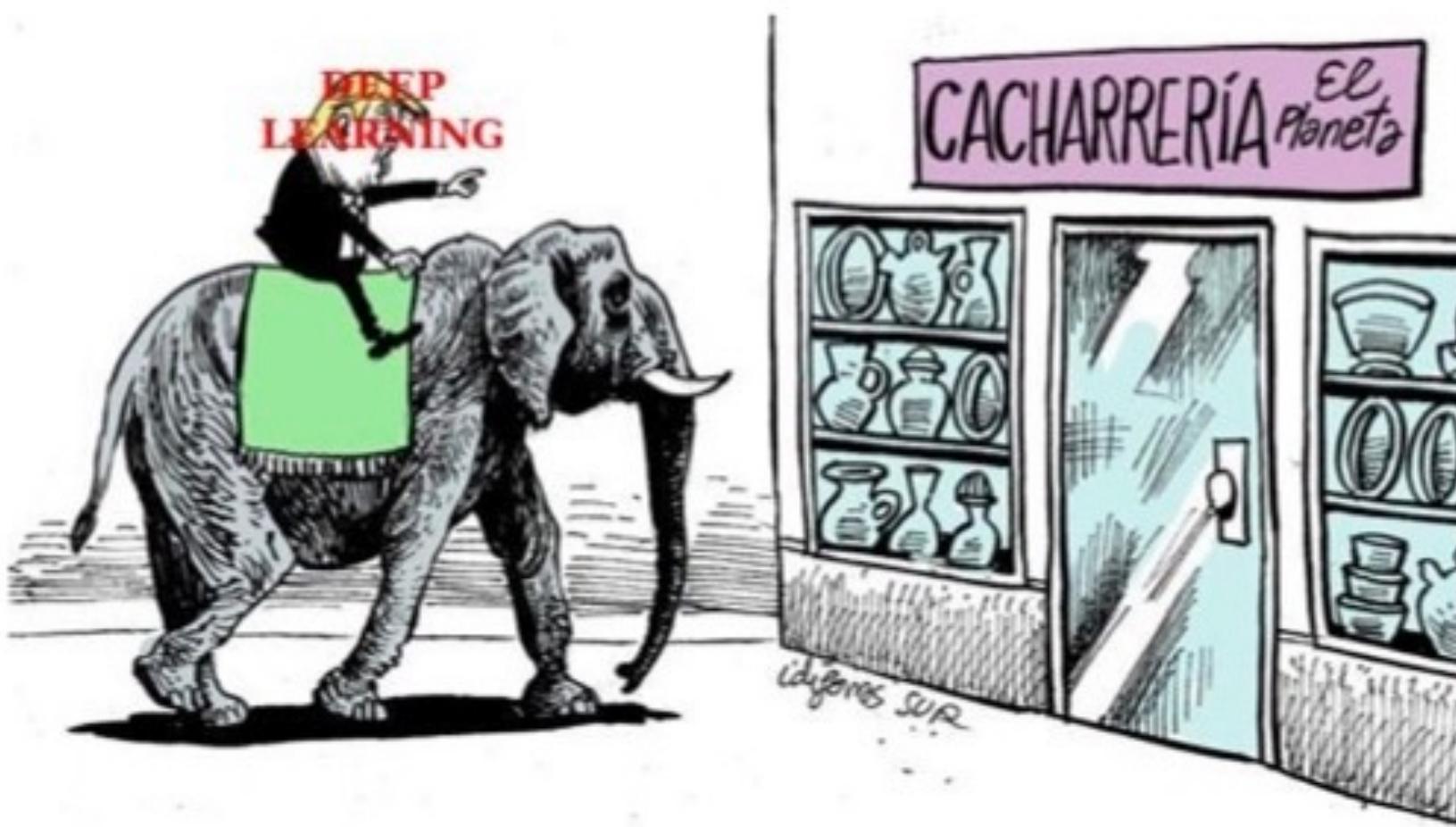
DNN

Concatenación

Inventario fijo

Selección de unidades

Técnicas de Síntesis: basadas en DNN



Técnicas de Síntesis: basadas en DNN

- Síntesis basada en redes neuronales profundas (*Deep neural network*, DNN)
 - Más potentes que los HMMs
 - Necesitan más datos y mayor capacidad computacional (hardware específico como GPU - *graphical processing units*)



Técnicas de Síntesis: basadas en DNN

Dada una secuencia de texto X en la entrada, se genera una secuencia acústica Y , según los parámetros del modelo Θ :

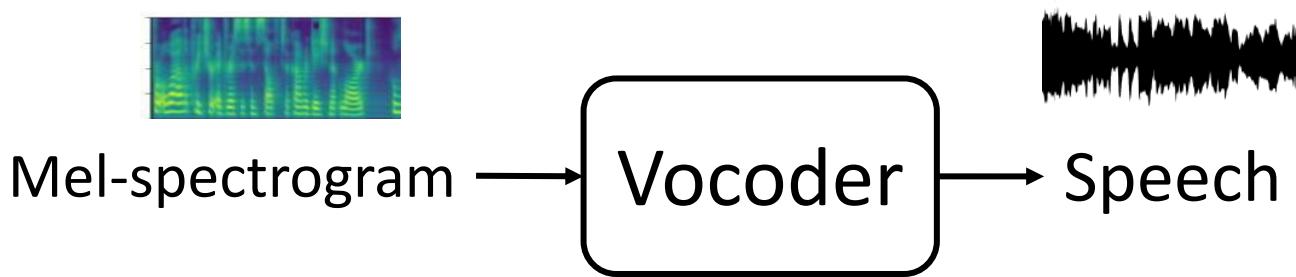
$$Y = \operatorname{argmax} P(Y|X, \Theta)$$

Dos familias:

- a) De espectograma a audio
- b) De texto a audio (sistemas end-to-end)

Técnicas de Síntesis: basadas en DNN

- a) De espectograma a audio: Se generan las características acústicas (espectrograma) y luego se usa un Vocoder neural



Un Vocoder (codificador de voz) encripta y comprime la señal de audio y viceversa. Esto se lograba tradicionalmente mediante técnicas de procesamiento de señales digitales. Un codificador de voz neuronal hace la codificación / decodificación utilizando una red neuronal.

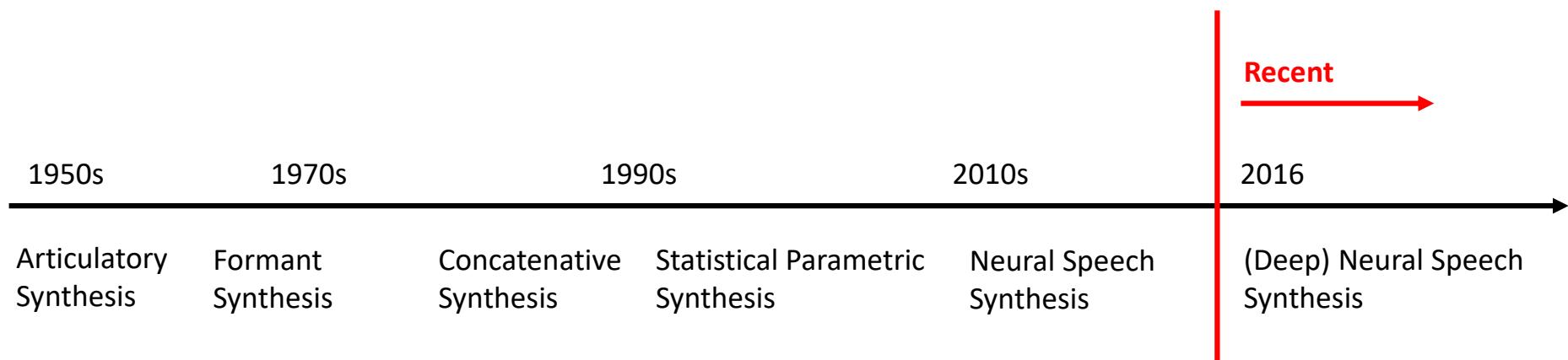
Técnicas de Síntesis: basadas en DNN

- b) De texto a audio (sistemas end-to-end): Bloques basados en redes neuronales (o con fronteras menos definidas). Se entrena con pares (*texto, audio*)
 - Modelo acústico:
 - Modelo de grafema a fonema
 - Modelo de segmentación
 - Modelos de duración de fonema y de frecuencia fundamental
 - Modelo de síntesis de audio (vocoder)

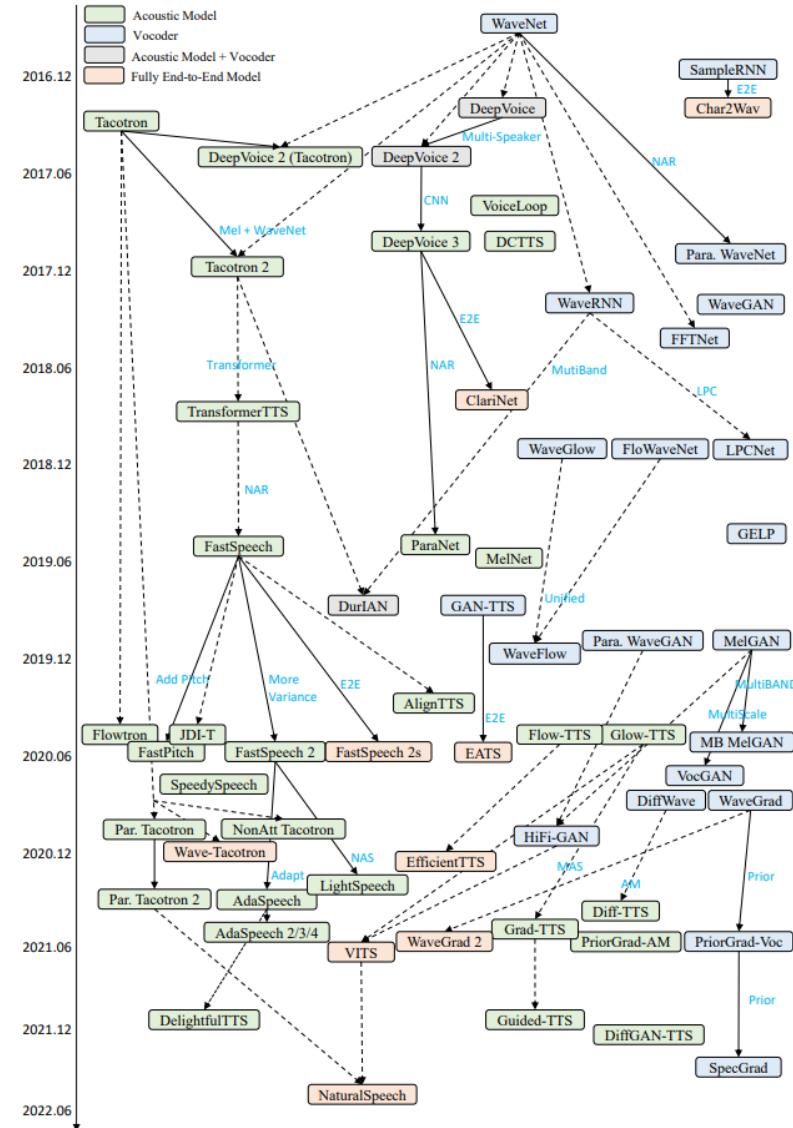
Técnicas de Síntesis: basadas en DNN

- De espectrograma a audio
 - Basadas en redes profundas: Familia Wave (**WaveNet**, WaveFlow, WaveGradNet)
 - Basadas en redes Generativas Adversarias (GAN): MelGAN, HiFi-GANGAN
- De texto a audio (sistemas end-to-end)
 - Familia DeepVoice
 - Familia Tacotron
 - ClariNet
 - Fast Speech
 - Triple M
 - **NaturalSpeech**
 - ...

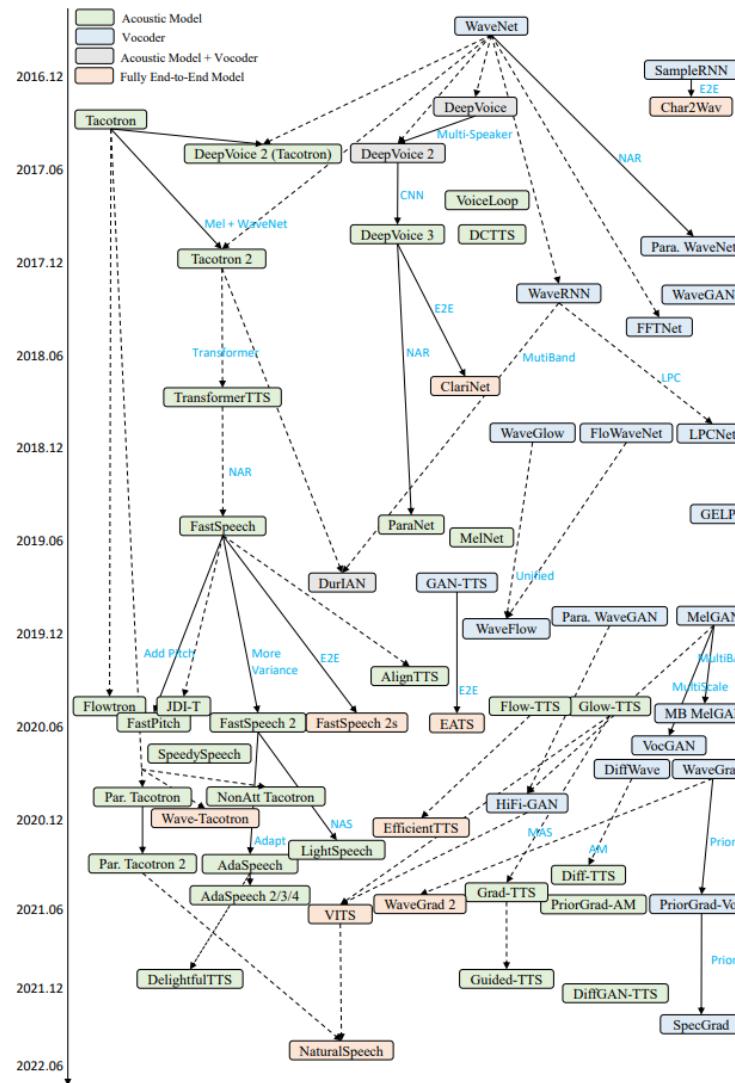
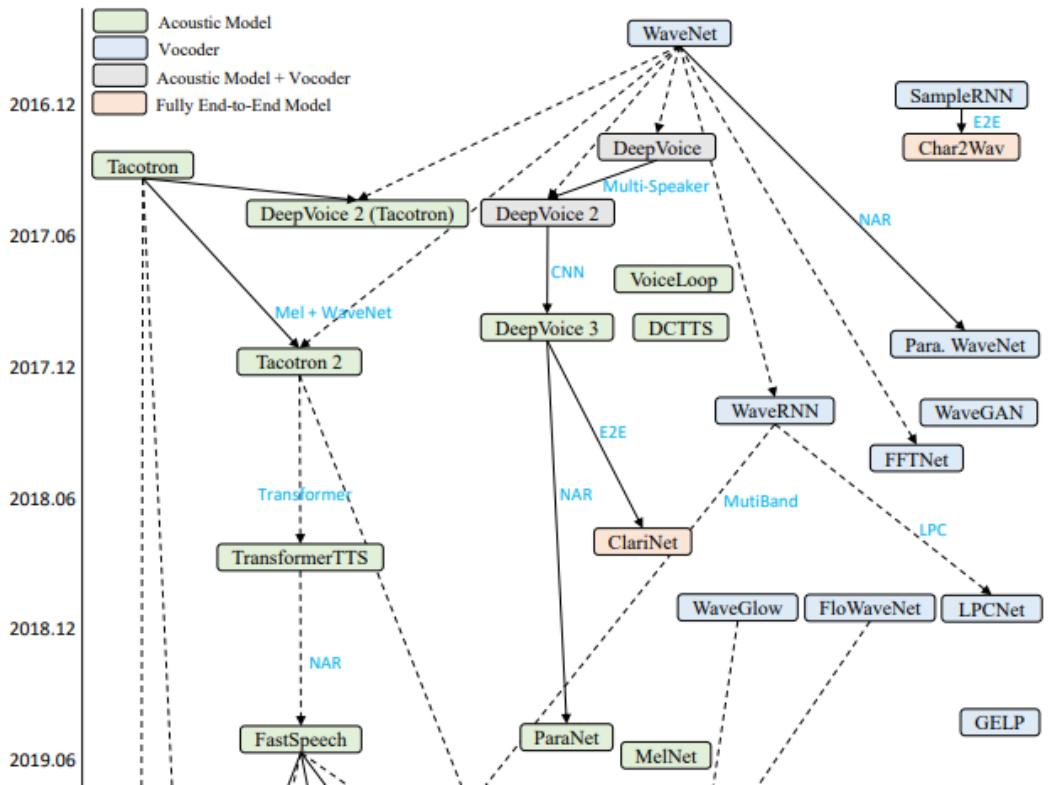
How “recent” this tutorial covers?



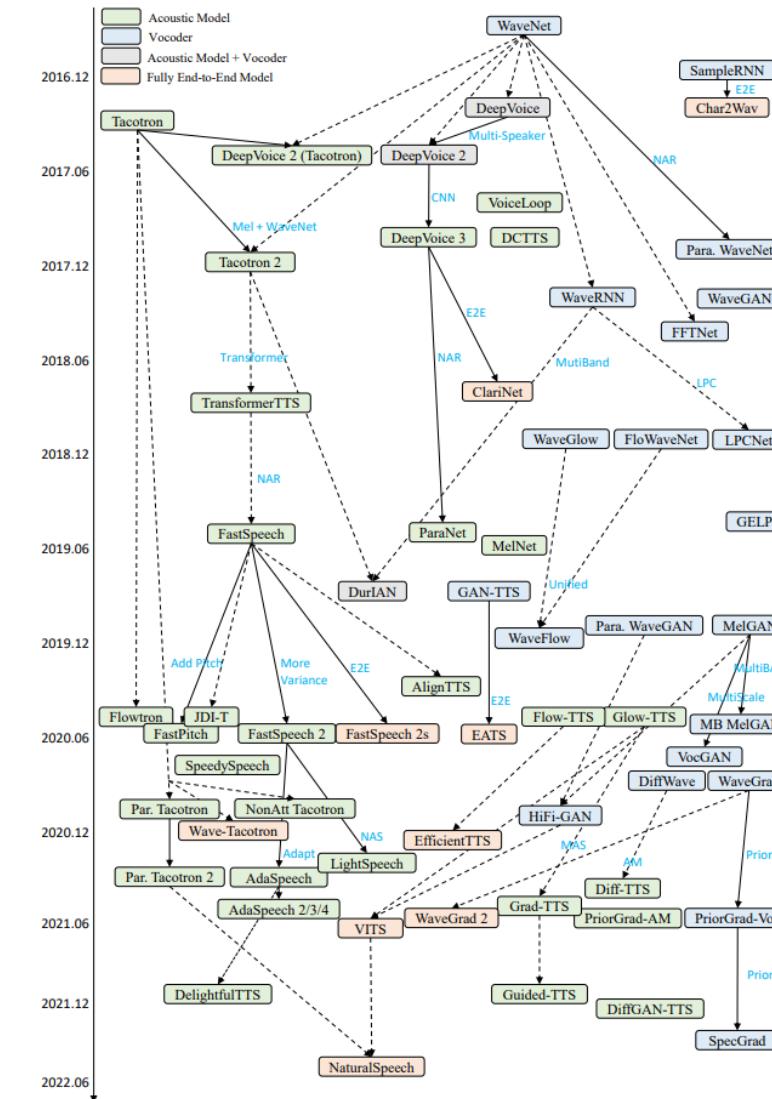
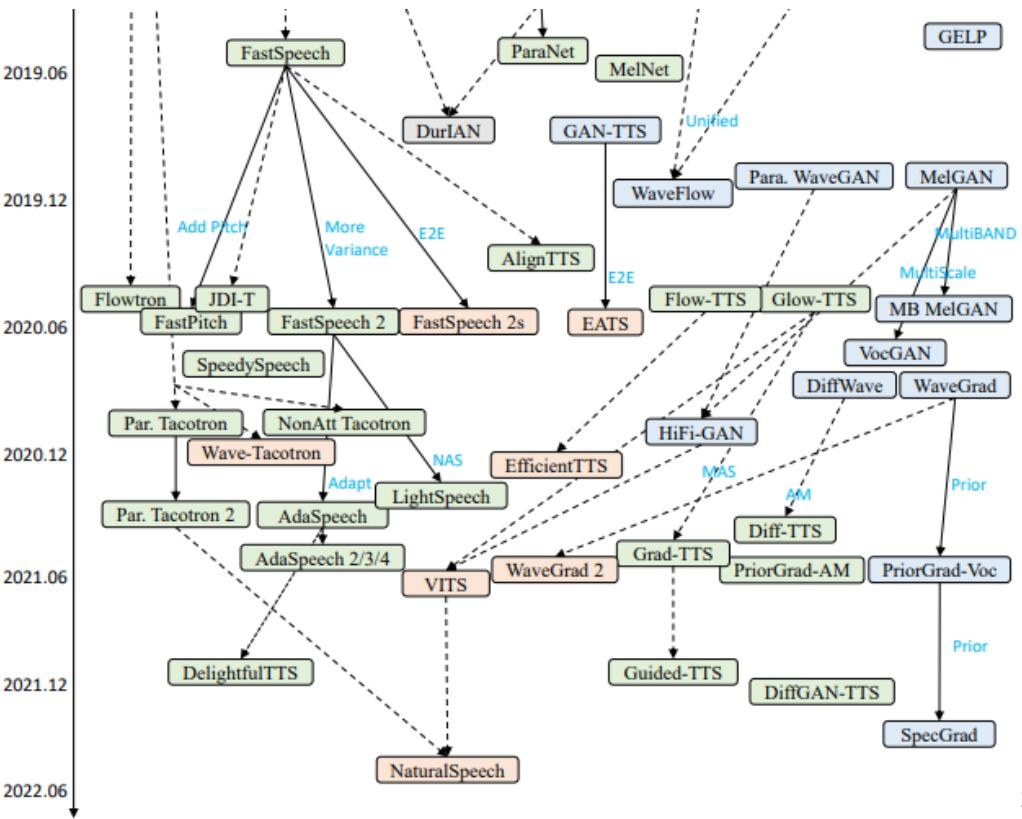
Recent advances



Recent advances



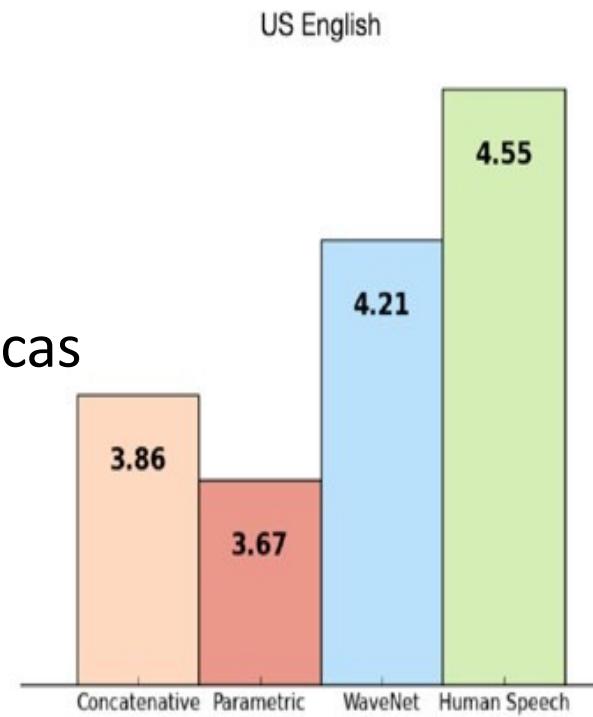
Recent advances



Técnicas de Síntesis: basadas en DNN

De espectrograma a audio

- **WaveNet**: red generativa
 - Utilizando sólo muestras de audio como entrada produce muestras de audio similares
 - Red convolucional que aprende qué combinaciones son realistas y cuáles no
 - Condicionada con características lingüísticas del texto de entrada, funciona como sistema TTS.
 - Características lingüísticas necesarias:
 - Conocimiento experto
 - Elaborar herramientas de análisis de texto
 - Crear un lexicon (guía de pronunciaciones)



Técnicas de Síntesis: basada en DNN

De texto a audio (sistema end-to-end)

NaturalSpeech 2022

Demo

[\(https://speechresearch.github.io/naturalspeech/\)](https://speechresearch.github.io/naturalspeech/)

- Entrenado con 44,000 horas de voz hablada y cantada.



Microsoft

Técnicas de Síntesis: basada en DNN

- Arquitectura del sistema NaturalSpeech

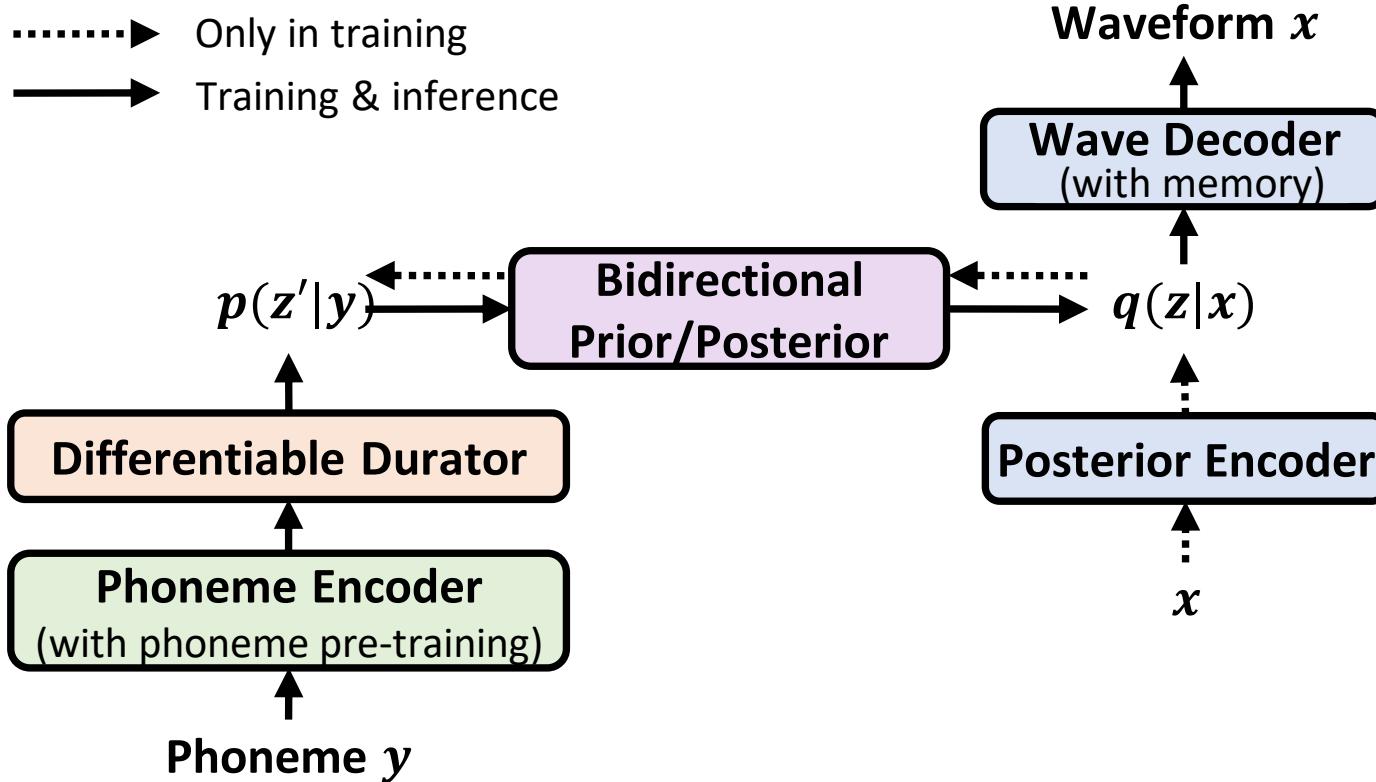


Figure 1: System overview of NaturalSpeech.

NaturalSpeech

TRAINING

- 1) In NaturalSpeech 2, a neural audio codec is first trained to convert a speech waveform into a sequence of latent vectors using a codec encoder, and then reconstruct the speech waveform from these latent vectors using a codec decoder.
- 2) After training the audio codec, the codec encoder is used to extract the latent vectors from the speech in the training set.
- 3) These vectors are then used as targets for the latent diffusion model, which is conditioned on prior vectors obtained from a phoneme encoder, a duration predictor, and a pitch predictor.

INFERENCIA

- 1) During the inference process, the latent diffusion model is first used to generate the latent vectors from the text/phoneme sequence. These latent vectors are then converted into a speech waveform using the codec decoder.

“Latent vectors”: el texto se codifica con embeddings. Estos embeddings se transforman con un modelo neural en una secuencia de vectores latentes en un espacio que es una representación aprendida que encapsula características significativas e interpretables del texto de entrada (tanto características lingüísticas como acústicas)

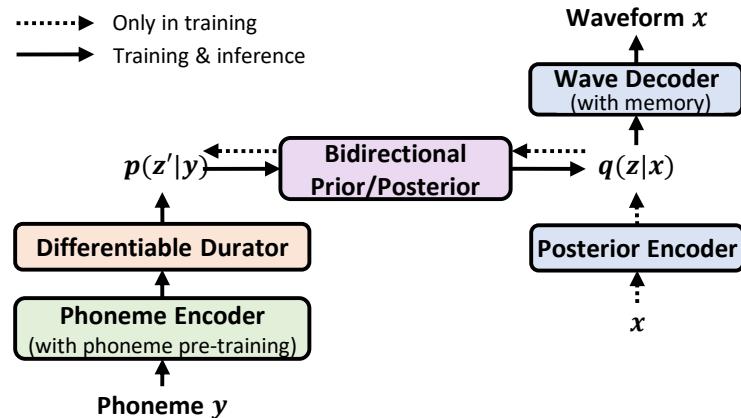


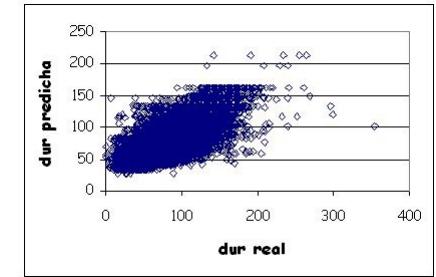
Figure 1: System overview of NaturalSpeech.

Evaluación

- Aspectos a evaluar
 - Inteligibilidad
 - Naturalidad
 - Adaptación a la tarea
- La calidad de la voz sintética depende de todos los componentes del sistema TTS
 - Preprocesado del texto, procesado lingüístico, generación de voz
 - Diferentes tipos de evaluación

Evaluación

- Tipos de evaluación
 - Medidas objetivas: comparar con una señal real
 - Repetibles
 - Rápidas
 - No dan idea de la calidad perceptual
 - Medidas subjetivas: preguntar a los usuarios su opinión
 - No repetibles
 - Mayor tiempo para obtener resultados
 - Dan idea sobre la calidad perceptual
 - Válidas para evaluar los aspectos buscados



Entzun eta eman iritzia

1)	2)	Zein nahiago duzu?:
1)	2)	Lehena aise nahiago dut
1)	2)	Lehena doi doia nahiago dut
1)	2)	Biak berdin (gaziki edo ongi)
1)	2)	Bigarrenra doi doia nahiago dut
1)	2)	Bigarrenra aise nahiago dut
1)	2)	Zein nahiago duzu?:

Evaluación

- Tipos de evaluación subjetiva
 - Inteligibilidad:
 - Test de opciones cerradas
 - Escritura de texto: escribir lo que se entiende
 - Naturalidad, calidad, parecido:
 - Mean Opinion Score (MOS): asignar puntuación del 1 al 5
(Voz humana: MOS entre 4.5 y 4.8)
 - Comparación de dos señales A/B
 - Comparación de dos señales con una referencia A/B/X
 - Pruebas de comprensión de textos
 - Retos (*Challenges*) <http://www.festvox.org/blizzard/>

Evaluación

- Tipos de evaluadores
 - Expertos
 - Motivados pero no representativos
 - Nativos
 - Difíciles de reclutar (posibilidad de recompensa)
 - Voluntarios
 - Contactados a través de la WEB
 - Fáciles de localizar
 - Poco control de las condiciones de la evaluación
 - Crowd-sourcing
 - Participación masiva
 - Detectar y descartar outliers



Evaluación

- Consideraciones prácticas
 - Definir el número de señales a evaluar
 - Diseñar cuidadosamente el test
 - Aleatorizar las señales
 - Motivar a los evaluadores
 - Asegurarse de que el número de evaluadores es suficiente
 - Explicar claramente qué se debe evaluar
 - Incluir señales naturales en el test
 - Comprobar la consistencia y fiabilidad de los evaluadores analizando sus respuestas

Evaluación

Escala MOS (Mean opinion score)

Valor	Significado
5	Excelente
4	Bueno
3	Decente
2	Pobre
1	Malo

Modelo	MOS
DeepVoice	3,94
DeepVoice 2	2,96
DeepVoice 3	3,78
Tacotron	3,82
Tacotron 2	4,52
ClariNet	4,22
FastSpeech	3,84
Triple M	4,57

Evaluación

Table 4: MOS and CMOS comparisons between NaturalSpeech and previous TTS systems.

System	MOS	CMOS
FastSpeech 2 [18] + HiFiGAN [17]	4.32 ± 0.15	-0.33
Glow-TTS [13] + HiFiGAN [17]	4.34 ± 0.13	-0.26
Grad-TTS [14] + HiFiGAN [17]	4.37 ± 0.13	-0.24
VITS [15]	4.43 ± 0.13	-0.20
NaturalSpeech	4.56 ± 0.13	0

*CMOS (Comparative Mean Opinion Score) is a well accepted method in the speech industry for comparing the voice quality of two TTS systems

Índice

- Introducción
- Estructura de un sistema TTS
- Análisis del texto
- Modelado prosódico
- Técnicas de síntesis
- Evaluación
- Trending topics

Trending Topics

- Bancos de voces y apps de comunicación personalizada



https://www.ted.com/talks/rupal_patel_synthetic voices as unique as fingerprints

The screenshot shows a web browser window for the Aholab website. The main content area is titled "Donate your voice". It includes a recording interface with a green square button labeled "Start Recording", a play button labeled "Play", and an upload button labeled "Upload". Below this is a section titled "Sentences" with a progress bar indicating "You have recorded 2/100 sentences". On the right side of the page, there is a "User menu" with options like "My account" and "Log out", a "Languages" section listing various languages with flags, and a "Sponsors" section featuring the Aholab logo.

The screenshots show the Aholab mobile application interface. The top part shows the login screen with fields for "Username" and "Password", and a "Login" button. The middle part shows a gesture drawing of a house. The bottom part shows the main synthesis interface with sections for "General Options" (including "Select model voice to synthesize" and dropdowns for "Default voice", "Pitch" (0.25 to 4), and "Speed" (0.25 to 4)), and buttons for "Done", "Discard", "Cancel", "Reset", and "Save".

Trending topics

- TTS expresivo y personalizado

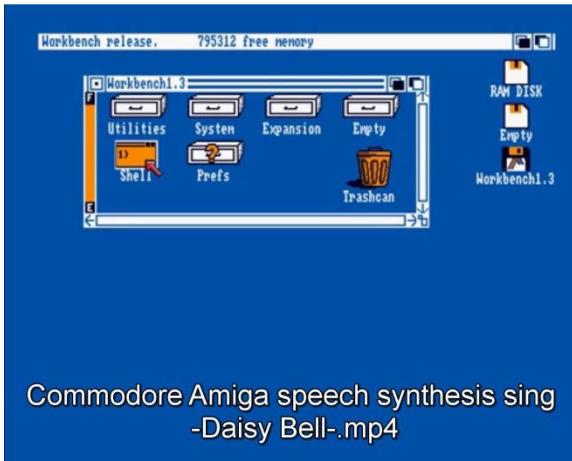


<http://www.idyacy.com/cgi-bin/bushomatic.cgi>

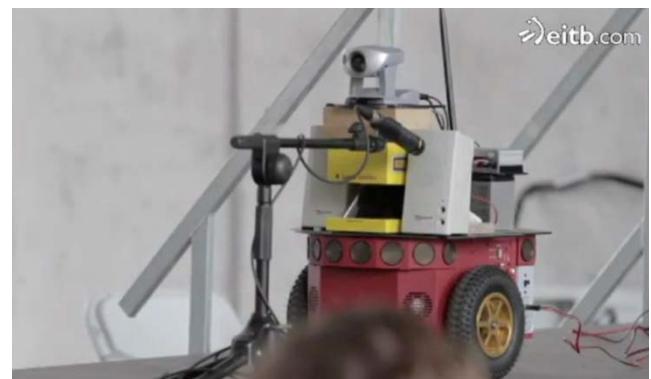
<http://www.oddcast.com/demos/tts/emotion.html>

Trending topics

- Síntesis de voz cantada



<http://host-d.oddcast.com/php/application> UI/doordId=373/clientId=1/
<http://www.sinsky.jp/>



E. Del Blanco, I. Hernández, E. Navas, X. Sarasola, D. Erro, “[Bertsokantari: a TTS based singing synthesis system](#)”
Proc. Interspeech, 2016

Trending topics

- Síntesis para lenguajes minoritarios
 - Hay más de 7000 idiomas en el mundo, pero los sistemas TTS más “populares” solo admiten unas docenas de idiomas
 - Existe una fuerte demanda comercial para ampliar los sistemas TTS a más idiomas
 - No ayuda: la falta de datos en idiomas minoritarios y el alto coste de la recopilación



Referencias

- Allen J., Hunnicut M.S., Klatt D.H. From text to speech: The MITTalk system, 1987, Cambridge, UK, University Press
- Arik, O., Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi (2017) DeepVoice: Real-time Neural Text-to-Speech <https://arxiv.org/abs/1702.07825>
- Arik, O., Chen, J., Peng, K., Ping, W., Zhou, Y., (2018) Neural Voice Cloning with a Few Samples <https://arxiv.org/pdf/1802.06006.pdf>
- Bisani, M., Hermann Ney. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. Speech Communication, Elsevier : North-Holland, 2008, 50 (5), pp.434. <10.1016/j.specom.2008.01.002>
- Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarikhani, A., Hu, Q., ... Zhang, H. (2017). Siri on-device deep learning-guided unit selection text-to-speech system. Interspeech, pp. 4011–4015.
- Chen, Y., Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, Nando de Freitas (2018) SAMPLE EFFICIENT ADAPTIVE TEXT-TO-SPEECH, <https://arxiv.org/pdf/1809.10460v1>
- Donahue J. et al., End-to-End Adversarial Text-to-Speech, arXiv:2006.03575, 2020
- Del Blanco, E. I. Hernández, E. Navas, X. Sarasola, D. Erro, “[Bertsokantari: a TTS based singing synthesis system](#)” Proc. Interspeech, 2016
- Erro, D., I. Sainz, E. Navas, I. Hernaez (2014) Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis , IEEE Jorunal of Selected Topics in Signal Processing, 8 (2)
- Fan, Y., Qian, Y., Xie, F., & Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based Recurrent Neural Networks. Proceedings of INTERSPEECH, (September), 1964–1968.
- Hen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.J. and Saurous, R.A. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In Proc. ICASSP <https://arxiv.org/pdf/1712.05884.pdf>
- Hirose, K., & Tao, J. (Eds.). (2015). Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis. Springer.
- King, S. (2011). An introduction to statistical parametric speech synthesis. Sadhana, 36(5), 837-852.
- King, S., Interspeech 2017 tutorial, Simon King, Oliver Watts, Srikanth Ronanki, Felipe Espic, Zhizheng Wu http://media.speech.zone/images/Interspeech2017_tutorial_Merlin_for_publication_watermarked_compressed_v2.pdf
- Moulines, E., F. Charpentier (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication, 9, 453–467
- Naihan Li et al., Neural Speech Synthesis with Transformer Network, arXiv:1809.08895, 2018
- Navas, E., I. Hernández, I. Luengo (2006) An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS, IEEE Trans. On SAP, 14(4)
- Navas, E., I. Hernández, I. Sainz (2008) Evaluation of automatic break insertion for an agglutinative and inflected language, Speech Communication, 50 (11-12)

Referencias

- Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio, 1–15.
- Oord, A., van den, Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., ... Hassabis, D. (2017). Parallel WaveNet: Fast High- Fidelity Speech Synthesis. <https://arxiv.org/pdf/1711.10433.pdf>
- Pierard, A. D. Erro, I. Hernández, E. Navas, T. Dutoit (2016) Surgery of Speech Synthesis Models to Overcome the Scarcity of Training Data, IBERSPEECH 2016
- Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., ... Miller, J. (2017). Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, 1–15. <http://arxiv.org/abs/1710.07654>
- Ping, W., Peng, K., Chen, J. (2018). ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech <https://arxiv.org/pdf/1807.07281.pdf>
- Prenger, R. et al., WaveGlow: A Flow-based Generative Network for Speech Synthesis, <arXiv:1811.00002, 2018>
- Ren, Y. et al., FastSpeech: Fast, Robust and Controllable Text to Speech, arXiv:1905.09263, 2019
- Ren, Y. et al., FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, <arXiv:2006.04558, 2020>
- Schultz, T., Michael Wand, Thomas Hueber, Krusienski Dean J., Christian Herff, Jonathan S. Brumberg, (2015) Biosignal-based Spoken Communication: A Survey, IEEE/ACM TASLP,25 (12)
- Tan, Xu, Tao Qin, Frank Soong, Tie-Yan Liu (2021) A Survey on Neural Speech Synthesis arXiv preprint arXiv:2106.15561
- Tan, Xu et al.(2022) NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality arXiv:2205.04421
- Taylor, P. (2009). Text-to-speech synthesis. Cambridge university press.
- Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, Proc. of ICASSP, pp.1315-1318, June 2000.
- Wu, Z., Watts, O., & King, S. (2016). Merlin : An Open Source Neural Network Speech Synthesis System. In 9th ISCA Speech Synthesis Workshop (SSW9) (pp. 218–223). Sunnyvale, CA, USA.
- Wu, Z. and S. King, “Investigating gated recurrent neural networks for speech synthesis,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2016.
- Zen, H., Senior, A., & Schuster, M. (2013, May). Statistical parametric speech synthesis using deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7962-7966). IEEE.
- Zen, H., Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. ICASSP, pp. 3844–3848.
- Zen, H. Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN (2015) <https://ai.google/research/pubs/pub4389>