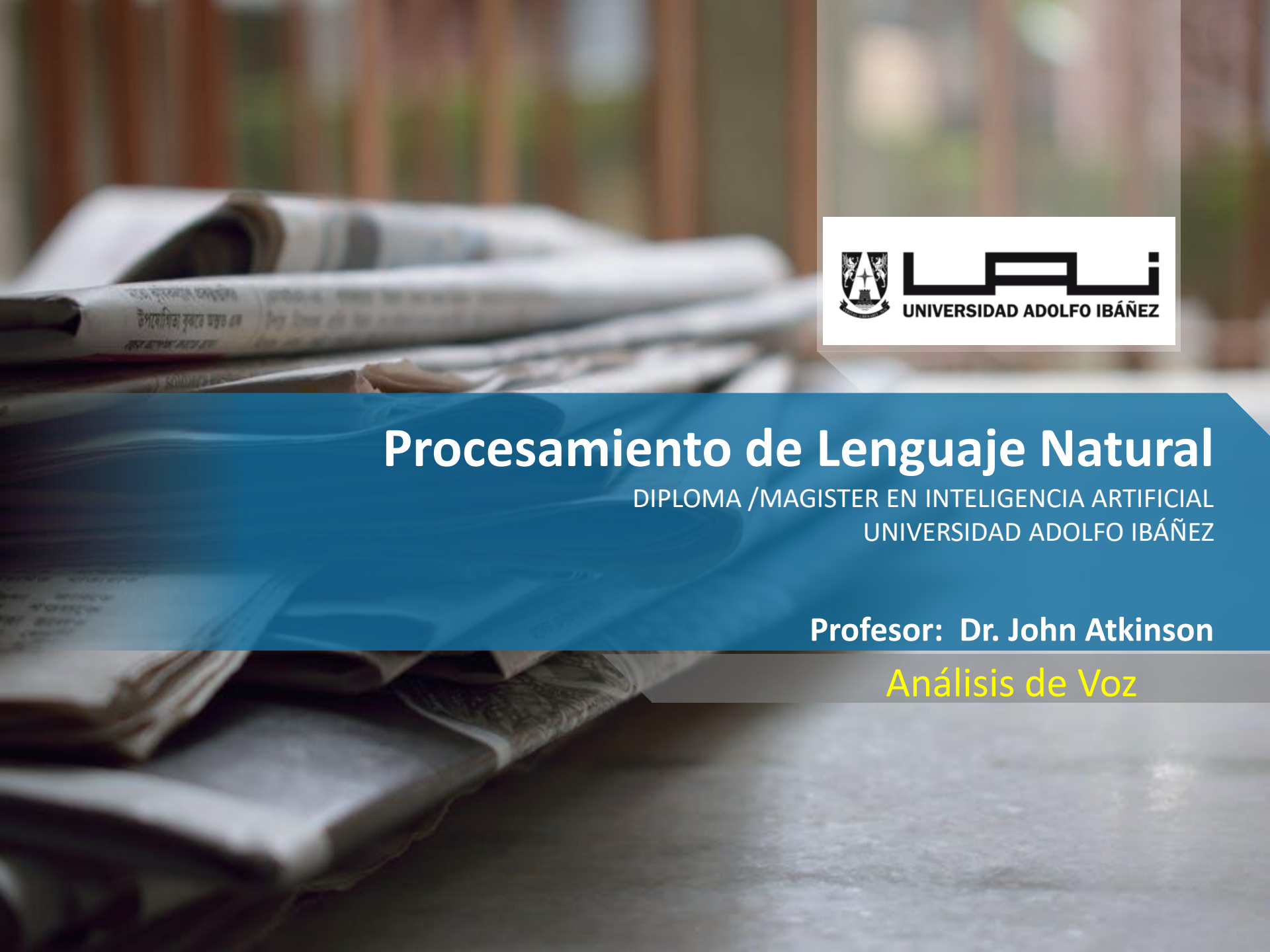


UNIVERSIDAD ADOLFO IBÁÑEZ

Procesamiento de Lenguaje Natural

DIPLOMA /MAGISTER EN INTELIGENCIA ARTIFICIAL
UNIVERSIDAD ADOLFO IBÁÑEZ

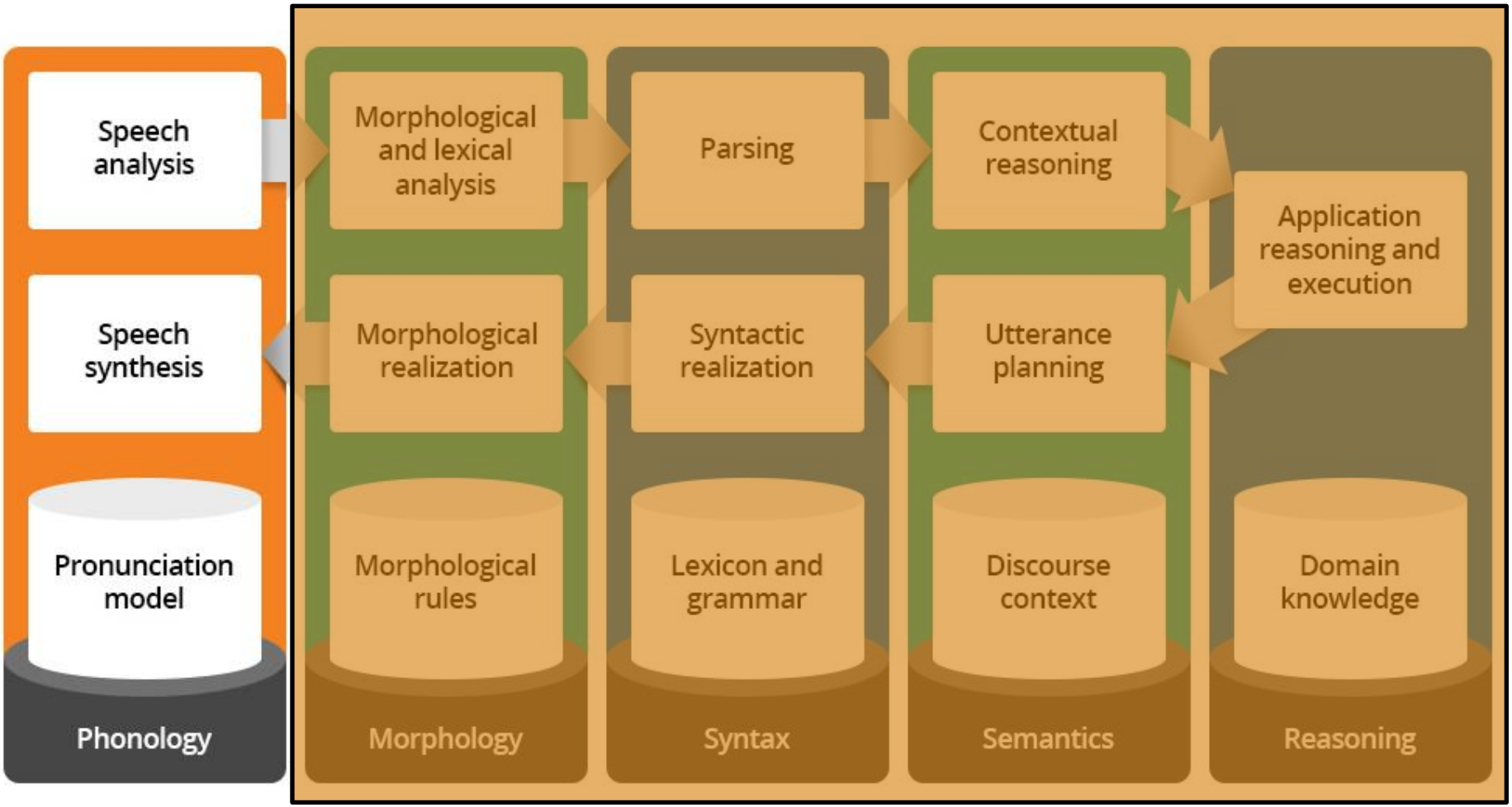
Profesor: Dr. John Atkinson

Análisis de Voz

OBJETIVO

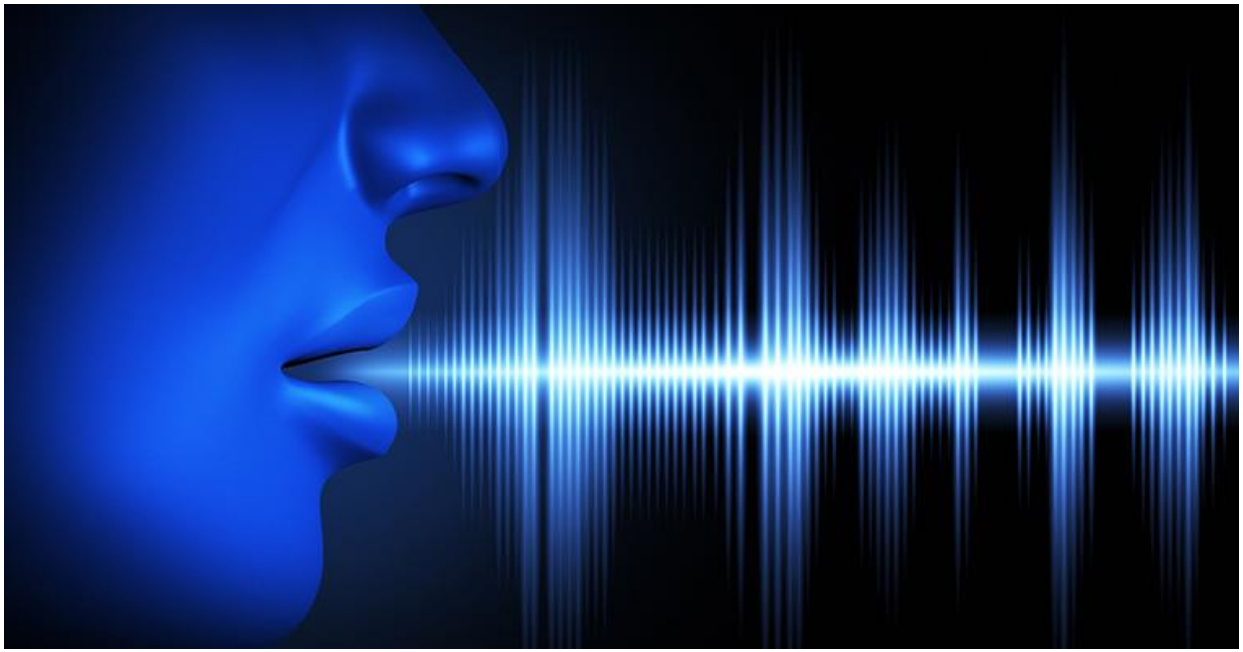
Entender los aspectos básicos conceptuales de las metodologías y tecnologías que permiten el reconocimiento y síntesis de voz.

ETAPAS EN NLP

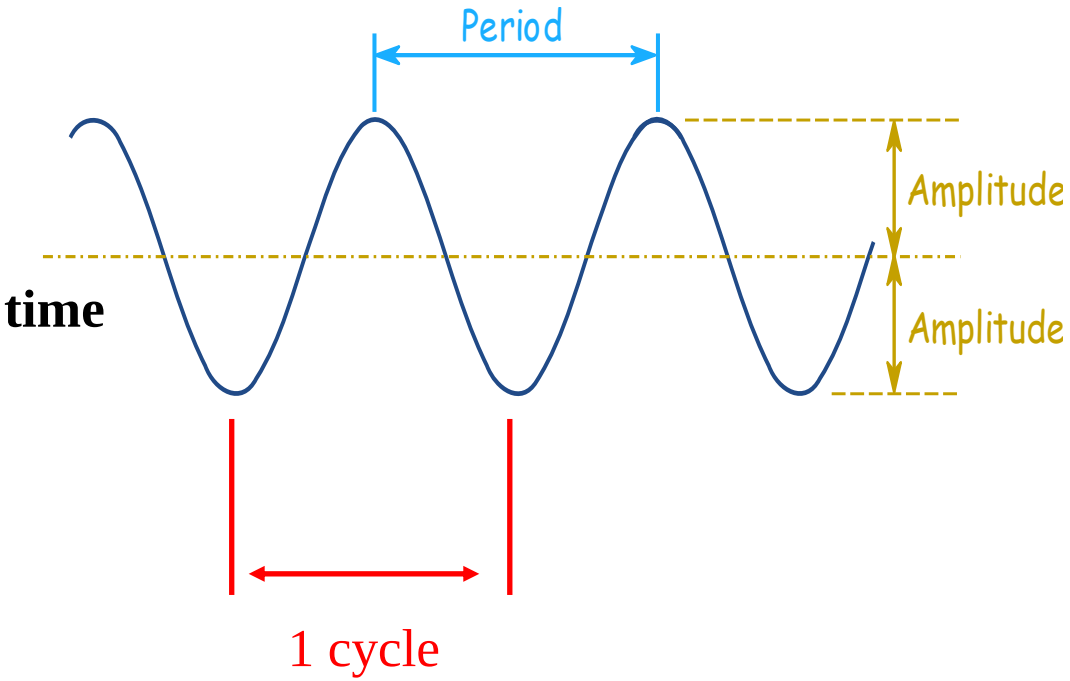


¿Qué pasa cuando hablamos?

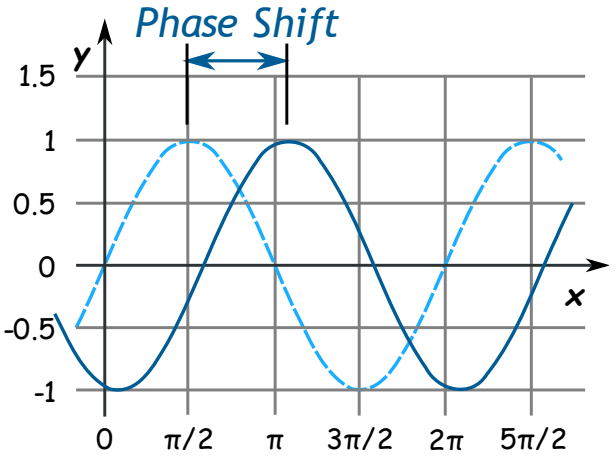
Transmitimos señales (*ondas*) desde una fuente (*emisor*) hacia un destino (*receptor*) a través de un medio (*canal*).



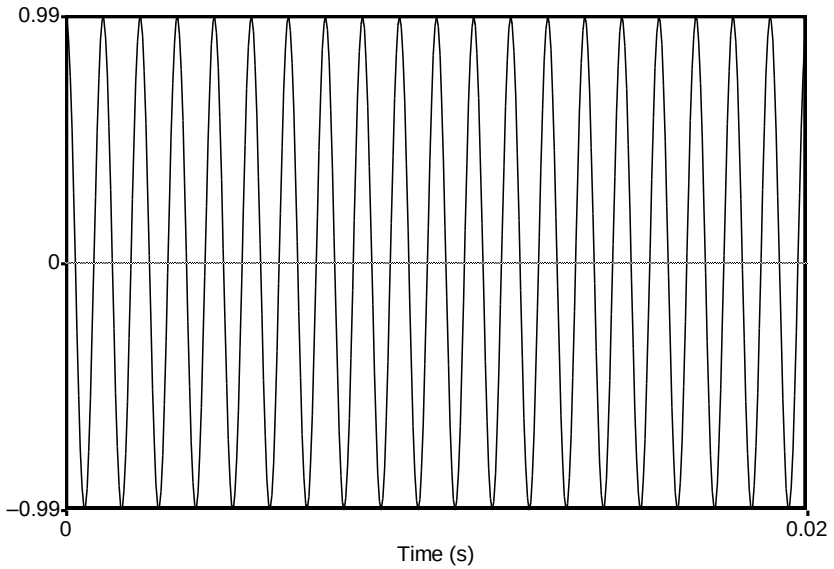
Conceptos: *Ondas Periódicas*



Frecuencia Fundamental
("ciclos por segundo" o Hz):
F0=1/Período



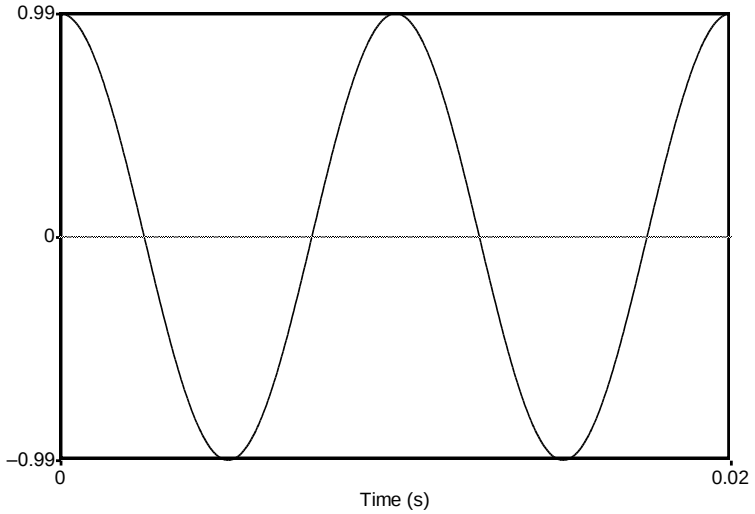
Ondas Periódicas



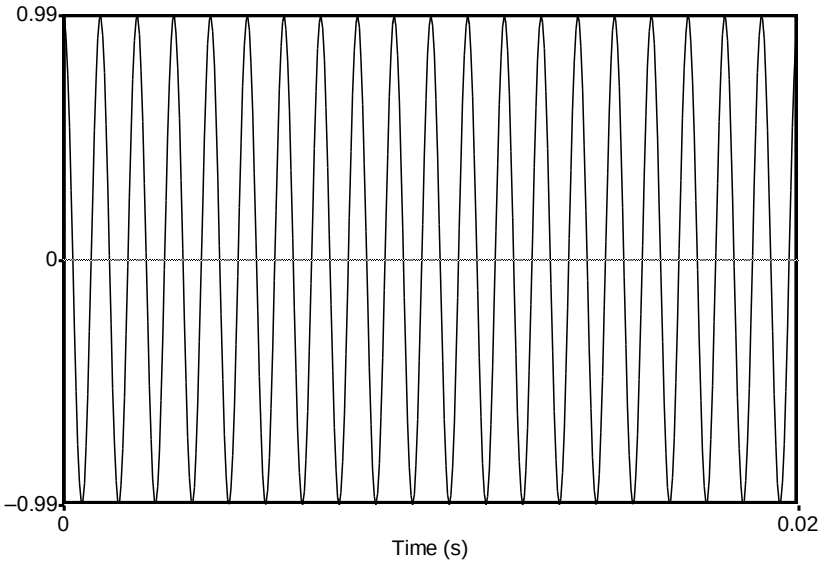
- Eje Y: Amplitud = cantidad de presión de aire en un instante
- Eje X: Tiempo. ¿F0?

  20 ciclos in .02 seg. = 1000 ciclos/seg ➔ F0 = 1000 Hz

Diferentes Frecuencias



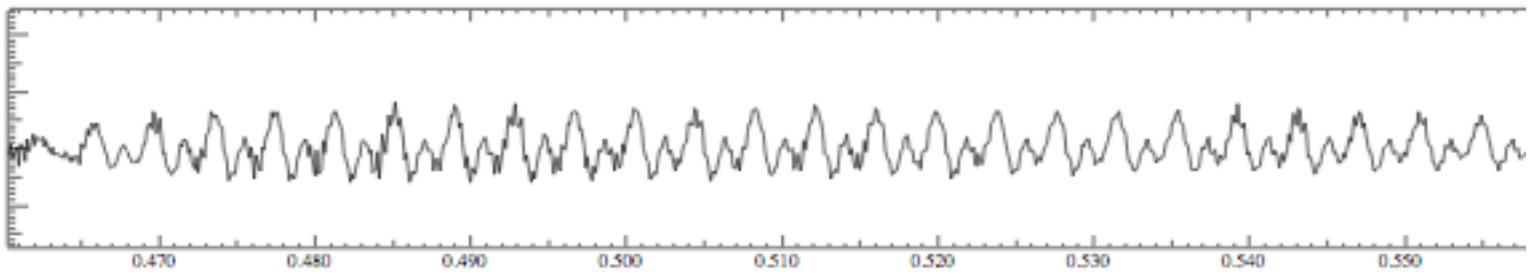
100 Hz



1000 Hz

Formas de Onda para Voz

Forma de onda de la *vocal* [iy] (en Inglés)

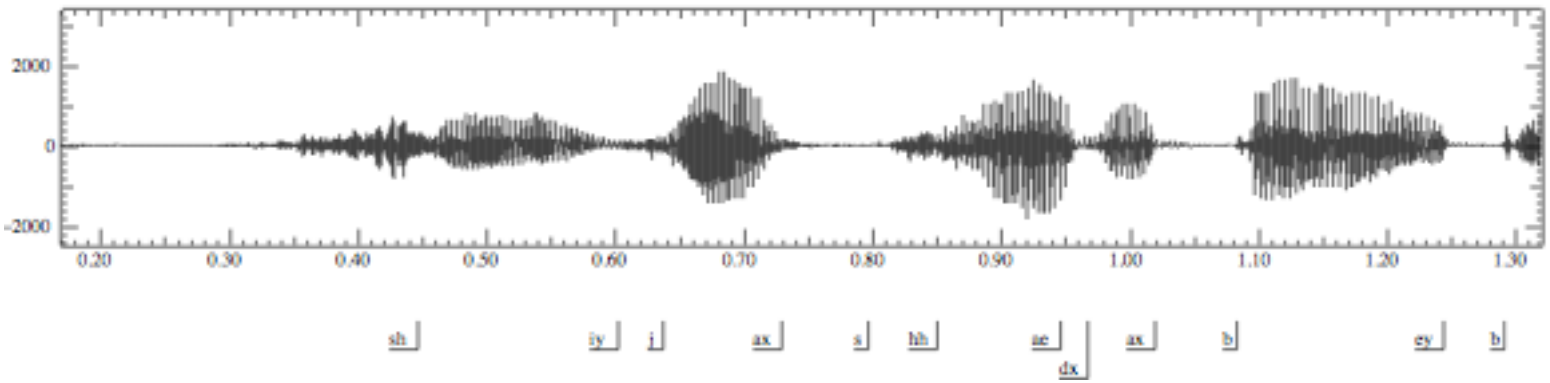


¿Qué podemos aprender de un registro de ondas?

- Vocal tiene 28 repeticiones en .11 segs, F0 es 28/.11 = ***255 Hz***.
- Esta es la velocidad a la cual se mueven las cuerdas vocales.

EJEMPLO

She just had a baby



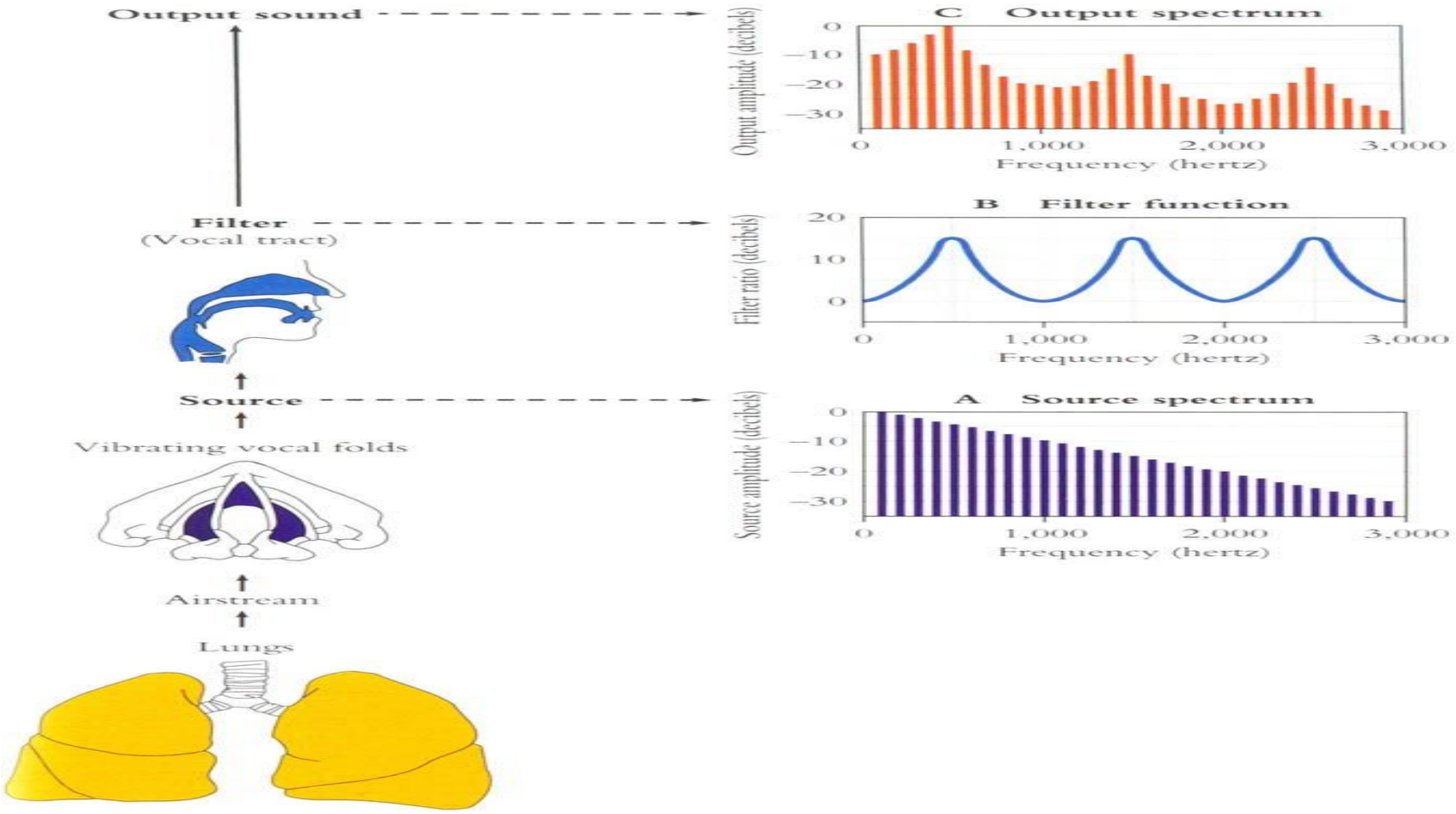
- Vocales son sonoras, largas y Fuertes.
- Combinaciones de ondas forman ***fonemas***.
- Varios ***peaks*** regulares en amplitud.
- Silencios de “cierre” (1.06 a 1.08 para primer [b], o 1.26 a 1.28 para segundo [b], etc)

¿Porqué los Peaks?

Razones de la *articulación*:

- Las vibraciones de las cuerdas vocales crean *armónicas* (expresadas por funciones *seno o coseno*)
- La boca es un *amplificador*:
  - ✓ Dependiendo de la forma de la boca, algunas armónicas se amplifican más que otras.

¿Cómo funciona en Humanos?



Formantes

- El *tracto vocal* actúa como "*amplificador*"; amplifica diferentes frecuencias.
- Las ***formantes*** son el resultado de diferentes formas de *tracto vocal*.
- Cada vez que las cuerdas vocales se abren y cierran, se expulsa aire desde los pulmones, actuando como "*llaves de pasada*" del el aire en el *tracto vocal*.
- Se debe ajustar las cavidades de resonancia con el fin de producir un número de diferentes frecuencias.

¿Y para qué serviría reconocer dichos *fonemas* y *formantes*?

Aplicaciones

- Dictado automático
- IVR (bancos, agencias, etc)
- Manos libres (en auto)
- Identificación del speaker
- Identificación de lenguaje
- Búsqueda en archivos de audio
- Muchas más..

Reconocimiento de Voz

- *Automatic Speech Recognition* (ASR) utiliza grandes vocabularios.
- ~20,000-64,000 palabras.
- Independiente del speaker vs. dependiente del speaker.
- Voz contínua vs palabras aisladas.

Evaluación: *Word Error Rate*

$$\textit{Word Error Rate} = \frac{100 * (\text{Insertions} + \text{Substitutions} + \text{Deletions})}{\text{Total Word in Correct Transcript}}$$

Ejemplo:

Referencia:

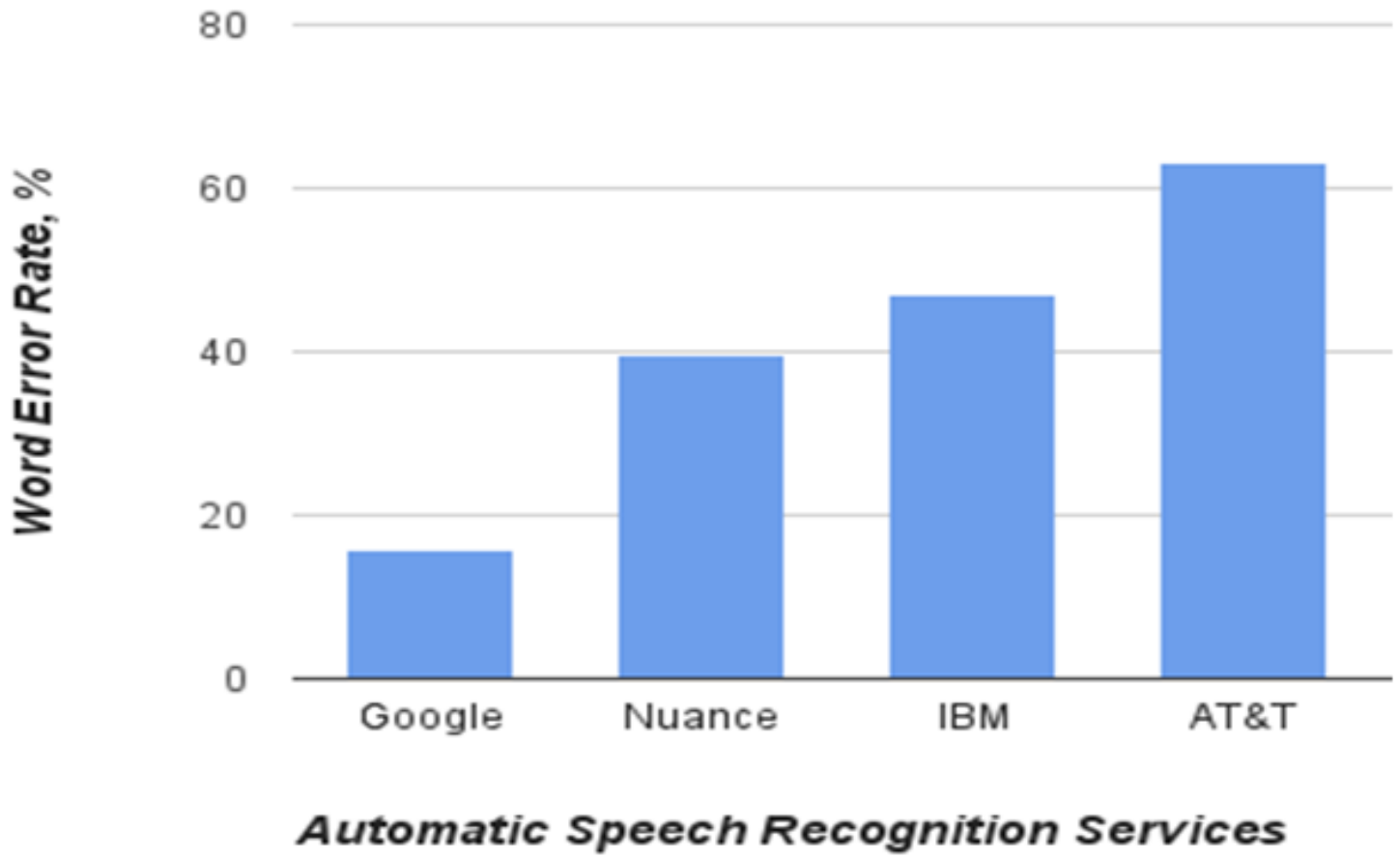
Quiero	******	Internet	Portátil	de promoción

Hipótesis:

Quiero	COMPRAR	un	Reptil	de promoción

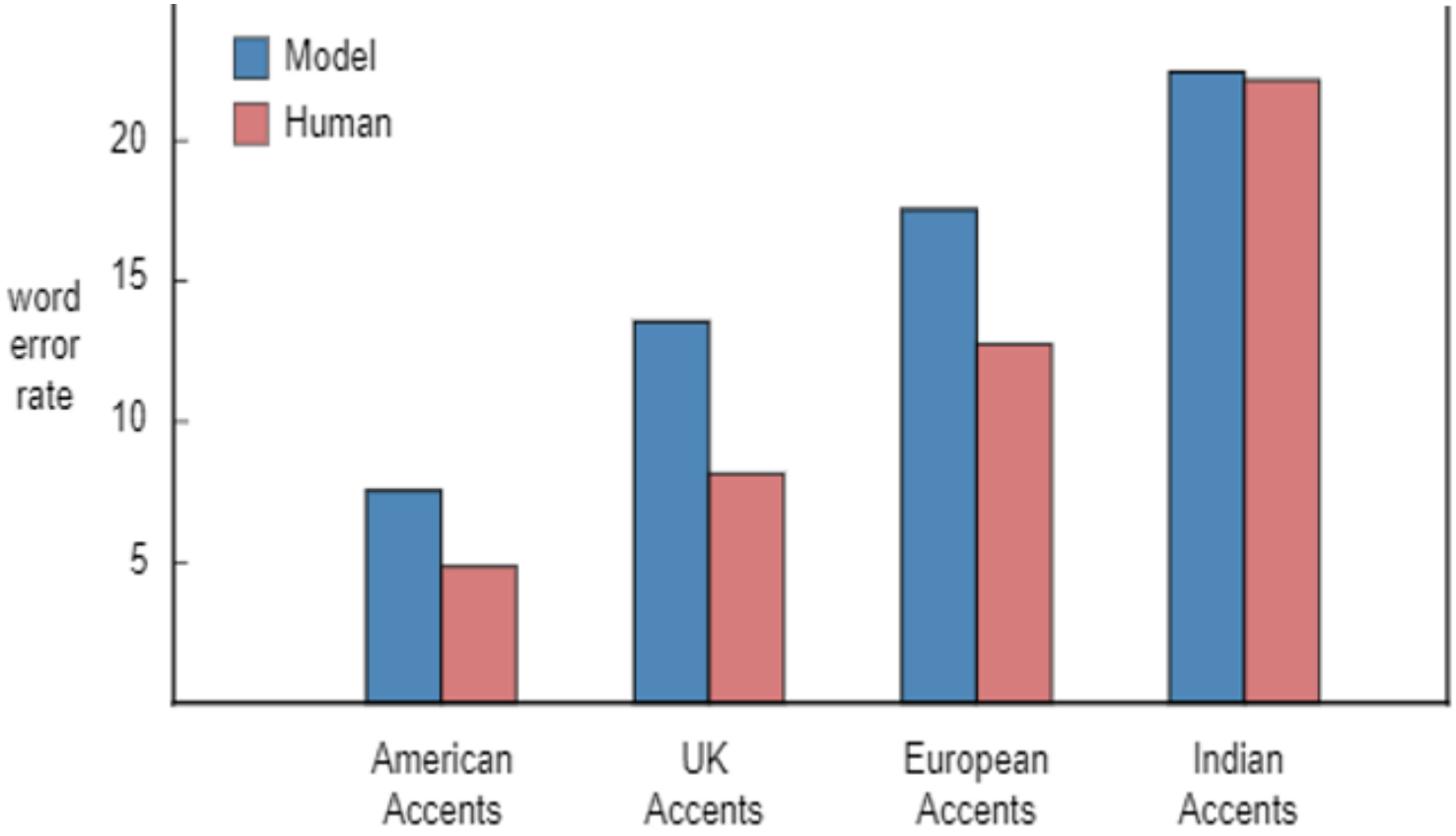
WER= 100*(1+2+0)/6 = 50%

Estado del Arte Comercial en ASR



¿Nos superarán las máquinas?

Modelos (ASR) vs Humanos

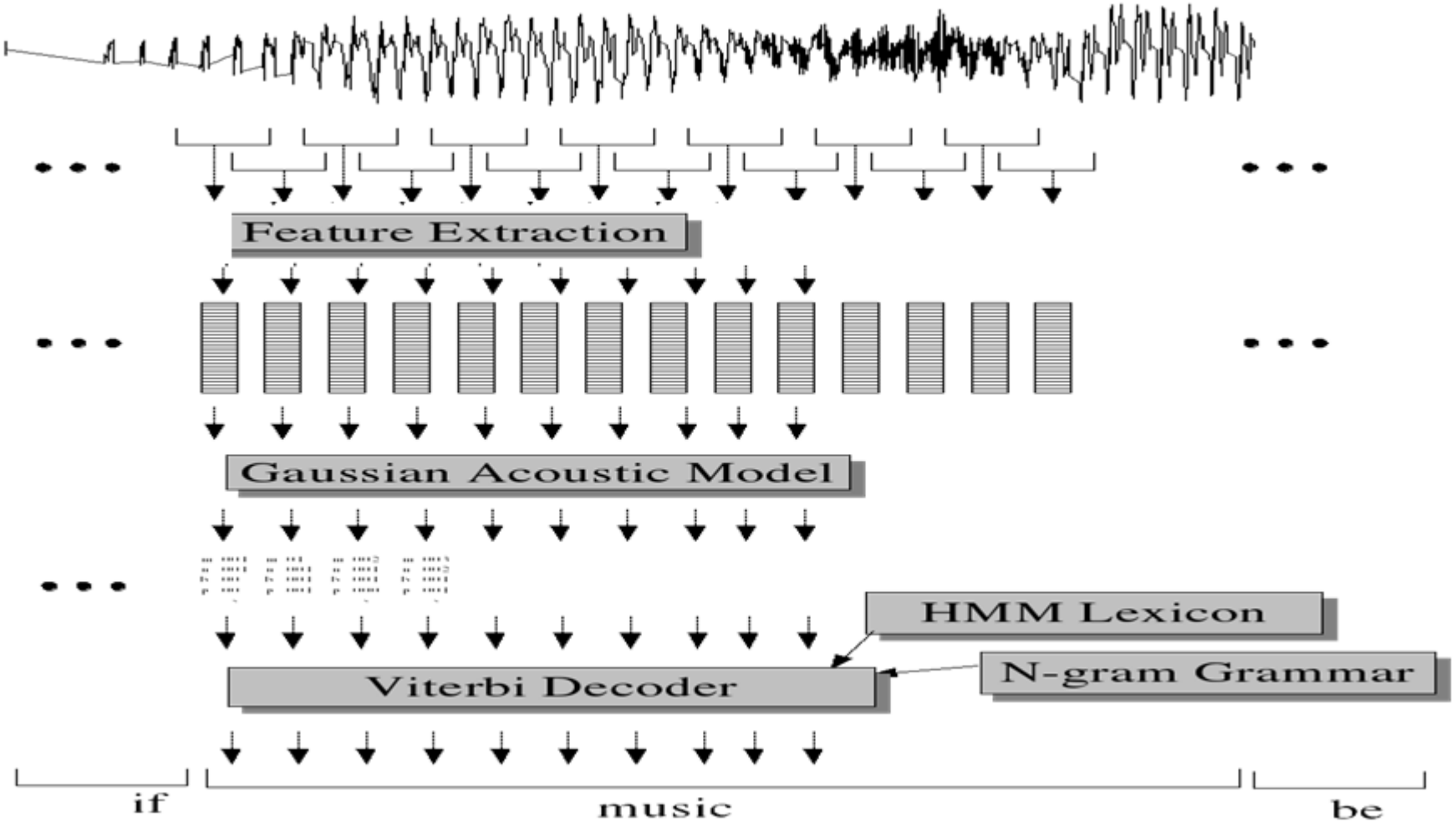


Diseño de ASR

1. Construir un *modelo (i.e., estocástico, neuronal, etc)* del proceso de conversion de “voz a palabras”
2. Recolectar muchas conversaciones, y transcribir todas las palabras para luego rotularlas.
3. Entrenar el modelo sobre las conversaciones rotuladas (*etiquetadas*).

Métodos usuales: aprendizaje automático, búsqueda heurística, modelos probabilísticos.

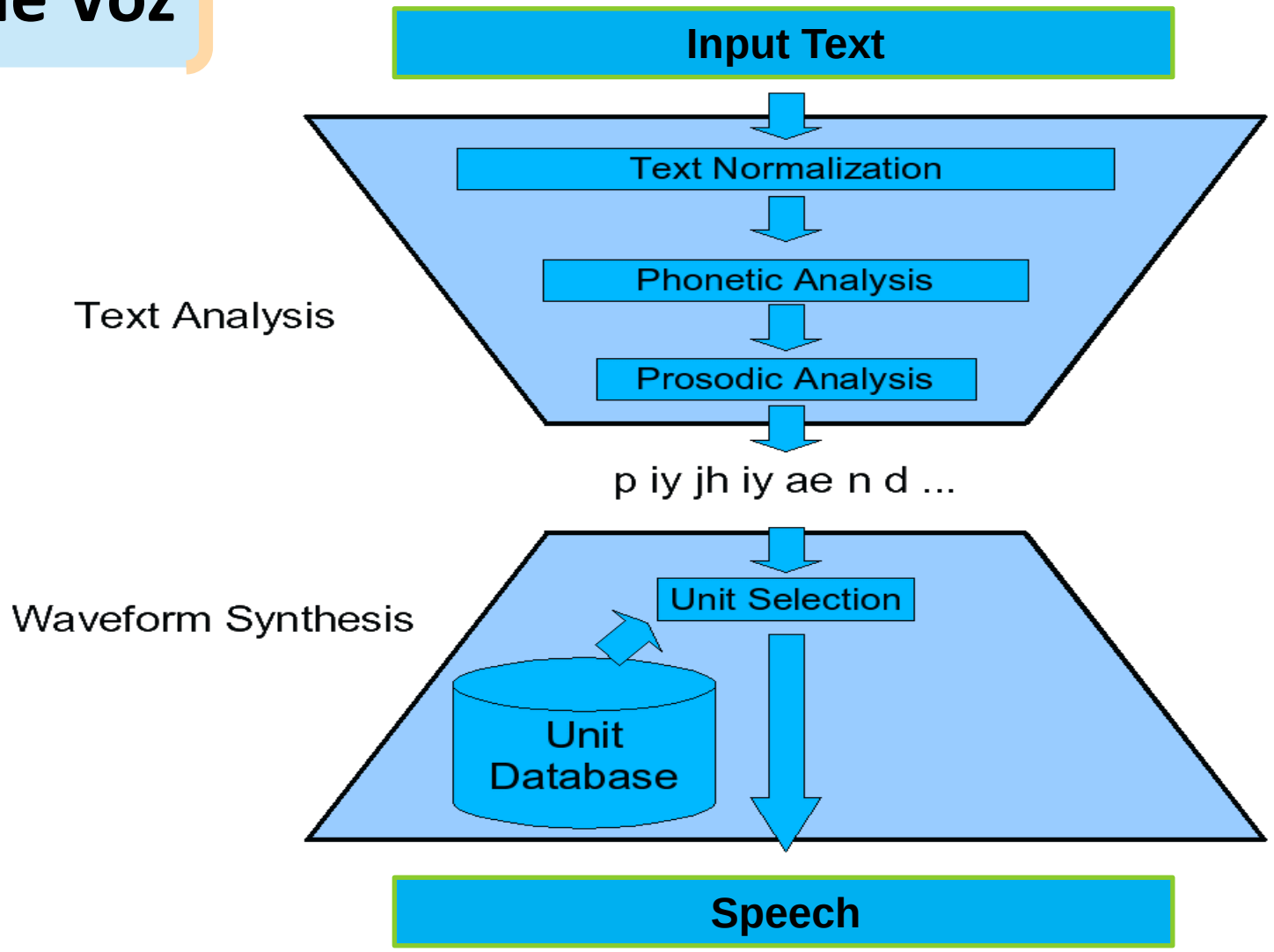
Arquitectura Típica de un ASR



El Modelo de *Canal con Ruido*

- Buscar a través del espacio de todas las oraciones posibles.
- Seleccionar la oración que es más probable dada *una forma de onda*.

Síntesis de Voz



RESUMEN

- El análisis de voz tiene muchas aplicaciones comerciales tanto en el reconocimiento (ASR) como en síntesis (TTS).
- El estado del arte de la tecnología ha avanzado significativamente reduciendo el WER.
- Problemas abiertos: ASR contínuo, TTS de voz y entonación “natural”, ruido ambiental, fonemas, etc


Tiempo de Ejercicios

Ejercicio

- ✓ Cargue en *Google Colab* el programa **`asr-tts`** que permite realizar reconocimiento y síntesis de voz, para un ejemplo simple.
- ✓ Siga instrucciones del profesor