Reconocimiento de Voz

Eduardo Morales, Enrique Sucar

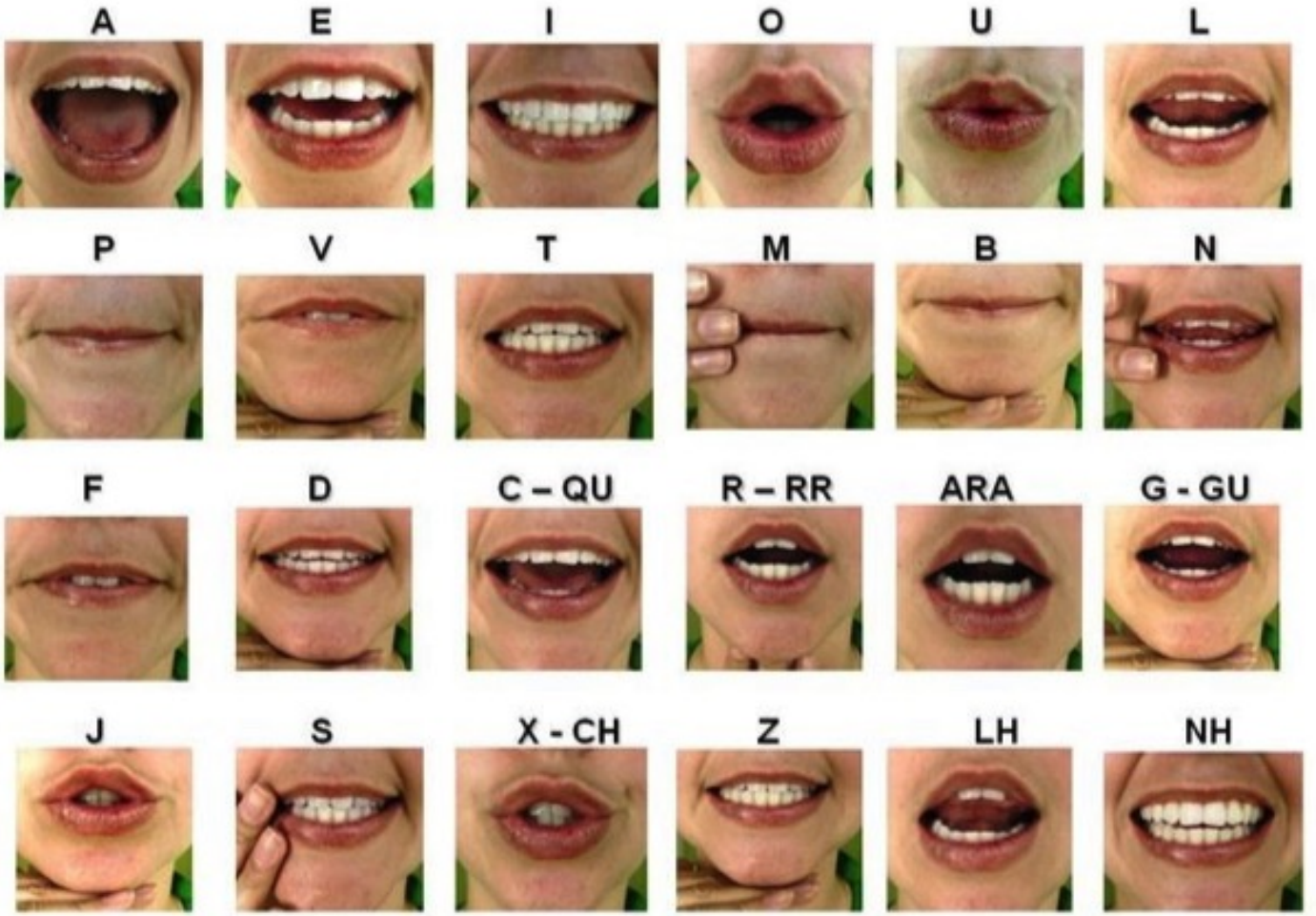
INAOE

Contenido

Fonética

- Es el estudio de los sonidos de la voz de los diferentes lenguajes en el mundo
- La pronunciación de las palabras se modela mediante una serie de sonidos básicos o *fonemas*
- Un fonema representa un sonido de un lenguaje – inglés, español, ...
- Los fonemas se dividen en dos clases principales: vocales y consonantes
- Existen alfabetos estándar para representar los fonemas como el IPA (International Phonetic Alphabet) y el ARPAbet.

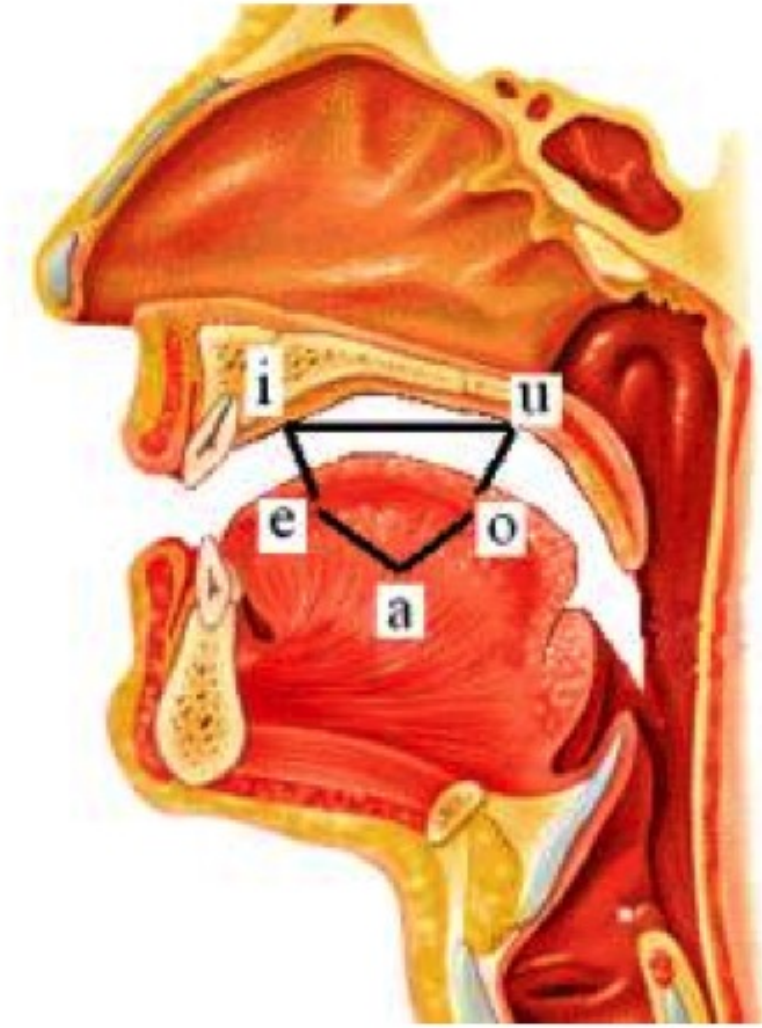
Fonemas en el Español



Fonética Articulatoria

- Es el estudio de como se producen los fonemas
- El sonido en los humanos se produce por el paso del aire generado en los pulmones por diferentes órganos: cuello, boca y nariz, esencialmente
- Un elemento importante son las *cuerdas vocales* que están en la laringe – son dos músculos que si se encuentran cercanos vibran y si están lejanos no vibran, lo que diferencía diversos fonemas
- La mayor parte de los sonidos se producen en la boca y algunos en nariz (nasales)
- Los sonidos se definen de acuerdo a la articulación de la boca

Órganos Vocales



Tipos de Articulaciones – Consonantes

Rasgo	Órganos	Ejemplos
Bilabial	Los dos labios	/p/, /b/, /m/
Labiodental	Labio inferior y dientes superiores	/f/
Interdental	Lengua entre los dientes	/z/
Dental	Lengua detrás de los dientes superiores	/t/, /d/
Alveolar	Lengua sobre la raíz de los dientes superiores	/s/, /l/, /r/, /rr/, /n/
Palatal	Lengua y paladar	/ch/, /y/, /ll/
Velar	Lengua y velo del paladar	/k/, /g/, /j/

Vocales

- Las vocales se diferencían también de acuerdo a las posiciones de las articulaciones
- Hay 3 factores principales: (i) altura de la parte más alta de la lengua, (ii) frente o atrás – de la parte más alta, y (iii) forma de los labios

Sílabas

- Una sílaba es, esencialmente, una vocal junto con algunas consonantes que la rodean y que están asociadas a la vocal
- La vocal central de la sílaba se la conoce como el núcleo
- Puede haber opcionalmente consonantes antes (onset) y después (coda) de la vocal
- La estructura de las sílabas define la *fonotáctica* de un lenguaje, estableciendo restricciones de que fonemas pueden seguir de otros
- Esto permite definir restricciones y también probabilidades de secuencias de fonemas (N-gram), lo que ayuda al reconocimiento

Categorías Fonológicas

- El sonido de los fonemas varía de acuerdo al contexto – coarticulación (fonemas antes y después) y otros factores
- La realizaciones de un fonema bajo diferentes contextos se conocen como *alófonos*
- Por ejemplo, algunos alófonos del fonema en inglés /t/: toucan, starfish, kitten, cat, butter, fruitcake, eight, past
- Otro factor que implica variaciones es el habla más coloquial y la velocidad con que se habla
- Una variación común es el “borrado”(deletion) de fonemas en particular al final de la palabra

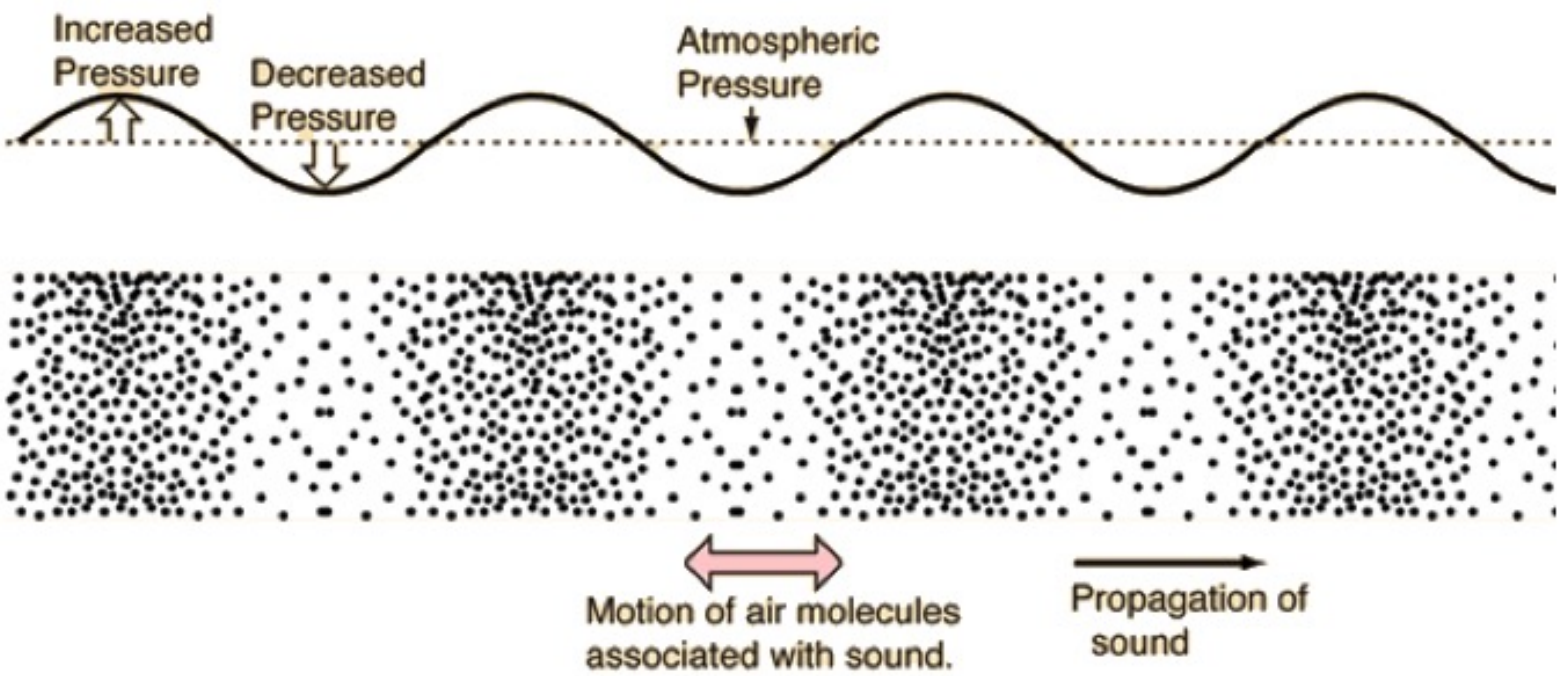
Factores de Variación Fonética

- Razón del habla: sílabas por segundo
- Frecuencia de las palabras o predecible (el borrado es más probable en palabras más frecuentes)
- Estado de ánimo del hablante
- Aspectos sociales del hablante: clase social, género, dialecto (lugar de origen)
- Contexto del hablante – situación social e interlocutor

Representación de las Señales Acústicas

- La entrada a nuestros oídos o un reconocedor de voz son ondas sonoras que son el producto de cambios de presión en el aire
- Dichas ondas se pueden representar mediante el cambio de presión del aire (magnitud) en el tiempo
- Dichas señales acústicas pasan por un convertidor análogo–digital para poder procesarlas en la computadora
- Esto implica un proceso de muestreo y cuantización

Señal Acústica



Muestreo

- Se mide la amplitud de la señal en ciertos tiempos de acuerdo a cierta *frecuencia de muestreo*
- La frecuencia de muestreo debe ser al menos dos veces mayor que la frecuencia mayor (Nyquist) o al menos dos muestras por ciclo
- La mayor parte de la voz humana tiene una frecuencia menor a 10,000 Hz (ciclos por segundo), por lo que bastaría una frecuencia de muestreo de 20,000 Hz
- En la práctica esta frecuencia es menor por diversos factores, normalmente 8,000 Hz para teléfonos y 16,000 Hz para micrófonos

Cuantización y Formatos

- El valor de la señal en cada muestra se discretiza normalmente en 8 o 16 bits
- Una vez que se muestrea y cuantiza la señal de voz se almacena utilizando normalmente algún formato estándar
- Antes de almacenarla se puede comprimir y también puede haber varios canales (estéreo)
- Algunos formatos comunes son el .wav (Microsoft), AIFF (Apple), AU (Sun)

Características de la Señal

- Como toda onda las propiedades básicas son su amplitud y frecuencia
- Aunque la señales acústicas no son una senoidal "pura", en particular en las vocales hay una frecuencia dominante (que depende de la frecuencia de vibración de las cuerdas vocales), conocida como *frecuencia fundamental* (F0)
- Además de la amplitud (valor) en una muestra, se utiliza la amplitud promedio en un periodo de tiempo, mediante el $RMS = \sqrt{1/N \sum_1^N X_i^2}$

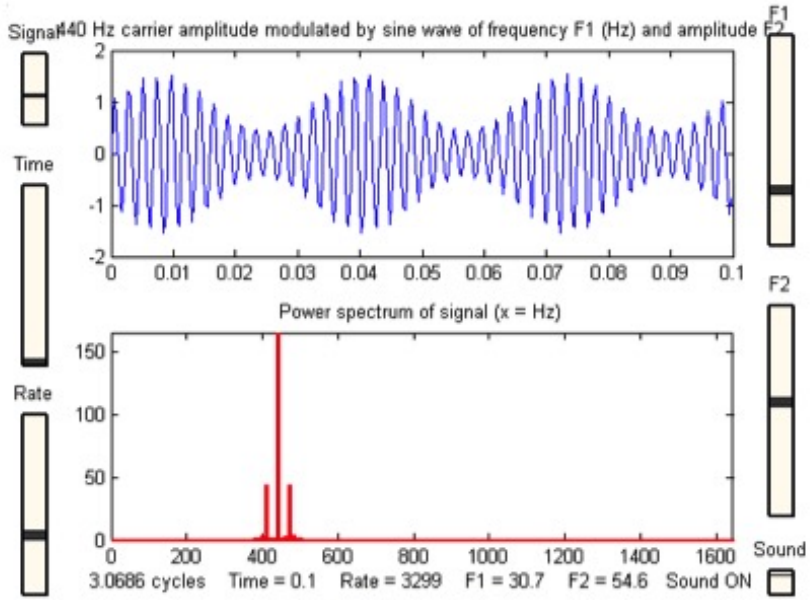
Pitch y Volumen

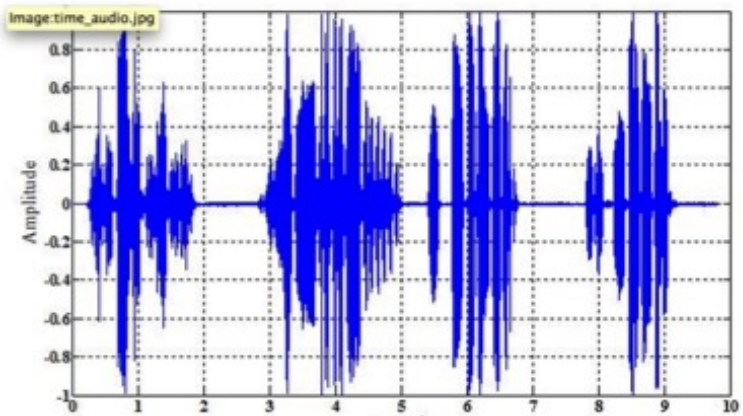
- El *pitch* es la percepción mental de la frecuencia fundamental
- En los humanos la percepción es lineal entre 100 y 1000 Hz, y para frecuencias mayores el pitch se correlaciona en forma logarítmica con la frecuencia
- Un modelo del pitch es la escala *mel*: $m = 1127 ln(1 + f/700)$
- El volumen es la escala perceptual de la potencia de la señal
- También es no-lineal: tenemos mayor resolución a menor potencia y depende de la frecuencia

Interpretación de los Fonemas

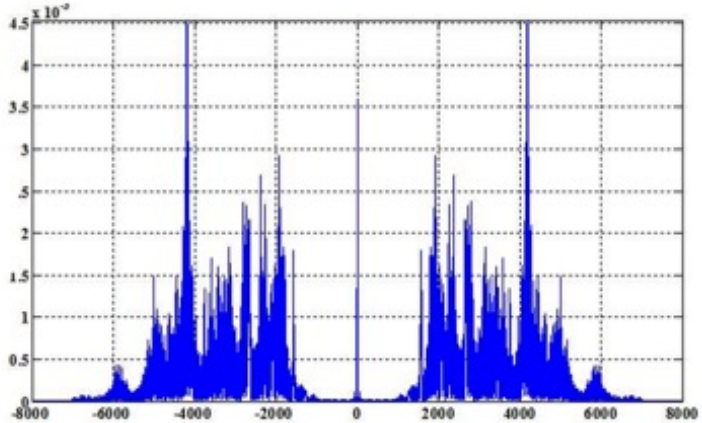
- A partir de la señal de voz se pueden distinguir *visualmente* algunos de los fonemas, en particular las vocales
- El reconocimiento de voz (en máquinas y humanos) se basa en una representación en base al espectro de frecuencia (Análisis de Fourier)
- El espectro representa la amplitud para cada unas de las componentes de frecuencia de la señal
- Ciertos *picos* en el espectro son característicos de ciertos fonemas – los fonemas tienden a tener una *firma espectral* característica

Ejemplos de Espectros





Time domain

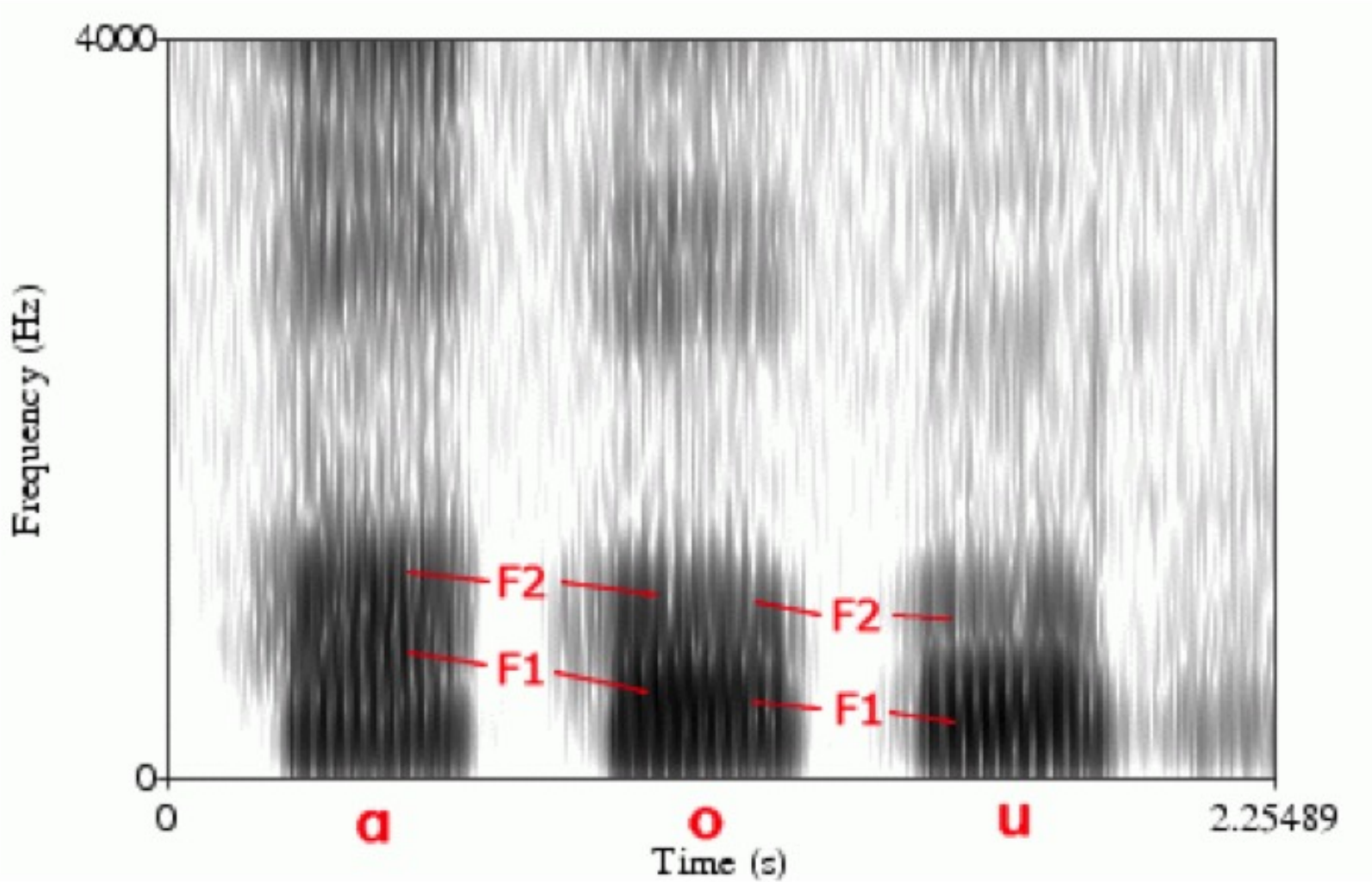


Frequency domain

Espectrograma

- Otra forma de ver el espectro es mediante un *espectrograma*: frecuencia vs. tiempo
- Cada pico o banda obscura en el espectrograma se conoce como *formante*
- Diferentes bandas o formantes son característicos de las diferentes vocales

Ejemplo de Formantes



Reconocimiento Automático del Habla

- El objetivo del reconocimiento automático del habla (ASR: automatic speech recognition) es transformar la señal acústica a una cadena de palabras.
- El problema en su forma general (cualquier hablante en cualquier ambiente) no está resuelto, aunque ha habido gran progreso recientemente que permite su uso en ciertos dominios
- Hay diversas aplicaciones de ASR: interacciíon humano-computadora, contestación automática en telefonía, interfaces multi-modeles, dictado, interfaces humano-robot, ayudantes inteligentes (SIRI, ALEXA, ...), etc.

Breve historia de ASR

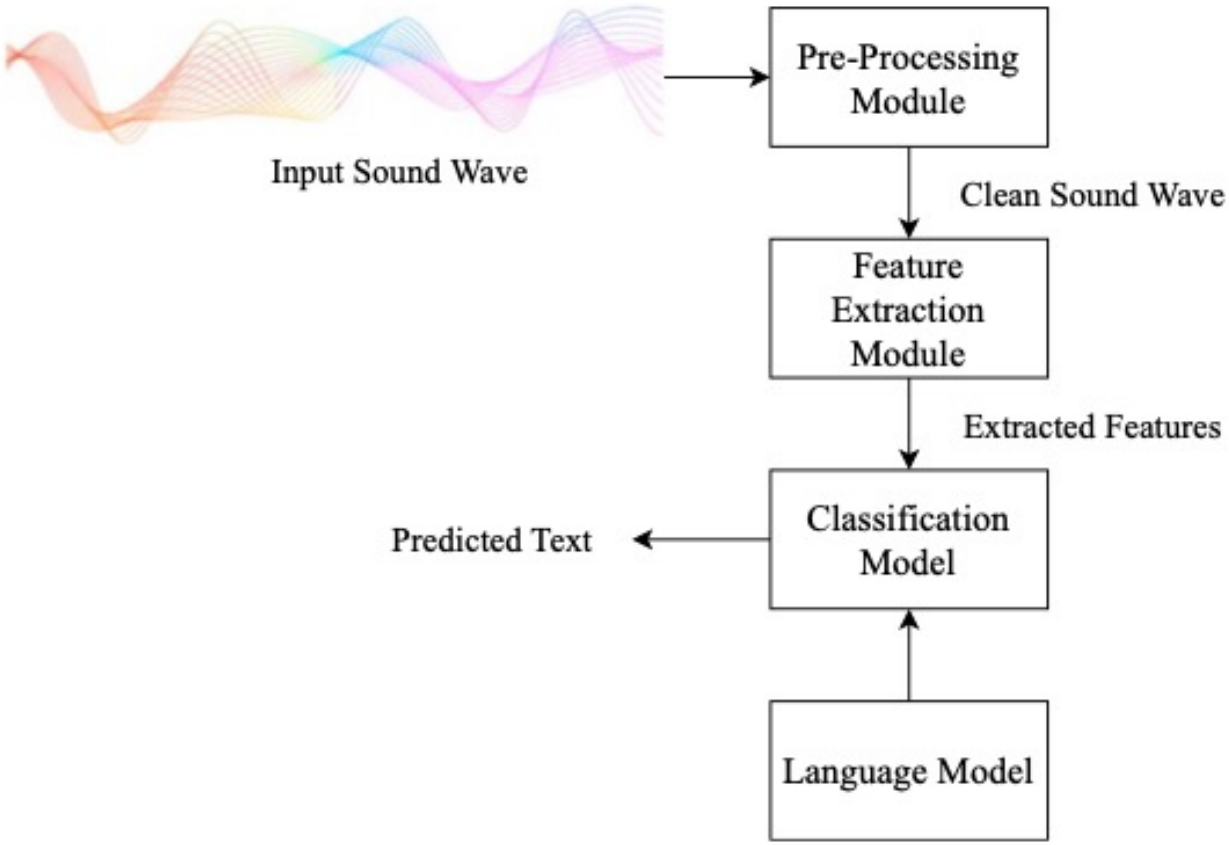
- Inicia en los 50s, en el que se inicia el tratar de reconocer y reproducir los lenguajes humanos.
- El primer sistema, *Audrey*, desarrollado en Laboratorios Bell, podía distinguir 8 dígitos diferentes; otro desarrollo de la época es del MIT que reconoce diversos dígitos de una persona.
- El área se expande en los 70s, hay un concurso de DARPA en el que surge el modelo del Pizarrón y también los Modelos Ocultos de Markov (que ganan el concurso).
- En los 80s se inventan los modelos de lenguaje (n-grama) para mejorar el reconocimiento.
- Recientemente ha surgido el uso de redes neuronales, en particular las redes recurrentes y LSTM.

Dimensiones

- Tamaño del vocabulario: de pocas palabras a 20,000-60,000 palabras
- Palabras aisladas vs. habla continua
- Lectura / dictado vs. conversiaciń entre personas
- Ruido ambiental
- Acento, tipo de hablante

La razón de errores depende de estos factores, de un 3 % en lectura de tipo Wall Street Journal ($< 20,000$ palabras) a un 20 % en conversación telefónica ($> 60,000$ palabras)

Arquitectura de un Sistema de RAV



Elementos

Hay entonces 3 elementos principales de un sistema de reconocimiento de voz:

- Extracción de características
- Clasificación
- Modelo del lenguaje

Extracción de Características

- Las observaciones para los clasificadores se extraen de la onda acústica, y consisten de un vector de características o *feature vector* que representa la información en una ventana temporal
- El más común en reconocimiento de voz son los **coeficientes cepstrales de la frecuencias de Mel** (MFCC)
- El primer paso, como comentamos antes, es el muestreo y cuantización de la señal acústica

MFCCs

MFCCs se calculan comúnmente de la siguiente forma:

1. Separar la señal en pequeños tramos.
2. A cada tramo aplicarle la Transformada de Fourier discreta y obtener la potencia espectral de la señal.
3. Aplicar el banco de filtros correspondientes a la Escala Mel al espectro obtenido en el paso anterior y sumar las energías en cada uno de ellos.
4. Tomar el logaritmo de todas las energías de cada frecuencia Mel

Vector de Características

- Normalmente se consideran los valores de los primeros 12 filtros
- Además se calcula derivada (velocidad) y doble derivada (aceleración) de cada coeficiente
- También se obtiene la *energía* de cada uno (valores, velocidad y aceleración)
- Esto da un vector de 39 características (valores reales) por muestra.
- Para utilizar esto como las observaciones una opción es transformar los valores en un conjunto de (256) valores discretos (codebook) mediante un proceso de agrupamiento. Otra opción es aproximar los valores continuos mediante distribuciones gaussianas.

Otras características

Se han utilizado otras formas de extracción de características:

- Codificación predictiva lineal (LPC) - representa la muestra actual como una combinación lineal de muestras pasadas
- Predicción perceptual lineal (PLP) - integra la energía en bandas críticas (mayor peso a frecuencias intermedias)
- Transformada Wavelet discreta (DWT) - toma en cuenta la información temporal

Clasíficación

Una vez extraídas las características de la señal, se utiliza algún esquema de clasificación para identificar a los fonemas. Veremos dos enfoques:

- Modelos Ocultos de Markov
- Redes Neuronales

Modelos Ocultos de Markov

- Una forma de ver el problema de reconocimiento de voz es considerar un *canal ruidoso*, considerando a la señal de voz como una versión ruidosa de la secuencia de palabras
- Bajo este punto de vista se puede atacar el problema bajo un enfoque bayesiano – encontrar las secuencia de palabras más probable dada la *evidencia* o señal acústica
- La señal acústica se puede ver como una secuencia de observaciones (muestras): $O = o_1, o_2, \ldots, o_t$
- La salida es una secuencia de palabras: $W = w_1, w_2, \ldots, w_n$
- Entonces lo que se busca es la secuencia de palabras más probable dada las observaciones: $W^* = argmax P(W \mid O)$

Modelo

- Usando el teorema de Bayes:
  $W^* = argmax P(W)P(O \mid W)$
- $P(W)$ es la probabilidad previa de la secuencia de palabras – **modelo del lenguaje**
- $P(O \mid W)$ se basa en la verosimilitud que proviene de las observaciones – **modelo acústico**
- Dado que es un proceso dinámico se puede modelar como un **Modelo Oculto de Markov** (HMM: Hidden Markov Model) mediante el cuál se combinan ambos factores

HMM

- Un Modelo Oculto de Markov (Hidden Markov model, HMM) es una cadena de Markov donde los estados no son directamente observables.
- Un HMM es un proceso doblemente estocástico: (i) un proceso escondido que no podemos observar, (ii) un segundo proceso que produce las salidas dado el primero.
- Por ejemplo, tenemos dos monedas caragadas, M_1 y M_2. M_1 tiene una mayor probabilidad de *águila*, mientras M_2 tiene mayor probabilidad de *sol*. Alguien secuencialmente lanza las dos monedas pero no sabemos cual. Sólo observamos la salida, *águila* o *sol*

Ejemplo

Parámetros:

$$\Pi = \begin{array}{cc} M_1 & M_2 \\ \hline 0.5 & 0.5 \end{array}$$

$$A = \begin{array}{c|cc} & M_1 & M_2 \\ \hline M_1 & 0.5 & 0.5 \\ M_2 & 0.5 & 0.5 \end{array}$$

$$B = \begin{array}{c|cc} & M_1 & M_2 \\ \hline H & 0.8 & 0.2 \\ T & 0.2 & 0.8 \end{array}$$

Cuadro: Probabilidades iniciales (Π), probabilidades de transición (A) y probabilidades de observación (B) del ejemplo.

Definición

Estados: $Q = \{q_1, q_2, ..., q_n\}$

Observaciones: $O = \{o_1, o_2, ..., o_m\}$

Vector de probabilidades iniciales: $\Pi = \{\pi_1, \pi_2, ..., \pi_n\}$, donde

$$\pi_i = P(S_0 = q_i)$$

Matriz de probabilidades de trancisión: $A = \{a_{ij}\}$,

$i = [1..n], j = [1..n]$, donde

$$a_{ij} = P(S_t = q_j \mid S_{t-1} = q_i)$$

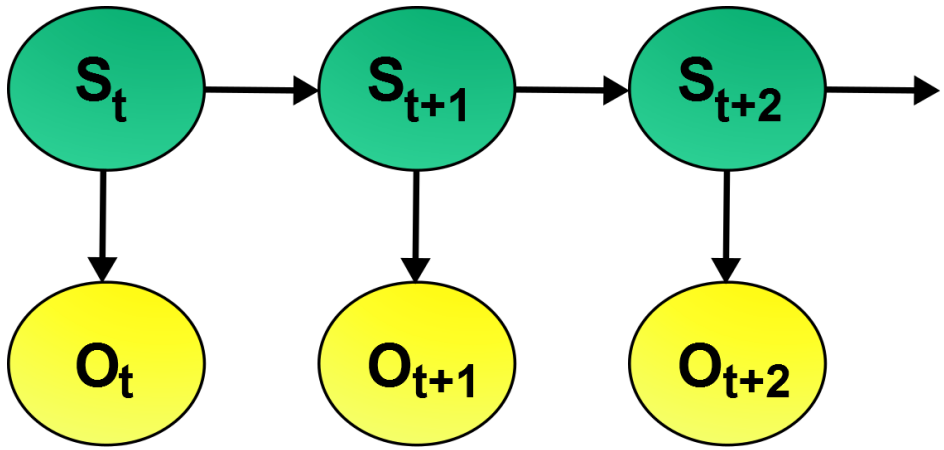
Matriz de probabilidades de observación: $B = \{b_{ij}\}$,

$i = [1..n], j = [1..m]$, donde

$$b_{ik} = P(O_t = o_k \mid S_t = q_i)$$

En forma compacta: $\lambda = \{A, B, \Pi\}$

Modelo Gráfico



Preguntas

1. *Evaluación*: dado un modelo, estimar la probabilidad de las observaciones.
2. *Secuencia óptima*: dada una secuencia de observaciones, encontrar la secuencia de estados más probables.
3. *Aprendizaje de los parámetros*: dado un conjunto de secuencias de observación, encontrar los parámetros del modelo.

Evaluación - método iterativo

- La idea básica del método *Forward*, es estimar las probabilidades por cada etapa de tiempo
- Calcular la probabilidad de una secuencia de observaciones hasta el tiempo t, en base a esto calcular para $t + 1$, ...
- Al final se obtiene la probabilidad de la secuencia completa.

Método iterativo

- Variable auxiliar *forward*:

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, S_t = q_i \mid \lambda) \quad (1)$$

- El método iterativo tiene 3 fases:
  - Inicialización – variables α para el tiempo inicial: $\alpha_1(i) = P(O_1, S_1 = q_i) = \pi_i b_i(O_1)$
  - Inducción – calcular $\alpha_{t+1}(i)$ en términos de $\alpha_t(i)$: $\alpha_t(j) = [\sum_i \alpha_{t-1}(i) a_{ij}] b_j(O_t)$
  - Terminación – $P(O \mid \lambda)$ se obtiene sumando α_T: $P(O) = \sum_i \alpha_T(i)$

Secuencia más probables

- La secuencia más probable Q dada la secuencia de observación O, de forma que se maximize $P(Q \mid O, \lambda)$
- Regla de Bayes: $P(Q \mid O, \lambda) = P(Q, O \mid \lambda)/P(O)$. Dado que $P(O)$ no depende Q, esto es equivalente a maximizar $P(Q, O \mid \lambda)$
- El método para obtener la secuencia óptima es el algoritmo de *Viterbi*

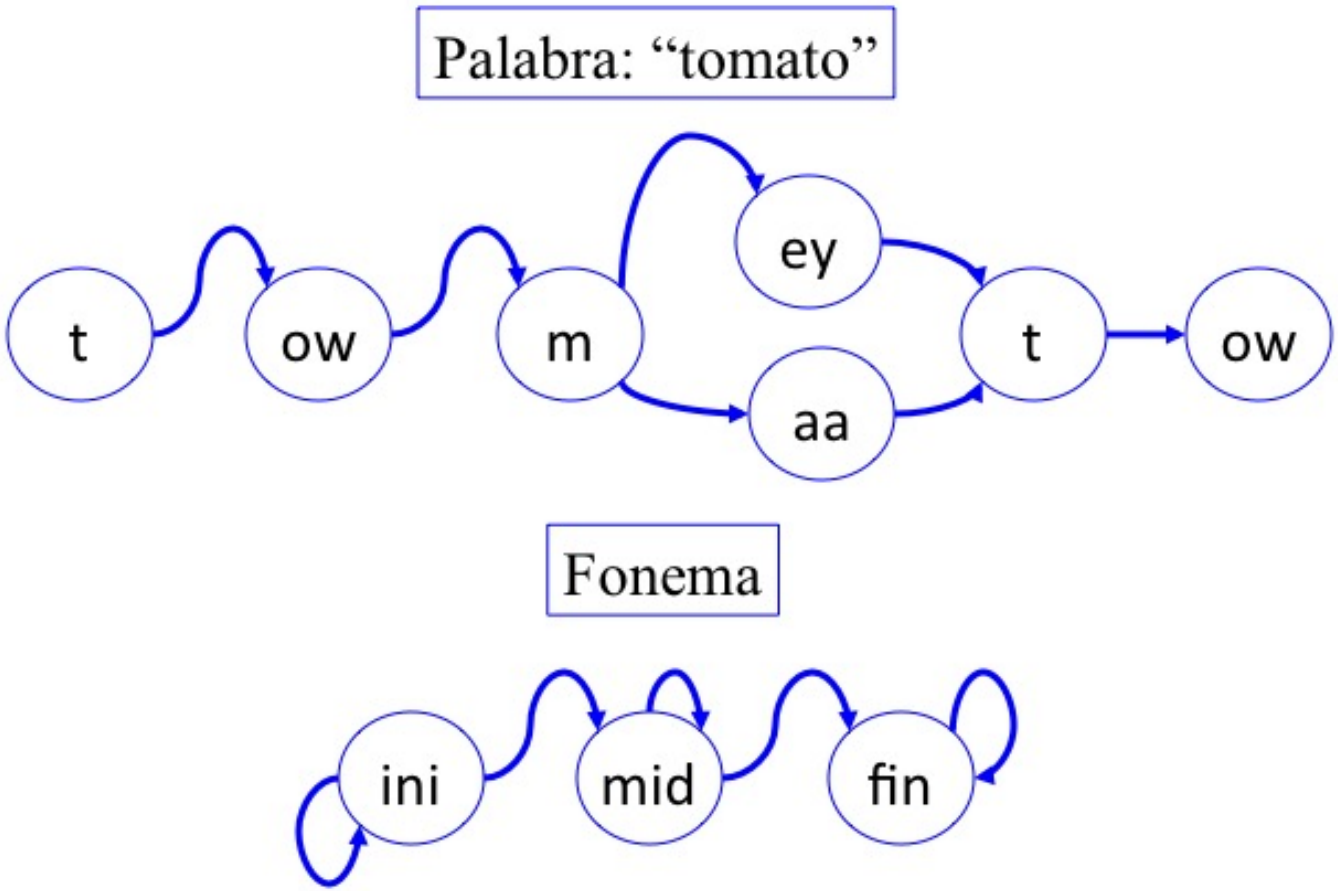
Aprendizaje de parámetros

- El método asume que la *estructura* del modelo es conocida: el número de estados y observaciones; por lo que se estiman sólo los parámetros:
- El algoritmo de Baum-Welch determina los parámetros del model, $\lambda = A, B, \Pi$, dado un número de secuencia de observaciones, $\mathbf{O} = O_1, O_2, ...O_K$
- Maximiza la probabilidad de las observaciones dado el modelo: $P(\mathbf{O} \mid \lambda)$

HMMs para ASR

- Se considera una estructura en que cada estado representa un *sub-fonema*: inicio, medio, fin; de forma que un fonema contiene 3 estados
- Estos modelos se pueden concatenar para obtener el modelo de una palabra
- Se considera una secuencia de transiciones de estado que sólo permite transiciones al siguiente estado o al mismo estado (HMM izquierda–derecha o Bakis)

Ejemplos de HMMs para ASR

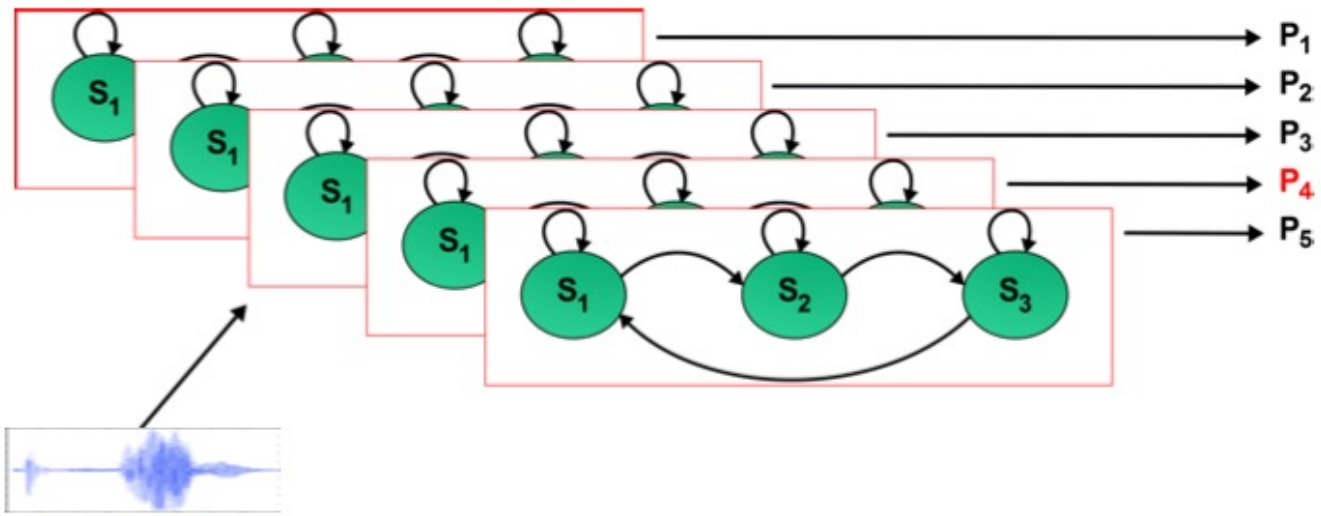


Modelo del HMM para ASR

- Q – estados que corresponden a subfonemas
- A – probabilidades de transición de un estado (sub-fonema) al siguiente (modelo del lenguaje)
- B – probabilidades de observación o *emisión*, expresan la probabilidad del vector de características de la muestra dado el sub-fonema

Reconocimiento

- Entrenamiento: se entrena un modelo (algoritmo de Baum-Welch) para cada fonema
- Reconocimiento: se obtiene la probabilidad de cada modelo (algoritmo Forward) y se selecciona el de mayor probabilidad



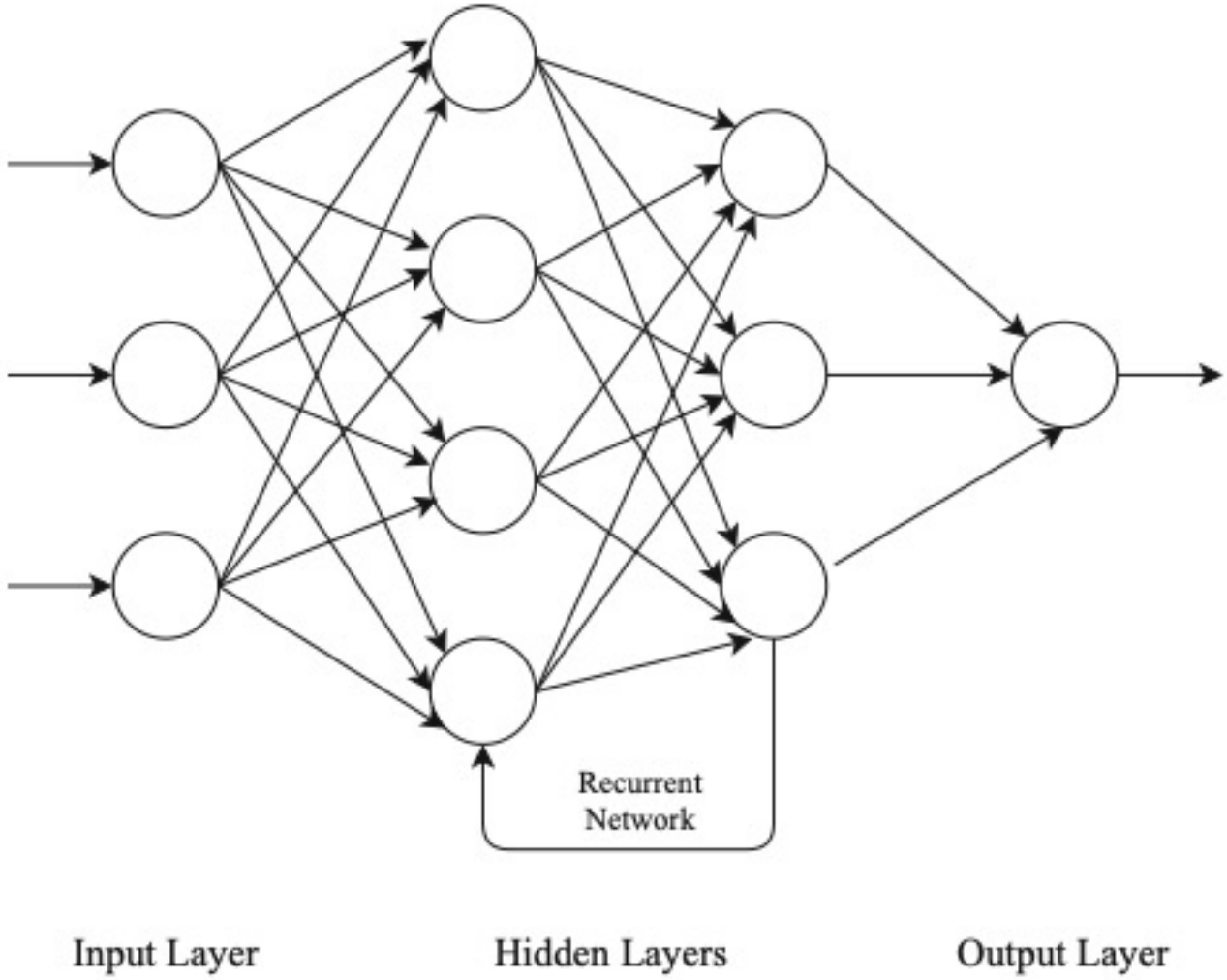
Redes Neuronales

- Recientemente se han aplicado redes neuronales (NNs) para modelado y reconocimiento de voz.
- Básicamente se entrena una NN para en base a las características de una ventana de la señal predecir el fonema.
- Se han utilizado para esto diferentes arquitecturas de redes neuronales:
  - Perceptrones multicapa
  - Mapas auto-organizativos (SOM)
  - Funciones de base radial (RBF)
  - Redes neuronales recorrentes
  - Redes neuronales convolucionales

Redes Neuronales Recurrentes

- Las redes neuronales recurrentes (RNN) tienen conexiones inversas (feedback) lo que permite guardar el estado interno.
- Para cierta secuencia de entrada, una RNN calcula el vector de salida y el vector para almacenar el estado interno.
- Utilizan celdas LSTM (Long-Short Term Memory) que les permiten tener memoria de corto y largo plazo.
- Para considerar no sólo el contexto pasado sino también el futuro, se utilizan RNNs bidireccionales (procesan el vector de entrada en ambas direcciones).

Redes Neuronales Recurrentes



Esquemas Híbridos

Otra alternativa es combinar diferentes tipos de modelos, por ejemplo:

- RBF-HMM – se utiliza RBF para extraer características que se dan como entrada a un HMM.
- CNN-LTSTM – se combinan CNNs (redes convolucionales) con LSTMs. Se tienen varias capas de CNNs combinadas con una capa posterior de LSTM bidireccional, seguida de una capa completamente conectada.
- FNN – combina sistema difusos con redes neuronales.

Modelos del Lenguaje

- Los modelos de lenguaje son un elemento importante para incrementar la eficacia de los sistemas de reconocimiento de voz.
- Básicamente incorporan restricciones estructurales del lenguaje a nivel palabras o fonemas.
- Para ello se pueden estimar las probabilidades de que una palabra (o fonema) pueda seguir después de otra (bigrama), después de otras dos (trigrama), etc.
- Los modelos de lenguaje ayudan a eliminar o al menos limitar las posibles ambigüedades de la señal de voz; al utilizarse en combinación con los clasificadores producen sistemas más robustos.

Datos

- Un elemento esencial para poder entrenar y probar los sistemas de reconocimiento de voz son los conjuntos de datos.
- Los datos consisten de grabaciones de voz junto con las transcripciones (textos) correspondientes.
- Estos conjuntos de datos deben ser amplios para incluir múltiples ejemplos de cada una de las palabras del vocabulario, así como diferentes hablantes y diferentes contextos.
- La mayoría de estos datos son en inglés, aunque existen también recursos en otros idiomas, y conjuntos multilingüe.

Ejemplos de datos para entrenar ASR

Dataset	Open-Source	Hours	Language
LibriSpeech	Yes	1000	English
HUB 5	No	2000	English
TIMIT	No	5.6	English
The CHiME-5	No	50.12	English
TED-LIUM	Yes	452	English
The Spoken Wikipedia [74]	Yes	1005	Multilingual
Common Voice	Yes	1900	Multilingual
CSTR VCTK [162]	Yes	09	English
AISHELL-1 [15]	Yes	170	Mandarin
Persian Consonant Vowel Combination (PCVC) [95]	Yes	–	Persian
Arabic Speech Corpus [49]	Yes	3.7	Arabic

Referencias

- D. Jurafsky, J. H. Martin, Speech and Language Processing, Prentice-Hall – Caps. 7 y 9
- Russel and Norvig, Cap. 23
- L.E. Sucar, Probabilistic Graphical Models, Springer, Cap. 5
- Rabiner, L.E.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Waibel A., Lee, K. (eds.) Readings in speech recognition, Morgan Kaufmann, 267-296 (1990)
- Kanungo, T.: Hidden Markov Models Software. http://www.kanungo.com/

Referencias

- J. Hinton et al., Deep neural networks for acousting modeling in speech recognition, IEEE SIGNAL PROCESSING MAGAZINE [82] NOVEMBER 2012
- Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, Imran Makhdoom, Automatic speech recognition: a survey, Multimedia Tools and Applications (2021) 80:9411–9457.