
quantil

Reconocimiento de voz (Speech to Text)

Carlos Andrés Reyes

13 de Abril del 2018

Vocabularios grandes y habla continua (LVCSR)

Objetivo: Dado un input acústico O encontrar, dentro de todas las frases pertenecientes a un lenguaje L, la más probable.

- Se divide el input acústico O en intervalos de tiempo (por ejemplo 10 milisegundos) y se trata como una secuencia símbolos individuales:

$$O = o_1, o_2, \ldots, o_t$$

. Cada observación o_1 es un vector de features acústicos, en su mayor parte descomposiciones espectrales.

- De manera similar se trata una frase como una secuencia de palabras:

$$W = w_1, w_2, \ldots, w_n$$

$$\hat{W} = \text{argmax}_{W \in \mathcal{L}} P(W \mid O) = \text{argmax}_{W \in \mathcal{L}} P(O \mid W)\, P(W)$$

Modelos

$P(W)$: Probabilidad previa, modelo de lenguaje.

- Modelos N-gram:

$$P(W) = \prod_{k=1}^{K} P(w_k | w_{k-1}, w_{k-2}, \ldots, w_{k-N+1})$$

$P(O \mid W)$: Verosimilitud de la observación, modelo acústico.

- Modelos ocultos de Markov.
- Modelos acústicos fonemas.

Fonemas y alófonos

- Son sonidos del habla que permiten distinguir palabras en una lengua. Los sonidos [p] y [b] son fonemas del español porque permiten distinguir palabras como pata y bata que tienen significado distinto y solo diferen en su pronunciación con respecto a estos dos sonidos.
- A cada fonema corresponden en el habla diferentes sonidos (alófonos) que varian según el sujeto que los pronuncie.
- El fonema es un modelo para los alófonos que se efectuan en el habla.
- Los fonemas son indivisible, no se dividen en unidades menores.

Fonemas y alófonos

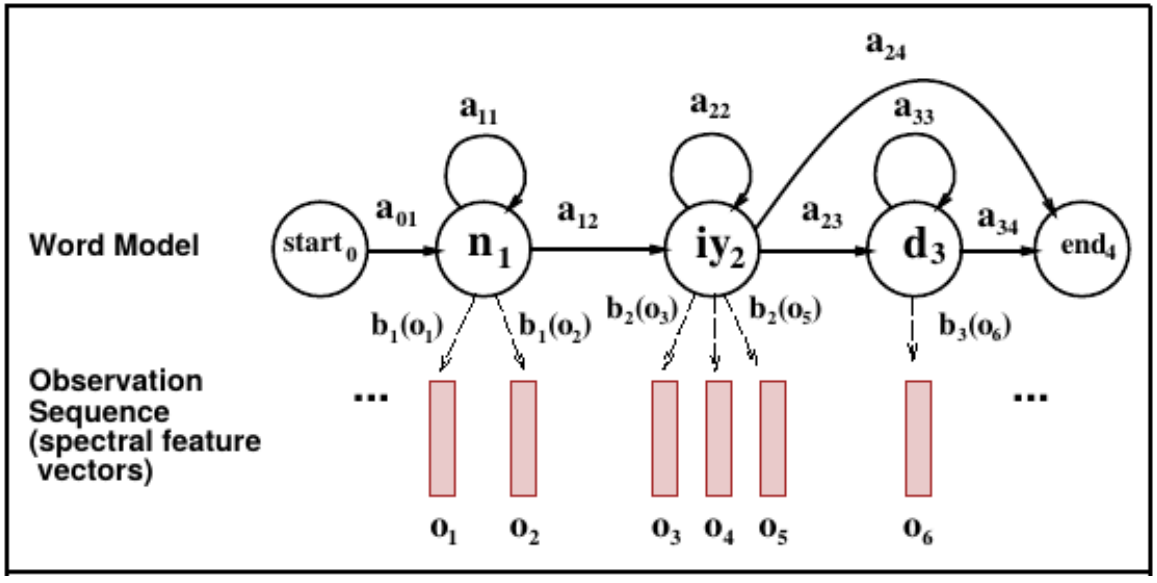
Definición formal: Sea $/f/$ un fonema que puede ser articulado como un conjunto de fonos $\{\phi_1, \phi_2, \ldots, \phi_n\}$ Se define la función $rasg(\cdot)$ que asigna a cada fono o fonema el conjunto de rasgos relevantes. Entonces se tiene:

$$\phi_j \in /f/ \Rightarrow rasg(/f/) \subset rasg(\phi_j)$$

Los rasgos están clasificados en las siguientes categorias

- Principales: silábico, consonático, aproximante, sonorante
- Laringales: Los estados de la glotis para cada sonido, e.g. sonoro, glotis extendida, constricción glotal.
- De modo: Continuante, nasal, estridente, lateral, relación retardada.
- Punto de articulación: Labial, coronal, dorsal, radical, laringal.

Modelo oculto de Markov (HMM)

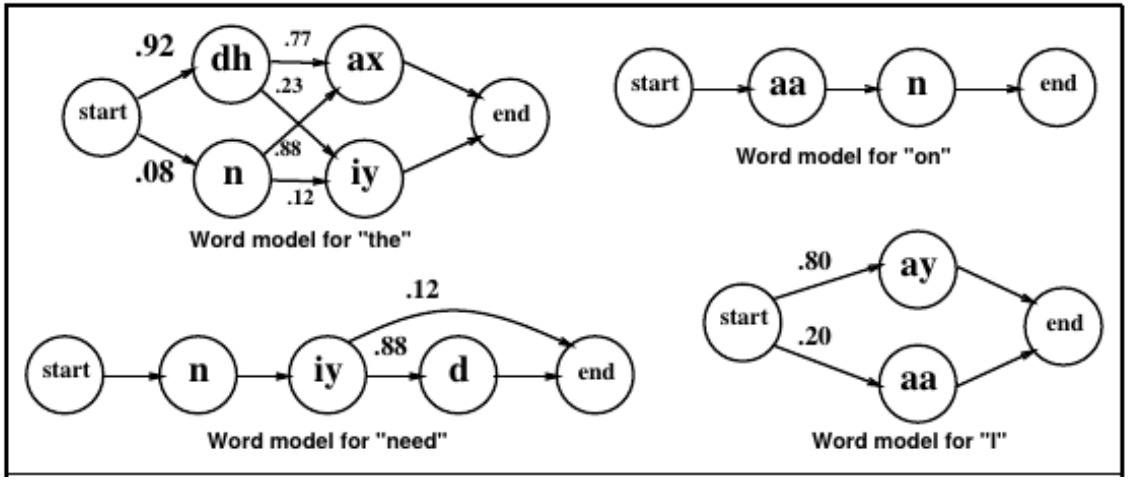


Parámetros que definen un modelo oculto de Markov

- **Estados:** $Q = q_1 q_2 \ldots q_N$
- **Probabilidades de transición:** $P(x_{t+1} = q_i \mid x_t = q_j) = a_{ij}$.
- **Verosimilitud de** o_t**:** $P(o = o_t \mid q = q_i) = b_i(o_t)$.

Encontrando la secuencia más probable

Buscamos la secuencia de estados más probable $q^* = (q_1 q_2 \dots q_T)$ dada una secuencia de observaciones $o = (o_1 o_2 \dots o_3)$.



I need	0.0016	need need	0.000047	# Need	0.000018
I the	0.00018	need the	0.012	# The	0.016
I on	0.000047	need on	0.000047	# On	0.00077
I I	0.039	need I	0.000016	# I	0.079
the need	0.00051	on need	0.000055		
the the	0.0099	on the	0.094		
the on	0.00022	on on	0.0031		
the I	0.00051	on I	0.00085		

Estructura de pronunciación

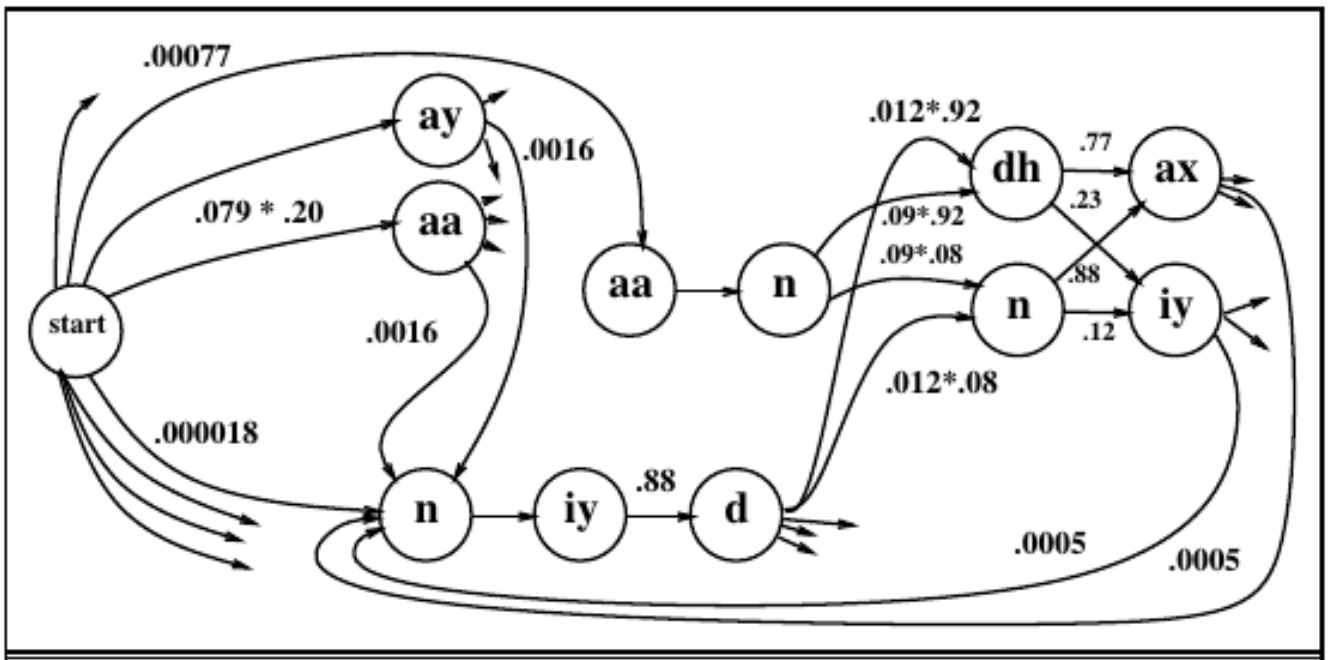


Figure 7.8 Single automaton made from the words *I*, *need*, *on*, and *the*. The arcs between words have probabilities computed from Figure 7.7. For lack of space the figure only shows a few of the between-word arcs.

Encontrando la frase más probable

Algoritmo de Viterbi Se define la matriz de Viterbi:

$$viterbi[t, i] = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1 q_2 \ldots q_{t-1}, q_t = i, o_1, o_2 \ldots o_t \mid \lambda)$$

Se llena la matriz de manera recursiva:

$$viterbi[t, j] = \max_i (viterbi[t-1; i] a_{ij}) b_j(o_t)$$

Se procede al estilo de "Hansel y Gretel" para escoger el camino más probable empezando al final.

Parámetros a entrenar

- Modelo de lenguaje: N-gram $P(w_k|w_{k-1}, w_{k-2}, \ldots, w_{k-N+1})$
- Verosimilitud de las observaciones $b_j(o_t)$
- Probabilidades de transición a_{ij}
- Léxico de pronunciación: Estructara del grafo de Markov

Insumos

- Archivos de audio de habla junto con su transcripción en palabras.
- Un corpus grande de texto para entrenar el modelo de lenguaje, incluyendo las transcripciones.
- Un corpus más pequeño de archivos de audio marcados fonéticamente. (Los subintervalos son anotados manualmente con sus respectivos fonemas.)

Structura gráfica HMM

- Se parte de un diccionario de pronunciación
- Cada fonema del diccionario se mapea a un estado del HMM
- Estimativo inicial de a_ij: Todos los estados son equiprobables.
- Estimativo inicial de $b_j(o_t)$: Se estiman a partir del corpus pequeño con anotación fonética. Los modelos gaussianos son menos sensibles a la inicialización.

Alineamento forzado de Viterbi

- Se toma como input las palabras correctas correspondientes al audio, y los vectores de features del audio.
- Se forza al algoritmo a a pasar por las palabras con las que se ha marcado el audio.
- Se necesita el algoritmo ya que las palabras tienen diferentes pronunciaciones y la duración de cada fonema no es fija.
- El resultado es un set de vectores de features acústicos marcados con los fonemas çorrectos". Estos se utilizan para reajustar los parámetros de $b_j(o_t)$.
- Para las probabilidades de transición a_{ij} se pueden tomar los conteos de las transiciones en el alineamiento forzoso.

Evaluación (Word Error Rate)

$$WER = 100\frac{(Inserciones + Supresiones + Substituciones)}{Total\ de\ Palabras}$$

REF:	i	***	**	UM	the	PHONE	IS	i	LEFT	THE	portable
HYP:	i	GOT	IT	TO	the	*****	FULLEST	i	LOVE	TO	portable
Eval:		I	I	S		D	S		S	S	

En el 2000 los mejores speech to text tenían un WER del 20 %.
Actualmente la mejor evaluación es de Google con 4.9 %.

GRACIAS