Departamento de Computación, FCEyN, UBA

# Procesamiento del Habla

Agustín Gravano

1er Cuatrimestre 2017

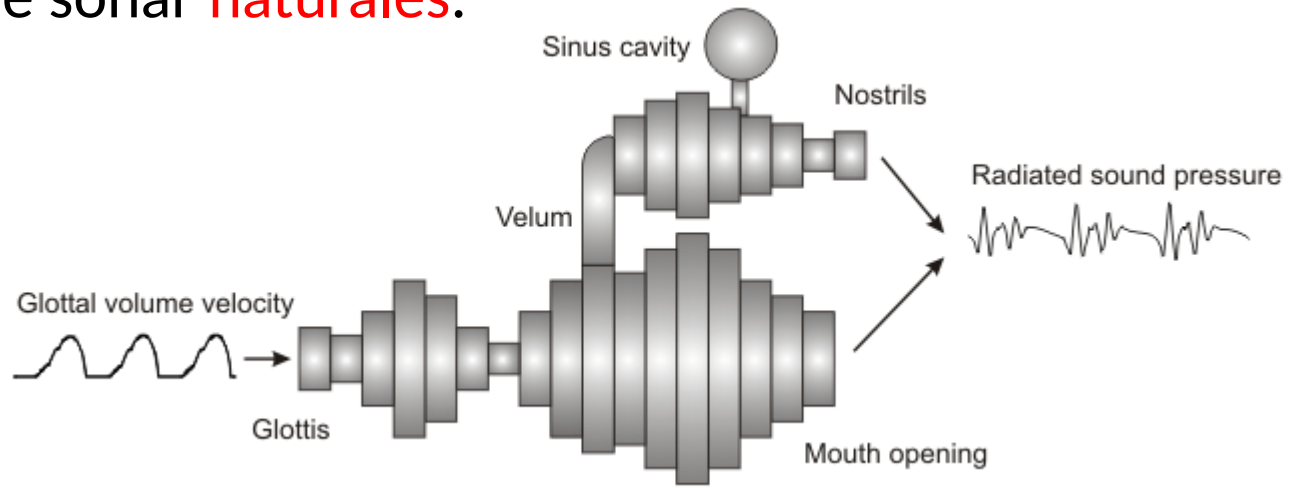
# Síntesis del Habla

# Tipos de Síntesis del Habla

- Síntesis articulatoria.
- Síntesis de formantes.
- Síntesis basada en HMM.
- Síntesis concatenativa.

# Síntesis Articulatoria

- Requiere precisos modelos mecánicos y acústicos de la producción del habla:
  - Vibración de las cuerdas vocales (sonidos sonoros).
  - Aspiración de aire (sonidos sordos).
  - Movimiento de los articuladores (lengua, labios, etc.).
- Cómputos (costosos) de la acústica del tracto vocal.
- Produce resultados inteligibles y con amplio control, pero todavía lejos de sonar naturales.



# Síntesis Articulatoria

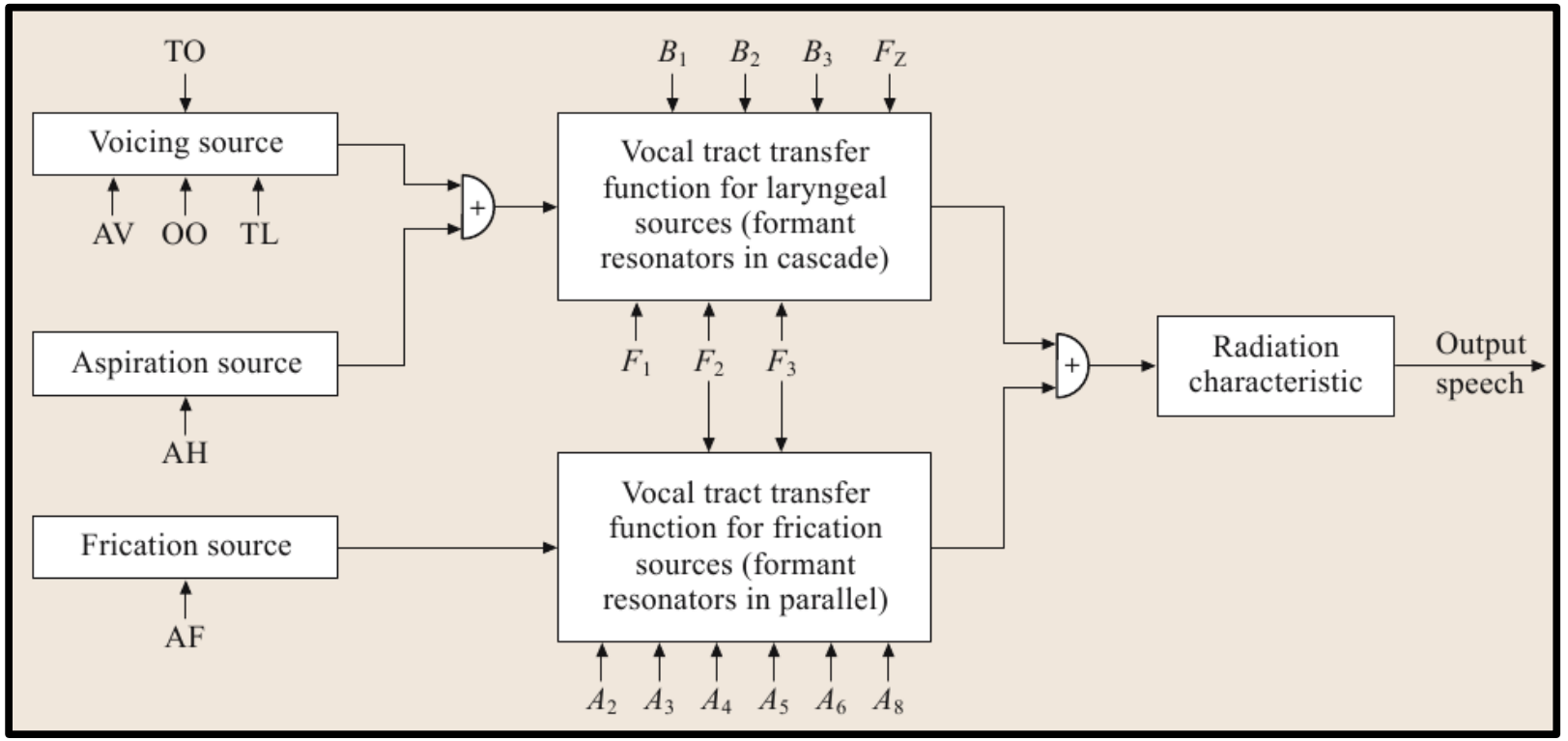
- Virtudes:
  - Control sobre la variabilidad prosódica → Muy útil para generar estímulos en experimentos de percepción (p.ej.).
  - Muy livianos.
  - Costo bajo de cambiar de idioma.

- http://www.vocaltractlab.de
  - tts-vocaltractlab-halt.avi – *"Nächster Halt: Hamburg."*
  - tts-vocaltractlab-zug.avi – *"Der Zug hat eine Stunde Verspätung."*

- Laboratorio de Sistemas Dinámicos, Depto. de Física
  - http://www.lsd.df.uba.ar

# Síntesis de Formantes

- Síntesis *paramétrica*, *acústica* o *de resonancia*.
- Encapsula el tracto vocal como una caja negra.
- Busca reproducir sus características de input/output.
- A una fuente de sonido le aplica una combinación de filtros.



# Síntesis de Formantes

- Parametric Artificial Talker, por Walter Lawrence (1953).
  - tts-pat.wav - *"What did you say before that? Tea or coffee? What have you done with it?"*
- Demos:
  - http://www.speech.kth.se/wavesurfer/formant/
  - http://www.asel.udel.edu/speech/tutorials/synthesis/Klatt.html
    Ejemplo: tts-klaat.wav, generado con estos comandos:

```
TIME=0;   AV=80; F1=500; F2=1500; F3=2250; F0=120
TIME=30;  AV=80; F1=650; F2=1700; F3=2500
TIME=50;  F2=1775
TIME=250; F0=90
TIME=500; F0=130; F1=800; F2=1900; F3=3000
END
```

- Virtudes = síntesis articulatoria (control, liviano, idioma).

# Ejercicio: Síntesis de Formantes con Klatt

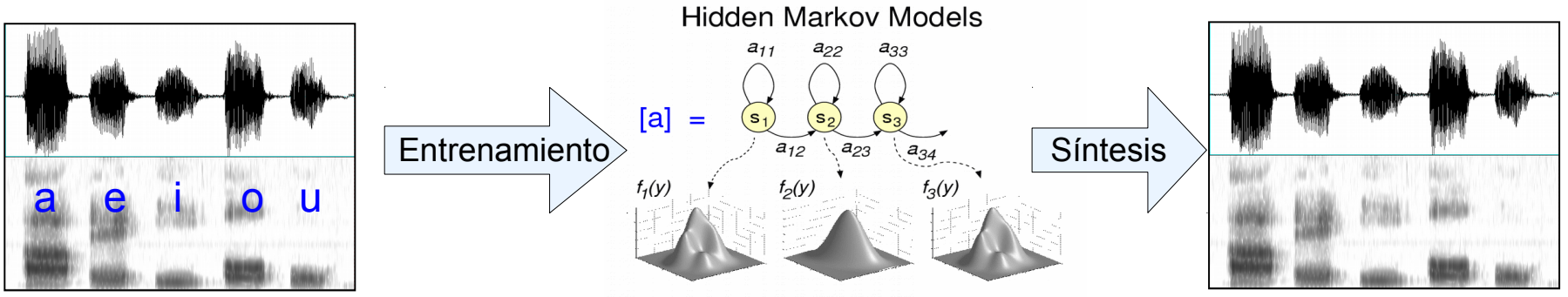
- Generar los siguientes sonidos usando la interfaz web del sintetizador de formantes Klatt (http://www.asel.udel.edu/speech/tutorials/synthesis/Klatt.html):
  - Las cinco vocales: /a/ /e/ /i/ /o/ /u/.
    - Con **Praat**, medir en aeiou.wav los formantes de cada vocal.

      *View & Edit*; Seleccionar vocal; *Formant > Get first/second/third formant*
    - Por ejemplo, este código genera una /a/:

      ```
      TIME=0; AV=80; F1=725; F2=1285; F3=2485
      TIME+500
      END
      ```
  - Diptongos (/iu/ /ui/ /ai/ /ia/, etc.).
  - La vocal /a/, ahora con tres contornos de entonación distintos (ej: ascendente, sostenido, descendente). Para esto, modificar F0 (frecuencia fundamental).
- Para aprender a manejar Klatt, leer las instrucciones de la página de arriba.
- Ayuda: para estos ejercicios no hace falta usar más variables que F0, F1, F2, F3, AV y TIME. Igual, experimenten con otras variables para ver qué efecto tienen.

# Síntesis Basada en HMM

- Hidden Markov Model (HMM+GMM)
  - Modelo probabilístico generativo usado en reconocimiento.
- Fono ↔ Patrón espectral.
- Reconocer un fono == Reconocer su patrón espectral.
- ¿Cómo usar HMM+GMM para síntesis?
  - Para sintetizar una secuencia de fonos [*xyz*], producimos el espectrograma de [*xyz*] según el HMM+GMM.




  - En la clase de Reconocimiento vamos a ver HMMs en mayor detalle.

# Síntesis Concatenativa

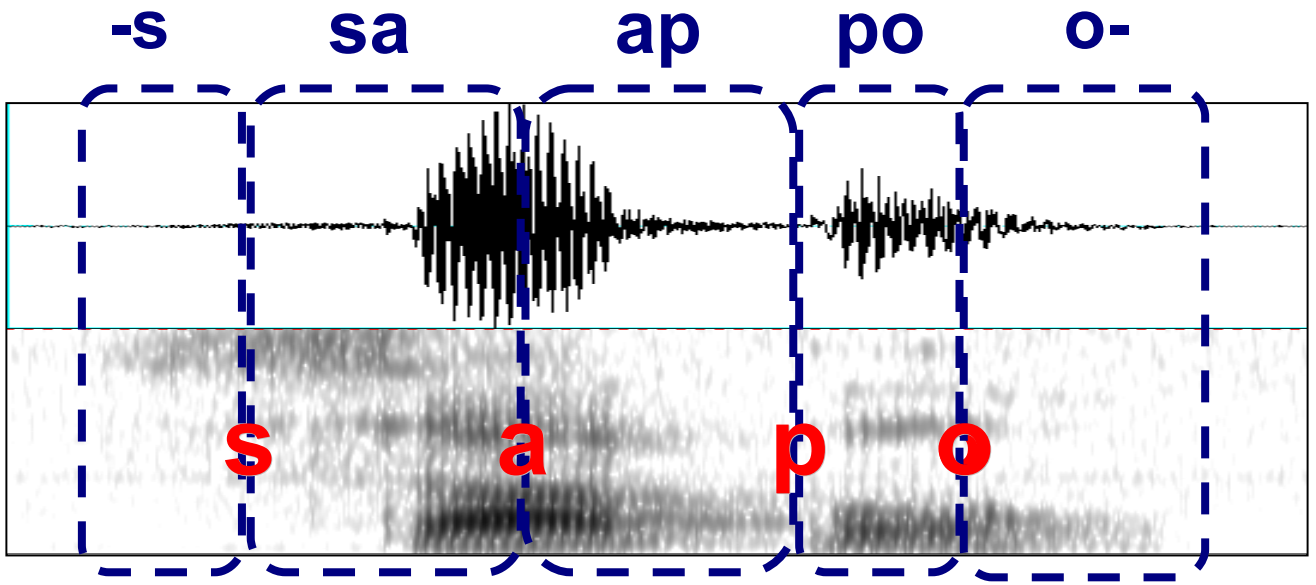
- 1936 – *Speaking Clock* del Reino Unido: tts-clock.wav
  *"At the third stroke, it will be eight fifty-seven, precisely."*
- 1981 – Hora oficial de la Argentina: tts-113.wav
- Tiene infinidad de usos en dominios específicos:
  *"El vuelo número N N N N de AEROLINEA con destino a CIUDAD se encuentra ESTADO."*
- Muy barato, fácil de implementar, de rápido desarrollo.

# Síntesis Concatenativa

1) Construir una base de datos de habla.
   a) Grabar a una persona diciendo oraciones preparadas.
   b) Recortar unidades de habla (frases, palabras, sílabas, sonidos, etc.).
2) Para cada nueva oración a sintetizar:
   a) Elegir las unidades de la base de datos.
   b) Concatenarlas.
   c) [Procesar el resultado para mejorar la calidad.]

# Difonos

- Para sintetizar cualquier oración (dominio abierto), necesitamos usar **unidades más chicas**.
- Difono: Desde la región estable de un fono hasta la región estable del siguiente fono.




- Los difonos capturan la coarticulación (influencia de un fono sobre sus vecinos) y evitan saltos en la señal.

# Síntesis Concatenativa

- Dominio abierto (sintetizar cualquier oración):
  - Unidad = difono (desde el medio de un fono hasta el medio del siguiente)
- Síntesis de difonos:
  - Guardar **una sola instancia** de cada difono.
  - **Modificar** la secuencia para cambiar la prosodia.
- Selección de unidades:
  - Guardar **varias instancias** de cada difono.
  - **Elegir** difonos cercanos a la prosodia deseada.

# Inventario de difonos

- Cantidad de difonos = O(fonos²)
- No todos los difonos son válidos en un lenguaje.
  - Restricciones fonotácticas.
  - Ejemplos en español: /pf/ /kg/ /pp/ ...
- Ejemplo:
  - Sistema en inglés (AT&T, Olive et al. 1998).
  - 43 fonos.
    - $43^2$ = 1849 difonos posibles.
    - Sólo 1162 difonos válidos.

# Frases, grabación y segmentación

- Tono, intensidad y duración constantes.
- Frases portadoras: aportan consistencia.

/ba/ → #tabama#

/pa/ → #tapama#

/sa/ → #tasama#

/-a/ → #ama#

/-e/ → #ema#

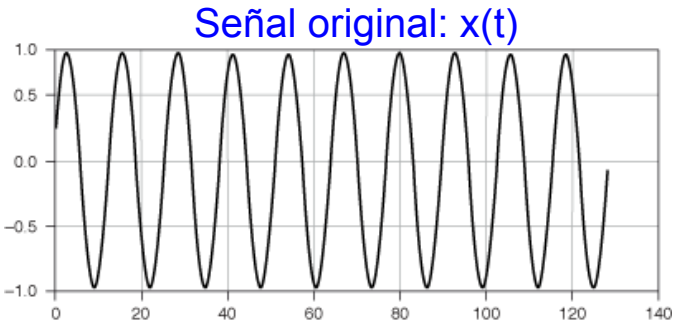
/-o/ → #oma#

# = silencio

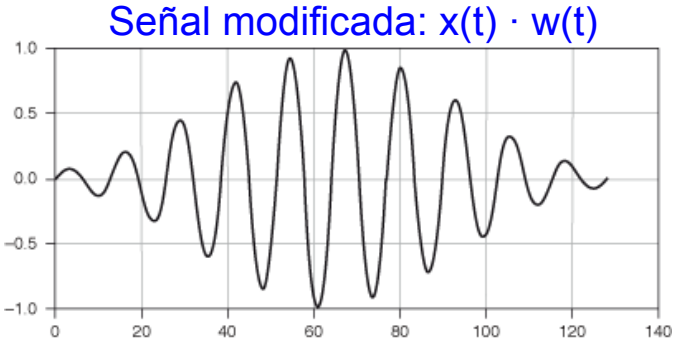
- Antes de grabar cada frase, escuchar una frase fija.
- Segmentación semi-automática de difonos.
  1) Alineación forzada con sistema de reconocimiento de habla.
  2) Corrección manual.

# Concatenación de difonos

- Concatenar difonos puede generar ruidos (*clicks*) causados por 3 tipos de discontinuidad:
  - De fase.
  - De tono (f0).
  - De espectro.
- *Smoothing* (suavizado) de las uniones.
  - *Hanning windowing*.







# Modificación de la prosodia

- Todos los difonos tienen la misma prosodia (f0, int, dur).
- La prosodia deseada se consigue con proc. de señales.
- La intensidad se puede modificar fácilmente.
- ¿Cómo modificar tono y duración?
  - Aumentar la duración de una señal disminuye el tono.
- TD-PSOLA:
  - *Time-Domain Pitch-Synchronous Overlap-and-Add*
  - Identificar ciclos básicos de la señal.
  - Para cambiar la duración: duplicar/borrar ciclos.
  - Para cambiar el tono: juntar o separar ciclos.
- Ver clase de Prosodia.

# Síntesis de difonos

- Base de datos pequeña: ~8MB para inglés (16Hz, 16 bits).
- Modificación de la prosodia con procesamiento de señales (e.g. TD-PSOLA): resultados poco naturales.
- Idea alternativa:
  - Guardar muchas instancia de cada difono; al sintetizar, elegir la instancia con las características prosódicas más parecidas a las deseadas.
  - Eso se conoce como síntesis con selección de unidades.

# Síntesis con selección de unidades

- Las frases a grabar deben contener múltiples instancias de cada difono.
  - Respetar la distribución de frecuencias del lenguaje.
  - En español, muchas instancias de */la/*; pocas de */pt/*.
- Varias horas de grabación.
- Segmentación semi-automática de fonemas.
  1) Alineación forzada con sistema de reconocimiento.
  2) Corrección manual.

# Síntesis con selección de unidades

- Síntesis = Encontrar en la base de datos la secuencia de difonos que mejor cumpla la especificación dada.
- ¿Qué significa “mejor”?
  - Costo del objetivo (*T*): Cuán bien respetan los difonos las características deseadas (prosodia, contexto, etc.).
  - Costo de unión (*J*): Cuán bien se concatenan los difonos adyacentes.

$$\hat{U} = \underset{U}{\operatorname{argmin}} \sum_i T(s_i, u_i) + \sum_i J(u_i, u_{i+1})$$

donde $s_i$ : especificación de la i-ésima unidad a sintetizar
$u_i$ : unidad de la base de datos

# Costo del objetivo $T(s, u)$

- ¿Cuánto se parecen la unidad *u* (de la base de datos) y la especificación objetivo *s*?
- Ejemplos de especificaciones de difonos:

```
/-t/, acentuado, principio frase, F0 alto, adverbio, ...
/la/, no acent., medio frase, F0 medio, artículo, ...
```

- Costo objetivo *T* = suma de *P* subcostos $T_p$:
  - Acentuación, posición en la frase y en la palabra, F0, duración, intensidad, POS de la palabra, etc.
  - Cada subcosto tiene un peso $w_p$

$$T(s,u) = \sum_{p=1}^{P} w_p \cdot T_p(s,u)$$

  - ¿Cómo se determinan los pesos $w_p$?

# Costo de unión $J(u_i, u_{i+1})$

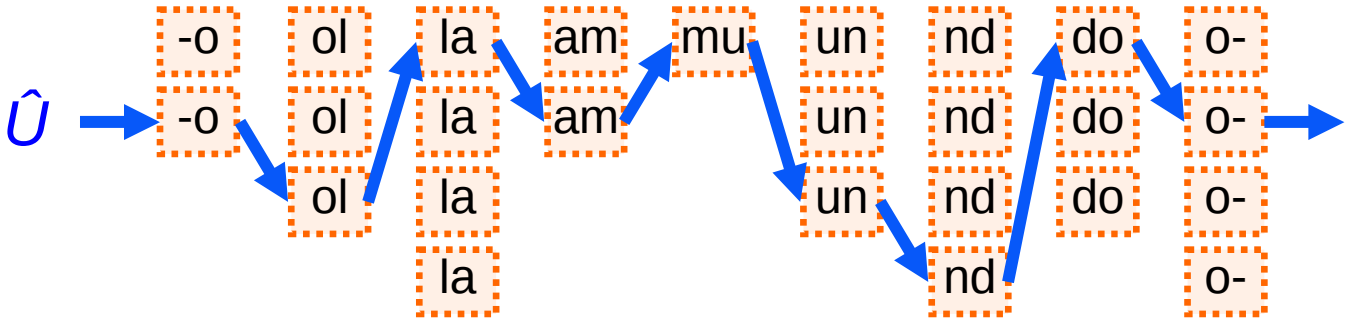
- ¿Cuán suave es la concatenación de dos unidades $u_i$ y $u_{i+1}$ (ambas de la base de datos)?
- Costo de unión $J$ = suma de $P$ subcostos $J_p$:
  - Intensidad, F0, atributos espectrales, etc.
  - Cada subcosto tiene un peso $w_p$.

$$J(u_i, u_{i+1}) = \sum_{p=1}^{P} w_p \cdot J_p(u_i, u_{i+1})$$

# Síntesis con selección de unidades

*S* = **-o ol la am mu un nd do o-**
(con sus especificaciones prosódicas y contextuales)

**Base de datos de difonos:**




$$\hat{U} = \underset{U}{\operatorname{argmin}} \sum_i T(s_i, u_i) + \sum_i J(u_i, u_{i+1})$$

# Síntesis con selección de unidades

- Algoritmo de Viterbi
  - Programación dinámica.
  - Encuentra en forma eficiente y exacta el camino con el costo más bajo.
  - Complejidad: $O(M * N^2)$.
    - $M$ = Longitud de la secuencia objetivo.
    - $N$ = Número de difonos en la base de datos.

# Síntesis con selección de unidades

- Resultados más naturales que con otras técnicas.
- Base de datos muy grande: O(GB).
- Búsqueda de difonos: cara computacionalmente.
  - Cuadrática en el tamaño de la base de datos!
  - Técnicas de optimización: e.g. clustering de difonos.
- La calidad puede ser muy mala cuando no hay buenos candidatos en la base de datos.
  - Problema: en interacciones humano-máquina suele ser muy molesto mezclar cosas muy buenas y muy malas.

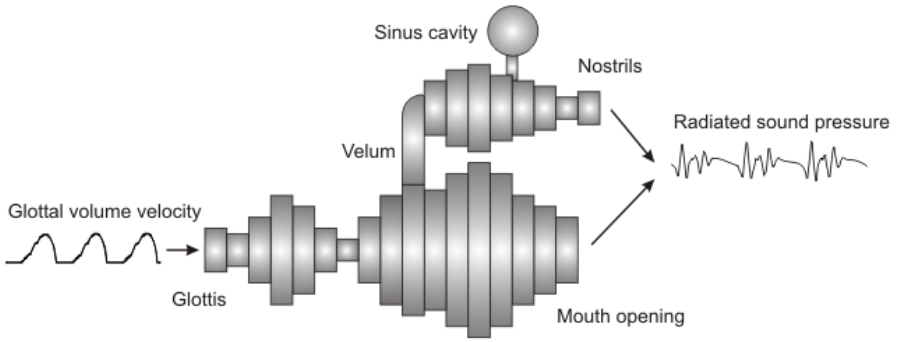
# Herramientas y Demos de TTS

- Herramientas para desarrollo:
  - Festival http://www.cstr.ed.ac.uk/projects/festival/
  - Mary TTS http://mary.dfki.de/

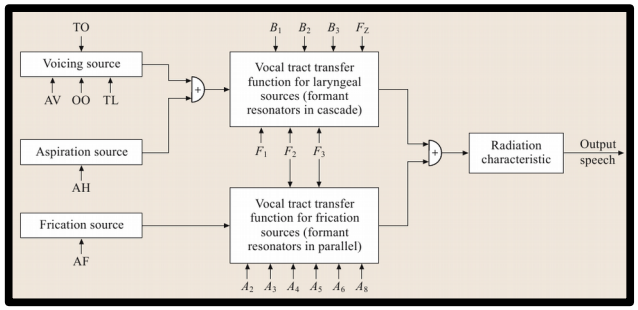
- Demos comerciales:
  - IBM BlueMix https://text-to-speech-demo.mybluemix.net
  - Cepstral http://www.cepstral.com/en/demos
  - LumenVox http://www.lumenvox.com/products/tts/
  - Google Chrome https://developer.chrome.com/extensions/examples/extensions/ttsdemo/ttsdemo.html

# Síntesis del Habla – Resumen

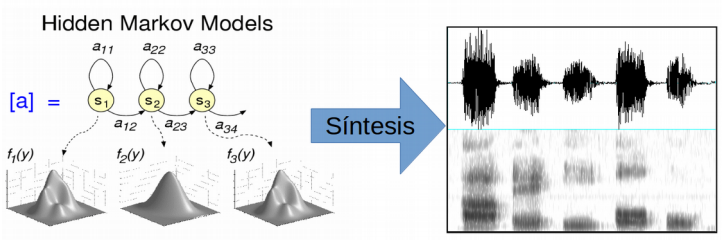
- Síntesis articulatoria:
  - Modelos del tracto vocal.




- Síntesis de formantes:
  - Fuentes y filtros.




- Síntesis basada en HMM.




- Síntesis concatenativa:
  - Difonos, TD-PSOLA, selección de unidades.