Facultad de Ingeniería
UNAM

Síntesis de Voz Utilizando Modelos Ocultos de Markov

DR. ABEL HERRERA CAMACHO

Laboratory
Tecnologías del Lenguaje

COLOQUIO DE LINGÜÍSTICA COMPUTACIONAL, CU, 2015

Outline

Speech Processing.

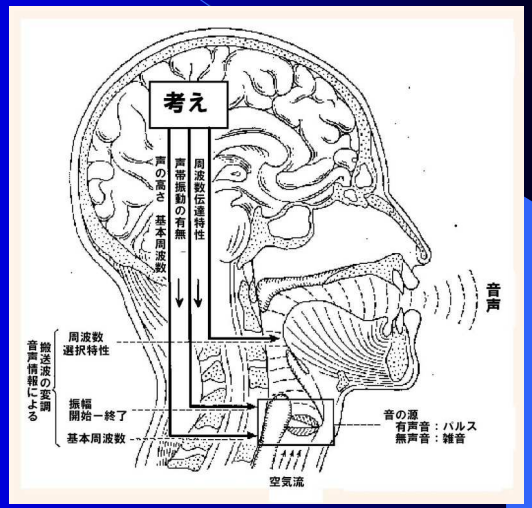
Speech Synthesis

HMM

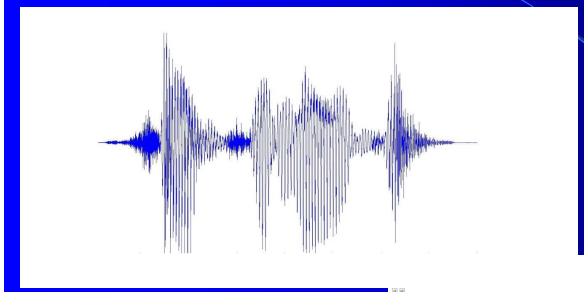
HTS Synthesizer.

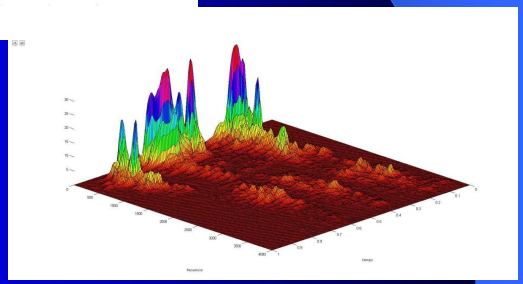
1. Speech Processing

F. Itakura, "Fundamentals of speech analysis and synthesis and its application to speech coding,"
IEICE FM06-2-1, July 2006

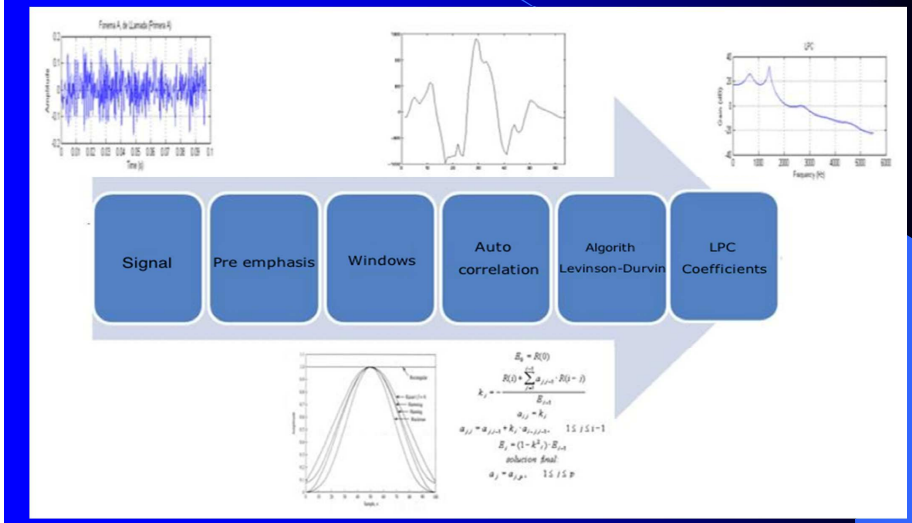


1. Speech Processing

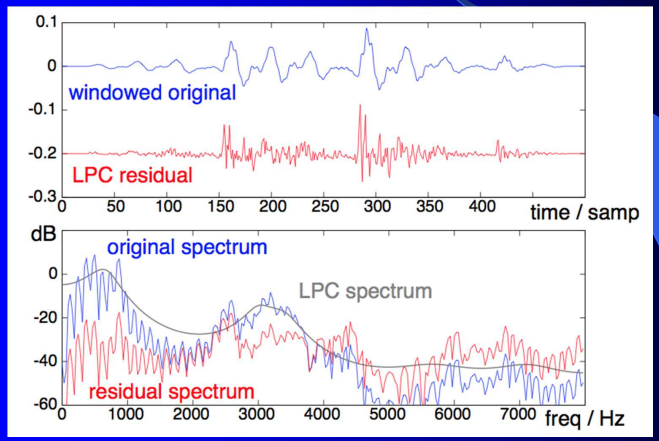




1. Speech Processing



1. Speech Processing



2. Speech Synthesis

- **Speech synthesis, TTS, is a process to generate speech from a machine without having all the phrases precorded.**
- **The speech synthesis is embedeed in the man-machine interaction.**
- **There are now synthesis applications, without other speech applications..**

2. Speech Synthesis

- **The first synthesizers were designed at they second half of the XX century.**
- **The strategy more used is the concatenation of acoustic units.**
- **The synthesizers have two parts:**
  - **Text analizer**
  - **Phonemes synthesizer**

2. Speech Synthesis

Text Analizer

0	900000	x^x-#+d=o1@1_0
900000	2074624	x^#-d+o1=s@1_3
2074624	2875493	#^d-o1+s=p@2_2
2875493	3318271	d^o1-s+p=a@3_1
3318271	3921548	o1^s-p+a=l@1_2
3921548	4373422	s^p-a+l=a1@2_1
4373422	4837008	p^a-l+a1=b@1_2
4837008	5518954	a^l-a1+b=r@2_1
5518954	6131516	l^a1-b+r=a@1_4
6131516	6413402	a1^b-r+a=s@2_3
6413402	7036105	b^r-a+s=e@3_2
7036105	7902170	r^a-s+e=n@4_1
7902170	8382171	a^s-e+n=u@1_2

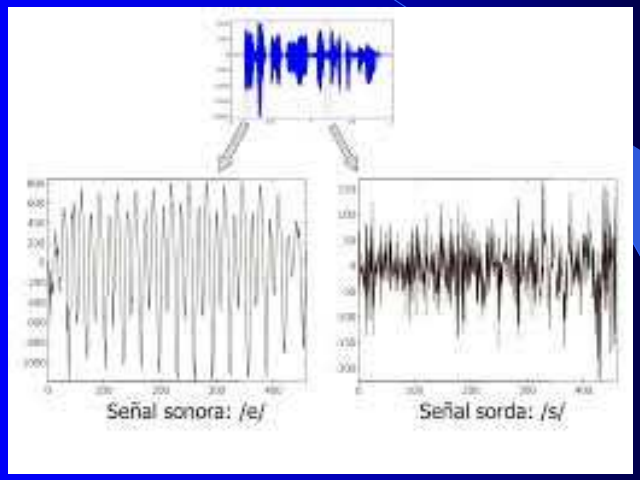
2. Speech Synthesis

Unit Sinthesizer

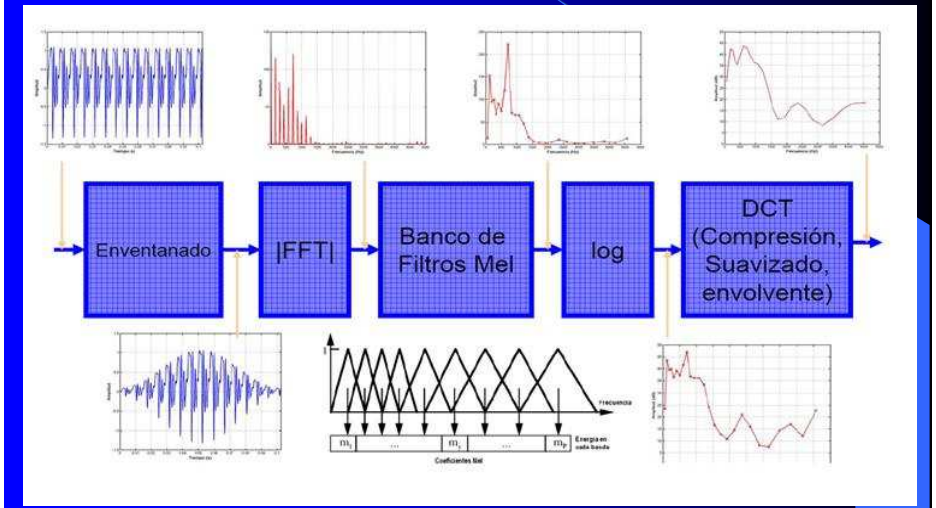
- **Unit Concatenation: phonemes or diphonemes**
- **Each information unit has the following minimum parameters:**

➢ **MFCC,**
➢ **F0**
➢ **Unit time duration**

3. HTS Synthesizer

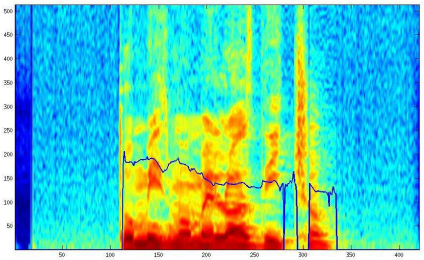


2. Speech Synthesis

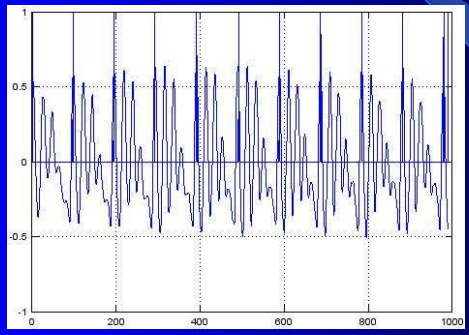


2. Speech Synthesis

- The best algorithm to obtain F0 is by Goncharoff and Gries.
- The goal is to obtain the discontinuity of F0 between segments.



2. Speech Synthesis



2. Speech Synthesis

- La concatenación de segmentos produce una voz de baja calidad, inteligible, monótona y no natural.
- Sintetizador por difonemas: usa TD-PSOLA, pero aún no es natural
- Festival, utiliza difonemas y árboles determinísticos que incorporan la prosodia, se aproxima a voz natural.

(see examples)

3. Sintetizador hts

HMM: $\lambda=(A,B,\pi)$

Continua.

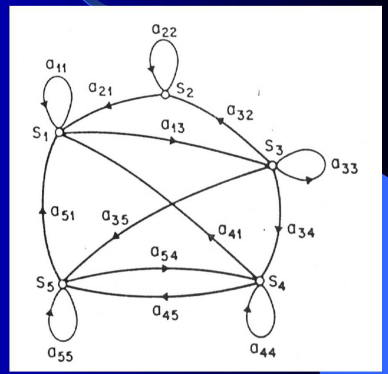
Topologías.

Duración de estados.

Dos problemas:

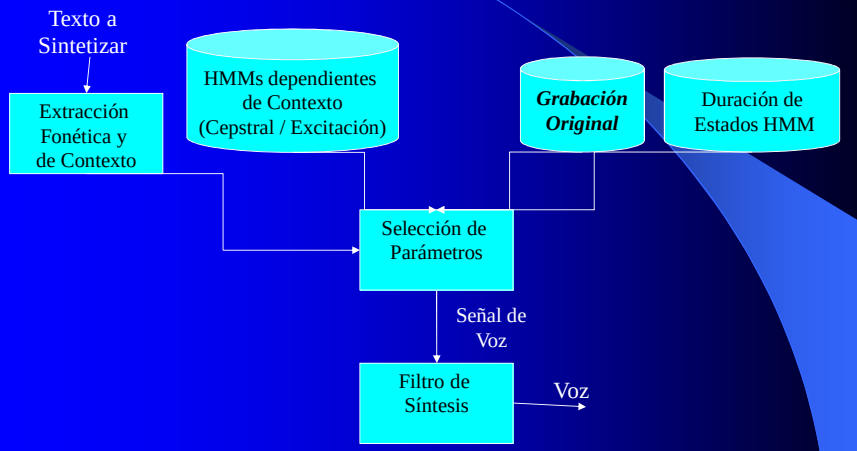
Obtención de parámetros

Obtención de trayectoria óptima



3. Sintetizador hts

3. HTS Synthesizer



3. HTS Synthesizer

- The HTK tools provide the HMM's training, generating phone clusters, where each cluster has very similar parameters.
- HTS uses these parameters (MelCepstral, F0 and duration) independently.
- HTS generates a selection tree that allows selecting the most likely set of parameters for each phone based on it's contexts.

3. HTS Synthesizer

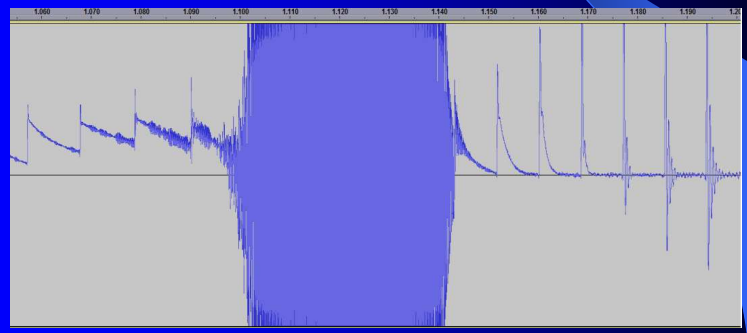
- HTK's parameters allow easy modification of the parameterization of the audio data.
- States: Each phone is split into a selected number of states. Each segment is analyzed and clustered independently.
- Frequency Warping: This parameter allows the use of Cepstral or Mel-Cepstral parameters.

3. HTS Synthesizer

- Gain: Log gain provides slightly better results, but at the cost of a much higher training time.
- Gamma: This parameter affects the reconstruction filter parameters placement of poles and zeroes. Best results were obtained using -1/3, increasing the number of cepstral parameters can render filter unstable.

3. HTS Synthesizer

Signal with 36 coefficients and Gamma=-1/3.



3. HTS Synthesizer

- Number of Cepstral Coefficients:With 12 coefficients the reconstructed signal is too distorted.
- With more coefficients the reconstructed signal is clearer, but with a high number of coefficients the filter can become unstable.
- Number of states per phone: No noticeable improvement was found above 5 states per phone.

3. HTS Synthesizer

- HMM techniques similar to the ones used in HTS have also been adapted into other synthetizers.
- For example, Festival uses HMMs to align a text transcription to the recorded data to extract the position of each phone or diphone.
- There is no smoothing at concatenation points between different segments.
- This technique is applied on synthesis modules such as Clustergen and Multisyn.
- (see examples)

Conclusiones

- HTS voices were created with both professional and non-professional speakers. Though we obtained good results from non-professional recordings, the professional recording generate more natural sounding voices and with better prosody.
- Due to the use of a impulse / noise filter, some buzziness is still present in the generated voices, though it is minimized when increasing Mel Cepstral Parameters.
- Other techniques might reduce this buzziness further, but at the cost of greatly increased processing and storage requirements

¡ Muchas gracias !

Dr. Abel Herrera Camacho
MsS. Fernando del Río

abelhc@hotmail.com
hitosan@Hotmail.com