Technion - Israel Institute of Technology
Dept. of Elec. Eng.

Signal and Image Processing Lab

winter 1994/5

Voice Conversion
using a
Glottal Excited Speech Model

Gilad Cohen*

*M. Sc. Thesis supervised by Prof. David Malah

Voice Conversion:

A technique to change or modify speaker individuality, i.e. , convert the speech of one speaker so that it sounds like that of another.

Application:

- Entertainment (Cartoon character voices)
- Providing speaker individuality to Synthesis-by-rule speech.
- Improve intelligibility of abnormal speech.
- Improving speech recognition systems trained on a “standard speaker”.

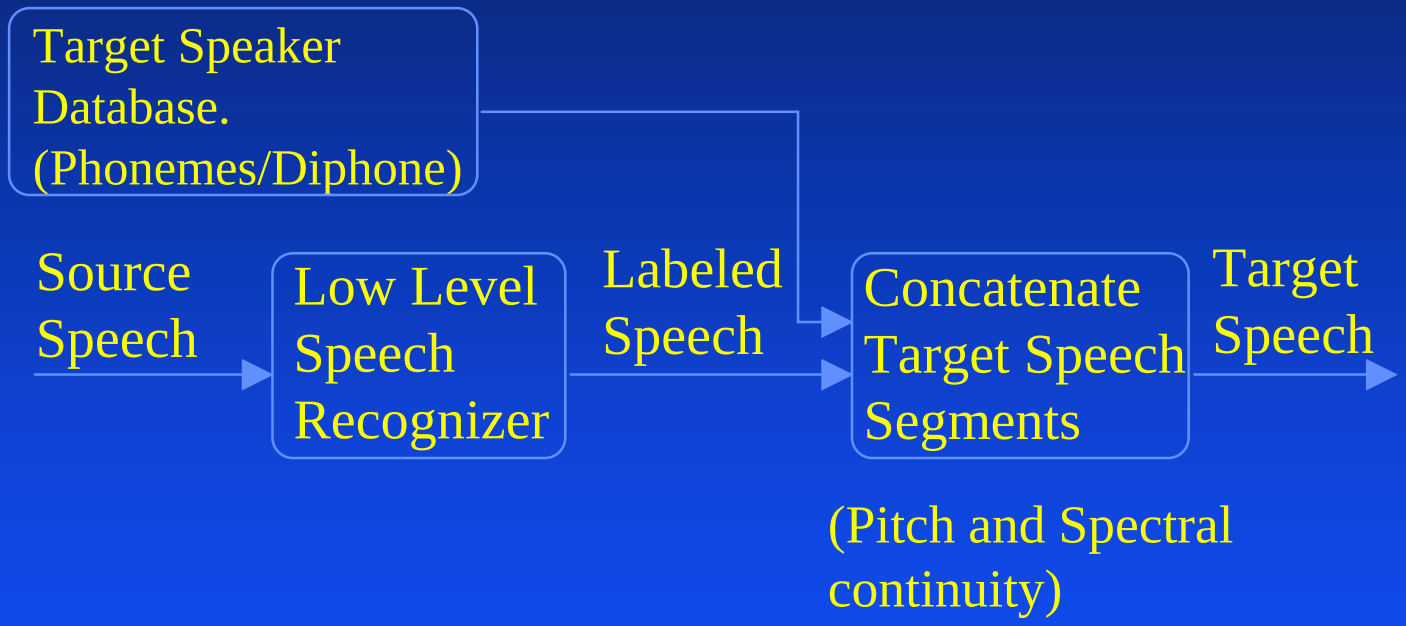
What distinguishes a speaker ?

- Factors related to physiology:
  - Acoustical characteristics of the glottal excitation.
  - Dimensions of the vocal tract.

- Factors related to the dynamics of speech:
  - Speaking rate.
  - Regional accent.
  - inflection (An alteration of pitch or tone).

Methods:

- Synthesis by concatenating small speech segments.

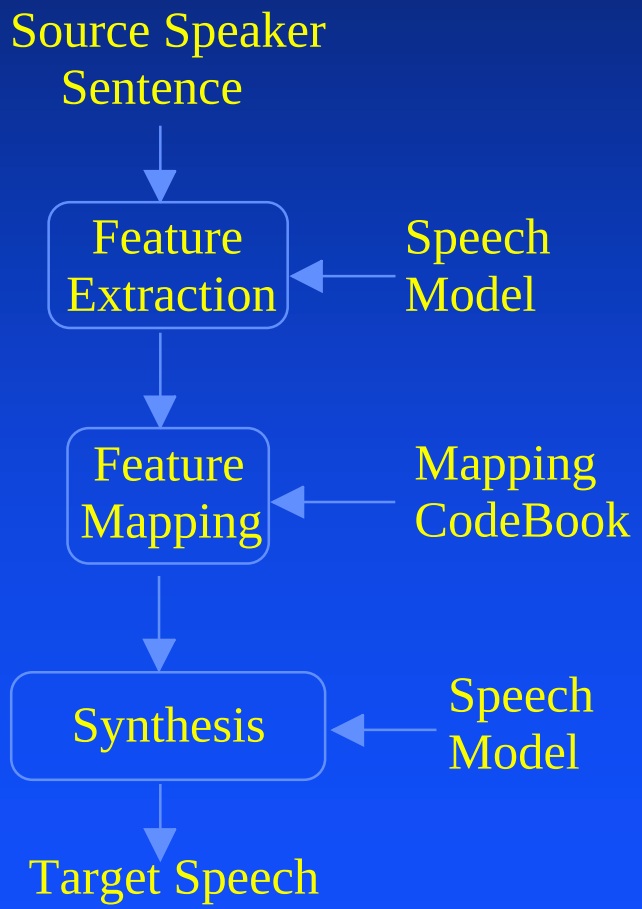


Methods (Cont'd):

- Using Speech Models.

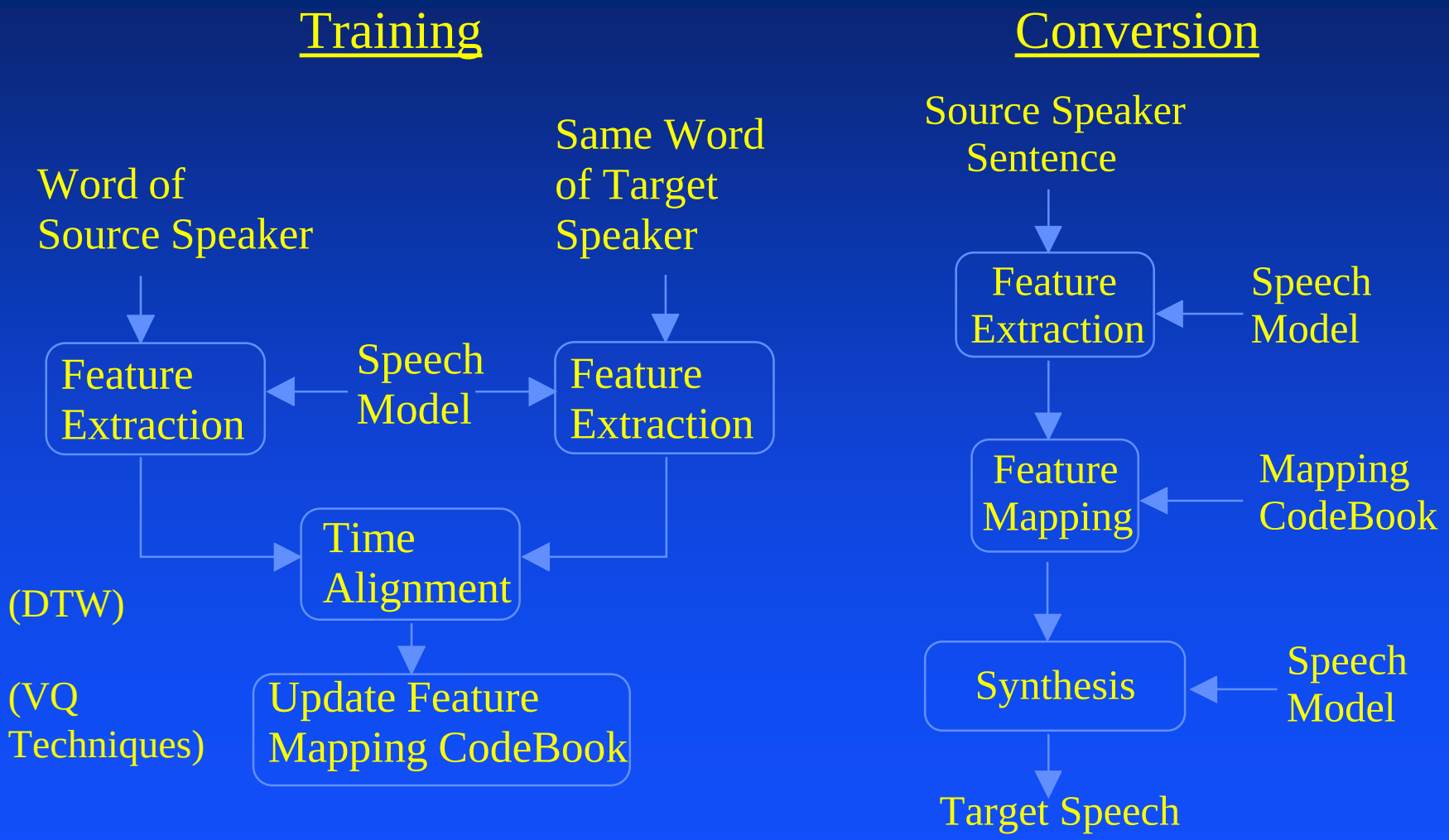
Training

Conversion



Methods (Cont'd):

- Using Speech Models.



Previous work: Using LPC VoCoder (noise/impulse train excitation). Very limited speech quality.

Current Model: Glottal Excited LPC Model

Significantly improves speech quality at low cost.

- Unvoiced sections (fixed length frames)

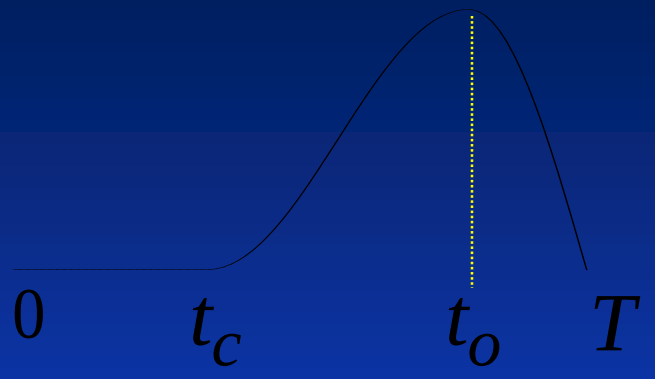
$$s(n) = e(n) * v(n)$$

- Voiced sections (pitch synchronous frame length)

$$s(n) = g(n) * v(n) * r(n)$$

- s -speech signal
- g- glottal pulse.
- v - vocal tract impulse response.
- r - lip radiation $\left(1 - \mu z^{-1}\right)$.
- e - white noise.

- Glottal Air Flow Model



$$g(t)=\begin{cases}0 & ; \quad 0\le t<t_c\\ \sin^2\left(\frac{\pi}{2}\frac{t-t_c}{t_o-t_c}\right) & ; \quad t_c\le t<t_o\\ \cos\left(\frac{\pi}{2}\frac{t-t_o}{T-t_o}\right) & ; \quad t_o\le t<T\end{cases}$$

3 timing parameters: $\{T, t_0/T, t_c/T\}$

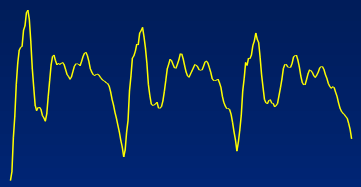
- Vocal Tract Model

$$V(z)=\frac{G}{1-\sum_{i=1}^{p} a_i z^{-i}}$$

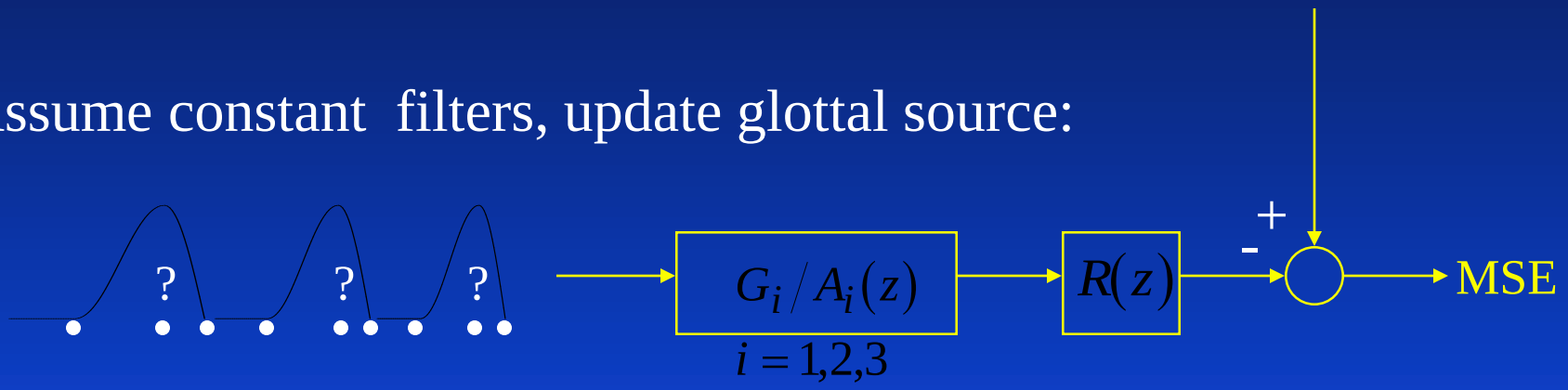
G - gain, $\{a_i\}$ - LPC parameters (p=10).

Analysis Stage - Voiced Sections

- For each 3-Pitch length window:

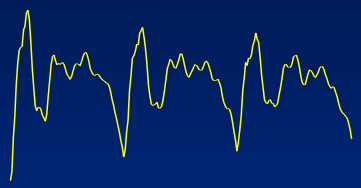


Assume constant filters, update glottal source:

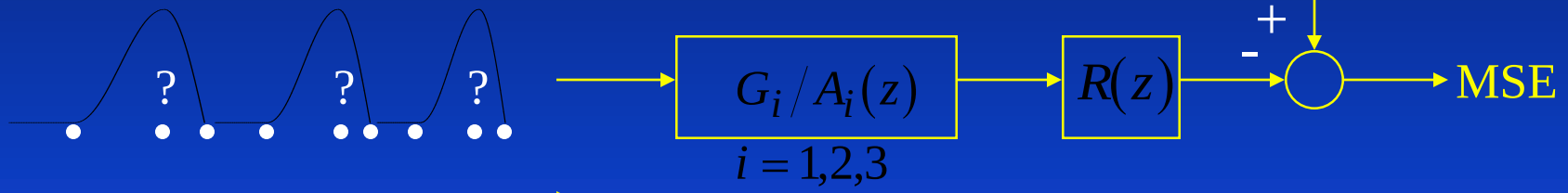


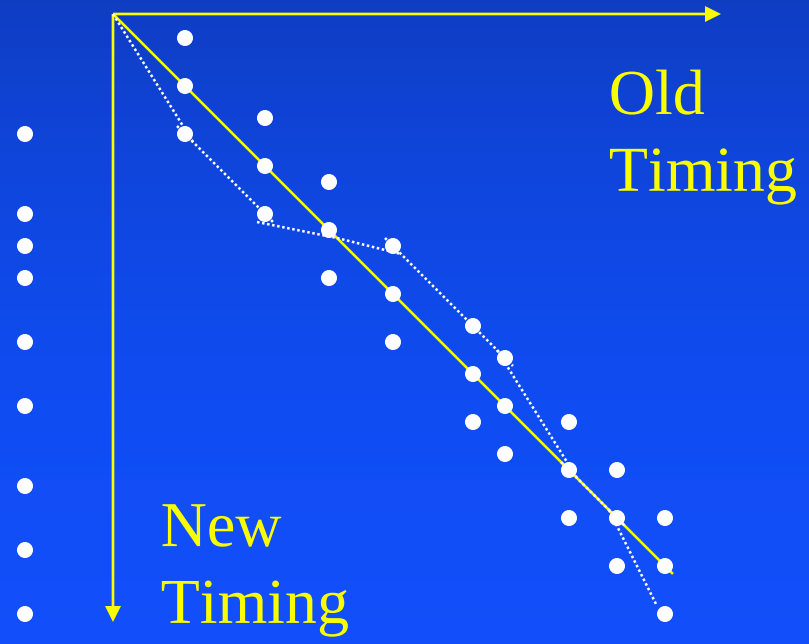
Analysis Stage - Voiced Sections

- For each 3-Pitch length window:



Assume constant filters, update glottal source:



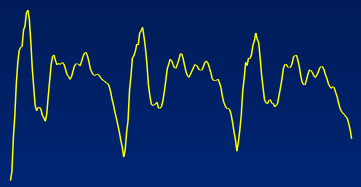


Find a path trough the lattice that gives best synthesis.

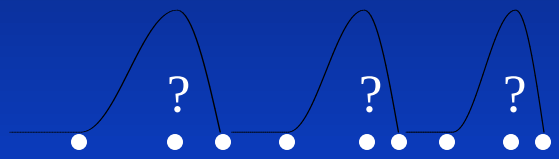
$\Rightarrow$ DTW problem with non-local cost.

Analysis Stage - Voiced Sections

- For each 3-Pitch length window:



Assume constant filters, update glottal source:



$G_i / A_i(z)$

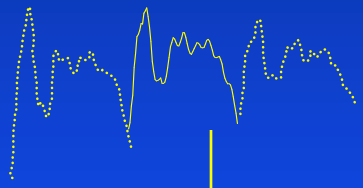
$i = 1,2,3$

$R(z)$

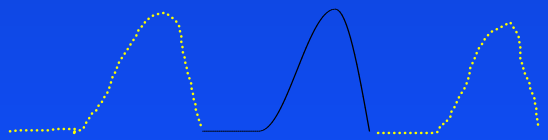
+

-

MSE



Assume constant glottal source, update filters:



$G_i / A_i(z) = ?$

$i = 1,2,3$

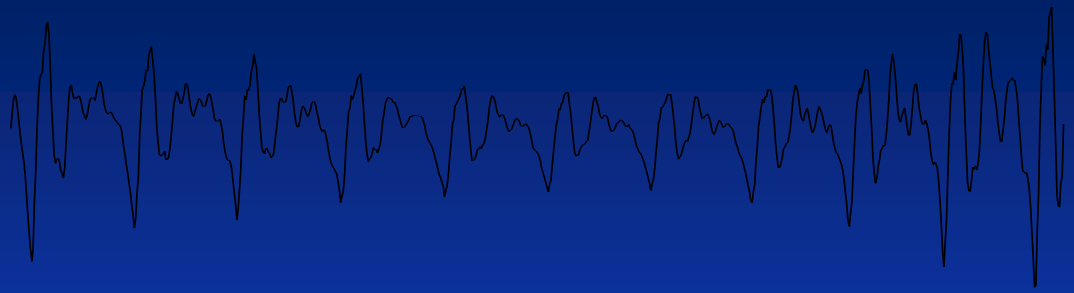
$R(z)$

+

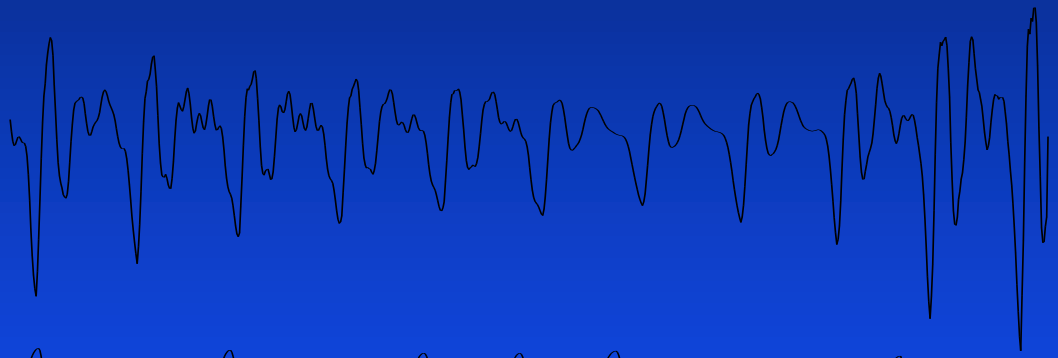
-

MSE

Up to now...

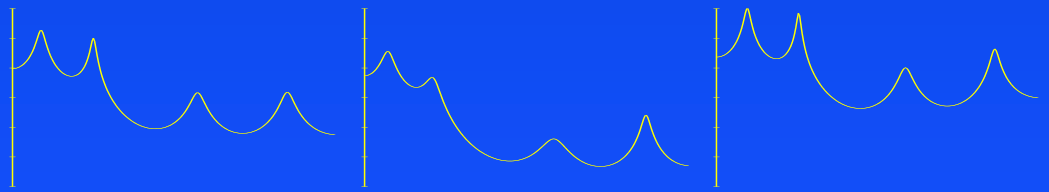


Original



Reconstructed

Glottal Excitation



LPC Spectrum