Lecture 17: Transformation of formants for voice conversion using artificial neural networks

Mark Hasegawa-Johnson

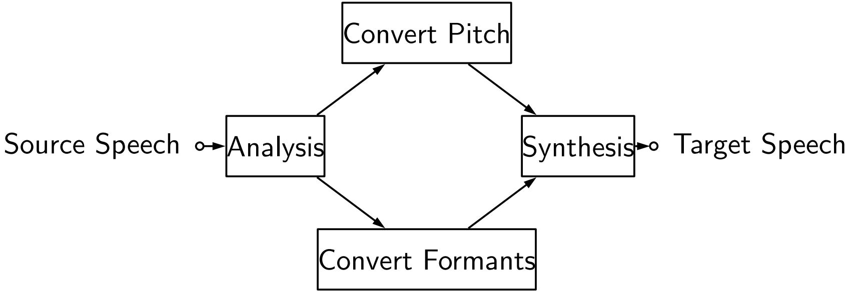
ECE 537, Fall 2022

1. Voice Conversion
2. Formant Synthesis: Spectral Envelope
3. Formant Synthesis: the Voice Source
4. Formant Analysis
5. Summary

Outline

1. Voice Conversion
2. Formant Synthesis: Spectral Envelope
3. Formant Synthesis: the Voice Source
4. Formant Analysis
5. Summary

Voice Conversion



Voice conversion generates a **target speech** that has the same text content as the **source speech**, but sounds as though produced by a particular target speaker.

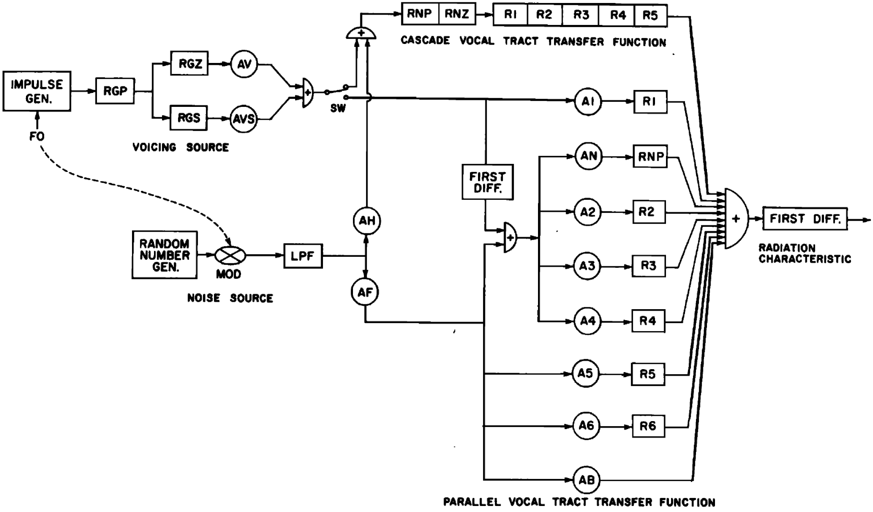
Usually, voice conversion is performed separately for **excitation parameters** and **spectral envelope parameters**.

	Excitation Parameters	**Envelope Parameters**
Formant Synthesis	F_0, V/UV, Gain, LF model parameters	F_1, F_2, F_3, F_4, B_1, B_2, B_3, B_4
LPC	$e[n]$, F_0, $\vec{\beta}$, Gain	$\vec{a}$
WORLD Synthesizer	Periodicity, Aperiodicity	Envelope
Factored Autoencoder	Pitch, Rhythm	Timbre

Outline

1. Voice Conversion
2. Formant Synthesis: Spectral Envelope
3. Formant Synthesis: the Voice Source
4. Formant Analysis
5. Summary

Formant Synthesis: Overview



Formant Synthesis: Envelope

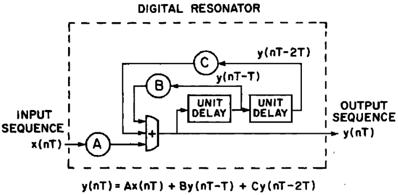
Formant synthesis computes speech by filtering an excitation, $e[n]$, through a transfer function, $h[n]$:

$$s[n] = h[n] * e[n]$$

The **transfer function**, $h[n]$, may include:

- **Regular formants (cascade synthesis):** appropriate for vowels, glides, and nasal consonants
  - **+Nasal Pole, Nasal Zero:** appropriate for nasal consonants
- **Selected formants (parallel synthesis):** appropriate for fricatives and plosives

The Formant Resonator

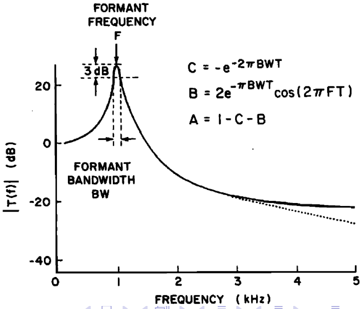


A formant resonator is:

$$R_k(z) = \frac{a_k}{1 - b_k z^{-1} - c_k z^{-2}},$$

which is implemented as:

$$y[n] = a_k x[n] + b_k y[n-1] + c_k y[n-2]$$



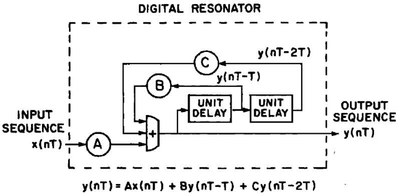
The Formant Resonator

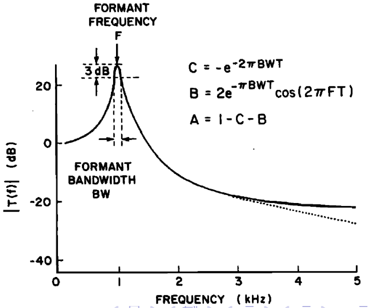
The filter parameters are related to the formant frequency, F_k, formant bandwidth, B_k, and sampling frequency $1/T$ by

$$c_k = -e^{-2\pi B_k T}$$

$$b_k = 2e^{-\pi B_k T}\cos(2\pi F_k T)$$

$$a_k = 1 - b_k - c_k$$

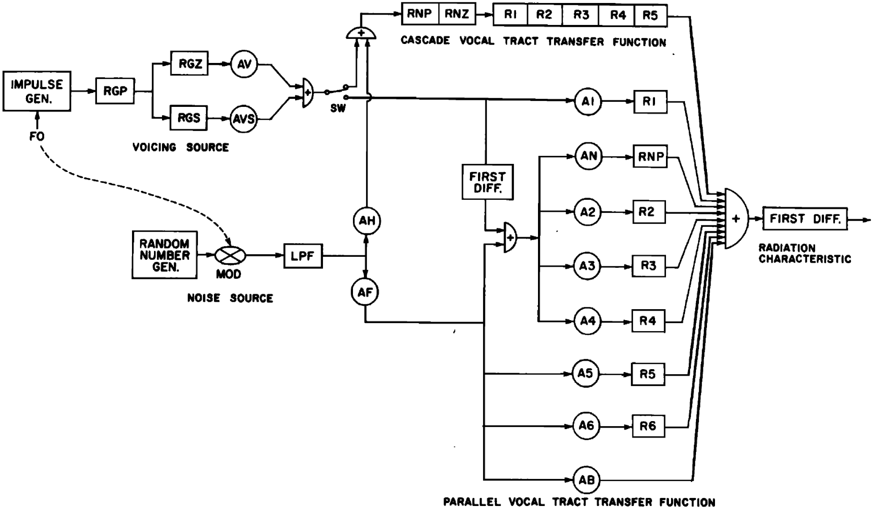




Outline

1. Voice Conversion
2. Formant Synthesis: Spectral Envelope
3. Formant Synthesis: the Voice Source
4. Formant Analysis
5. Summary

Formant Synthesis: Overview



Formant Synthesis: Excitation

$$s[n] = h[n] * e[n]$$

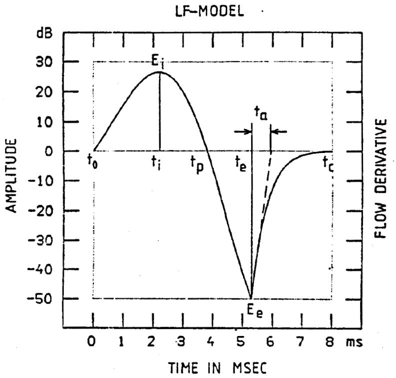
The **excitation signal**, $e[n]$, may include:

- **Regular voicing:** a parametric model of the air pressure immediately above the glottis (proportional to $u_g'(t)$, the derivative of the volume velocity through the glottis)
- **Sinusoidal/breathy voicing:** a parametric model of $u_g'(t)$ when the glottis doesn't close completely
- **Aspiration:** turbulent noise at the glottis, filtered by the whole vocal tract.
- **Frication:** turbulent noise at a supraglottal constriction, filtered by only part of the vocal tract

Regular Voicing: The LF Model

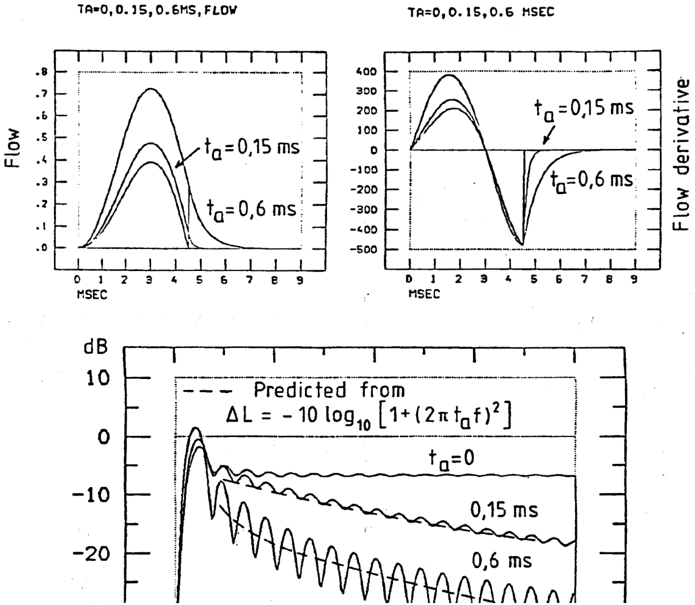
The LF (Liljencrants-Fant) model is a parametric model of $e(t) = u_g'(t)$, the derivative of volume velocity through the glottis. From time 0 to time t_e, $u_g'(t)$ is an unstable oscillation. At time t_e, the vocal folds start to collide, and start to slow down.

$$u_g'(t) = \begin{cases} E_0 e^{\alpha t} \cos(\omega_g t) & t < t_e \\ \frac{E_0}{\epsilon t_a}\left(1 - e^{\epsilon(t_c - t)}\right) & t > t_e \end{cases}$$



(c) Fant, Liljencrants & Lin, 1985.

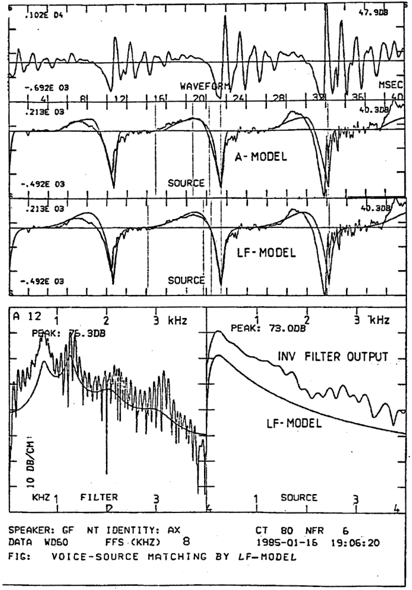
http://www.speech.kth.se/qpsr



(c) Fant, Liljencrants & Lin, 1985. `http://www.speech.kth.se/qpsr`

Shape of the LF model is determined by T_0 (the pitch period) plus four other parameters:

- E_e, amplitude of excitation
- t_e, time of the excitation
- time from upward-going zero-crossing, t_c, to downward-going zero-crossing, t_p
- slope of the return part, $\frac{E_e}{t_a}$



(c) http://www.speech.kth.se/qpsr

Outline

1. Voice Conversion
2. Formant Synthesis: Spectral Envelope
3. Formant Synthesis: the Voice Source
4. Formant Analysis
5. Summary

How do we find formant frequencies and bandwidths?

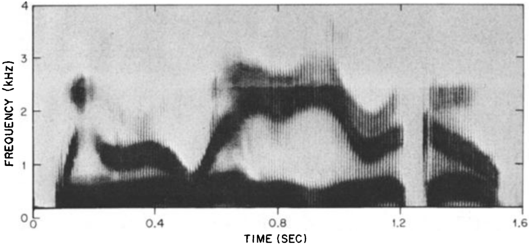
Basically, the formant frequencies and bandwidths are the roots of the LPC polynomial:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{\prod_{i=1}^{p}(1 - p_k z^{-1})}$$

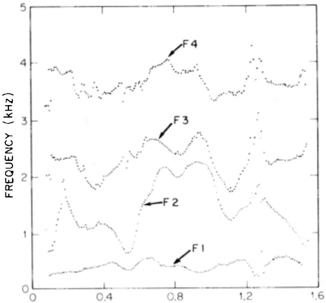
$$F_k = \frac{1}{2\pi T}\angle p_k$$
$$B_k = -\frac{1}{\pi T}\ln|p_k|$$

Utterance: "Why were you away a year ago?" Notice that formant tracking fails during the /g/.

Atal and Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," 1971; (c) Acoustical Society of America.



(a)

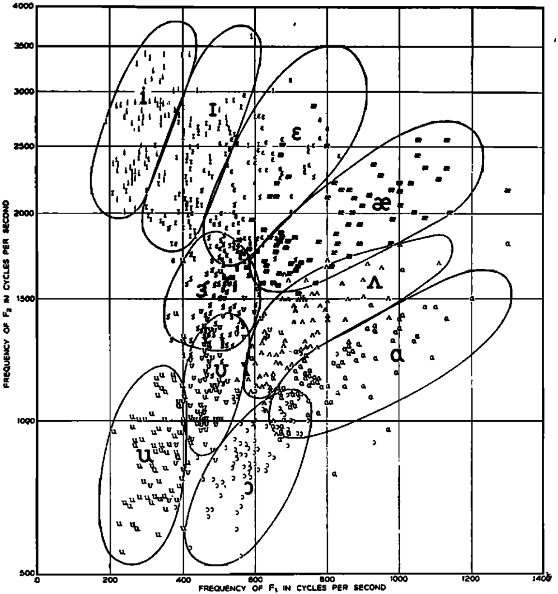


A few complications (but not many)

- Formant tracks are unreliable during consonants, creaky voice, & breathy voice.
- Use dynamic programming to find the most likely formant tracks during consonants, creaky voice, & breathy voice.
- A good implementation: `http://praat.org`.

Formant frequencies determine the vowel. Inside each ellipse, people with longer jaws (e.g., men) typically have lower formants, and vice versa.

Peterson and Barney, 1952.



Outline

1. Voice Conversion
2. Formant Synthesis: Spectral Envelope
3. Formant Synthesis: the Voice Source
4. Formant Analysis
5. Summary

Summary

- Voice conversion usually separates excitation and envelope
- Envelope can be modeled using a formant synthesizer
- Excitation can be modeled using the LF model
- Formant analysis finds the roots of LPC