Speech Analysis and Synthesis by Linear Prediction of the Speech Wave

B.S. Atal
Suzanne L.Hanauer

Presented by

Tuneesh Kumar Lella

Agenda

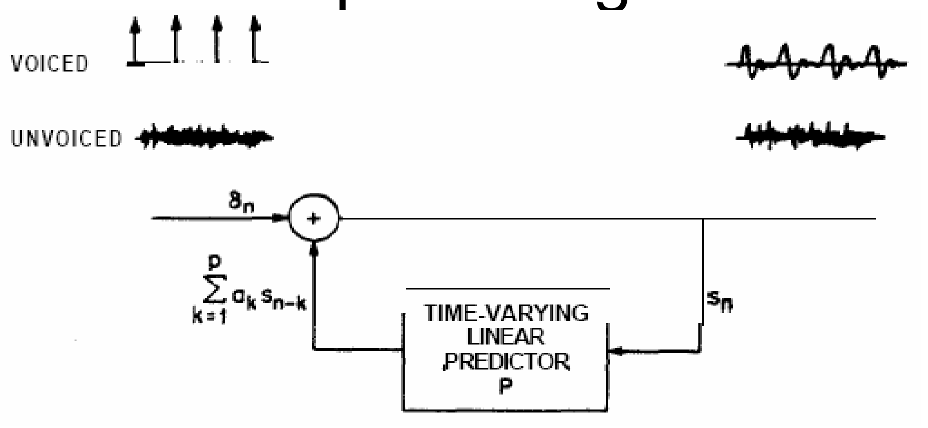
- Motivation
- Model of speech wave
- Speech Analysis
- Speech synthesis
- Applications
- Discussion

Motivation

- Efficient representation of speech signals in terms of less number of slowly varying parameters
- Spectral analysis - Not efficient
  - Needs long speech segments
  - Little information between pitch harmonics

Model of Speech Wave

- Speech sounds are produced by acoustical excitation of human vocal tract
- Representation of speech signal



- Output at nth sampling instant (where a_ks are predictor coefficients) is $s_n = \sum_{k=1}^{p} a_k s_{n-k} + \delta_n$

Transfer Function

Recall: the definition of Z-transform $$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}$$

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{K} a_k x(n-k)$$

$$\Big\downarrow \text{ZT}$$

$$E(z) = X(z)[1 - \sum_{k=1}^{P} a_k z^{-k}]$$

or

$$X(z) = E(z)H(z), H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}$$

Filter (Transfer Function)

Z-Transform of the Speech Wave

- Glottal volume flow together with radiation

$$\frac{K_1 K_2 (1 - z^{-1})}{(1 - z_a z^{-1})(1 - z_b z^{-1})}$$

- It is approximated as

$$\frac{K_1 K_2}{[1 + (1 - z_a) z^{-1}](1 - z_b z^{-1})}$$

Number of Predictor Coefficients

- Number of coefficients ‘p’ determined by
  - Number of resonances and anti-resonances
  - Nature of glottal volume function
  - Radiation
- Mostly used ‘p’ value is 12

Model Parameters

- Hence speech wave can be represented by
  - Predictor coefficients (a_k)
  - Pitch period
  - RMS values of speech samples
  - A binary parameter (speech-voiced or unvoiced)

Speech Analysis

- Samples of voiced speech are linearly predictable from the past ‘p’ samples
- Prediction error

$$E_n = s_n - \hat{s}_n \qquad \hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}.$$

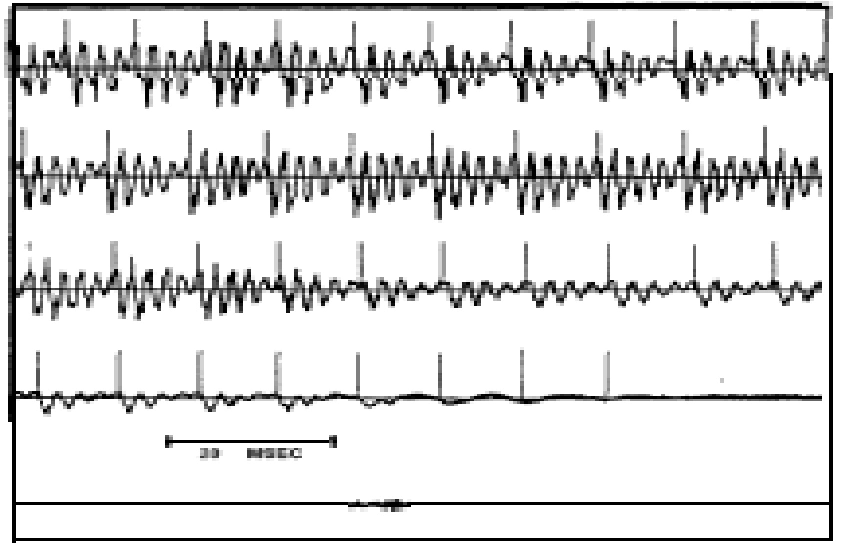
- Mean squared prediction error is

$$\langle E_n^2 \rangle_{av} = \langle (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 \rangle_{av}.$$

- Coefficients a_k are selected such that mean square error is minimum

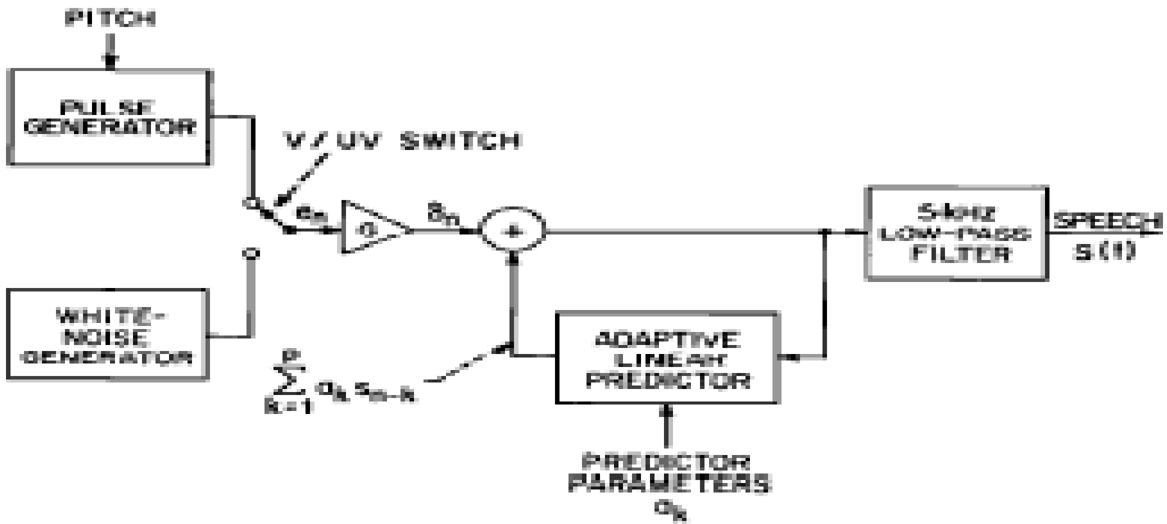
Pitch Analysis

- Positions of pitch pulses can be found using prediction errors En
- The pulses are the vertical lines in the figure



Speech Synthesis

- Block diagram of speech synthesizer



Synthesizer Control Parameters

- Control parameters reset at every pitch period for voiced speech and once every 10msec for unvoiced speech
- To ensure stability of recursive filter, autocorrelation is used for prediction of predictor coefficients

A Little Bit of Calculus

Recall: How to minimize a function $f(a_1,a_2,\ldots,a_P)$?

Answer:
$$\frac{\partial f}{\partial a_i} = 0, i = 1,2,\ldots,P$$

Here
$$f(a_1,\ldots,a_P) = \sum_{n=1}^{M}[x(n) - \sum_{k=1}^{P} a_k x(n-k)]^2$$

$$\frac{\partial f}{\partial a_i} = 0 \Rightarrow \sum_{n=1}^{M} \underline{x(n)x(n-i)} = \sum_{k=1}^{P} a_k \sum_{n=1}^{M} \underline{x(n-i)x(n-k)} \qquad (1)$$

auto-correlation

Autocorrelation Method

$$\sum_{n=1}^{M} x(n)x(n-i) = \sum_{k=1}^{P} a_k \sum_{n=1}^{M} x(n-i)x(n-k)$$

Optimal LPC given by

$$\boxed{r_n(i) = \sum_{k=1}^{P} a_k r_n(|i-k|)}$$

Hence, we can compute the predictor coefficients from samples of autocorrelation function and vice-versa

Synthesized Speech Signal

- Amplitude of nth synthesized sample ‘sn’

$$s_n = q_n + v_n = q_n + g u_n$$

Where q_n is from linear predictor and v_n is contributed by excitation from current segment

$$q_n = \sum_{k=1}^{p} a_k q_{n-k}, \quad 1 \leq n \leq M \qquad u_n = \sum_{k=1}^{p} a_k u_{n-k} + e_n, \quad 1 \leq n \leq M$$

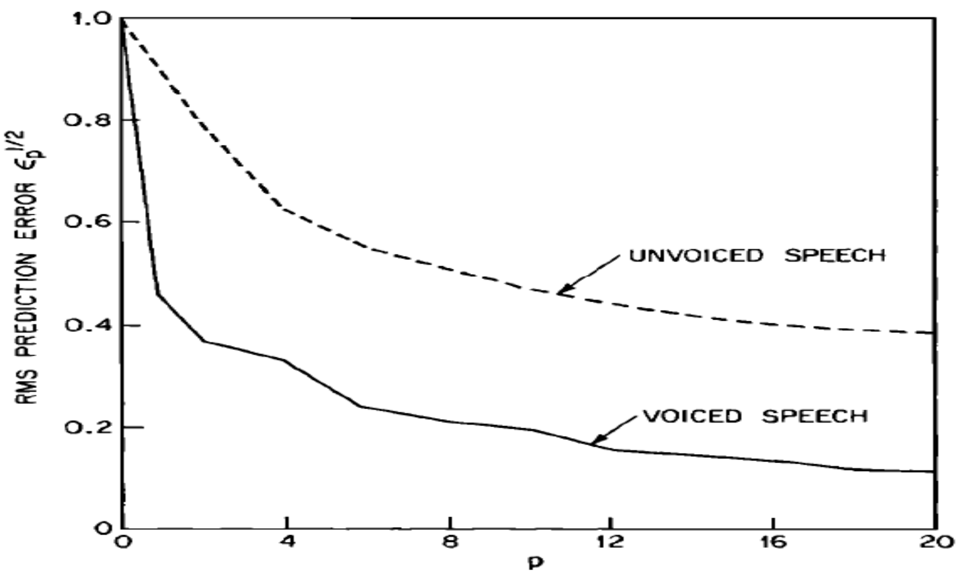
- The mean squared values of speech samples P_s is

$$P_s = \frac{1}{M} \sum_{n=1}^{M} (q_n + g u_n)^2 = \overline{(q_n + g u_n)^2} \quad \Rightarrow \quad g^2 \overline{u_n^2} + 2g\overline{q_n u_n} + \overline{q_n^2} - P_s = 0.$$

- Equation is solved to find ‘g’

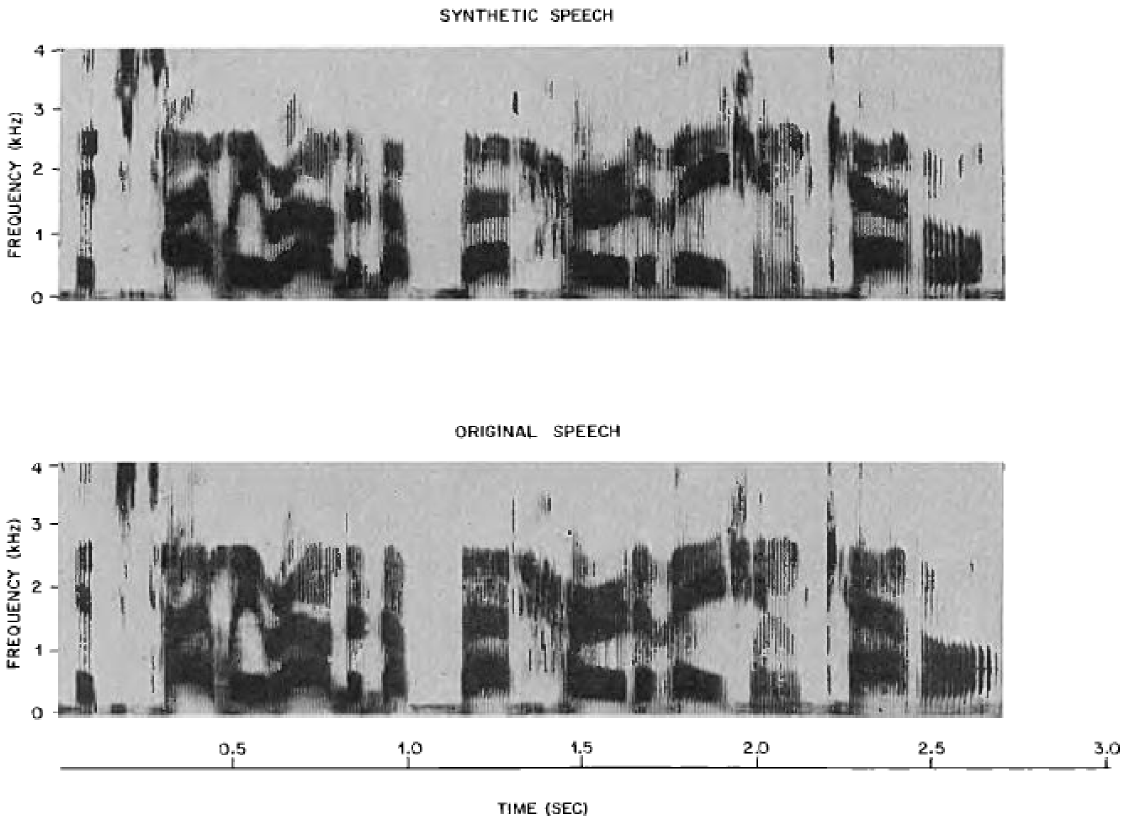
Computer Simulation of Analysis-Synthesis System

- Speech wave low pass filtered to 5KHz and then sampled at frequency of 10KHz
- Optimal value for p is found to be 12



Comparison of synthetic and original speech signals

The uttered sentence is “Its time we rounded up that herd of Asian cattle”



Applications

- Digital storage and transmission of speech
- Separation of spectral envelope and fine structure
- Formant analysis
- Re-forming the speech signals

Digital Storage and Transmission of Speech

- Efficient coding method for synthesizing control information needed
- Encoding predictor coefficients should ensure stability of linear filter
- Direct quantization not efficient for predictor coefficients
- Efficient method - quantize frequencies and bandwidths of poles
- Pitch (6bits), RMS values(5 bits), voiced-unvoiced (1bit) and poles (60bits)-72bits in total

Separation of Spectral Envelope and Fine Structure

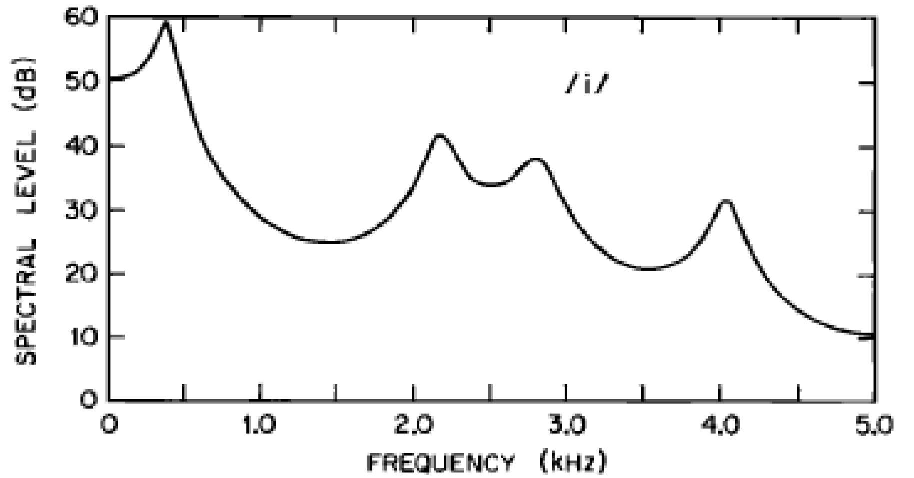
- Fine structure is contributed by the source
- Spectral envelope is the power spectrum of the impulse response of linear filter
- Relation between Spectral Envelope G(f) and predictor coefficients a_k is expressed as

$$G(f)=1/\left|1-\sum^{p} a_k e^{-2\pi j f k/f_s}\right|^2,$$

- Spectral samples of G(f), spaced $f_s/2p$ apart, are sufficient for reconstruction of spectral envelope

Spectral Envelope

- Spectral envelope for the vowel 'I' in "we" spoken by a female speaker at F0=200Hz

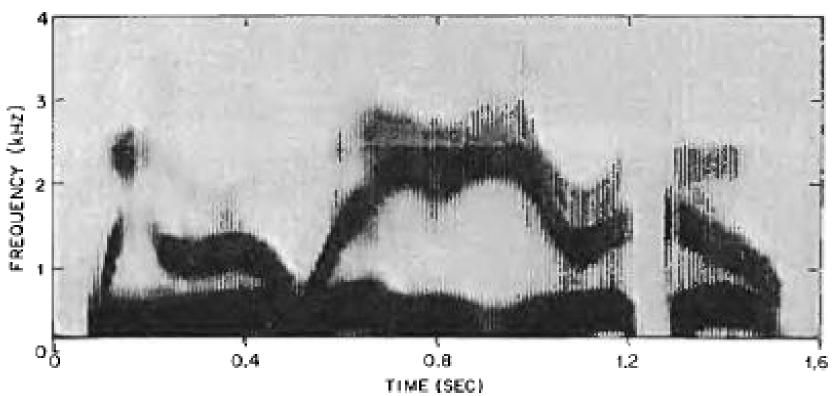


Formant Analysis

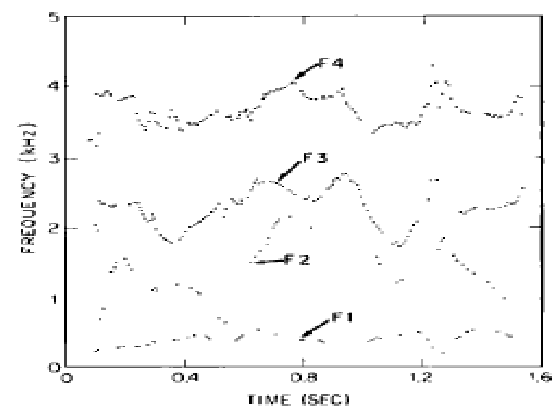
- Objective is to determine complex natural frequencies of vocal tract
- Poles contributed by source fall on real axis or they have a relatively small peak
- Magnitude of spectral peak of a pole is compared to a threshold to determine whether the pole is natural frequency of vocal tract

Formant Analysis

- Formant frequencies for the utterance “we were away a year ago” by male speaker



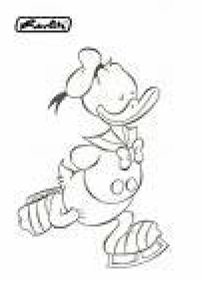
Wideband sound spectrogram



Formants obtained by computer program

Re-forming the Speech Signals

- Synthesis procedure allows independent control of spectral envelope, relative durations, pitch and intensity
- Speaking rate may be altered
- Recovery of “helium speech”



Conclusions

- Problems encountered with Fourier analysis were removed
- Speech signal is synthesized by a single recursive filter
- Synthesized speech has no perceptible degradation in quality
- Synthesis parameters encoded efficiently
- Computationally very fast

Thank You