Outline

• Introduction of Speech Production
  ▪ Speech production mechanism
  ▪ Three major kind of signals in human speech
    • Voiced, Unvoiced, Silence
  ▪ Pitch
• Realization
  ▪ Block Diagram
  ▪ Voiced/Unvoiced and Pitch detection
  ▪ Prediction Coefficients
• Linear Prediction using Autocorrelation Method
• Results and discussion
  ▪ Prediction order, language, gender
Speech Production Mechanism

- Muscles in lungs expel the air inside
- Air flows thru the vocal cords
- Human controls vocal tract shape
Vocal Cords

- Air flows thru vibrating vocal cords produce quasi-periodic impulses
  - Vowels: a, e, i, o, u
- Air flows thru resting vocal cords produce white noises
  - Fricatives: s, sh, ch, f
- Vocal tract act as linear time-invariant system
- Nasal Cavity acts as zeros in a system
  - Nasal sounds: m, n, ng
- Human control vocal tract shape to produce different sounds
- Artificial Larynx Demo
3 major kinds of signals in human speech

- Voiced: quasi-periodic structure, formants shown in spectrogram
- Unvoiced: noise-like signal, high-energy content
- Silence: low-energy, noise-like
Analysis in Spectrum

- Quasi-periodic structure in time causes harmonic appears in the spectrogram
- Can use linear prediction model to estimate formants
- Inverse filtering
- Use autocorrelation to estimate pitch
Block Diagram of a LPC Vocoder

- Pitch detection
- Voice/Unvoiced selector
- Linear Prediction
V/UV selector and Pitch Estimator

- Use prediction coefficients to design inverse filter
- After inverse filtering, the harmonic structure is explicit
- Perform autocorrelation, detect peak in designated range
- If peak value is above 4 times STD, Voiced, if not, unvoiced
Linear Prediction using Autocorrelation Method

- See speech as a stationary process in a short time
- Use Linear Prediction to model formants in speech
- Error signal is the residue, can be modeled as periodic impulses or white Gaussian noise.

\[ s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + Ge(n) \]  

\[ \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \]

\[ E[e^2(n)] = E\left[ \left( s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 \right] \]

- Use Durbin-Levinson Recursion to get coefficients
Programming Flow Chart

1. Extract Frame
2. Pitch Detection
   - Voice/U V
   - Impulse chain
   - White noise
3. Get linear prediction parameters
4. Synthesis Using parameters and input signals
5. Overlap-adding every frame
6. Final Frame?
   - Y
   - Play
   - N
Results

- **English Female**
  - Original 🎧
  - Synthesize using pitch = 90 🎧
  - Synthesize using only noise 🎧
  - Synthesize using prediction order = 12 🎧
  - Synthesize using prediction order = 3 🎧

- **English Male**
  - Original 🎧
  - Synthesize using pitch = 90 🎧
  - Synthesize using only noise 🎧
  - Synthesize using prediction order = 12 🎧
  - Synthesize using prediction order = 3 🎧

- **Chinese Female** ‘你好 謝謝你今天的幫忙’
  - Original 🎧
  - Synthesize using pitch = 90 🎧
  - Synthesize using only noise 🎧
  - Synthesize using prediction order = 12 🎧
  - Synthesize using prediction order = 3 🎧
Results

- English Female ‘How are you, thanks for your help today.’
  - Original 🎧
  - Synthesize using pitch = 90 🎧
  - Synthesize using only noise 🎧
  - Synthesize using prediction order = 12 🎧
  - Synthesize using prediction order = 3 🎧
Discussion

• Lousy performance when come across voiced fricative (v, z) and nasal sound.
  ▪ Due to incorrect modeling
  ▪ Can use other algorithms to model residue better
• Lousy performance when decrease prediction order
• Seriously affected by incorrect pitch detection
• For a child-like voice, performance is worse than male and female
• Seems ok to change language
Reference