Voice modeling

- Source filter models
- Linear Prediction
  chapt. 2.5.2
Speech production mechanism

- Speech is an acoustic sound pressure wave created when air is expelled from the lungs through the trachea and vocal tract.

- Acoustic wave passes through the vocal tract: its frequency content is altered by the resonance of the vocal tract (acoustic tube).

Resonances of the vocal tract = **Formants**
The speech signal

Speech sounds in the spectrogram

Voiced: oscillation of the vocal chords in a quasi-periodic manner causes a periodic excitation of the vocal tract [vowels, nasals]

Unvoiced: are generated by forming a constriction at some point in the vocal tract (usually towards the mouth end), and forcing air through the constriction → turbulence [fricatives]

(Plosive): result from making a complete closure, building up pressure behind the closure and abruptly releasing it.
Voice production

- Source sound
  - Air pumped by lungs through vocal folds

- Two types of signals
  - Voiced signals: quasi-periodic oscillation of vocal folds
  - Airflow is a sequence of “puffs”
    - Typical oscillation frequencies: 110-200-300Hz for males-females-children
  - Unvoiced signals: vocal folds do not oscillate
    - /s/ - vocal folds completely open
    - /h/ - vocal folds partially open
Voice production

Filter

- System composed of vocal tract, oral cavity, nasal cavity
- Pharynx: can change geometry by moving larynx or soft palate
- Oral cavity: can change geometry by moving tongue, palate, cheeks, lips, jaw
- Nasal cavity: fixed geometry

Effects

- Resonances (called formants)
- Restrictions produce turbulent flow (noise), e.g. /s/ or /f/
Voice production

- Vocal folds vibrating at two different frequencies
  - A1 (110 Hz), A2 (220 Hz)

- Same vocal tract (B)

- Result: same vowel at two different pitches (C1, C2)

On the other hand: same pitch and shifted vocal tract resonances produce “Donald Duck” effect (speech in helium)
Tuning of vocal tract resonances by sopranos
- As pitch $f_0$ increases, mouth more open:
- first formant increases and matches $f_0$

Overtone singing (two pitches perceived)
- E.g., “two cavities” technique:
- resonance emphasizes strongly one harmonic component

“Quintina” in Sardinia religious singing (tenores)
- The coincidence of the harmonics of 4 real voices produces perception of a 5th virtual voice (Our Lady)

https://www.youtube.com/watch?v=UHTF1-IhuC0
The speech signal

- Elements of the speech signal:
  - spectral resonances (formants, moving)
  - periodic excitation (voicing, pitched)
  - pitch contour
  - noise excitation (fricatives, unvoiced, no pitch)
  - transients (stop-release bursts)
  - amplitude modulation (nasals, approximants)
Methods to synthesize speech

- **Concatenative synthesis** uses different length prerecorded samples derived from natural speech.
  - probably the easiest way to produce intelligible and natural sounding synthetic speech.
  - usually limited to one speaker and one voice
  - usually require more memory capacity than other methods.

- **Formant synthesis** models the transfer function of vocal tract → based on source-filter-model.
  - widely used synthesis method during last decades
  - Based on the source-filter-model of speech
  - provides infinite number of sounds → very flexible

- **Articulatory synthesis** attempts to model the human speech production system directly.
  - involves models of the human articulators and vocal cords.
  - The articulators are usually modeled with a set of area functions between glottis and mouth.
  - excitation parameters may be glottal aperture, cord tension, and lung pressure.
  - allow accurate modeling of transients due to abrupt area changes, whereas formant synthesis models only spectral behavior.
  - still too complicated for high quality implementations
History

- Mechanical models (late 1700s)
  - Speaking machine (1791) by Wolfgang von Kempelen (also known as the inventor of the Turk...)
  - Bettered by Sir Charles Wheatstone (late 1800s)
- Airflow passed through reed
- Resonances changed by manipulating a tube
History

Electrical synthesizers (1930s)

- “Voder” (Bell Labs: Dudley, Riesz, Watkins, 1939)
- two excitation signals: noise, impulse train
- resonant filters in parallel
- resonances controlled by keyboard, pitch pedal
History

- Concatenative synthesis
  - most used technique today in Text-To-Speech (TTS) systems
  - segments of speech recorded and concatenated (trade-off between segment length and available memory)
    - words: $\sim 10^5$ ; syllables: $\sim 10^4$ ; phonemes: 40-50.
    - Diphones (phonemes with transitions to adjacent phonemes): $\#(\text{phonemes})^2 \sim 10^3$
- Problems: concatenation (PSOLA), prosody, personalization to individual voices
Vowels

Formants:
- spectral picks of the sound spectrum
- characteristic of the different vowels
- the two first formants, $F_1$ and $F_2$, are enough to disambiguate the vowel
American vs. Italian vowel formant frequencies
Italian vowels
Formants in a wide-band spectrogram
Formant synthesis

- Realizes a source-filter model (vocal folds – vocal tract)
  - Depending on whether voiced or unvoiced speech has to be simulated → two different paths are used

voiced signal

\[ S(z) = g_v X(z) \cdot [G(z) \cdot V(z) \cdot R(z)] \]

unvoiced signal

\[ S(z) = g_u X(z) \cdot [V(z) \cdot R(z)] \]
Source signals: glottal pulse

- It's the airflow at the glottis produced by vocal folds
- Model: filter G shapes airflow waveform
- Example: FIR model

\[ g_{\text{FIR}}[n] = \begin{cases} 
\frac{1}{2} \left[ 1 - \cos \left( \frac{\pi n}{N_1} \right) \right], & 0 \leq n \leq N_1, \\
\cos \left( \frac{\pi (n-N_1)}{2N_2} \right), & N_1 \leq n \leq N_1 + N_2 \\
0 & \text{elsewhere.}
\end{cases} \]

- Example: IIR model

\[ G_{\text{IIR}}(z) = \frac{1}{\left[ 1 - \exp(-c/F_s)z^{-1} \right]^2} \]
Filter

- Action of the vocal tract and lip radiation, transforming airflow at the glottis into acoustic pressure at the lips

- **Vocal tract**: $k$ resonant filters, one per formant (e.g., 2nd order resonators)
  - 3 formants enough, 5 formants needed for good quality
  - Structures: **cascade** or **parallel**

$$V_{\text{casc}}(z) = g \prod_{i=1}^{K} V_i(z), \quad V_{\text{par}}(z) = \sum_{i=1}^{K} a_i \cdot V_i(z)$$

- Only formant frequency as control information
- One amplitude control for all formant
- Good quality for vowels
- Not so good for nasals
- Controlling the bandwidth and gain for each formant
- Good for nasals, fricative
- More parameters to control

- **Lip radiation**: modeled as a load that converts the airflow signals at the lips into an outgoing pressure wave
  - acts approximately as a derivator

$$R(z) = 1 - \rho z^{-1}$$
A single frequency formant can be modeled with a two-pole resonator. Formant frequency and bandwidth to be specified.

\[
H(z) = \frac{b_0}{1 + a_1 z^{-1} + a_2 z^{-2}}
\]

\[
a_1 = -2r \cos(\omega_c) \quad a_2 = r^2 \quad b_0 = (1-r)\sqrt{1-2r \cos(2\omega_c) + r^2}
\]

\[r = e^{-\frac{\pi B}{F_s}} \quad \text{bandwidth}
\]

\[\omega_c = \frac{2\pi f_c}{F_s} \quad \text{center frequency}\]
Formant synthesis example

- spectra of two pulse trains with fundamental frequencies at 150 Hz and 250 Hz
- first three formants of the vowel /a/
- spectra of the two output signals obtained by filtering the pulse trains through a parallel combination of the three formants
Parallel synthesis example

a) pulse signal (pitch=250 Hz)

b) 2nd order cells configured for vowel /i /

c) pulse wave convolved by 2nd order cells

d) result when pulse waveform pitch=100 Hz
Problem!

Given a signal modeled with a source-filter model how to extract the spectral envelope?
The source-filter model

- Notional separation of:
  - **source**: excitation, fine time-frequency structure
  - **filter**: resonance, broad spectral structure

- More a modeling approach than a model
Digital audio effects based on source-filter processing

- If we have the source-filter model of a signal → digital audio effect based on this model

![Diagram of source-filter processing]
Spectral envelope

A **Spectral envelope** is a smoothing of a spectrum, which tends to leave aside the spectral lines structure while preserving the general form of the spectrum.

If signal contains harmonic partials only

- Spectral envelop is the curve that passes through the partial peaks
  - peaks values have to be retrieved
  - interpolation scheme to complete the curve in between the peaks

If the sound contains inharmonic partials or a noisy part

- the notion of a spectral envelope becomes completely dependent on the definition of what belongs to the **source** and what belongs to the **filter**.
Spectral envelope estimation

Estimation

- The **channel vocoder** uses frequency bands and performs estimations of the amplitude of the signal inside these bands and thus the spectral envelope.

- **Linear prediction** estimates an all-pole filter that matches the spectral content of a sound. When the order of this filter is low, only the formants are taken, hence the spectral envelope.

- **Cepstrum** techniques perform smoothing of the logarithm of the FFT spectrum (in decibels) in order to separate this curve into its slow varying part (the spectral envelope) and its quickly varying part (the source signal).
Envelope: Channel vocoder

- in time domain:
  - bank of bandpass filters
  - \( \rightarrow \) channel vocoder
    - octave spaced
    - equally spaced

- in frequency domain:
  - circular convolution
The cepstrum

- **Original motivation:** Assume a source-filter model:

  - Excitation source $g[n]$:
  - Resonance filter $H(e^{i\omega})$

- **Define ‘Homomorphic deconvolution’:**
  - source-filter convolution: $g[n]*h[n]$
  - FT → product: $G(e^{i\omega}) \cdot H(e^{i\omega})$
  - log → sum: $\log G(e^{i\omega}) + \log H(e^{i\omega})$
  - IFT: $\rightarrow$ separate fine structure: $c_g[n] + c_h[n]$
    = deconvolution

- **Definition:**
  Real cepstrum $c_n = \text{idft}(\log|\text{dft}(x[n])|)$
Stages in cepstral deconvolution

- Original waveform has excitation fine structure convolved with resonances
- DFT shows harmonics modulated by resonances
- Log DFT is *sum* of harmonic ‘comb’ and resonant bumps
- IDFT separates out resonant bumps (low quefrency) and regular, fine structure (‘pitch pulse’)
- Selecting low-n cepstrum separates resonance information (deconvolution / ‘liftering’)
Linear prediction

Consider a general system that describes a source-filter model:

\[ x(z) \xrightarrow{g} H(z) \rightarrow s(z) \]

The output is expressed as a linear combination of \( p \) past samples and \( q+1 \) input values:

\[ s[n] = g x[n] + \sum_{k=1}^{q} b_k x[n-k] + \sum_{k=1}^{p} a_k s[n-k] \]

**Linear prediction analysis** works on an approximation of this system: **ALL POLE MODEL**

\[ S(z) = g H_{LP}(z) X(z) \]

\[ H_{LP}(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \]

The output can be predicted using a linear combination of \( p \) past samples plus a weighted input:

\[ s[n] = g x[n] + \sum_{k=1}^{p} a_k s[n-k] \]
Linear prediction

Deconvolution problem: given $s$, estimate $x$ and $H$

\[ S(z) = gH(z)X(z), \quad \text{with} \quad H(z) = \frac{1 + \sum_{k=1}^{q} b_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \]

Linear prediction

- additional hypothesis: $H$ is an “all-pole” filter

\[ S(z) = gH_{LP}(z)X(z), \quad \text{with} \quad H_{LP}(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \]

- means that the sample $s[n]$ is “predicted” by a linear combination of past samples (plus current input)

\[ s[n] = gx[n] + \sum_{k=1}^{p} a_k s(n - k) \]
Linear prediction

Estimate of filter coefficients

\[ \{ \tilde{a}_k \} \ (k = 1, \ldots, p) \]

\[ \tilde{s}[n] = \sum_{k=1}^{p} \tilde{a}_k s(n - k) \]

\[ \tilde{S}(z) = P(z)S(z), \]

\[ E(z) = A(z)S(z), \]

\[ A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \]

if the speech signal obeys the all pole model and if

\[ \tilde{a}_k = a_k \]

then

\[ e[n] = g x[n], \]

\[ A(z) = [H_{LP}(z)]^{-1} \]
LP analysis/synthesis

- Prediction of the current output sample $s[n]$
  \[ \tilde{S}(z) = P(z)S(z) \]
  
  - with $P(z)$ prediction filter

- Residual or prediction error
  \[ e[n] = s[n] - \tilde{s}[n] \]

- Comparing with all pole model:
  
  - If $\tilde{a}_k = a_k$
    
    - $e[n] = g \times x[n]$
    
    - $A(z) = \frac{1}{H_{LP}(z)}$

- Prediction filter estimated by minimizing the error energy of the residual:
  \[ E\{e\} = \sum_{m=-\infty}^{+\infty} e^2[m] \]
Inverse filter

- $A(z)$ inverse or whitening filter
  - $E(z) = A(z) S(z)$
  - $S(z) = H_{LP}(z) E(z)$

- if we assume that the prediction error has white (flat) spectrum
  - all-pole filter $H_{LP}(z)$ completely characterizes the spectrum of $s[n]$
  - roots of $A(z) = $ poles of $H_{LP}(z)$: formants

Prediction filter estimated by minimizing the error energy of the residual:

$$E\{e\} = \sum_{m=-\infty}^{+\infty} e^2[m]$$

- Minimizing the residual energy is equivalent to finding a best spectral fit in the frequency domain
LP example

LP example for the female utterance “la” with prediction order $p = 50$

Prediction error: strong pitch peaks $\Rightarrow$ computing $f_0$ for PSOLA
LP Applications

- Analysis-synthesis (coding, transmission):

\[ S(z) = \frac{E(z)}{A(z)} \]

- hence can reconstruct by filtering \( e[n] \) with \( \{a_i\} \)s
- whitened, decorrelated, minimized \( e[n] \) are easy to quantize
- .. or can model \( e[n] \) e.g. as simple pulse train

- Recognition/classification
  - LP fit responds to spectral peaks (formants)
  - can use for recognition (convert to cepstra?)

- Modification
  - separating source and filter supports cross-synthesis
  - pole / resonance model supports ‘warping’
  - (e.g. male → female)
Example LP

- LP example for the female utterance “la” with prediction order $p = 50$
  - target signal $s[n]$ (dotted line)
  - unit variance residual $x[n]$ (solid line): pitch marks
  - magnitude spectrum $|S(f)|$ (thin line)
  - estimated spectral envelope $|g \cdot H_{LP}(f)|$ (thick line)
Example LP (cont.)

LP filter spectra for different prediction orders for the female utterance “la” plotted with different offsets.
LP spectral matching

$p = 4, 8$
LP spectral matching

- \( p = 16, 32 \)
LP spectral matching

- \( p = 64, 128 \)
Example

vowel /i/

A. Input Signal Multiplied by a Hamming Window

B. Input Spectrum

C. Inverse Filter Spectrum

D. Reciprocal of Inverse Filter Spectrum

E. Output Signal

F. Output Spectrum
Example

Fricative /v/
Filter order

- vowel /ae/
Linear Predictive Coding (LPC)

- Compression degree is strongly depended on the order $p$ of the LP analysis
- Excitation signal is not coded
  - only pitch and voiced/unvoiced flag
Linear Predictive Coding (LPC)

Coding
- Linear prediction and estimation of excitation signal
- “Best” matching excitation signal chosen from a codebook
- Transmitted: LP coefficients and index of excitation signal

Decoding
- Excitation signal taken from the codebook
- Speech resynthesized using LP filter

Order of the LP filter:

\[
p = \begin{cases} 
F_s + 4, & \text{for voiced speech,} \\
F_s, & \text{for unvoiced speech,}
\end{cases}
\]
Vocoding effect

“Cross-synthesis” between two sounds $x_1$,
- Spectrum of $x_1$ is shaped with spectral envelope of $x_2$
- Typically $x_2$ is a voice signal

Used to whiten $x_1[n]$

Used to filter the residual of $x_1[n]$

https://ccrma.stanford.edu/~jos/SpecEnv/Application_Example_Cross_Synthesis.html
GSM Speech Encoder

- **GSM**: Global System for Mobile communication
- **Short Term Prediction (STP)**: LPC analysis filter
- **Long Term analysis**: for pitch, gain etc.
- **Regular pulse excitation (RPE)**: highly simplified and compressed

**Pre-processing**

- **Hamming Window**
- **Segmentation**: 20ms
- **Pre-emphasis**

**STP**

- **Short Term Prediction**
- **LPC Inverse Filter**

**LTP**

- **Long Term Prediction**
- **Gain, pitch**

**Regular pulse excitation (RPE)**

- **LAR coefficients**
- **LPF**

**MUX**
For ‘communications quality’:
- 8 kHz sampling (4 kHz bandwidth)
- ~10th order LPC (up to 5 pole pairs)
- update every 20-30 ms → 300 - 500 param/s
additional material
Klatt's history of speech synthesis: A Few Highlights

The **VODER**, the **FIRST EVER** electronic speech synthesis, demonstrated at the New York World's Fair, 1939, by Bell Laboratories

The **only MONOTONE SYNTHESIZER** ever made! Designed to make research on speech acoustics very easy. Yet, the man-in-the-street assumes computers talk with flat intonation. The `Pattern Playback' contributed greatly to our understanding of speech acoustics. Haskins Laboratories, New York City, 1951

First song in synthetic speech, "**Bicycle Built for Two**" with synthetic piano. This song was reprised by the `decorticated' **Hal** in `**2001: A Space Odyssey**'. Bell Laboratories, by Louis Gerstman and Max Mathews, 1961

Can you tell the **REAL** from the **ARTIFICIAL**? These short phrases were copied from natural recordings by John Holmes in 1961. It is quite hard to tell the real from the synthetic. However, weeks and weeks of labor was required to produce each phrase. (Archive #7)

http://www.cs.indiana.edu/rhythmmsp/ASA/highlights.html
**SPEAK-N-SPELL** toy. The first mass produced high-tech speech product. The system announces words to be spelled using a simple keyboard. Popular for 3-8 year olds in the early 1980s.

**MODERN SPEECH SYNTHESIS BY RULE.** Dennis Klatt devoted many years to development of **MITalk**, a research system that converted ordinary printed text into intelligible speech synthesized entirely ``by rule``.

The commercial version of this system, **DECTalk**, comes standardly with the following **voice options**:

- A. `Perfect Paul' --talking at about 300 words/minute (Arch. #36)
- B. `Beautiful Betty’ (Archive #35B)
- C. `Huge Harry ‘(Archive #35C)
- D. `Kit the Kid’ (Archive #35D)
- E. `Whispering Wendy’ (Archive #35E)
CELP (Code Excited Linear Prediction)

adaptive codebook (periodic)

random codebook (noise, pulse)

Feedback (analysis by synthesis)

LSP parameter

LPC synthesis

perceptual error

gain

input
Synthesis model for vocoder

\[ \text{(random)} \]

\[ \text{synthesis filter} \]

\[ \Sigma \]

\[ \text{gain} \]

\[ \text{pitch interval} \]
Synthesis model for multi-pulse

pitch interval

gain

amplitude and position of pulse

Σ

synthesis filter
Synthesis model for regular multi-pulse

- pitch interval
- gain
- amplitude of regular pulse

Σ synthesis filter
Synthesis model for CELP

- pitch interval
- gain
- selection of code vector
- synthesis filter
Vector Quantizer: 2-D
Synthesis model of VSELP

Pitch interval

Gain

Polarity of base vector

Synthesis filter
Synthesis model for CS-CELP

pitch interval

selection of vector pair

gain

∑
synthesis filter
Synthesis model of ACELP

pitch interval

gain

selection of unit pulse position

Σ

synthesis filter