EE E6820: Speech & Audio Processing & Recognition

Lecture 5: Speech modeling

Dan Ellis <dpwe@ee.columbia.edu>
Michael Mandel <mim@ee.columbia.edu>

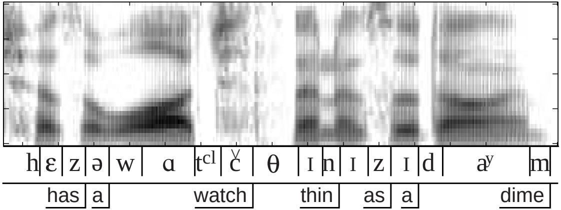
Columbia University Dept. of Electrical Engineering
http://www.ee.columbia.edu/∼dpwe/e6820

February 19, 2009

1. Modeling speech signals
2. Spectral and cepstral models
3. Linear predictive models (LPC)
4. Other signal models
5. Speech synthesis

Outline

The speech signal



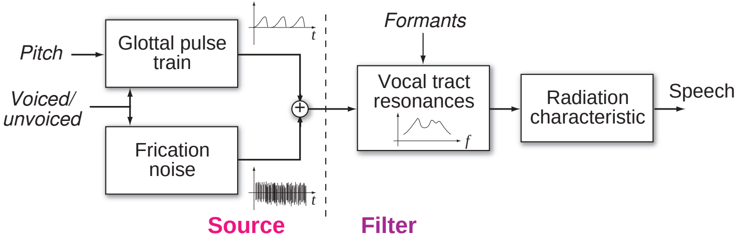
Elements of the speech signal

- spectral resonances (formants, moving)
- periodic excitation (voicing, pitched) + pitch contour
- noise excitation
- transients (stop-release bursts)
- amplitude modulation (nasals, approximants)
- timing!

The source-filter model

Notional separation of

- source: excitation, fine time-frequency structure
- filter: resonance, broad spectral structure



More a modeling approach than a single model

Signal modeling

- Signal models are a kind of representation
  - to make some aspect explicit
  - for efficiency
  - for flexibility
- Nature of model depends on goal
  - classification: remove irrelevant details
  - coding/transmission: remove perceptual irrelevance
  - modification: isolate control parameters
- But commonalities emerge
  - perceptually irrelevant detail (coding) will also be irrelevant for classification
  - modification domain will usually reflect 'independent' perceptual attributes
  - getting at the abstract information in the signal

Different influences for signal models

- Receiver
  - see how signal is treated by listeners

    $\longrightarrow$ cochlea-style filterbank models ...
- Transmitter (source)
  - physical vocal apparatus can generate only a limited range of signals ...

    $\longrightarrow$ LPC models of vocal tract resonances
- Making explicit particular aspects
  - compact, separable correlates of resonances

    $\longrightarrow$ cepstrum
  - modeling prominent features of NB spectrogram

    $\longrightarrow$ sinusoid models
  - addressing unnaturalness in synthesis

    $\longrightarrow$ Harmonic+noise model

Application of (speech) signal models

- Classification / matching
  Goal: highlight important information
  - speech recognition (lexical content)
  - speaker recognition (identity or class)
  - other signal classification
  - content-based retrieval
- Coding / transmission / storage
  Goal: represent just enough information
  - real-time transmission, *e.g.* mobile phones
  - archive storage, *e.g.* voicemail
- Modification / synthesis
  Goal: change certain parts independently
  - speech synthesis / text-to-speech (change the words)
  - speech transformation / disguise (change the speaker)

Outline

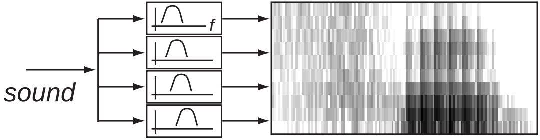
Spectral and cepstral models

- Spectrogram seems like a good representation
  - long history
  - satisfying in use
  - experts can 'read' the speech
- What is the information?
  - intensity in time-frequency cells
  - typically 5ms $\times$ 200 Hz $\times$ 50 dB

$\longrightarrow$ Discarded detail:
  - phase
  - fine-scale timing
- The starting point for other representations

Short-time Fourier transform (STFT) as filterbank

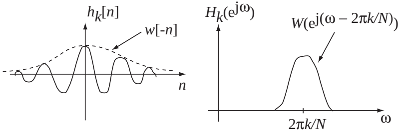
View spectrogram rows as coming from separate bandpass filters



Mathematically:

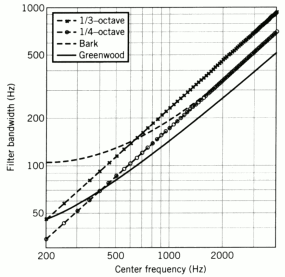
$$X[k, n_0] = \sum_n x[n] w[n - n_0] \exp\left(-j \frac{2\pi k(n - n_0)}{N}\right)$$
$$= \sum_n x[n] h_k[n_0 - n]$$

where $h_k[n] = w[-n] \exp\left(j \frac{2\pi kn}{N}\right)$



Spectral models: which bandpass filters?

- Constant bandwidth? (analog / FFT)
- But: cochlea physiology & critical bandwidths
  - $\longrightarrow$ implement ear models with bandpass filters & choose bandwidths by *e.g.* CB estimates
- Auditory frequency scales
  - constant 'Q' (center freq / bandwidth), mel, Bark, ...

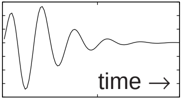


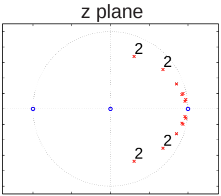
Gammatone filterbank

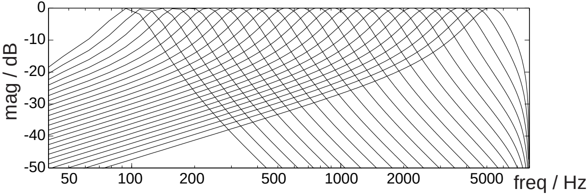
- Given bandwidths, which filter shapes?
  - match inferred temporal integration window
  - match inferred spectral shape (sharp high-freq slope)
  - keep it simple (since it's only approximate)

$\longrightarrow$ Gammatone filters
  - 2N poles, 2 zeros, low complexity
  - reasonable linear match to cochlea

$$h[n] = n^{N-1} e^{-bn} \cos(\omega_i n)$$

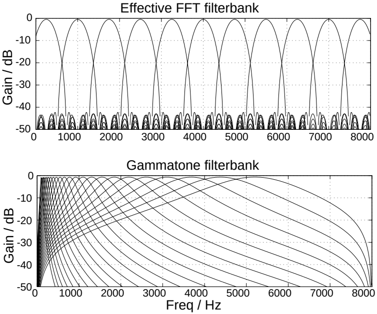




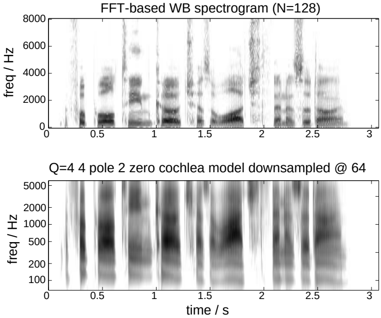


Constant-BW vs. cochlea model

Frequency responses



Spectrograms



Magnitude smoothed over 5-20 ms time window

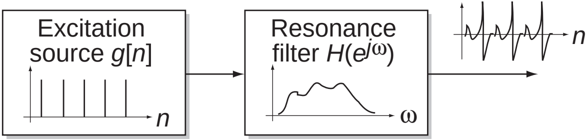
Limitations of spectral models

- Not much data thrown away
  - just fine phase / time structure (smoothing)
  - little actual 'modeling'
  - still a large representation
- Little separation of features

  e.g. formants and pitch
- Highly correlated features
  - modifications affect multiple parameters
- But, quite easy to reconstruct
  - iterative reconstruction of lost phase

The cepstrum

- Original motivation: assume a source-filter model:



- Define 'Homomorphic deconvolution':

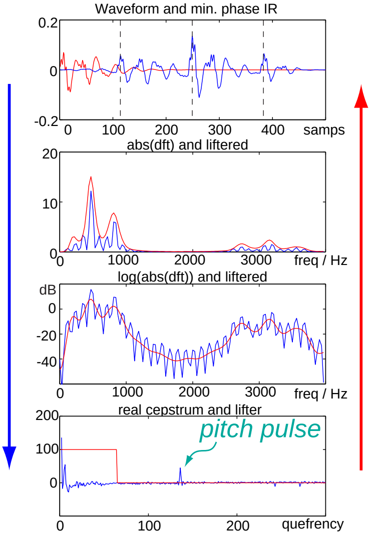
source-filter convolution	$g[n] * h[n]$
FT $\rightarrow$ product	$G(e^{j\omega})H(e^{j\omega})$
log $\rightarrow$ sum	$\log G(e^{j\omega}) + \log H(e^{j\omega})$
IFT $\rightarrow$ separate fine structure	$c_g[n] + c_h[n]$
$=$ deconvolution	

- Definition

  Real cepstrum $c_n = \mathrm{idft}\left(\log\left|\mathrm{dft}(x[n])\right|\right)$

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')



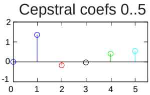
Properties of the cepstrum

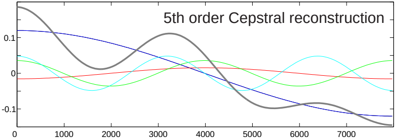
- Separate source (fine) from filter (broad structure)
  - smooth the log magnitude spectrum to get resonances
- Smoothing spectrum is filtering along frequency

  i.e. convolution applied in Fourier domain

  $\rightarrow$ *multiplication* in IFT ('liftering')
- Periodicity in time $\rightarrow$ harmonics in spectrum $\rightarrow$ 'pitch pulse' in high-n cepstrum
- Low-n cepstral coefficients are DCT of broad filter / resonance shape

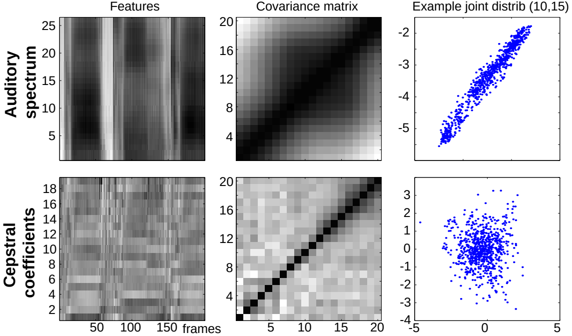
$$c_n = \int \log \left| X(e^{j\omega}) \right| (\cos n\omega + \cancel{j \sin n\omega})\, d\omega$$





Aside: correlation of elements

- Cepstrum is popular in speech recognition
  - feature vector elements are decorrelated



  - c_0 'normalizes out' average log energy
- Decorrelated pdfs fit diagonal Gaussians
  - simple correlation is a waste of parameters
- DCT is close to PCA for (mel) spectra?

Outline

Linear predictive modeling (LPC)

- LPC is a very successful speech model
  - it is mathematically efficient (IIR filters)
  - it is remarkably accurate for voice (fits source-filter distinction)
  - it has a satisfying physical interpretation (resonances)
- Basic math
  - model output as linear function of prior outputs:

$$s[n] = \left(\sum_{k=1}^{p} a_k s[n-k] \right) + e[n]$$

    ... hence "linear prediction" (p^{th} order)
  - $e[n]$ is excitation (input), AKA prediction error

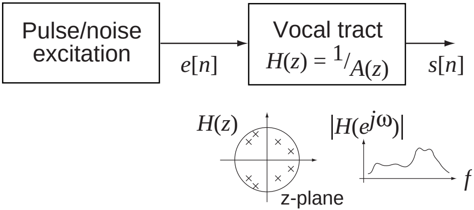
$$\Rightarrow \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}$$

... all-pole modeling, 'autoregression' (AR) model

Vocal tract motivation for LPC

- Direct expression of source-filter model

$$s[n] = \left(\sum_{k=1}^{p} a_k s[n-k] \right) + e[n]$$



- Acoustic tube models suggest all-pole model for vocal tract
- Relatively slowly-changing
  - update $A(z)$ every 10-20 ms
- Not perfect: Nasals introduce zeros

Estimating LPC parameters

- Minimize short-time squared prediction error

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

- Differentiate w.r.t. a_k to get equations for each k:

$$0 = \sum_{n} 2 \left(s[n] - \sum_{j=1}^{p} a_j s[n-j] \right) (-s[n-k])$$

$$\sum_{n} s[n]s[n-k] = \sum_{j} a_j \sum_{n} s[n-j]s[n-k]$$

$$\phi(0,k) = \sum_{j} a_j \phi(j,k)$$

- where $\phi(j,k) = \sum_{n=1}^{m} s[n-j]s[n-k]$ are correlation coefficients
    - p linear equations to solve for all a_js ...

Evaluating parameters

- Linear equations $\phi(0,k) = \sum_{j=1}^{p} a_j \phi(j,k)$
- If $s[n]$ is assumed to be zero outside of some window

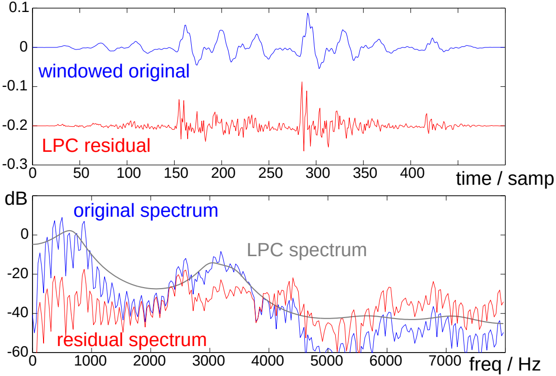
$$\phi(j,k) = \sum_n s[n-j]s[n-k] = r_{ss}(|j-k|)$$

  - $r_{ss}(\tau)$ is autocorrelation
- Hence equations become:

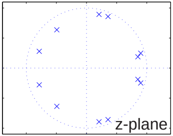
$$\begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix} = \begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(2) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

- Toeplitz matrix (equal antidiagonals)
  $\rightarrow$ can use Durbin recursion to solve
- (Solve full $\phi(j,k)$ via Cholesky)

LPC illustration



Actual poles



Interpreting LPC

- Picking out resonances
  - if signal really was source + all-pole resonances, LPC should find the resonances
- Least-squares fit to spectrum
  - minimizing $e^2[n]$ in time domain is the same as minimizing $E^2(e^{j\omega})$ by Parseval

  $\rightarrow$ close fit to spectral peaks; valleys don't matter
- Removing smooth variation in spectrum
  - $\frac{1}{A(z)}$ is a low-order approximation to $S(z)$
  - $\frac{S(z)}{E(z)} = \frac{1}{A(z)}$
  - hence, residual $E(z) = A(z)S(z)$ is a 'flat' version of S
- Signal whitening:
  - white noise (independent $x[n]$s) has flat spectrum

  $\rightarrow$ whitening removes temporal correlation

Alternative LPC representations

- Many alternate p-dimensional representations
  - coefficients $\{a_j\}$
  - roots $\{\lambda_j\}$: $\prod \left(1 - \lambda_j z^{-j}\right) = 1 - \sum a_j z^{-1}$
  - line spectrum frequencies...
  - reflection coefficients $\{k_j\}$ from lattice form
  - tube model log area ratios $g_j = \log\left(\frac{1-k_j}{1+k_j}\right)$
- Choice depends on:
  - mathematical convenience / complexity
  - quantization sensitivity
  - ease of guaranteeing stability
  - what is made explicit
  - distributions as statistics

LPC applications

- Analysis-synthesis (coding, transmission)
  - $S(z) = \frac{E(z)}{A(z)}$ hence can reconstruct by filtering $e[n]$ with $\{a_j\}$s
  - whitened, decorrelated, minimized $e[n]$s are easy to quantize

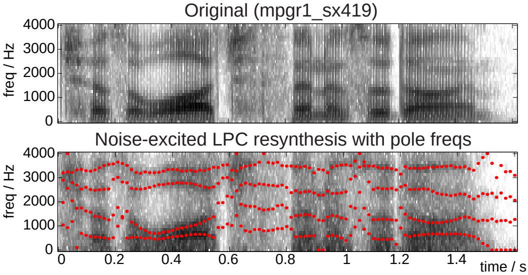
  ... or can model $e[n]$ *e.g.* as simple pulse train
- Recognition / classification
  - LPC fit responds to spectral peaks (formants)
  - can use for recognition (convert to cepstra?)
- Modification
  - separating source and filter supports cross-synthesis
  - pole / resonance model supports 'warping'

    e.g. male $\rightarrow$ female

Aside: Formant tracking

- Formants carry (most?) linguistic information
- Why not classify → speech recognition?

  e.g. local maxima in cepstral-liftered spectrum pole frequencies in LPC fit
- But: recognition needs to work in *all* circumstances
  - formants can be obscured or undefined

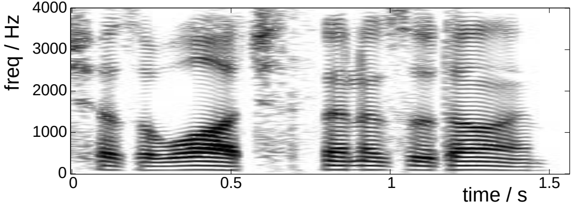


→ need more graceful, robust parameters

Outline

Sinusoid modeling

- Early signal models required low complexity
  e.g. LPC
- Advances in hardware open new possibilities. . .
- NB spectrogram suggests harmonics model



  - 'important' info in 2D surface is set of tracks?
  - harmonic tracks have ∼smooth properties
  - straightforward resynthesis

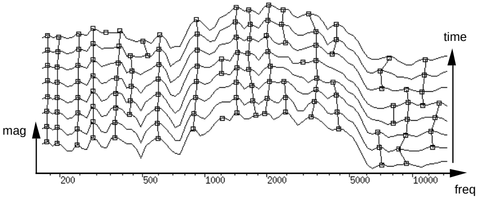
Sine wave models

- Model sound as sum of AM/FM sinusoids

$$s[n] = \sum_{k=1}^{N[n]} A_k[n] \cos(n\,\omega_k[n] + \phi_k[n])$$

  - A_k, ω_k, ϕ_k piecewise linear or constant
  - can enforce harmonicity: $\omega_k = k\omega_0$

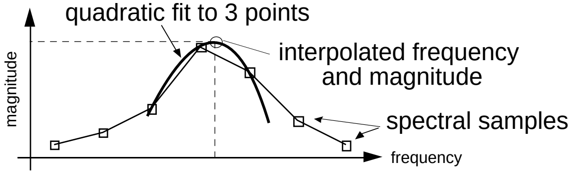
- Extract parameters directly from STFT frames:



  - find local maxima of $|S[k, n]|$ along frequency
  - track birth/death and correspondence

Finding sinusoid peaks

- Look for local maxima along DFT frame
  i.e. $|s[k-1,n]| < |S[k,n]| > |S[k+1,n]|$
- Want exact frequency of implied sinusoid
  - DFT is normally quantized quite coarsely
  e.g. 4000 Hz / 256 bands = 15.6 Hz/band



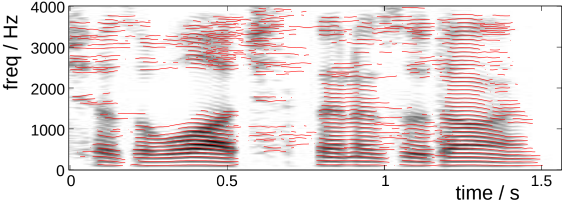
  - may also need interpolated unwrapped phase
- Or, use differential of phase along time (pvoc):

$$\omega = \frac{a\dot{b} - b\dot{a}}{a^2 + b^2} \qquad \text{where } S[k,n] = a + jb$$

Sinewave modeling applications

- Modification (interpolation) and synthesis
  - connecting arbitrary ω and ϕ requires cubic phase interpolation (because $\omega = \dot{\phi}$)
- Types of modification
  - time and frequency scale modification

    ... with or without changing formant envelope
  - concatenation / smoothing boundaries
  - phase realignment (for crest reduction)
- Non-harmonic signals? OK-ish

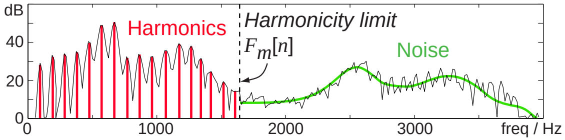


Harmonics + noise model

- Motivation to improve sinusoid model because
  - problems with analysis of real (noisy) signals
  - problems with synthesis quality (esp. noise)
  - perceptual suspicions
- Model

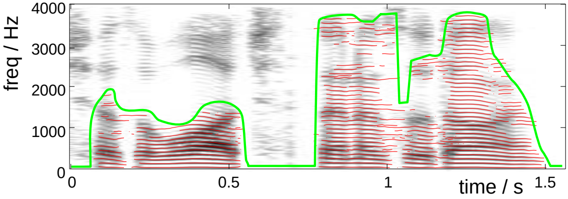
$$s[n] = \sum_{k=1}^{N[n]} \underbrace{A_k[n] \cos(nk\omega_0[n])}_{\text{Harmonics}} + \underbrace{e[n](h_n[n] * b[n])}_{\text{Noise}}$$

  - sinusoids are forced to be harmonic
  - remainder is filtered and time-shaped noise
- 'Break frequency' $F_m[n]$ between H and N

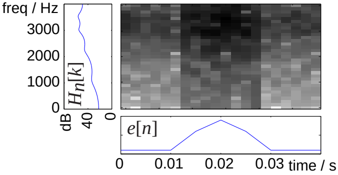


HNM analysis and synthesis

Dynamically adjust $F_m[n]$ based on 'harmonic test':



Noise has envelopes in time $e[n]$ and frequency H_n

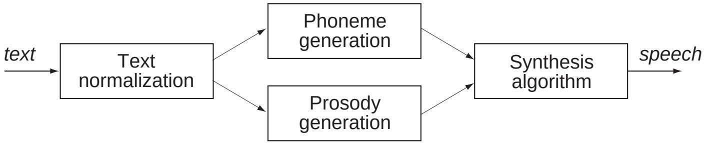


- reconstruct bursts / synchronize to pitch pulses

Outline

Speech synthesis

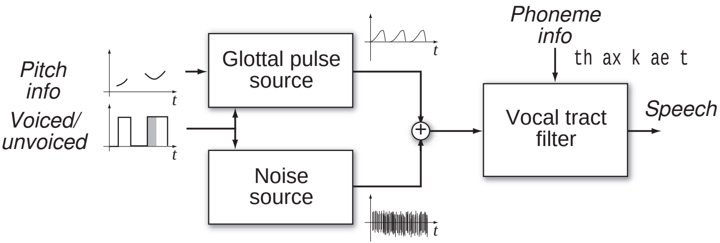
- One thing you can do with models
- Synthesis easier than recognition?
  - listeners do the work
  - ... but listeners are very critical
- Overview of synthesis



  - normalization disambiguates text (abbreviations)
  - phonetic realization from pronunciation dictionary
  - prosodic synthesis by rule (timing, pitch contour)
  - ... all control waveform generation

Source-filter synthesis

Flexibility of source-filter model is ideal for speech synthesis

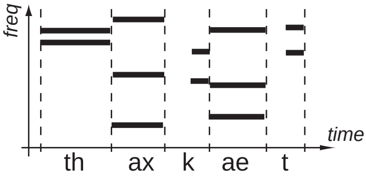


Excitation source issues

- voiced / unvoiced / mixture ([th] *etc.*)
- pitch cycles of voiced segments
- glottal pulse shape → voice quality?

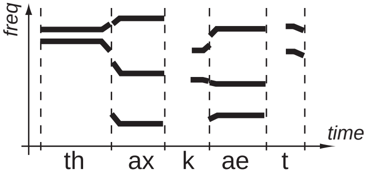
Vocal tract modeling

Simplest idea: store a single VT model for each phoneme



- but discontinuities are very unnatural

Improve by smoothing between templates



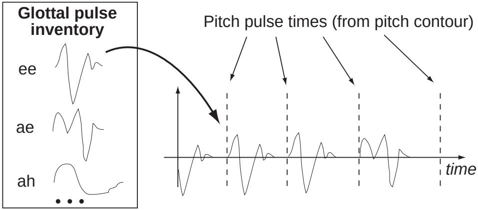
- trick is finding the right domain

Cepstrum-based synthesis

- Low-n cepstrum is compact model of target spectrum
- Can invert to get actual VT IR waveforms:

$$c_n = \mathrm{idft}(\log |\mathrm{dft}(x[n])|)$$
$$\Rightarrow h[n] = \mathrm{idft}(\exp(\mathrm{dft}(c_n)))$$

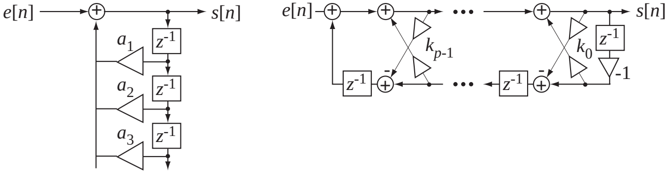
- All-zero (FIR) VT response
  $\rightarrow$ can pre-convolve with glottal pulses



  - cross-fading between templates OK

LPC-based synthesis

- Very compact representation of target spectra
  - 3 or 4 pole pairs per template
- Low-order IIR filter → very efficient synthesis
- How to interpolate?
  - cannot just interpolate a_j in a running filter
  - but lattice filter has better-behaved interpolation

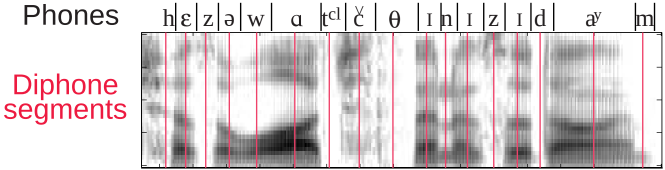


- What to use for excitation
  - residual from original analysis
  - reconstructed periodic pulse train
  - parametrized residual resynthesis

Diphone synthesis

- Problems in phone-concatenation synthesis
  - phonemes are context-dependent
  - coarticulation is complex
  - transitions are critical to perception

→ store transitions instead of just phonemes

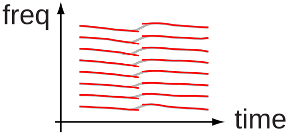


  - ∼ 40 phones ⇒ ∼ 800 diphones
  - or even more context if have larger database

- How to splice diphones together?
  - TD-PSOLA: align pitch pulses and cross fade
  - MBROLA: normalized multiband

HNM synthesis

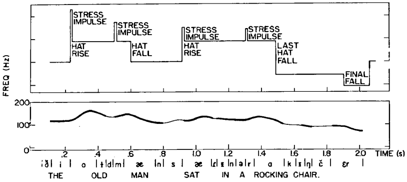
- High quality resynthesis of real diphone units + parametric representation for modification
  - pitch, timing modifications
  - removal of discontinuities at boundaries
- Synthesis procedure
  - linguistic processing gives phones, pitch, timing
  - database search gives best-matching units
  - use HNM to fine-tune pitch and timing
  - cross-fade A_k and ω_0 parameters at boundaries



- Careful preparation of database is key
  - sine models allow phase alignment of all units
  - larger database improves unit match

Generating prosody

- The real factor limiting speech synthesis?
- Waveform synthesizers have inputs for
  - intensity (stress)
  - duration (phrasing)
  - fundamental frequency (pitch)
- Curves produced by superposition of (many) inferred linguistic rules
  - phrase final lengthening, unstressed shortening, ...



- Or learn rules from transcribed elements

Summary

- Range of models
  - spectral, cepstral
  - LPC, sinusoid, HNM
- Range of applications
  - general spectral shape (filterbank) $\rightarrow$ ASR
  - precise description (LPC + residual) $\rightarrow$ coding
  - pitch, time modification (HNM) $\rightarrow$ synthesis
- Issues
  - performance vs computational complexity
  - generality vs accuracy
  - representation size vs quality

Parting thought

not all parameters are created equal...

References

Alan V. Oppenheim. Speech analysis-synthesis system based on homomorphic filtering. *The Journal of the Acoustical Society of America*, 45(1):309–309, 1969.

J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4): 561–580, 1975.

Bishnu S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

J.E. Markel and AH Gray. *Linear Prediction of Speech*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 34(4):744–754, 1986.

Wael Hamza, Ellen Eide, Raimo Bakis, Michael Picheny, and John Pitrelli. The IBM expressive speech synthesis system. In *INTERSPEECH*, pages 2577–2580, October 2004.