Speech Coding Based on Physiological Models of Speech Production

by Juergen Schroeter & M. Mohan Sondhi

(from *Advances in Speech Signal Processing*, Furui & Sondhi, eds.)

Discussion led by Tina Bennett

Outline

- Motivation (*Section 1* - 1 slide)
- Model Components (*Section 2* - 6 slides)
- Deriving Parameters (*Section 3* - 4 slides)
- Results and Future Directions – circa 1992 (*Sections 3.3 & 4* – 1 slide)

Motivation Arguments

- No source-filter model → instead, interaction between *excitation* and *impedence* (vocal/nasal tract)
- Thus, more natural speech produced

  → better parameters for coding, pitch changes natural and automatic, interpolation of parameters

Model components

2.1 Geometry of Vocal & Nasal Tracts

- Vocal Tract: *25 (now 35) years of articulatory modeling* (x-ray data)
  - Their own work based on Mermelstein (1973)
  - see Figure 1 and Table 1 for parameters
- Nasal Tract: fixed filter is sufficient
  - Nasal tract w/model of sinus maxillaries can be efficiently modeled by Helmholtz resonator coupled to nasal tract

Model components

2.2 Wave Propagation in the Tracts

- Assumptions:
  - *linear wave equation applies*

    → exception: just before & after plosives
  - *tract can be approximated as variable-area tube*

    → but what effects of curvature?

    (very little for tube w/*constant* radius)
  - *the motion is planar*

    → reasonable for 3500 Hz or less (tube opening 1.9 inches max), but otherwise?

Model components

2.2 Wave Propagation in the Tracts

- Use *chain matrix* to compute acoustical properties of the tube:
  - relates pressure & volume velocities at one end of the tube to the other (i.e. P_{out}, U_{out} - lips/nose end; P_{in}, U_{in} - glottis end)
- See matrix formula, p. 237
  - Tube is approximated by concatenating 10-20 constant segments (easier to compute K for constant cross section)
  - Can account for losses and walls (see details in paper)
  - Note that all linear properties can be derived (transfer function, input impedance) → equations 2 & 3, p. 238

Model components
2.3 Modeling the Excitation

- Two types (ignoring clicks)
  - *Voiced excitation:* lung air flow interrupted by glottal vibration (i.e. voiced sounds)
  - *Turbulent excitation:* flow through narrow constriction (i.e. fricatives & aspiration)
  - May be combined (i.e. voiced fricatives)

Model components

2.3 Modeling the Excitation

- Modeling Voiced Excitation
  - Two-mass model (Ishizaka & Flanagan, 1972)
    - Figure 2, p. 239
    - Equations 4 through 8b (pp. 239-240)
  - Improvements from incorporation of additional losses (e.g. vertical phasing)
  - Lung pressure assumed to be constant (no breathy voice!)
  - Glottal "chink" (affects F_0; helpful for /h/-to-vowel transitions)
  - Try parametric model of glottal area function instead
    - Advantages: independent control of acoustic features; precise positioning of glottal closures
    - Disadvantages: need more parameters & higher update rate

Model components

2.3 Modeling the Excitation

- Modeling Turbulent Excitation
  - Frication (constriction in vocal tract)
    - Figure 3, equations 10-13, pp. 242-243
      - Series noise location critical (different depending on which fricative), but not for volume velocity source – equation 14, p. 243
  - Aspiration (turbulence at glottis)
    - Same approach

Deriving parameters

- Note 1 to many mapping problem (derivation from LPC vector fails here)
- Their approach: (Figure 4, p. 247)
  - Analysis-by-synthesis & Hooke-and-Jeeves optimization methods
  - Parameters iteratively adjusted to minimize cost function

Deriving parameters

3.1.1 More about the Cost Function

- Cost function: evaluate match between signals (original & synthetic)
- For cost function, compute the following: (for original and synthetic speech)
    - autocorrelations normalized by residual energy
    - tenth-order LPC vectors
    - autocorrelations of LPC vectors
- Alleviates non-uniqueness problem

Deriving parameters
3.1.1 More about the Cost Function

- Four components (3 similarity factors, 1 smoothness penalization): (see equations 17-23, pp. 249-250)

  1) symmetrized likelihood ratio between LPC vectors (original and synthesized)

  2) comparison of energy (original and synthesized)

  3) comparison of time derivatives of glottal excitation (original and synthesized)

  4) parameter distance between frames: smoothness constraint (penalizes large changes in movement)

- These are combined and weighted (based on voicing)
- Joint optimization (tract & glottal params.) worked best

Deriving parameters

3.2 Initialization of Tract Optimization

- Problem: will converge to local (possibly bad) minimum
- Solution: use a codebook for starting shape of vocal tract
  - must cover space of natural speech
  - must ignore spectral tilt
  - see Figure 5 for sample comparison of formant spaces (natural speech vs. using codebook)

Results & Future *(circa 1992)*

- Figure 7: original utterance compared with two synthesized versions
  - optimizing parameters of Mermelstein articulatory model & optimizing tract areas
  - area optimization wins
- Benefits of articulatory model (e.g. interpolation) argue for further investigation
  - improve model for acoustic properties from tracts (incorporate variable losses, more than one fricative noise source)
  - parametric models for glottal opening
  - strategies for coding glottal and tract parameters

Note: References are not noted here; please see the paper!