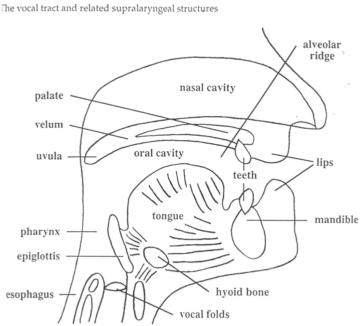
Modeling speech using pole-zero models

Christian H. Kasess

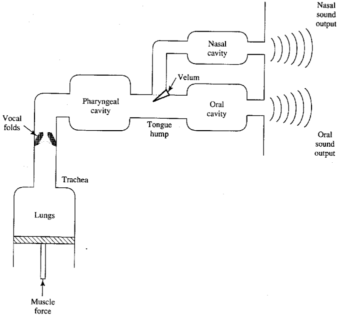
Acoustics Research Institute

25.10.2012

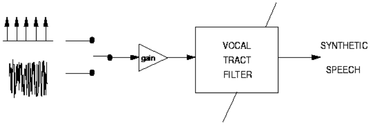


http://pegasus.cc.ucf.edu/ cnye/vocal tract pic.htm

- Roughly divided into three cavities
  - Pharyngeal
  - Oral
  - Nasal
- Oral vowel production
  - Nasal section closed off by velum
- Nasals and nasalized vowels
  - Nasal section coupled
- Laterals (e.g. /l/)
  - Airflow on one (or both) sides of the tongue
  - Generates side branches



A block diagram of human speech production.



The engineering model for speech synthesis.

http://health.tau.ac.il/Communication Disorders/noam

- Glottis acts as source (pulse train)
- Vocal tract acts as 'slowly' varying linear filter

- Source and filter often assumed independent
  - Glottal opening and closing changes VT filter
- Glottal pulse is not ideal pulse
- Effect of glottis not linear
- Still the source-filter model is useful
  - Commonly used in phonetics
  - Model parameters can be used for speaker recognition
  - Useful for formant tracking

- All-pole model captures resonances or formants
- Autoregressive model (AR), linear predictive coding (LPC)

$$y(n) = \sum_{i=1}^{p} a_i y(n-i) + x(n)$$

- Works well with vowels
- Easy to estimate
  - Solve the Yule-Walker equations (Toeplitz) with the Levinson-Durbin algorithm

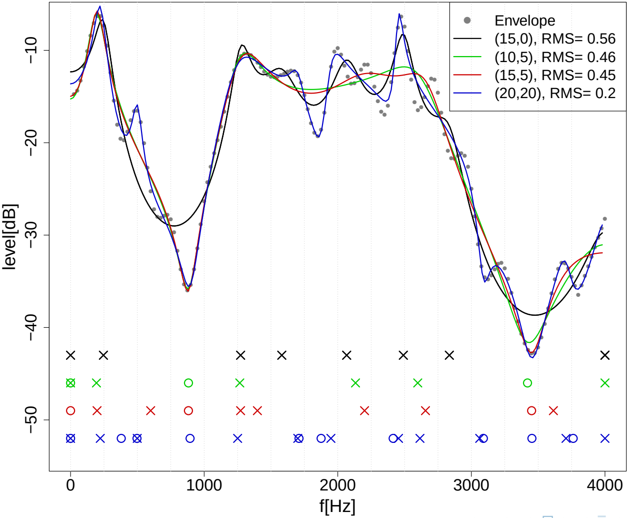
$$\gamma(n) = \sum_{i=1}^{p} a_i \gamma(n-i) + \sigma_x^2 \delta_{n,0}$$

- Direct link to simple physical model

Correlation function...$\gamma(i) = E[y(n)y(n-i)]$

- Nasal spectra show spectral dips
  - Oral cavities and paranasal cavities act as resonators
  - Side branches cause decrease in energy
  - Pole-zero model more efficient
- Problems with pole-zero models
  - Trickier to estimate
  - Requires in general non-linear methods
  - Correspondence to physical model more difficult

All-pole vs. pole-zero model ctd.



- Auto Regressive Moving Average (ARMA)

$$y(n) - \sum_{k=1}^{p} a_k y(n-k) = \sum_{j=0}^{q} b_j x(n-j) \tag{1}$$

- Pole-zero model

$$\hat{y}(\omega) = \frac{\sum\limits_{j=0}^{q} b_j e^{-i\omega k}}{\sum\limits_{k=0}^{p} a_k e^{-i\omega k}} \hat{x}(\omega) = \frac{B\left(e^{-i\omega}, \theta\right)}{A\left(e^{-i\omega}, \theta\right)} \hat{x}(\omega) \tag{2}$$

- Estimation in general a non-linear problem

Time domain

- Not suitable for perceputal frequency scales

Spectral domain

- Perceputal frequency scales can be included
- Logarithmic spectrum can be used
- Spectral envelope needs to be extracted
  - Harmonics for voiced segments due to glottis
  - Envelope represents VT transfer function (+ glottal pulse)

- Linear spectrum
  - Assumptions about phase are necessary (minimum phase)
  - Speech signal is not minimum phase (glottis)
- Log spectrum

$$\theta = \text{argmin}_{\theta'} \sum_{k=0}^{K-1} \left| \log \hat{y}(\omega_k) - \log \frac{B\left(e^{i\omega_k}, \theta'\right)}{A\left(e^{i\omega_k}, \theta'\right)} \right|^2$$

  - Perceptually relevant
- Log amplitude spectrum

$$\theta = \text{argmin}_{\theta'} \sum_{k=0}^{K-1} \left| \log |\hat{y}(\omega_k)| - \log \left| \frac{B\left(e^{i\omega_k}, \theta'\right)}{A\left(e^{i\omega_k}, \theta'\right)} \right| \right|^2$$

  - Phase ignored, minimum phase system easy to obtain
- Cepstral domain
  - Computationally efficient (only for linear frequency)

- Estimate numerator and denominator separately
- Recursive Methods
  - Do not necessarily converge to local minimum
- Non-linear optimization
  - Newton method
    - Calculation of Hessian necessary
    - Numerically expensive and potentially unstable
  - Gauss-Newton method
    - Hessian approximated through first derivatives
    - Convergence issues
  - Quasi-Newton
    - Approximate Hessian (or its inverse) using iterative scheme
    - Numerically stable and inexpensive

- Postitions of poles and zeros
  - Number of complex and real poles/zeros needs
  - Multiplicity
- Quadratic factors
  - Multiplicity
- Polynomial coefficients
  - Only number of poles and zeros

- Substitute non-linear problem with a linear one
- Steiglitz-McBride (1965, 1977)

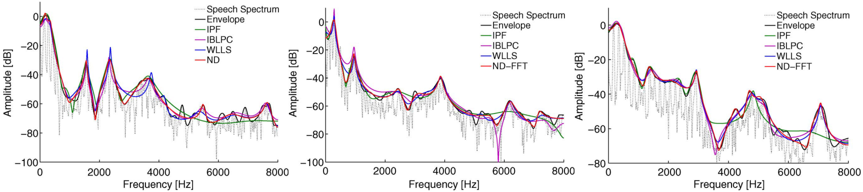
$$\begin{aligned}\theta_i &= \text{argmin}_{\theta'} \sum_{k=0}^{K-1} \left| \hat{y}(\omega_k) \frac{A\left(e^{i\omega_k},\theta'\right)}{A\left(e^{i\omega_k},\theta_{i-1}\right)} - \frac{B\left(e^{i\omega_k},\theta'\right)}{A\left(e^{i\omega_k},\theta_{i-1}\right)} \right|^2 \\ &= \text{argmin}_{\theta'} \sum_{k=0}^{K-1} \left| \hat{y}(\omega_k) - \frac{B\left(e^{i\omega_k},\theta'\right)}{A\left(e^{i\omega_k},\theta'\right)} \right|^2 \left| \frac{A\left(e^{i\omega_k},\theta'\right)}{A\left(e^{i\omega_k},\theta_{i-1}\right)} \right|^2\end{aligned}$$

- More general: Weighted linear least squares (WLLS)

$$\theta_i = \text{argmin}_{\theta'} \sum_{k=0}^{K-1} W(\omega_k, \theta_{i-1}) \left| \hat{y}(\omega_k) A\left(e^{i\omega_k}, \theta'\right) - B\left(e^{i\omega_k}, \theta'\right) \right|^2$$

Marelli and Balazs 2010

- Logarithmic amplitude spectrum
- Estimation of polynomial coefficients
- Quasi-Newton with line search
  - Gradient calculated analytically
  - Broyden-Fletcher-Goldfarb-Shanno (BFGS) method
  - Iterative approximation of the inverse Hessian (rank-one updates)
  - Line search along gradient
- Initialized using the WLLS method

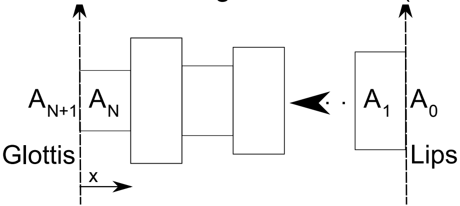


- New method shows lowest error
- Fewer iterations for polynomial representation

Summary Pole-zero

- Efficient representation for laterals, nasals, ...
- Different estimation schemes
- Newton-like method gives good results
- Speaker verification improved as compared to LPC only (Enzinger et al. 2011)
- Important questions
  - What is an appropriate degree for the polynomials?
  - Should the glottal source be corrected?
  - What about physiological constraints?

- Vocaltract as a segmented tube (Wakita 1973, Fant 1960)



- Two equations per segment m (volume velocity)

$$\begin{aligned} p_m(x) &= \frac{\rho c}{A_m}\left(u_m^+ exp(-ikx) + u_m^- exp(ikx)\right) \\ u_m(x) &= u_m^+ exp(-ikx) - u_m^- exp(ikx) \end{aligned} \tag{3}$$

- Volume velocity and pressure are matched at boundaries
- Lossless model (no friction or viscosity, below 4000 Hz ...)

- Transfer function $u_{lips}/u_{glottis} = u_0/u_N$

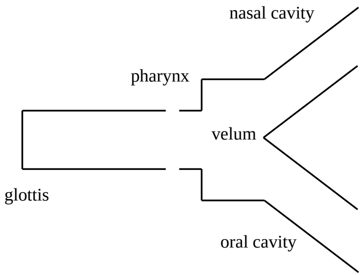
$$\hat{A}(\mu, z) = z^{N/2}(1\ 0) \prod_{m=N}^{0} \frac{1}{1-\mu_m} \begin{pmatrix} 1 & \mu_m \\ \mu_m z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (4)$$

- Correspondence requires fixed segment length (related to f_s)
- specific boundary conditions required (e.g. N=2)

$$\hat{A}(\mu, z) \propto 1 + (\mu_0\mu_1 + \mu_1\mu_2)z^{-1} + \mu_0\mu_2 z^{-2}$$

- For μ_0 or $\mu_N = \pm 1$ reflection coefficients are calculated by recursive algorithm (Markel and Gray, 1976)

m-th reflection coefficient $\mu_m := \frac{A_m - A_{m+1}}{A_m + A_{m+1}}$ and $z := \exp i2\pi\frac{f}{f_s} = \exp i2\pi f \frac{c}{2l}$



- Nasal tract is added
- Each tract is modeled as segmented tube
- For nasals: nasal tract open, oral tract closed
- Vocaltract model has pole-zero characteristic
  - Transfer function given as $f(\mu, z) = \frac{\hat{B}(\mu,z)}{\hat{A}(\mu,z)}$

- No direct way from pole-zero to branched-tube model
- Numerator polynomial appears also in denominator
  - Pole-zero model has $2N + M + L$ coefficients
  - Two-tube model has $N + M + L + 1$ parameters
  - Numerator can be calculated precisely
- Current estimation methods
  - Estimate pole-zero model
  - Apply step-down to numerator and
  - Minimize error with respect to either
    - denomiator polynomial (Lim and Lee 1996) or
    - signal filtered with numerator(Schnell 2003)
  - Gives precedence to zeros

- Estimate all parameters at once
- Use a Bayesian approach to model inversion
- Include prior assumptions about vocal tract smoothness
  - Reflection coefficients close to zero imply a smooth tract
- Sigmoidal parameter transform $\mu_m \rightarrow \theta_m$
  - Restricts reflection coefficients to $(-1, 1)$
- Estimation is based on the log smoothed spectral envelope

$$y(\omega) := \ln G(\omega) = f(\theta, \omega) + \epsilon(\omega)\,. \tag{5}$$

G...envelope, f...transfer function B/A, ϵ...error, θ...transformed μ

$$y(\omega) := \ln G(\omega) = f(\theta, \omega) + \epsilon(\omega)$$

Law of Bayes

$$p(\theta, \lambda | y) \propto p(y|\theta, \lambda)\, p(\theta)\, p(\lambda) = p(y, \theta, \lambda) \tag{6}$$

Under normality assumptions

$$\begin{aligned} p(y|\theta, \lambda) &= \mathcal{N}(y|f(\theta), \Sigma) \\ p(\theta) &= \mathcal{N}\left(\theta|\eta_\theta, \Pi_\theta^{-1}\right) \\ p(\lambda) &= \mathcal{N}\left(\lambda|\eta_\lambda, \Pi_\lambda^{-1}\right). \end{aligned} \tag{7}$$

Covariance of error ϵ is defined as

$$\Sigma^{-1} = g(\lambda) = I_n \exp \lambda \tag{8}$$

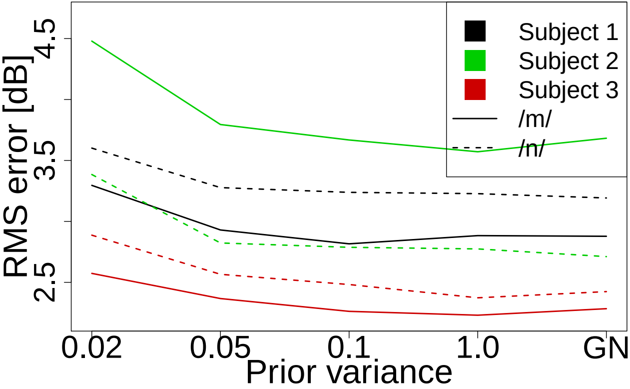
Under a variational approach

$$p\left(\theta, \lambda | y\right) = q(\theta, \lambda) = q(\theta)q(\lambda) \tag{9}$$

with

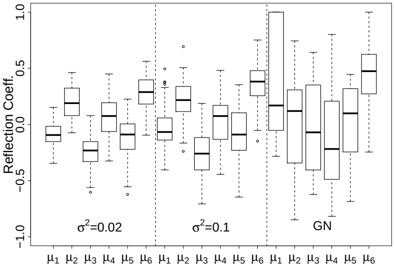
$$\begin{aligned} q(\theta) &= \mathcal{N}\left(\theta | \mu_\theta, \Sigma_\theta\right) \\ q(\lambda) &= \mathcal{N}\left(\lambda | \mu_\lambda, \Sigma_\lambda\right). \end{aligned} \tag{10}$$

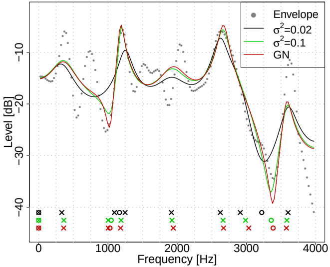
- Iterate λ and θ alternatively
- Use unscented transform for calculating the integrals
- Posterior distribution based on Laplace approximation
  - Find maximum of $q(\theta)$ ($q(\lambda)$) using non-linear optimization
  - Variance follows from 2nd order derivative (approximated by Jacobian)



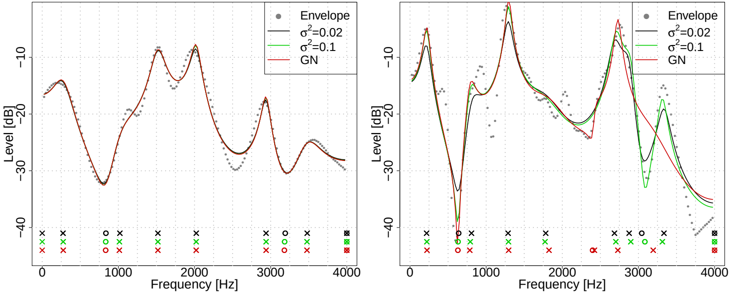
- RMS levels off for higher prior variances
- Simple optimization comparable to Bayesian estimation

Effect of priors I

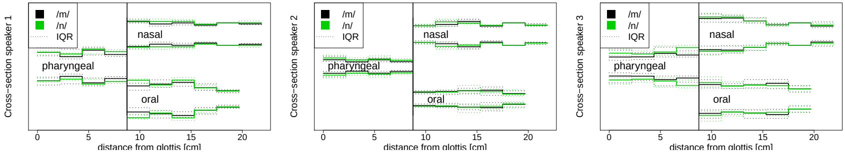




- Less variance for Bayesian scheme
- Effect of tighter priors
  - Spectral features are not always captured as well



- Sometimes the effect of priors is neglgible
- Using the Bayesian scheme may result in fitting different zeros



- Smallest variance in nasal tube
- Differences between /n/ and /m/ in all three branches
    - Differences not what is to be expected
    - Model too simple to capture the nasals properly

The new method

- uses simultaneous estimation of naso-pharyngal and oral section
- applies smoothness priors within a variational Bayesian approach
- does not build on a separate pole-zero estimation

Results show:

- Application to recorded speech data yields in general good spectral fits
- Tradeoff between prior variance and accuracy
- The Bayesian method is more robust against varying initial conditions than a standard optimizer

- Pole-zero models are more efficient for certain types of phonemes
- Non-linear optimization gives best results
- Applications in coding and speaker identification

Physiological models

- Physiological models constrain the solution
- Number of parameters is given naturally
- Other asumptions necessary e.g. terminations ...
- A glottal model is needed
- Different models for e.g. lateral or nasal

- Tracking algorithm
- Glottal excitation model
- Using anatomically motivated priors
  - important if a more complex nasal tract model is included
- Implementing Webster-Horn equation
  - uses conical instead of cylindrical elements
- Impedance models for glottis and lips (nostrils)
- Lossy model for friction and heat conduction
  - exponential decaying term

- G. Fant. Acoustic theory of speech production, with calculation based on X-ray studies of Russian articulations. Mouton De Gruyter, 1960.
- J. Flanagan. Speech analysis, synthesis, and perception. Springer, Berlin, 1972.
- K. Friston and J. Mattout and N. Trujillo-Barreto and J. Ashburner and W. Penny. Variational free energy and the Laplace approximation. Neuroimage, 34, 220–234, 2006.
- I.-T. Lim and B.G. Lee. Lossy Pole-Zero Modeling for Speech Signals. IEEE Trans. Speech Audio Processing, 4(2), 1996.
- D. Marelli and P. Balazs. On Pole-Zero Model Estimation Methods Minimizing a Logarithmic Criterion for Speech Analysis. IEEE, IEEE Trans. Audio Speech Lang. Process., 18(2):237–248, 2010.
- J.D. Markel and A.H. Gray, Jr.. Linear Prediction of Speech. Springer, Berlin, 1976.
- K. Schnell. Rohrmodelle des Sprechtraktes. Analyse, Parameterschätzung und Syntheseexperimente. PhD thesis, Universität Frankfurt, 2000.
- H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Trans. on Aud. and Electroacoustics, AU-21(5):417–427, 1972.
- E. Enzinger, P. Balazs, D. Marelli and T. Becker. A Logarithmic Based Pole-Zero Vocal Tract Model Estimation for Speaker Verification, ICASSP 2011
- Steiglitz, K., and L.E. McBride. A Technique for the Identification of Linear Systems, IEEE Trans. Automatic Control, Vol. AC-10, pp.461-464, 1965.