Modeling speech using pole-zero models

Christian H. Kasess
Acoustics Research Institute

25.10.2012
The vocal tract

- 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
Source-filter model

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... \( \gamma(i) = E[y(n)y(n - i)] \)
Pole-zero models

- 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.

Kasess (ARI)  Vocal tract modeling
Pole-zero models

- **Auto Regressive Moving Average (ARMA)**

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

- **Pole-zero model**

  \[ \hat{y}(\omega) = \frac{\sum_{j=0}^{q} b_j e^{-i\omega k}}{\sum_{k=0}^{p} a_k e^{-i\omega k}} \hat{x}(\omega) = \frac{B(e^{-i\omega}, \theta)}{A(e^{-i\omega}, \theta)} \hat{x}(\omega) \]

- **Estimation in general a non-linear problem**
Time or frequency?

Time domain
- Not suitable for perceptual frequency scales

Spectral domain
- Perceptual 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)

Kasess (ARI)
Vocal tract modeling
SPL 2012
Spectral error measures

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

- **Log spectrum**

\[
\theta = \arg\min_{\theta'} \sum_{k=0}^{K-1} \left| \log \hat{y}(\omega_k) - \log \frac{B(e^{i\omega_k}, \theta')}{A(e^{i\omega_k}, \theta')} \right|^2
\]

- Perceptually relevant

- **Log amplitude spectrum**

\[
\theta = \arg\min_{\theta'} \sum_{k=0}^{K-1} \left| \log |\hat{y}(\omega_k)| - \log \frac{B(e^{i\omega_k}, \theta')}{A(e^{i\omega_k}, \theta')} \right|^2
\]

- Phase ignored, minimum phase system easy to obtain

- **Cepstral domain**
  - Computationally efficient (only for linear frequency)
Optimization Methods

- 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
PZ representation

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

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

\[
\theta_i = \arg\min_{\theta'} \sum_{k=0}^{K-1} \left| \hat{y}(\omega_k) \frac{A(e^{i\omega_k}, \theta')}{A(e^{i\omega_k}, \theta_{i-1})} - \frac{B(e^{i\omega_k}, \theta')}{A(e^{i\omega_k}, \theta_{i-1})} \right|^2
\]

- More general: Weighted linear least squares (WLLS)

\[
\theta_i = \arg\min_{\theta'} \sum_{k=0}^{K-1} W(\omega_k, \theta_{i-1}) \left| \hat{y}(\omega_k) A(e^{i\omega_k}, \theta') - B(e^{i\omega_k}, \theta') \right|^2
\]
- 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
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?
Segmented tube model

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

\[
A_{N+1} \quad A_N \quad \cdots \quad A_1 \quad A_0
\]

- Glottis
- Lips

- Two equations per segment \( m \) (volume velocity)

\[
\begin{align*}
p_m(x) &= \frac{\rho c}{A_m} (u_m^+ \exp(-ikx) + u_m^- \exp(ikx)) \\
u_m(x) &= u_m^+ \exp(-ikx) - u_m^- \exp(ikx)
\end{align*}
\]  

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

Kasess (ARI)
One-tube Model

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

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

(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 $\mu_0$ 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
Vocal tract model has pole-zero characteristic

\[ f(\mu, z) = \frac{\hat{B}(\mu, z)}{\hat{A}(\mu, z)} \]
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
  - denominator polynomial (Lim and Lee 1996) or
  - signal filtered with numerator (Schnell 2003)
- Gives precedence to zeros
New Ansatz

- 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

$$y(\omega) := \ln G(\omega) = f(\theta, \omega) + \epsilon(\omega).$$  \hspace{1cm} (5)

$G$...envelope, $f$...transfer function $B/A$, $\epsilon$...error, $\theta$...transformed $\mu$
New Ansatz

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

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

\[
\begin{align*}
p(y|\theta, \lambda) &= \mathcal{N}(y|f(\theta), \Sigma) \\
p(\theta) &= \mathcal{N}(\theta|\eta_\theta, \Pi_\theta^{-1}) \\
p(\lambda) &= \mathcal{N}(\lambda|\eta_\lambda, \Pi_\lambda^{-1}) .
\end{align*}
\tag{7}
\]

Covariance of error \( \epsilon \) is defined as

\[
\Sigma^{-1} = g(\lambda) = I_n \exp \lambda \tag{8}
\]
Under a variational approach

$$p(\theta, \lambda|y) = q(\theta, \lambda) = q(\theta)q(\lambda)$$  \hspace{1cm} (9)

with

$$q(\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$$
$$q(\lambda) = \mathcal{N}(\lambda|\mu_\lambda, \Sigma_\lambda).$$ \hspace{1cm} (10)

- Iterate $\lambda$ and $\theta$ 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
Less variance for Bayesian scheme

Effect of tighter priors
  - Spectral features are not always captured as well
• Sometimes the effect of priors is negligible
• 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 assumptions 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


