Linear Prediction of Voiced Speech

Jake Gunther

February 2013

Department of Electrical & Computer Engineering
speech coding

speech coding has (at least) two goals

1. low bit rate
   - find the most compact representation of speech
   - analysis

2. quality of synthesized speech
   - faithful reconstruction
   - natural qualities
   - synthesis
anatomy

http://vimeo.com/13591762

http://www.youtube.com/watch?v=9Tlpkdq8a8c
acoustic theory of human speech production

1. volume of air expelled from the lungs

2. oscillations in taught vocal folds

3. chopping up the exiting air into quasi-periodic puffs of air (glottal flow)

4. excites resonances in the vocal tract (vocal tract = organs between the vocal folds and the lips/nostrils)

5. air volume is radiated at the lips and converted into a propagating speech pressure waveform (lips = time derivative)
top: glottal pulse

bottom: glottal derivative pulse
types of sounds in English

• voiced (quasi-periodic glottal flow)
  – /a/, /e/, /i/, /o/, /u/ (all the vowels)

• nasals (quasi-periodic glottal flow)
  – /l/, /m/, /n/, /ng/, /r/

• fricatives (quasi-periodic glottal flow)
  – /th/ (voiced), /v/, /z/

• noise-like (air forced through constrictions in the vocal tract)
  – /f/, /h/, /s/, /th/ (unvoiced)

• plosives (sudden release of air blocked, e.g. by tongue or lips)
  – /k/, /p/, /t/
  – /b/, /d/, /g/, /j/
DSP models for speech

\[ u_g(n) \rightarrow V(z) \rightarrow L(z) \rightarrow s(n) \]

\[ V(z) = \frac{G}{B(z)}, \quad B(z) = 1 - A(z), \quad A(z) = \sum_{i=1}^{q} a(i) z^{-i}, \quad L(z) = 1 - z^{-1} \]

- \( u_g(n) \) is the glottal source signal
- \( V(z) \) is an all-pole vocal tract filter (models spectral structure)
- \( L(z) \) is the lip radiation
- \( s(n) \) is the radiated speech signal
source-filter model

\[ u(n) \rightarrow V(z) \rightarrow s(n) \]

\[ S(z) = V(z)L(z)U_g(z) = V(z)U(z) = \frac{GU(z)}{B(z)} \]

\[ U(z) = L(z)U_g(z) \quad \text{(glottal flow derivative)} \]

parameters of the source-filter model

- \( G, a(1), \ldots, a(q) \) (vocal tract filter) - filter order and bit allocation

- representation of \( u(n) \) (source excitation) - compact representation needed
analysis & synthesis

- analysis - estimate the parameters for recorded speech
- transmit/store - keep the parameters and discard the speech waveform
- synthesis - compute speech signal given parameters
  - recreate \( u(n) \) and play through \( V(z) \)
two-state model

- unvoiced speech

\[ u(n) = \text{Gaussian white noise} \]

- voiced speech

\[ u(n) = \sum_{i=1}^{M} c_i \delta(n - iN_0 - n_0), \quad \text{or} \]

\[ u(n) = \sum_{i=1}^{M} c_i g(n - iN_0 - n_0) \]

much greater naturalness when \( g(t) \) a GFD pulse
two interpretations for $u(n) \leftrightarrow U(z)$

$B(z)S(z) = GU(z), \quad B(z) = 1 - A(z)$ or

$$s(n) = \sum_{i=1}^{q} a(i) s(n - i) + Gu(n)$$

- $B(z) = 1 - A(z)$ is a prediction-error filter
- $B(z)S(z) = GU(z)$ is the prediction error

1. $U(z)$ is the excitation in the source-filter model (synthesis)

2. $U(z)$ is the error (residual) when an ideal predictor is applied to the speech signal $s(n)$ (analysis)

this tells us how to estimate parameters
model for clean speech

- analysis interval of \( m + q \) samples (20 ms)

- \( s(1), s(2), \cdots, s(m + q) \) speech samples

\[
s_1 = S_2 a + u \quad \text{or} \quad S b = u,
\]

where

\[
S = [s_1 | S_2] = \begin{bmatrix}
    s(q + 1) & s(q) & \cdots & s(1) \\
    s(q + 2) & s(q + 1) & \cdots & s(2) \\
    \vdots & \vdots & \ddots & \vdots \\
    s(q + m) & s(q + m - 1) & \cdots & s(m)
\end{bmatrix}, \quad \text{(Toeplitz)}
\]

\[
a = \begin{bmatrix} a_1 \\ \vdots \\ a_q \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -a \end{bmatrix}, \quad u = \begin{bmatrix} u(1) \\ \vdots \\ u(m) \end{bmatrix}
\]
model for noisy speech

\[ s(n) = \sum_{i=1}^{q} a(i)s(n - i) + Gu(n), \]

\[ x(n) = s(n) + w(n), \quad w(t) \text{ is white noise} \]

\[ \mathbf{s} = [s(1), \ldots, s(m + q)]^T \quad \text{and} \]

\[ \mathbf{x} = [x(1), \ldots, x(m + q)]^T, \]

\[
\mathbf{X} = \begin{bmatrix}
x(q + 1) & x(q) & \cdots & x(1) \\
x(q + 2) & x(q + 1) & \cdots & x(2) \\
\vdots & \vdots & \ddots & \vdots \\
x(q + m) & x(q + m - 1) & \cdots & x(m)
\end{bmatrix}, \quad \text{(Toeplitz)}
\]
previous work: least-norm approaches

\[
\min_b \|Sb\|_p \quad \text{s.t. } e^T b = 1, \quad e^T = [1, 0, \cdots, 0]
\]

- objective minimizes the prediction error \( Sb = u \)
- minimum mean-squared error linear prediction, \( p = 2 \), least-squares = ML estimation when \( u(n) \) is white Gaussian noise, tends toward dense residuals
- \( p = 1 \), ML estimation when \( u(n) \) is white Laplacian, tends toward sparse residuals
- Huber norm when \( u(n) \) is Gaussian in the middle and Laplacian in the tails

\[
\|u\|_H = \sum_i \varphi_H(u_i), \quad \varphi_H(u) = \begin{cases} 
\frac{u^2}{2}, & |u| < c, \\
|u| - \frac{c^2}{2}, & |u| > c,
\end{cases}
\]
• for voiced speech residual \( u(n) \) should be sparse \((q = 0)\), relax to \(q = 1\)

• GFD pulses not sparse but \( \frac{\partial^2}{\partial t^2} \text{GFD} \) is sparse

\[
\min_b \| \Delta^2 Sb \|_1 \quad \text{s.t.} \quad e^T b = 1,
\]

where \( \Delta^2 \) is an operator that approximates the second

• linear prediction in noise

\[
\min_b \| Xb \|_2 \quad \text{s.t.} \quad e^T b = 1,
\]

• equivalent to

\[
\min_{u, b} \| u \|_2 \quad \text{s.t.} \quad (X - [u \mid 0])b = 0, \quad e^T b = 1.
\]

all error assumed to be in the first column of \( X \)
new idea

• most previous approaches characterize $u(n)$ statistically through the choice of the norm

• can the glottal pulse shape be exploited?

• mathematical models of glottal pulse available (non-linear)

• previous attempts have estimated the predictor $b$ along with the parameters in mathematical model of glottal pulse
  – disadvantage: very complicated (non-convex) optimization
  – advantage: flexibility of model

• new idea
  – given a codebook of example glottal pulses, construct a dictionary (linear model) from shifted replicas of the codebook pulses
  – the prediction residual is a sparse linear combination of dictionary pulses
dictionary construction

• let \( g = [g_1, \cdots, g_n]^T \) be an glottal pulse

• construct a \( m \times (m + n - 1) \) Toeplitz dictionary matrix

\[
D = \begin{bmatrix}
g_n & \cdots & g_1 & 0 & 0 & 0 \\
0 & \ddots & \ddots & 0 & 0 \\
0 & 0 & \ddots & \ddots & 0 \\
0 & 0 & 0 & g_n & \cdots & g_1
\end{bmatrix}
\]

• for a codebook \( \{g_1, \cdots, g_K\} \), construction dictionary

\[
D = [D_1, \cdots, D_k], \quad u = Dc
\]

• dictionary representation for the prediction residual

\[
Sb - Dc = 0
\]

where \( c \) is sparse (has only a few non-zero elements)
analysis = parameter estimation

\[
\begin{align*}
\min_{b,c} & \quad \|c\|_1 \quad \text{s.t.} \quad \|Sb - Dc\|_2 \leq \varepsilon_{\text{fit}}, \quad e^T b = 1
\end{align*}
\]

- estimate glottal source parameters together with predictor
- location of non-zeros in \(c\) encodes glottal period and phase
- \(\varepsilon_{\text{fit}}\) trades sparsity (bit rate) for accuracy
results
refinement step

\[
\begin{align*}
\min_{b_0, c_0} & \quad \| S b_0 - D_0 c_0 \|_2 \\
\text{s.t.} & \quad e^T b_0 = 1
\end{align*}
\]
residuals
optimization of glottal pulse

\[
\begin{align*}
\min_{b,c,h} & \quad \|c\|_1 \quad \text{s.t.} \quad \|h - g\|_2 \leq \varepsilon_{gp}, \quad e^T b = 1, \\
& \quad \|Sb - D(h)c\|_2 \leq \varepsilon_{fit},
\end{align*}
\]

- the product \(D(h)c\) makes this problem bilinear (non-convex)

- let \(H = D(h)\), then \(Hc = Ch\)

\[
C = \begin{bmatrix}
  c_n & \cdots & c_1 \\
  \vdots & \ddots & \vdots \\
  c_{n+m-1} & \cdots & c_m
\end{bmatrix} \quad (m \times n, \text{Toeplitz}) \quad (1)
\]
linearize

• define $\alpha(h, c) = D(h)c = Hc = Ch$, and linearize about $(h', c')$,

$$
\alpha(h, c) \approx \alpha(h', c') + \left[ \begin{array}{c} \nabla_h f(h', c') \\ \nabla_c f(h', c') \end{array} \right]^T \left[ \begin{array}{c} h - h' \\ c - c' \end{array} \right] \\
= H'c' + C'(h - h') + H'(c - c') \\
= H'c + C'h - H'c'
$$

• linearized optimization problem

$$
\min_{b,c,h} \|c\|_1 \quad \text{s.t.} \quad \|h - g\|_2 \leq \varepsilon_{gp}, \quad e^T b = 1, \\
\|Sb - H'c - C'h + H'c'\|_2 \leq \varepsilon_{fit}
$$

• need to iterate
results

![Graph 1](chart1.png)

![Graph 2](chart2.png)

![Graph 3](chart3.png)
linear prediction in noise

\[
\min_{b,c,s} \left\| c \right\|_1 \quad \text{s.t.} \quad \left\| s - x \right\|_2 \leq \varepsilon_w, \quad e^T b = 1, \\
\left\| S b - D c \right\|_2 \leq \varepsilon_{\text{fit}}
\]

linearize

\[
\min_{b,c,s} \left\| c \right\|_1 \quad \text{s.t.} \quad \left\| s - x \right\|_2 \leq \varepsilon_w, \quad e^T b = 1, \\
\left\| S' b + B' s - S' b' - D c \right\|_2 \leq \varepsilon_{\text{fit}}.
\]

where

\[
B = \begin{bmatrix}
b_q & \cdots & b_0 & 0 & 0 & 0 \\
0 & \ddots & \ddots & 0 & 0 \\
0 & 0 & \ddots & \ddots & 0 \\
0 & 0 & 0 & b_q & \cdots & b_0
\end{bmatrix}
\]
putting it all together

\[
\begin{align*}
\min_{b, c, s, h} & \quad \|c\|_1 \quad \text{s.t.} \quad \|Sb - Dc\|_2 \leq \varepsilon_{\text{fit}}, \quad e^T b = 1, \quad \|s - x\|_2 \leq \varepsilon_w, \quad \|h - g\|_2 \leq \varepsilon_{\text{gp}} \\
\text{linearize} & \\
\min_{b, c, s, h} & \quad \|c\|_1 \quad \text{s.t.} \quad \|h - g\|_2 \leq \varepsilon_{\text{gp}}, \\
& \quad e^T b = 1, \quad \left\| \begin{bmatrix} S^T & B^T & -H^T & -C^T \end{bmatrix} \begin{bmatrix} b - b' \\ s \\ c - c' \\ h \end{bmatrix} \right\|_2 \leq \varepsilon_{\text{fit}} \\
\text{solve} & \\
\min_{b, c, s, h} & \quad \|s - x\|_2 + \gamma_c \|c\|_1 + \gamma_b \|b\|_1, \quad \text{s.t.} \\
& \quad e^T b = 1, \quad \left\| \begin{bmatrix} S^T & B^T & -H^T & -C^T \end{bmatrix} \begin{bmatrix} b - b' \\ s \\ c - c' \\ h \end{bmatrix} \right\|_2 \leq \varepsilon_{\text{gp}}
\end{align*}
\]
results

![Graph 1](Error Amplitude vs Sample Index)

- **Error Amplitude**
- Sample Index: 20, 40, 60, 80, 100, 120, 140, 160
- **Least-Squares Clean Speech**
- **Least-Squares Noisy Speech**
- **Estimated Speech**

![Graph 2](Pulse Amplitude vs Sample Index)

- **Pulse Amplitude**
- Sample Index: 10, 20, 30, 40, 50, 60
- **Codebook g**
- **Estimated g**

![Graph 3](Coefficient Amplitude vs Coefficient Index)

- **Coefficient Amplitude**
- Coefficient Index: 50, 100, 150, 200
- **Codebook g (Initialization)**
- **Estimated g**
contributions

- joint estimation of linear predictor and excitation parameters
- convex optimization (efficient solutions available)
- parameter $\varepsilon_{\text{fit}}$ trades sparsity for accuracy
- glottal pulse shape exploited through sparse dictionary representation of prediction error
- estimate glottal pulse
- estimate signal when noise is present
- technique can be applied whenever prediction residual is structured