Sinusoidal Modeling

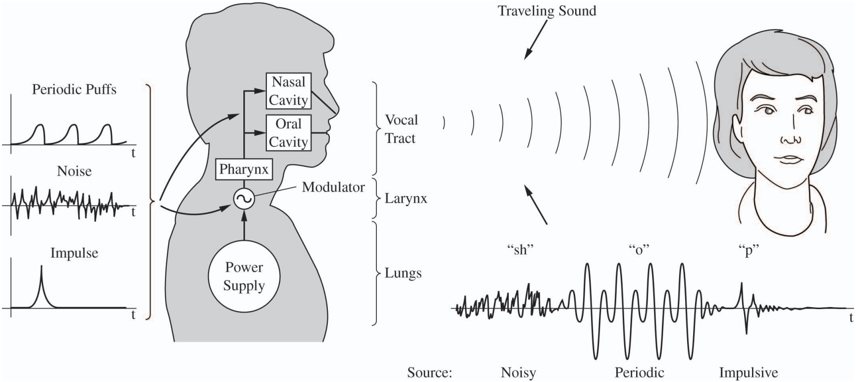
Yannis Stylianou

University of Crete, Computer Science Dept., Greece,

yannis@csd.uoc.gr

SPCC 2015

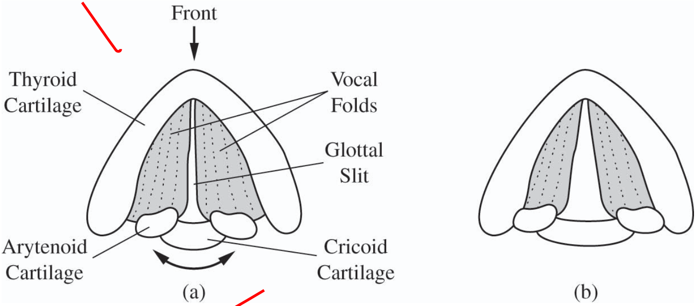
A Simple view



Used after permission: [1]

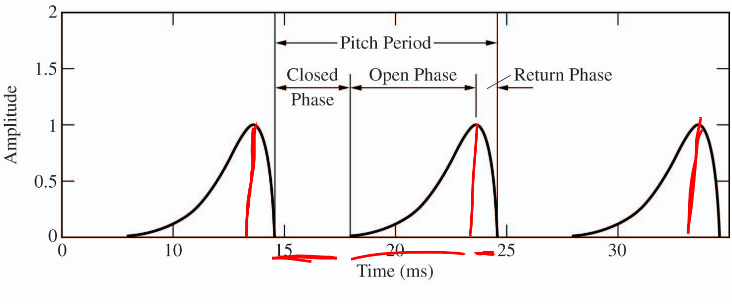
Downward-looking into the larynx: Vocal folds

Left: Voicing, **Right:** Breathing

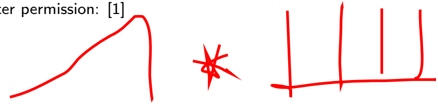


Used after permission: [1]

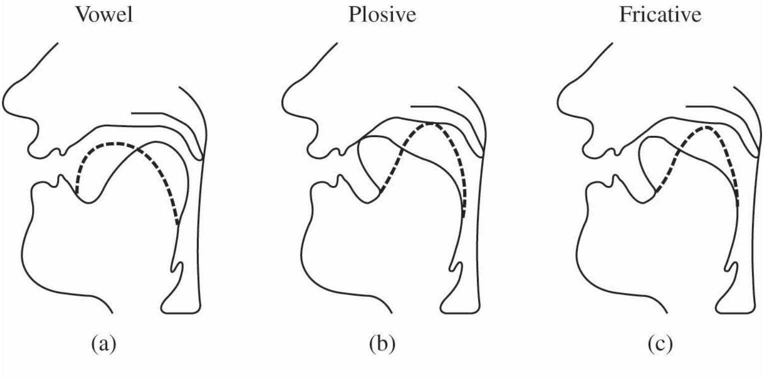
Glottal airflow velocity



Used after permission: [1]



Vocat tract shapes



Used after permission: [1]

Example

Let the excitation of vocal tract, $h[n]$, be:

$$u[n] = g[n] \star p[n]$$

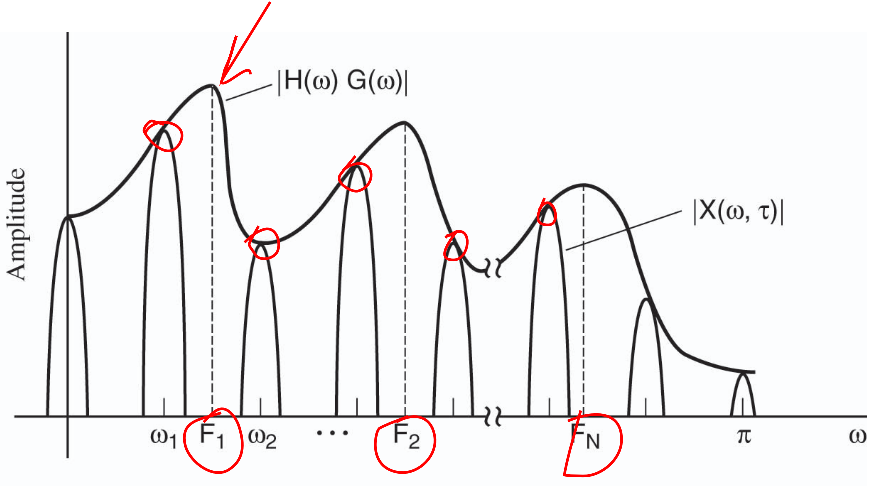
then, the output speech, $x[n, \tau]$, is given by:

$$x[n, \tau] = w[n, \tau]\{h[n] \star (g[n] \star p[n])\}$$

and

$$X(\omega, \tau) = \frac{1}{P} \sum_{k=-\infty}^{\infty} H(\omega_k) G(\omega_k) W(\omega - \omega_k, \tau)$$

Harmonics and Formants



Used after permission: [1]

Sinusoidal Models

- Sinusoidal Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi f_k t}$$

- Special case, Harmonic Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi k f_0 t}$$

- Estimation of parameters (Linear approach):

$$\gamma = \mathcal{F}_{f_k} \mathbf{s}$$

- Model Evaluation, Mean-Squared Error (MSE):

$$\epsilon = \int_{w} |s(t) - x(t)|^2 \, dt$$

Sinusoidal Models

- Sinusoidal Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi f_k t}$$

- Special case, Harmonic Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi k f_0 t}$$

- Estimation of parameters (Linear approach):

$$\gamma = \mathcal{F}_{f_k} \mathbf{s}$$

- Model Evaluation, Mean-Squared Error (MSE):

$$\epsilon = \int_w |s(t) - x(t)|^2 \, dt$$

Sinusoidal Models

- Sinusoidal Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi f_k t}$$

- Special case, Harmonic Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi k f_0 t}$$

- Estimation of parameters (Linear approach):

$$\gamma = \mathcal{F}_{f_k} \mathbf{s}$$

- Model Evaluation, Mean-Squared Error (MSE):

$$\epsilon = \int_{w} |s(t) - x(t)|^2 \, dt$$

Sinusoidal Models

- Sinusoidal Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi f_k t}$$

- Special case, Harmonic Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi k f_0 t}$$

- Estimation of parameters (Linear approach):

$$\gamma = \mathcal{F}_{f_k} \mathbf{s}$$

- Model Evaluation, Mean-Squared Error (MSE):

$$\epsilon = \int_{w} |s(t) - x(t)|^2 \, dt$$

Source-Filter with Sinusoids

- Source:

$$u(t) = \sum_{k=-K(t)}^{K(t)} \alpha_k(t) e^{j\phi_k(t)}$$

where:

$$\phi_k(t) = \int_0^t \Omega_k(\sigma) d\sigma + \phi_k$$

- Filter: $h(t, \tau)$ with Fourier Transform (FT):

$$H(t, \Omega) = M(t, \Omega) e^{j\Phi(t,\Omega)}$$

Source-Filter with Sinusoids

- Source:

$$u(t) = \sum_{k=-K(t)}^{K(t)} \alpha_k(t) e^{j\phi_k(t)}$$

where:

$$\phi_k(t) = \int_0^t \Omega_k(\sigma) d\sigma + \phi_k$$

- Filter: $h(t, \tau)$ with Fourier Transform (FT):

$$H(t, \Omega) = M(t, \Omega) e^{j\Phi(t,\Omega)}$$

Source-Filter with Sinusoids

$$s(t) = \sum_{k=1-K(t)}^{K(t)} A_k(t) e^{j\theta_k(t)}$$

where:

$$\begin{aligned} A_k(t) &= \alpha_k(t) M\left[t, \Omega_k(t)\right] \\ \theta_k(t) &= \phi_k(t) + \Phi\left[t, \Omega_k(t)\right] \\ &= \int_0^t \Omega_k(\sigma) d\sigma + \Phi\left[t, \Omega_k(t)\right] + \phi_k \end{aligned}$$

Sinusoidal Analysis: Stationarity Assumption

We assume stationarity inside the analysis window:

$$\begin{aligned} A_k(t) &= A_k \\ \Omega_k(t) &= \Omega_k \\ K(t) &= K \end{aligned}$$

which leads to:

$$\theta_k(t) = \Omega_k(t - t_l) + \theta_k$$

and to:

$$s(t) = \sum_{k=-K}^{K} A_k e^{j\theta_k} e^{j\Omega_k(t-t_l)} \quad t_l - \frac{T}{2} \leq t \leq t_l + \frac{T}{2}$$

In discrete time:

$$s[n] = \sum_{k=-K}^{K} \gamma_k e^{j\omega_k n} \quad -\frac{N_w - 1}{2} \leq n \leq \frac{N_w - 1}{2}$$

Sinusoidal Analysis: Stationarity Assumption

We assume stationarity inside the analysis window:

$$
\begin{aligned}
A_k(t) &= A_k \\
\Omega_k(t) &= \Omega_k \\
K(t) &= K
\end{aligned}
$$

which leads to:

$$\theta_k(t) = \Omega_k(t - t_l) + \theta_k$$

and to:

$$s(t) = \sum_{k=-K}^{K} A_k e^{j\theta_k} e^{j\Omega_k(t-t_l)} \quad t_l - \frac{T}{2} \leq t \leq t_l + \frac{T}{2}$$

In discrete time:

$$s[n] = \sum_{k=-K}^{K} \gamma_k e^{j\omega_k n} \quad -\frac{N_w - 1}{2} \leq n \leq \frac{N_w - 1}{2}$$

Sinusoidal Analysis: Stationarity Assumption

We assume stationarity inside the analysis window:

$$\begin{aligned} A_k(t) &= A_k \\ \Omega_k(t) &= \Omega_k \\ K(t) &= K \end{aligned}$$

which leads to:

$$\theta_k(t) = \Omega_k(t - t_l) + \theta_k$$

and to:

$$s(t) = \sum_{k=-K}^{K} A_k e^{j\theta_k} e^{j\Omega_k(t-t_l)} \quad t_l - \frac{T}{2} \leq t \leq t_l + \frac{T}{2}$$

In discrete time:

$$s[n] = \sum_{k=-K}^{K} \gamma_k e^{j\omega_k n} \quad -\frac{N_w - 1}{2} \leq n \leq \frac{N_w - 1}{2}$$

Sinusoidal Analysis: Stationarity Assumption

We assume stationarity inside the analysis window:

$$
\begin{aligned}
A_k(t) &= A_k \\
\Omega_k(t) &= \Omega_k \\
K(t) &= K
\end{aligned}
$$

which leads to:

$$\theta_k(t) = \Omega_k(t - t_l) + \theta_k$$

and to:

$$s(t) = \sum_{k=-K}^{K} A_k e^{j\theta_k} e^{j\Omega_k(t-t_l)} \quad t_l - \frac{T}{2} \leq t \leq t_l + \frac{T}{2}$$

In discrete time:

$$s[n] = \sum_{k=-K}^{K} \gamma_k e^{j\omega_k n} \quad -\frac{N_w - 1}{2} \leq n \leq \frac{N_w - 1}{2}$$

Mean-Squared Error

Given the original measured waveform, $y[n]$ and the synthetic speech waveform, $s[n]$, estimate the unknown parameters A_k^l, ω_k^l, and θ_k^l by minimizing the MSE criterion:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n] - s[n]|^2$$

which can be written as:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n]|^2 + N_w \sum_{k=1}^{K^l} \left(\left| Y(\omega_k^l) - \gamma_k^l \right|^2 - |Y(\omega_k^l)|^2 \right)$$

which can be reduced further to:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n]|^2 - N_w \sum_{k=1}^{K^l} |Y(\omega_k^l)|^2$$

Mean-Squared Error

Given the original measured waveform, $y[n]$ and the synthetic speech waveform, $s[n]$, estimate the unknown parameters A_k^l, ω_k^l, and θ_k^l by minimizing the MSE criterion:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n] - s[n]|^2$$

which can be written as:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n]|^2 + N_w \sum_{k=1}^{K^l} \left(\left| Y(\omega_k^l) - \gamma_k^l \right|^2 - |Y(\omega_k^l)|^2 \right)$$

which can be reduced further to:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n]|^2 - N_w \sum_{k=1}^{K^l} |Y(\omega_k^l)|^2$$

Mean-Squared Error

Given the original measured waveform, $y[n]$ and the synthetic speech waveform, $s[n]$, estimate the unknown parameters A_k^l, ω_k^l, and θ_k^l by minimizing the MSE criterion:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n] - s[n]|^2$$

which can be written as:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n]|^2 + N_w \sum_{k=1}^{K^l} \left(\left| Y(\omega_k^l) - \gamma_k^l \right|^2 - |Y(\omega_k^l)|^2 \right)$$

which can be reduced further to:

$$\epsilon^l = \sum_{n=-(N_w-1)/2}^{n=(N_w-1)/2} |y[n]|^2 - N_w \sum_{k=1}^{K^l} |Y(\omega_k^l)|^2$$

Karhunen-Loève Expansion

- Karhunen-Loève expansion allows constructing a random process from harmonic sinusoids with uncorrelated complex amplitudes.
- Estimated power spectrum should not vary "too much" over consecutive frequencies.

Following the above necessary constraints, for unvoiced speech, and for a window width to be *at least* 20ms, an 100 Hz harmonic structure provides good results.

Karhunen-Loève Expansion

- Karhunen-Loève expansion allows constructing a random process from harmonic sinusoids with uncorrelated complex amplitudes.
- Estimated power spectrum should not vary “too much” over consecutive frequencies.

Following the above necessary constraints, for unvoiced speech, and for a window width to be *at least* 20ms, an 100 Hz harmonic structure provides good results.

Karhunen-Loève Expansion

- Karhunen-Loève expansion allows constructing a random process from harmonic sinusoids with uncorrelated complex amplitudes.
- Estimated power spectrum should not vary “too much” over consecutive frequencies.

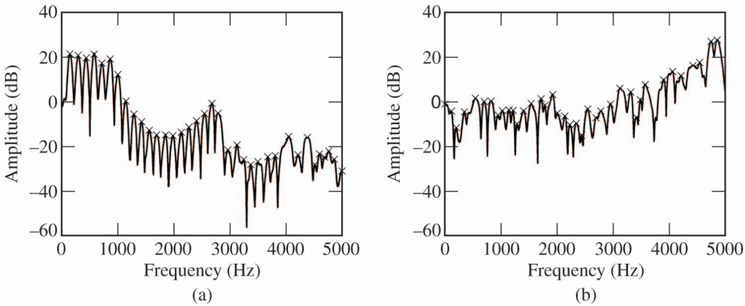
Following the above necessary constraints, for unvoiced speech, and for a window width to be *at least* 20ms, an 100 Hz harmonic structure provides good results.

Karhunen-Loève Expansion

- Karhunen-Loève expansion allows constructing a random process from harmonic sinusoids with uncorrelated complex amplitudes.
- Estimated power spectrum should not vary "too much" over consecutive frequencies.

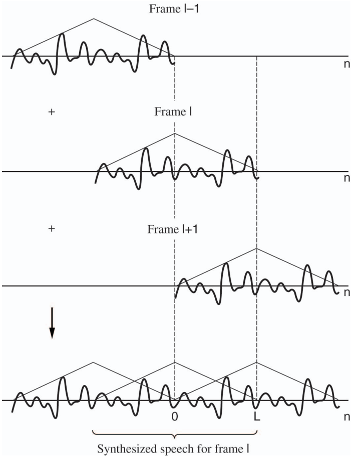
Following the above necessary constraints, for unvoiced speech, and for a window width to be *at least* 20ms, an 100 Hz harmonic structure provides good results.

Sinusoidal Analysis: Peak picking

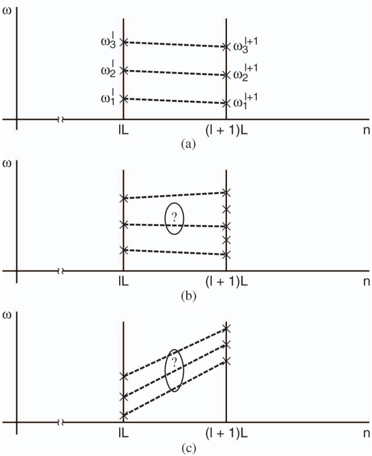


Used after permission: [1]

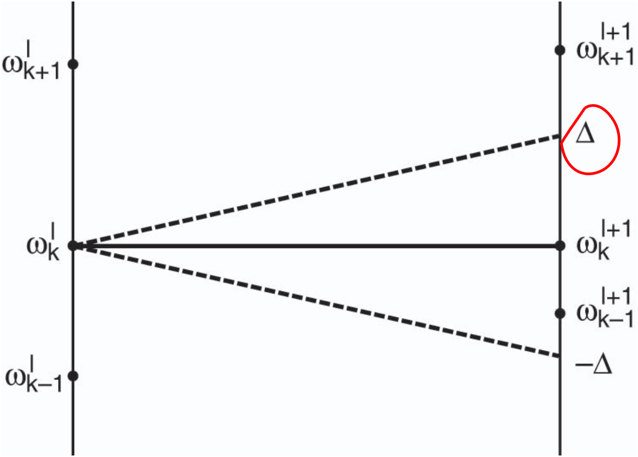
Sinusoidal Synthesis: OLA, a simple solution



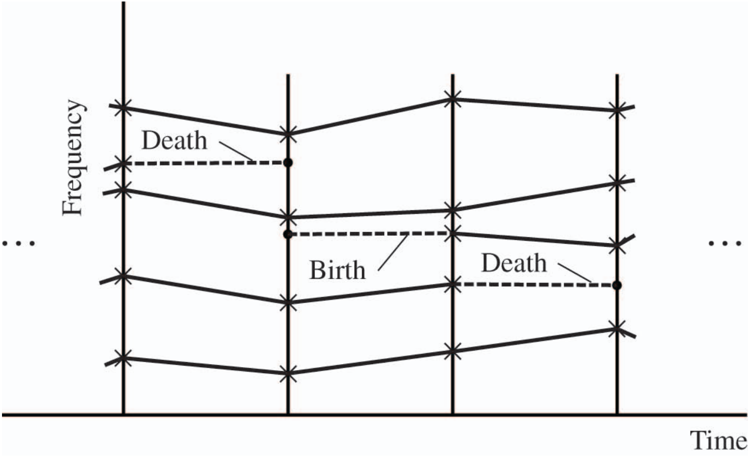
Problem of frequency matching



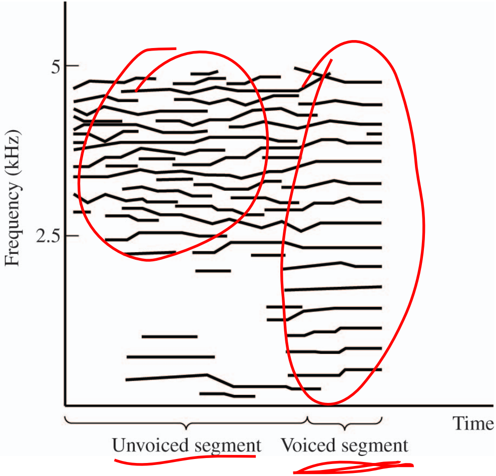
Frame-to-Frame Peak Matching



The birth/death process



A birth/death process in speech



Amplitude and Phase Interpolation

- Linear amplitude interpolation model:

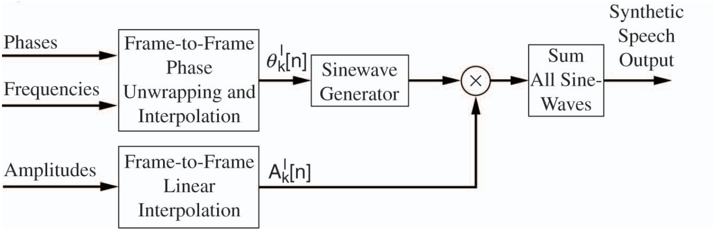
$$A_k^l[n] = A_k^l + \left(A_k^{l+1} - A_k^l\right)\left(\frac{n}{L}\right) \quad n = 0, 1, 2, \cdots, L-1$$

- Cubic Phase interpolation model

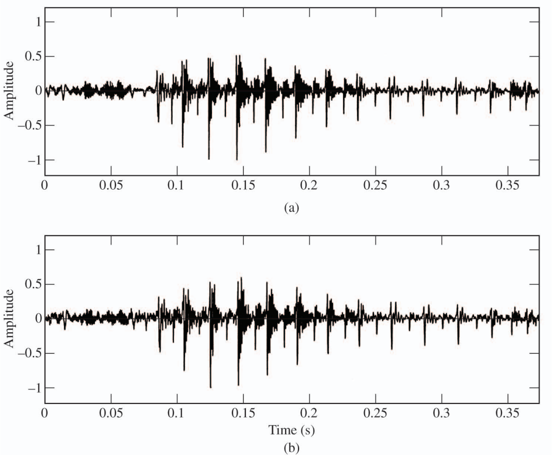
$$\theta(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3$$

$+ 2\pi M$

Block diagram of the Synthesis System



Synthesis: Reconstruction Example



Used after permission: [1]

Harmonic model

- Sinusoidal Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi f_k t}$$

- Special case, Harmonic Model:

$$x(t) = \sum_{k=-K}^{K} \gamma_k e^{j2\pi k f_0 t}$$

- Estimation of parameters (Linear approach):

$$\gamma = \mathcal{F}_{f_k} \mathbf{s}$$

- Model Evaluation, Mean-Squared Error (MSE):

$$\epsilon = \int_{w} |s(t) - x(t)|^2 \, dt$$

Brief overview of HNM

- HNM is a pitch-synchronous harmonic plus noise representation of the speech signal.
- Speech spectrum is divided into a low and a high band delimited by the so-called maximum voiced frequency.
- The low band of the spectrum (below the maximum voiced frequency) is represented solely by harmonically related sine waves.
- The upper band is modeled as a noise component modulated by a time-domain amplitude envelope.
- HNM allows high-quality copy synthesis and prosodic modifications.

HNM IN EQUATIONS

- Harmonic part:

$$h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j\, k\omega_0(t)\, t}$$

- Noise part:

$$n(t) = e(t)\left[v(\tau, t) \star b(t)\right]$$

- Speech:

$$s(t) = h(t) + n(t)$$

Periodic part

- HNM$_1$: Sum of exponential functions without slope

$$h_1[n] = \sum_{k=-L(n_a^i)}^{L(n_a^i)} a_k(n_a^i) e^{j2\pi k f_0(n_a^i)(n-n_a^i)}$$

- HNM$_2$: Sum of exponential function with complex slope

$$h_2[n] = \Re\left\{ \sum_{k=1}^{L(n_a^i)} A_k(n) \exp^{j2\pi k f_0(n_a^i)(n-n_a^i)} \right\}$$

where

$$A_k(n) = a_k(n_a^i) + (n - n_a^i) b_k(n_a^i)$$

with $a_k(n_a^i), b_k(n_a^i)$ to be complex numbers (amplitude and slope respectively). $\Re$ denotes taking the real part.

PERIODIC PART

- HNM$_1$: Sum of exponential functions without slope

$$h_1[n] = \sum_{k=-L(n_a^i)}^{L(n_a^i)} a_k(n_a^i) e^{j2\pi k f_0(n_a^i)(n-n_a^i)}$$

- HNM$_2$: Sum of exponential function with complex slope

$$h_2[n] = \Re \left\{ \sum_{k=1}^{L(n_a^i)} A_k(n) \exp^{j2\pi k f_0(n_a^i)(n-n_a^i)} \right\}$$

where

$$A_k(n) = a_k(n_a^i) + (n - n_a^i) b_k(n_a^i)$$

with $a_k(n_a^i), b_k(n_a^i)$ to be complex numbers (amplitude and slope respectively). $\Re$ denotes taking the real part.

PERIODIC PART *continuing*

- HNM_3: Sum of sinusoids with time-varying real amplitudes

$$h_3[n] = \sum_{k=0}^{L(n_a^i)} a_k(n) cos(\varphi_k(n))$$

where

$$\begin{aligned} a_k(n) &= c_{k0} + c_{k1}(n - n_a^i)^1 + \cdots + c_{kp}(n - n_a^i)^{p(n)} \\ \varphi_k(n) &= \epsilon_k + 2\pi\, k\zeta\,(n - n_a^i) \end{aligned}$$

where $p(n)$ is the order of the amplitude polynomial, which is, in general, a time-varying parameter.

Residual (Noise) part

The non-periodic part is just the *residual* signal obtained by subtracting the periodic-part (harmonic part) from the original speech signal in the time-domain

$$r[n] = s[n] - h[n]$$

where $h[n]$ is either $h_1[n], h_2[n]$, or $h_3[n]$.

Amplitudes and phases estimation

Having f_0 estimated for voiced frames, amplitudes and phases are estimated by minimizing the criterion:

$$\epsilon = \sum_{n=n_a^i-N}^{n_a^i+N} w^2[n](s[n] - \hat{h}[n])^2$$

where $n_a^i = n_a^{i-1} + P(n_a^{i-1})$, and $P(n_a^{i-1})$ denotes the pitch period at n_a^{i-1}.

- for HNM_1 and HNM_2, this criterion has a quadratic form and is solved by inverting an over-determined system of linear equations.
- For HNM_3, however,a non-linear system of equations has to be solved.

The residual signal $r[n]$ is estimated by

$$\hat{r}[n] = s[n] - \hat{h}[n]$$

Amplitudes and phases estimation

Having f_0 estimated for voiced frames, amplitudes and phases are estimated by minimizing the criterion:

$$\epsilon = \sum_{n=n_a^i-N}^{n_a^i+N} w^2[n](s[n] - \hat{h}[n])^2$$

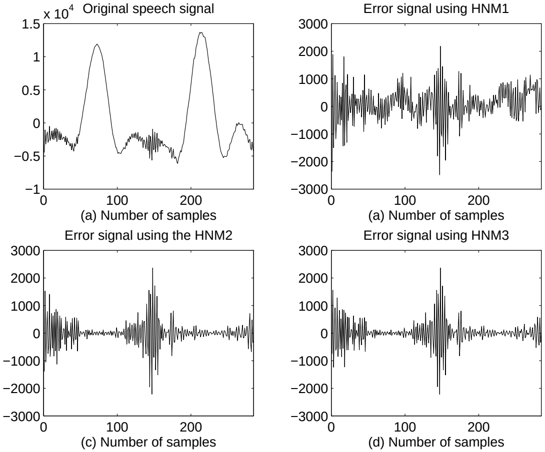
where $n_a^i = n_a^{i-1} + P(n_a^{i-1})$, and $P(n_a^{i-1})$ denotes the pitch period at n_a^{i-1}.

- for HNM_1 and HNM_2, this criterion has a quadratic form and is solved by inverting an over-determined system of linear equations.
- For HNM_3, however,a non-linear system of equations has to be solved.

The residual signal $r[n]$ is estimated by

$$\hat{r}[n] = s[n] - \hat{h}[n]$$

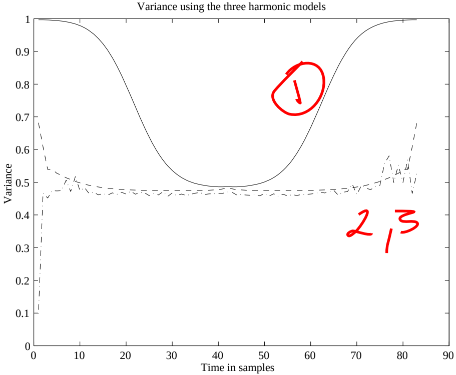
Time domain characteristics of $\hat{r}[n]$



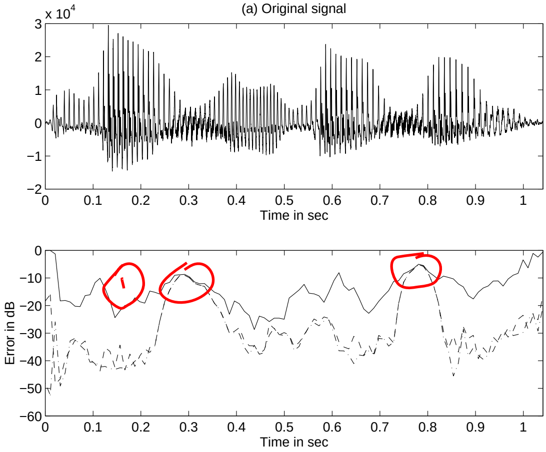
Variance of the residual signal

The variance of the residual signal is given as:

$$E(\mathbf{r}\mathbf{r}^h) = \mathbf{I} - \mathbf{W}\mathbf{P}(\mathbf{P}^h\mathbf{W}^h\mathbf{W}\mathbf{P})^{-1}\mathbf{P}^h\mathbf{W}^h$$



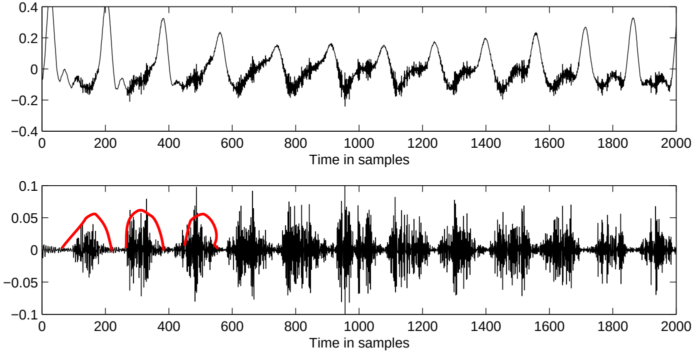
Modeling error



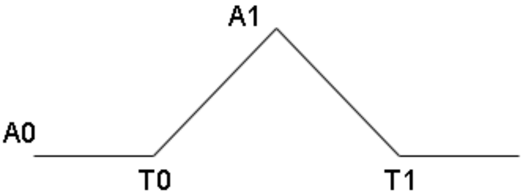
Modeling the residual signal

- Full bandwidth representation using a low-order (10th) AR filter
- Time-domain characteristics of the residual signal can be modeled using various functions

Time domain energy modulation



Triangular envelope



Signal envelope

There are many ways to obtain the "envelope" of a signal, as:

- Hilbert Transform (analytic signal)
- Low-pass local energy (energy envelope):

$$e[n] = \frac{1}{2N+1} \sum_{k=-N}^{N} |r[n-k]|$$

where $r[n]$ denotes the residual signal.

Signal envelope

There are many ways to obtain the "envelope" of a signal, as:

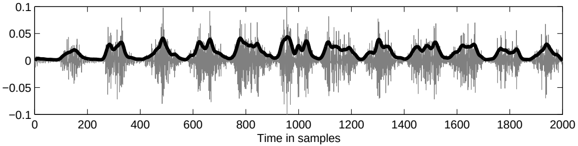
- Hilbert Transform (analytic signal)
- Low-pass local energy (energy envelope):

$$e[n] = \frac{1}{2N+1} \sum_{k=-N}^{N} |r[n-k]|$$

where $r[n]$ denotes the residual signal.

Example of energy envelope

Example of Energy Envelope, with $N = 7$

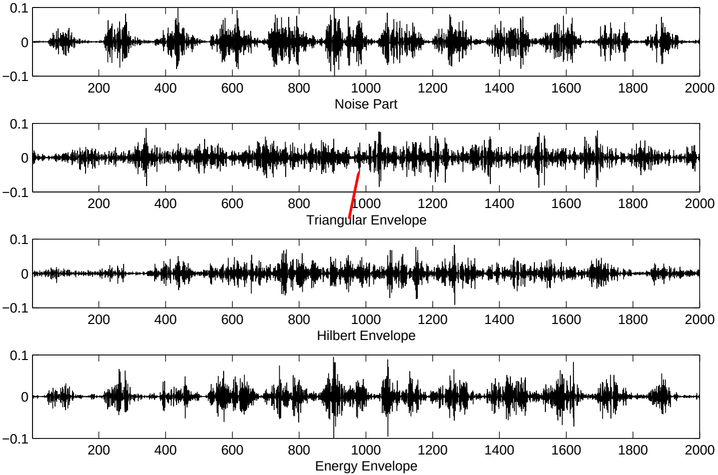


The energy envelope can be efficiently parameterized with a few Fourier coefficients:

$$\hat{e}[n] = \sum_{k=-L_e}^{L_e} A_k e^{j2\pi k(f_0/f_s)n}$$

where L_e is set to be 3 to 4

Comparing time-domain modulations



Sound Examples using HNM_1

Examples with HNM_1 and Maximum Voiced frequency fixed at 4000 Hz

	Original	Triangular	Hilbert	Energy
Male				
Male				
Female				
Female				

Sound Examples using HNM_2

Examples with HNM_2 and Maximum Voiced frequency fixed at 4000 Hz

	Original	Triangular	Hilbert	Energy
Male				
Male				
Female				
Female				

HNM$_2$

We recall that the HNM$_2$ is given by:

$$s(t) = \left(\sum_{k=-L}^{L} (a_k + t b_k) e^{2\pi j k f_0 t} \right) w(t)$$

or in frequency domain:

$$S(f) = \sum_{k=-L}^{L} \left(a_k W(f - k f_0) + j b_k W'(f - k f_0) \right)$$

where $W(f)$ is the Fourier Transform of window $w(t)$

Frequency Domain properties of HNM$_2$

Let $\vec{a}_k$ and $\vec{b}_k$ denote the vectors corresponding respectively to the complex a_k and b_k and

let's decompose $\vec{b}_k$ into two components:

- one collinear to $\vec{a}_k$, and
- one perpendicular to $\vec{a}_k$.

Thus, $\vec{b}_k$ is given by

$$\vec{b}_k = \rho_{1,k}\vec{a}_k + \rho_{2,k}\vec{a}_k^{\perp},$$

where

$$\begin{aligned} \vec{a}_k^{\perp} &= (-a_k^I, a_k^R)^T \\ \rho_{1,k} &= \frac{\langle \vec{a}_k, \vec{b}_k \rangle}{|\vec{a}_k|^2} \\ \rho_{2,k} &= \frac{\langle \vec{a}_k^{\perp}, \vec{b}_k \rangle}{|\vec{a}_k|^2} \end{aligned}$$

$\langle .,. \rangle$ denotes the inner product between two vectors.

Frequency Domain properties of HNM_2

Let $\vec{a}_k$ and $\vec{b}_k$ denote the vectors corresponding respectively to the complex a_k and b_k and
let's decompose $\vec{b}_k$ into two components:

- one collinear to $\vec{a}_k$, and
- one perpendicular to $\vec{a}_k$.

Thus, $\vec{b}_k$ is given by

$$\vec{b}_k = \rho_{1,k}\vec{a}_k + \rho_{2,k}\vec{a}_k^{\perp},$$

where

$$\vec{a}_k^{\perp} = (-a_k^I, a_k^R)^T$$

$$\rho_{1,k} = \frac{\langle \vec{a}_k, \vec{b}_k \rangle}{|\vec{a}_k|^2}$$

$$\rho_{2,k} = \frac{\langle \vec{a}_k^{\perp}, \vec{b}_k \rangle}{|\vec{a}_k|^2}$$

$\langle . , . \rangle$ denotes the inner product between two vectors.

LET'S LOOK AT THE k^{th} COMPONENT

- The *kth* component can be written as:

$$S_k(f) = a_k[W(f - kf_0) - \rho_{2,k} W'(f - kf_0) + j\rho_{1,k} W'(f - kf_0)]$$

- For small values of $\rho_{2,k}$, using a first order approximation of the Taylor series of $W(f)$, we have:

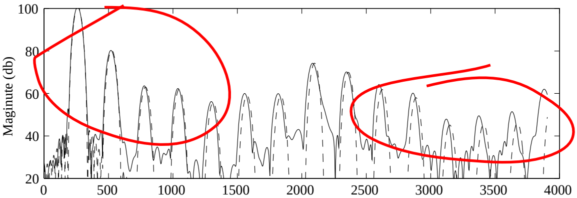
$$W(f - kf_0) - \rho_{2,k} W'(f - kf_0) \approx W(f - kf_0 - \rho_{2,k})$$

- and then:

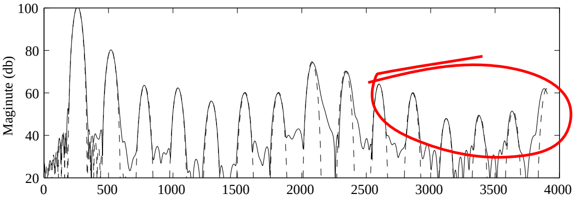
$$S_k(f) \approx a_k[W(f - kf_0 - \rho_{2,k}) + j\rho_{1,k} W'(f - kf_0)]$$

- Then, HNM$_2$ becomes Quasi-Harmonic Model (QHM).

SENSITIVITY ON f_0



(a) Frequency in Hz



(b) Frequency in Hz

T. F. Quatieri, *Discrete-Time Speech Signal Processing*.
Engewood Cliffs, NJ: Prentice Hall, 2002.

THANK YOU

for your attention on this part