Modeling of Speech Signal for Analysis Purposes

or Mathematical modeling of jitter and shimmer

Yannis Stylianou

University of Crete, Computer Science Dept., Multimedia Informatics Lab
yannis@csd.uoc.gr

Limsi, France, 2008 August 13th

Multimedia Informatics Lab

- 4 Professors:
  1. T. Mouchtaris (Audio and Speech Processing)
  2. Y. Stylianou (Speech and Signal Processing)
  3. P. Tsakalides (Signal Processing and Sensor Networks)
  4. G. Tziritas (Image and Video Processing)
- 3 Post Docs, 8 Ph.D. Students and many students in M.Sc. degree
- Strong connections with a Research Center: FORTH.

My current research topics

- Speech Processing
  - Voice Quality Assessment
  - Algorithms for Speech Pathology
  - Non-linear speech modeling and processing
  - Inverse Filtering
  - Voice Transformation
  - Multimodal User identification
- Music Signal Processing
- Marine Mammals Acoustics

Multimedia Informatics Lab

Selected Recent Projects:

- FP6-IST NoE SIMILAR: Human-computer interaction similar to the way humans do it.
- FP6-IST Strep PISTE: Personalized, Immersive Sports TV Experience
- FP6-Marie Curie TOK: Collaborative Signal Processing for Efficient Wireless Sensor Networks
- GSRT Wireless Sensor Networks: Theory and Applications in Structural Health Monitoring
- GSRT AKMON: Advanced Algorithms for Voice Quality Assessment
- GSRT TV++: Multimedia processing for Broadcast News

Industrial Partners: France Telecom, British Telecom, FORTH-Net

Modeling

Modeling for ...

- Coding
- Modifications
- Synthesis and Analysis

MODELING

Modeling for ...

- Coding
- Modifications
- Synthesis and Analysis

Modeling

Modeling for ...

- Coding
- Modifications
- Synthesis and Analysis

Synthesis: Jitter and Shimmer

Defining Jitter and Shimmer

Definition (Jitter)

Jitter is defined as perturbations of the glottal source signal that occur during vowel phonation and affect the glottal pitch period.

Definition (Shimmer)

Shimmer is defined as perturbations of the glottal source signal that occur during vowel phonation and affect the glottal energy.

Defining Jitter and Shimmer

Definition (Jitter)

Jitter is defined as perturbations of the glottal source signal that occur during vowel phonation and affect the glottal pitch period.

Definition (Shimmer)

Shimmer is defined as perturbations of the glottal source signal that occur during vowel phonation and affect the glottal energy.

Some Estimators ...

Let $u[n]$ be a pitch period sequence.
Absolute jitter:

$$\frac{1}{N-1}\sum_{n=1}^{N-1}|u(n+1)-u(n)|$$

Let $u[n]$ be a peak amplitude sequence of N samples.
Absolute Shimmer:

$$\frac{1}{N-1}\sum_{n=1}^{N-1}|u(n+1)-u(n)|$$

Some Estimators ...

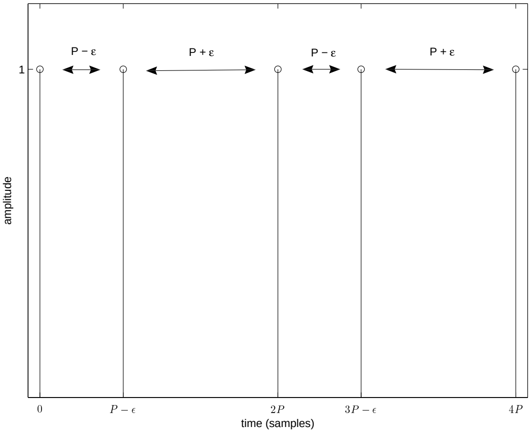
Let $u[n]$ be a pitch period sequence.
Absolute jitter:

$$\frac{1}{N-1}\sum_{n=1}^{N-1}|u(n+1)-u(n)|$$

Let $u[n]$ be a peak amplitude sequence of N samples.
Absolute Shimmer:

$$\frac{1}{N-1}\sum_{n=1}^{N-1}|u(n+1)-u(n)|$$

Jitter: Aperiodicity through periodicity[1]



In mathematical terms

- We model the glottal impulse train as:

$$p[n] = \sum_{k=-\infty}^{+\infty} \delta[n-(2k)P] + \sum_{k=-\infty}^{+\infty} \delta[n+\epsilon-(2k+1)P]$$

- We may show that its power spectrum is then:

$$\begin{aligned} |P(\omega)|^2 &= \frac{2}{P^2}(1+\cos[(\epsilon-P)\omega])\left[\delta_{l\omega_0}(\omega)+\delta_{(l+\frac{1}{2})\omega_0}(\omega)\right] \\ &= H(\epsilon,\omega)+S(\epsilon,\omega) \end{aligned}$$

In mathematical terms

- We model the glottal impulse train as:

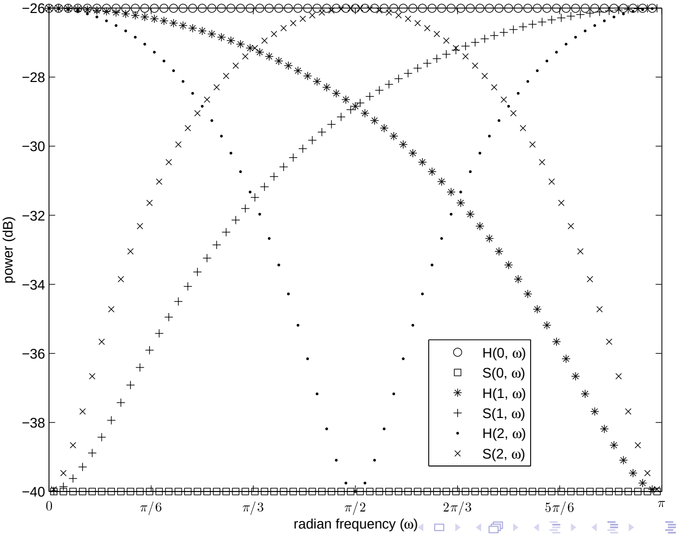
$$p[n] = \sum_{k=-\infty}^{+\infty} \delta[n-(2k)P] + \sum_{k=-\infty}^{+\infty} \delta[n+\epsilon-(2k+1)P]$$

- We may show that its power spectrum is then:

$$\begin{aligned} |P(\omega)|^2 &= \frac{2}{P^2}\left(1+\cos\left[(\epsilon-P)\omega\right]\right)\left[\delta_{l\omega_0}(\omega)+\delta_{(l+\frac{1}{2})\omega_0}(\omega)\right] \\ &= H(\epsilon,\omega)+S(\epsilon,\omega) \end{aligned}$$

Examples of power spectrum

On synthetic glottal signal



Experiments

Goal: discriminate pathological from normal voices, based on jitter

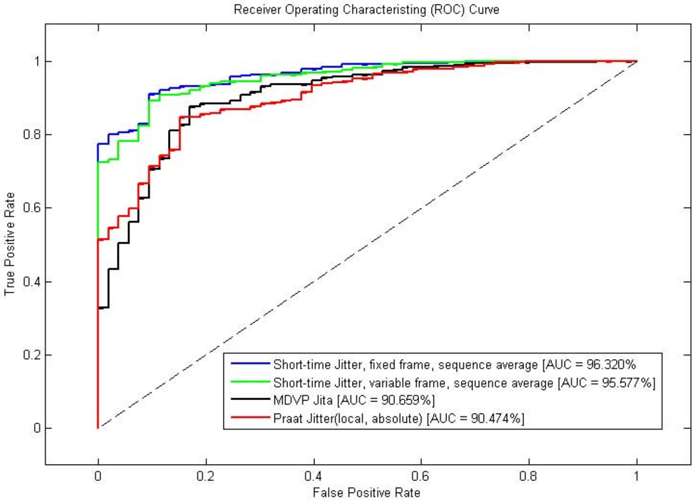
- Database: Massachusetts Eye and Ear Infirmary (MEEI) [2]
  - Sustained vowels,
  - 53 subjects with normal voice,
  - 657 subjects with a wide variety of pathological conditions
- Jitter estimation methods:
  - PRAAT2007 (P. Boersma and D. Weenink) [3]
  - Multi-Dimensional Voice Program (MDVP), (Kay-Pentax elemetrics, 2007) [4]
  - Our approach [1]

Experiments

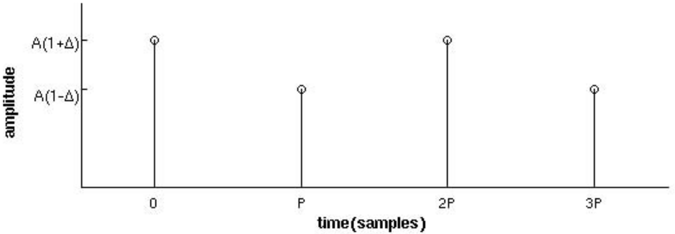
Goal: discriminate pathological from normal voices, based on jitter

- Database: Massachusetts Eye and Ear Infirmary (MEEI) [2]
  - Sustained vowels,
  - 53 subjects with normal voice,
  - 657 subjects with a wide variety of pathological conditions
- Jitter estimation methods:
  - PRAAT2007 (P. Boersma and D. Weenink) [3]
  - Multi-Dimensional Voice Program (MDVP), (Kay-Pentax elemetrics, 2007) [4]
  - Our approach [1]

Results in ROC curves



Shimmer: Aperiodicity through periodicity[1]



In mathematical terms

- We model the glottal impulse train as:

$$g[n] = A(1+\Delta)\delta_{(2k)P}[n] + A(1-\Delta)\delta_{(2k+1)P}[n]$$

- We may show that its Fourier Transform is then:

$$G(\omega) = A\left[(1+\Delta) + (1-\Delta)e^{-j2\pi\frac{\omega}{\omega_0}}\right]\frac{\omega_0}{4\pi}\sum_{k=-\infty}^{+\infty}\delta(\omega - k\frac{\omega_0}{2})$$

- Splitting

$$\begin{aligned} G(l\omega_0) &= A\frac{\omega_0}{2\pi} \\ G((l+1/2)\omega_0) &= A\frac{\omega_0}{2\pi}\Delta \end{aligned}$$

In mathematical terms

- We model the glottal impulse train as:

$$g[n] = A(1+\Delta)\delta_{(2k)P}[n] + A(1-\Delta)\delta_{(2k+1)P}[n]$$

- We may show that its Fourier Transform is then:

$$G(\omega) = A\left[(1+\Delta) + (1-\Delta)e^{-j2\pi\frac{\omega}{\omega_0}}\right]\frac{\omega_0}{4\pi}\sum_{k=-\infty}^{+\infty}\delta(\omega - k\frac{\omega_0}{2})$$

- Splitting

$$\begin{aligned} G(l\omega_0) &= A\frac{\omega_0}{2\pi} \\ G((l+1/2)\omega_0) &= A\frac{\omega_0}{2\pi}\Delta \end{aligned}$$

In mathematical terms

- We model the glottal impulse train as:

$$g[n] = A(1+\Delta)\delta_{(2k)P}[n] + A(1-\Delta)\delta_{(2k+1)P}[n]$$

- We may show that its Fourier Transform is then:

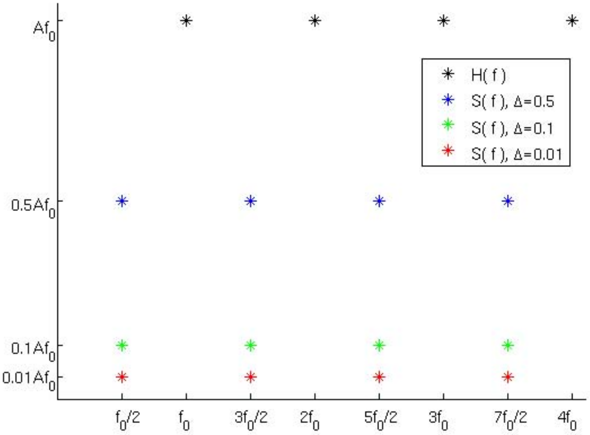
$$G(\omega) = A\left[(1+\Delta) + (1-\Delta)e^{-j2\pi\frac{\omega}{\omega_0}}\right]\frac{\omega_0}{4\pi}\sum_{k=-\infty}^{+\infty}\delta(\omega - k\frac{\omega_0}{2})$$

- Splitting

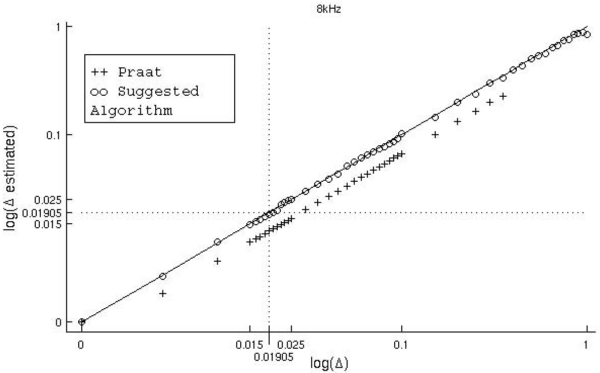
$$\begin{aligned} G(l\omega_0) &= A\frac{\omega_0}{2\pi} \\ G((l+1/2)\omega_0) &= A\frac{\omega_0}{2\pi}\Delta \end{aligned}$$

Examples of spectrum

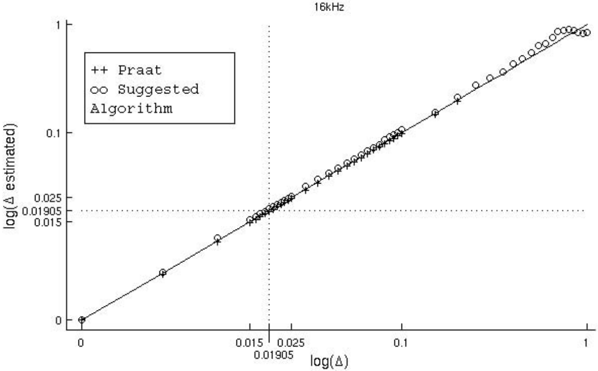
On synthetic glottal signal



Experiment at 8kHz

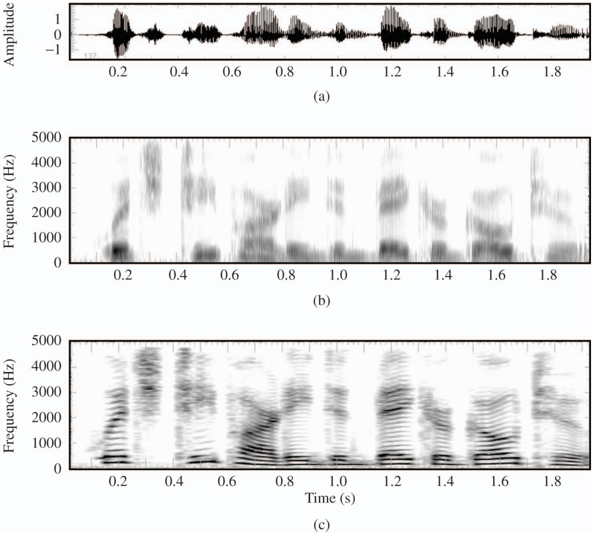


Experiment at 16kHz

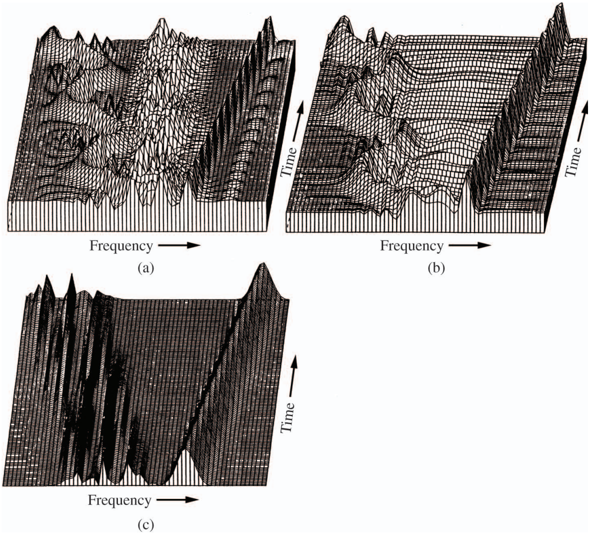


Analysis: Jitter and Shimmer

Short-Time Fourier Transform



Time-Frequency Distributions [5]



Modeling the Periodic part of Speech

- Sum of simple exponential functions

$$h_1(t) = \Re\left\{\sum_{k=1}^{L} a_k e^{j2\pi k \frac{f_0}{f_s} t}\right\}$$

- Sum of exponential functions with complex slope (HNM_2[6])

$$h_2(t) = \Re\left\{\sum_{k=1}^{L} A_k(t) \exp^{j2\pi k \frac{f_0}{f_s} t}\right\}$$

where

$$A_k(t) = a_k + t\, b_k$$

Modeling the Periodic part of Speech

- Sum of simple exponential functions

$$h_1(t) = \Re\left\{\sum_{k=1}^{L} a_k e^{j2\pi k \frac{f_0}{f_s} t}\right\}$$

- Sum of exponential functions with complex slope (HNM_2[6])

$$h_2(t) = \Re\left\{\sum_{k=1}^{L} A_k(t) \exp^{j2\pi k \frac{f_0}{f_s} t}\right\}$$

where

$$A_k(t) = a_k + t\, b_k$$

REVISITING HNM$_2$

We recall that the periodic part of HNM$_2$ is given by:

$$s(t) = \left(\sum_{k=-L}^{L} A_k(t) e^{2\pi jkf_0 t} \right) w(t)$$

with $A_k(t) = a_k + tb_k$, or in frequency domain:

$$S(f) = \sum_{k=-L}^{L} \left(a_k W(f - kf_0) + jb_k W'(f - kf_0) \right)$$

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

Time-Domain Properties of HNM_2

- Instantaneous Amplitude:

$$m_k(t) = |a_k + tb_k| = \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2}$$

- Instantaneous Phase:

$$\begin{aligned}\phi_k(t) &= 2\pi k f_0 t + \angle(a_k + tb_k) \\ &= 2\pi k f_0 t + atan\frac{a_k^I + tb_k^I}{a_k^R + tb_k^R}\end{aligned}$$

- Instantaneous Frequency:

$$\begin{aligned}f_k(t) &= \frac{1}{2\pi}\phi_k'(t) \\ &= kf_0 + \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{m_k^2(t)}\end{aligned}$$

Time-Domain Properties of HNM_2

- Instantaneous Amplitude:

$$m_k(t) = |a_k + tb_k| = \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2}$$

- Instantaneous Phase:

$$\begin{aligned} \phi_k(t) &= 2\pi k f_0 t + \angle(a_k + tb_k) \\ &= 2\pi k f_0 t + atan\frac{a_k^I + tb_k^I}{a_k^R + tb_k^R} \end{aligned}$$

- Instantaneous Frequency:

$$\begin{aligned} f_k(t) &= \frac{1}{2\pi}\phi_k'(t) \\ &= kf_0 + \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{m_k^2(t)} \end{aligned}$$

Time-Domain Properties of HNM_2

- Instantaneous Amplitude:

$$m_k(t) = |a_k + tb_k| = \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2}$$

- Instantaneous Phase:

$$\begin{aligned} \phi_k(t) &= 2\pi k f_0 t + \angle(a_k + tb_k) \\ &= 2\pi k f_0 t + atan\frac{a_k^I + tb_k^I}{a_k^R + tb_k^R} \end{aligned}$$

- Instantaneous Frequency:

$$\begin{aligned} f_k(t) &= \frac{1}{2\pi}\phi_k'(t) \\ &= kf_0 + \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{m_k^2(t)} \end{aligned}$$

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

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

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

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

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:

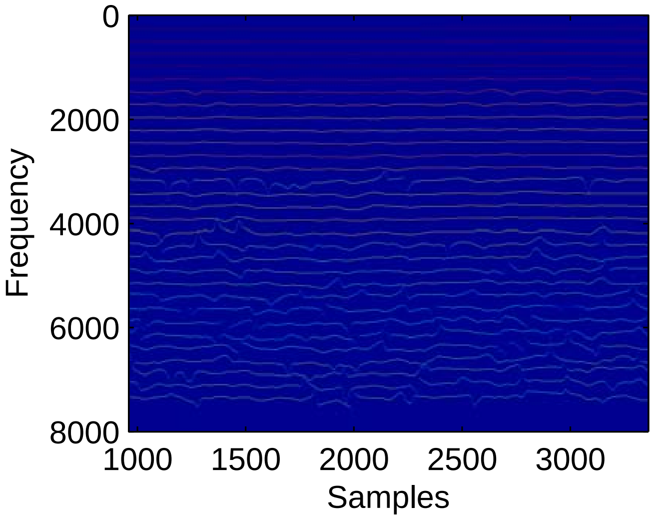
$$W(f-kf_0)-\rho_{2,k}W'(f-kf_0) \approx W(f-kf_0-\rho_{2,k})$$

- and then:

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

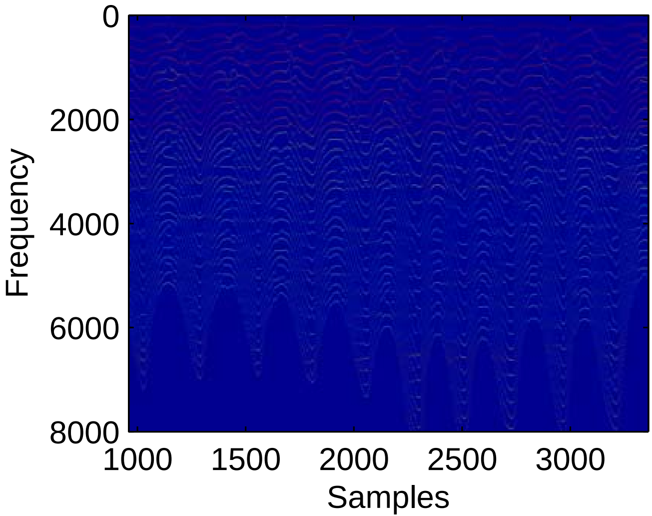
Time-Frequency Analysis using HNM_2

Healthy voice



TIME-FREQUENCY ANALYSIS USING HNM_2

PATHOLOGIC VOICE



Sinusoidal model

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) cos[\theta_k(t)]$$

where

$$A_k(t) = \underbrace{a_k(t)}_{\text{excitation}} \cdot \underbrace{M_k(t)}_{\text{vocal track}}$$

and

$$\theta_k(t) = \underbrace{\phi_k(t)}_{\text{excitation}} + \underbrace{\Phi_k(t)}_{\text{vocal track}}$$

$$\phi_k(t) = 2\pi k \int_0^t f_0(\tau) d\tau + \phi_k$$

Jitter and Shimmer

Jitter:

$$f_0(t) = f_0 - \delta sin(\pi f_0 t)$$

Shimmer:

$$a_k(t) = a_k[1 + \gamma_k \, cos(\pi f_0 t)]$$

so then:

$$s(t) = \sum_{k=-K}^{K} A_k[1 + \gamma_k cos(\pi f_0 t)]e^{j(2\pi k f_0 t + \delta_k cos(\pi f_0 t) + \theta_k)} w(t)$$

and by writing: $e^{j\delta_k cos(\pi f_0 t)} \approx 1 + j\delta_k cos(\pi f_0 t)$, then:

$$s(t) \approx \sum_{k=-K}^{K} A_k e^{j\theta_k}[1 + (\gamma_k + j\delta_k) cos(\pi f_0 t)]e^{j2\pi k f_0 t} w(t)$$

Jitter and Shimmer

Jitter:

$$f_0(t) = f_0 - \delta sin(\pi f_0 t)$$

Shimmer:

$$a_k(t) = a_k[1 + \gamma_k cos(\pi f_0 t)]$$

so then:

$$s(t) = \sum_{k=-K}^{K} A_k[1 + \gamma_k cos(\pi f_0 t)] e^{j(2\pi k f_0 t + \delta_k cos(\pi f_0 t) + \theta_k)} w(t)$$

and by writing: $e^{j\delta_k cos(\pi f_0 t)} \approx 1 + j\delta_k cos(\pi f_0 t)$, then:

$$s(t) \approx \sum_{k=-K}^{K} A_k e^{j\theta_k}[1 + (\gamma_k + j\delta_k) cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

Jitter and Shimmer

Jitter:

$$f_0(t) = f_0 - \delta sin(\pi f_0 t)$$

Shimmer:

$$a_k(t) = a_k[1 + \gamma_k \, cos(\pi f_0 t)]$$

so then:

$$s(t) = \sum_{k=-K}^{K} A_k[1 + \gamma_k cos(\pi f_0 t)] e^{j(2\pi k f_0 t + \delta_k cos(\pi f_0 t) + \theta_k)} w(t)$$

and by writing: $e^{j\delta_k cos(\pi f_0 t)} \approx 1 + j\delta_k cos(\pi f_0 t)$, then:

$$s(t) \approx \sum_{k=-K}^{K} A_k e^{j\theta_k}[1 + (\gamma_k + j\delta_k) cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

Jitter and Shimmer in HNM$_2$

Suggesting:

$$s(t) = \sum_{k=-K}^{K} [a_k + b_k cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

and by letting $b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k$, then:

$$s(t) = \sum_{k=-K}^{K} a_k [1 + (\rho_{1,k} + j\rho_{2,k}) cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

comparing to what we would like to have:

$$s(t) \approx \sum_{k=-K}^{K} A_k e^{j\theta_k} [1 + (\gamma_k + j\delta_k) cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

Jitter and Shimmer in HNM_2

Suggesting:

$$s(t) = \sum_{k=-K}^{K} [a_k + b_k cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

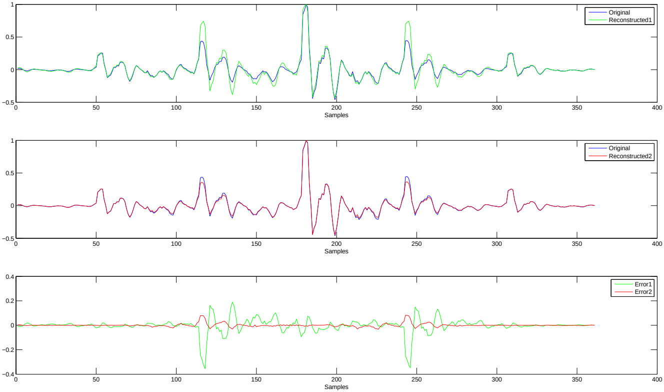
and by letting $b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k$, then:

$$s(t) = \sum_{k=-K}^{K} a_k [1 + (\rho_{1,k} + j\rho_{2,k}) cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

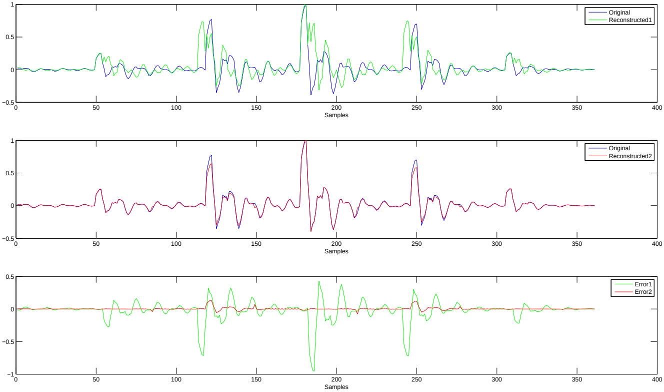
comparing to what we would like to have:

$$s(t) \approx \sum_{k=-K}^{K} A_k e^{j\theta_k} [1 + (\gamma_k + j\delta_k) cos(\pi f_0 t)] e^{j2\pi k f_0 t} w(t)$$

Modeling Shimmer



Modeling Jitter



Acknowledgments

- I wish to thank T. Quatieri and Prentice Hall for gave me the permission to use figures from Tom's book[7].
- My students, Miltos Vassilakis and Yannis Pantazis for their work on jitter and shimmer and on HNM_2.

M. Vasilakis and Y. Stylianou, "A mathematical model for accurate measurement of jitter," in *MAVEBA 2007*, (Florence, Italy), 2007.

K. Elemetrics, "Disordered Voice Database (Version 1.03)," 1994.

P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.6.24) [Computer program]," 2007.

K. Elemetrics, "Multi-Dimensional Voice Program (MDVP) [Computer program]," 2007.

L. Cohen, *Time-Frequency Analysis*.
Englewood Cliffs, NJ: Prentice-Hall, 1995.

Y. Stylianou, "Modeling speech based on harmonic plus noise models.," in *Nonlinear Speech Modeling and Aplications* (G. Chellot, A. Esposito, M. Faundez, and M. M, eds.), pp. 244–260, Springer-Verlag, 2005.

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