CS578 - Speech Signal Processing

Lecture : Harmonic Models of Speech

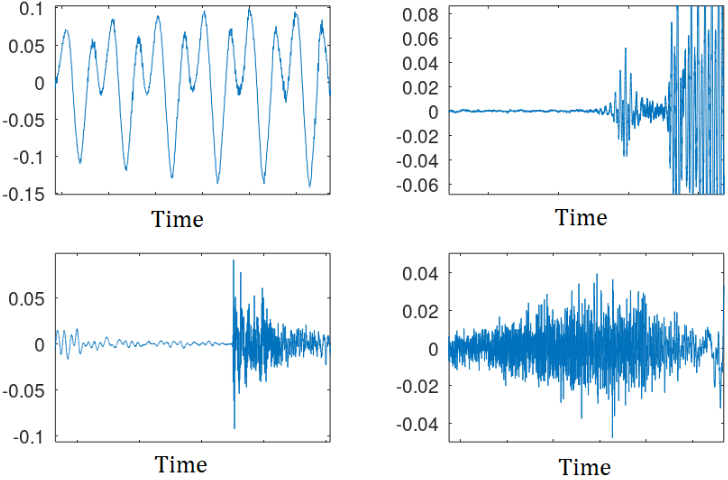
George P. Kafentzis

University of Crete, Computer Science Dept., Speech Signal Processing Lab
kafentz@csd.uoc.gr

Univ. of Crete

Outline

Motivation



Are they similar??

Motivation

- Standard sinusoidal model (McAulay and Quatieri, 1986 [1]) treats all speech components equally
- Voiced frames: sum of sinusoids
- Unvoiced frames: sum of sinusoids (under the Karhunen-Loeve expansion assumption)
- Is it the best way to treat them?
- **Decomposition!**

Motivation

- Standard sinusoidal model (McAulay and Quatieri, 1986 [1]) treats all speech components equally
- Voiced frames: sum of sinusoids
- Unvoiced frames: sum of sinusoids (under the Karhunen-Loeve expansion assumption)
- Is it the best way to treat them?
- **Decomposition!**

Motivation

- Standard sinusoidal model (McAulay and Quatieri, 1986 [1]) treats all speech components equally
- Voiced frames: sum of sinusoids
- Unvoiced frames: sum of sinusoids (under the Karhunen-Loeve expansion assumption)
- Is it the best way to treat them?
- **Decomposition!**

Motivation

- Standard sinusoidal model (McAulay and Quatieri, 1986 [1]) treats all speech components equally
- Voiced frames: sum of sinusoids
- Unvoiced frames: sum of sinusoids (under the Karhunen-Loeve expansion assumption)
- Is it the best way to treat them?
- **Decomposition!**

Motivation

- Standard sinusoidal model (McAulay and Quatieri, 1986 [1]) treats all speech components equally
- Voiced frames: sum of sinusoids
- Unvoiced frames: sum of sinusoids (under the Karhunen-Loeve expansion assumption)
- Is it the best way to treat them?
- **Decomposition!**

Outline

Background

Mentioning just a few works for speech analysis...

BACKGROUND

- Multi-Band Excitation Vocoder (Griffin et al.1988 [2])
  - $S(\omega) = H(\omega)E(\omega)$
  - $E(\omega)$ is represented by an f_0, a V/UV decision for each harmonic, and the phase of each voiced harmonic
  - Parameters are estimated by comparing the original vs the synthetic speech spectrum
  - Voiced portion is synthesized in time domain while unvoiced part is synthesized in frequency domain

Background

- Multi-Band Excitation Vocoder (Griffin et al.1988 [2])
  - $S(\omega) = H(\omega)E(\omega)$
  - $E(\omega)$ is represented by an f_0, a V/UV decision for each harmonic, and the phase of each voiced harmonic
  - Parameters are estimated by comparing the original vs the synthetic speech spectrum
  - Voiced portion is synthesized in time domain while unvoiced part is synthesized in frequency domain

Background

- Multi-Band Excitation Vocoder (Griffin et al.1988 [2])
  - $S(\omega) = H(\omega)E(\omega)$
  - $E(\omega)$ is represented by an f_0, a V/UV decision for each harmonic, and the phase of each voiced harmonic
  - Parameters are estimated by comparing the original vs the synthetic speech spectrum
  - Voiced portion is synthesized in time domain while unvoiced part is synthesized in frequency domain

Background

- Multi-Band Excitation Vocoder (Griffin et al.1988 [2])
  - $S(\omega) = H(\omega)E(\omega)$
  - $E(\omega)$ is represented by an f_0, a V/UV decision for each harmonic, and the phase of each voiced harmonic
  - Parameters are estimated by comparing the original vs the synthetic speech spectrum
  - Voiced portion is synthesized in time domain while unvoiced part is synthesized in frequency domain

Background

- Multi-Band Excitation Vocoder (Griffin et al.1988 [2])
  - $S(\omega) = H(\omega)E(\omega)$
  - $E(\omega)$ is represented by an f_0, a V/UV decision for each harmonic, and the phase of each voiced harmonic
  - Parameters are estimated by comparing the original vs the synthetic speech spectrum
  - Voiced portion is synthesized in time domain while unvoiced part is synthesized in frequency domain

Background

Multi-band Excitation Vocoder (Griffin et al.1988 [2])

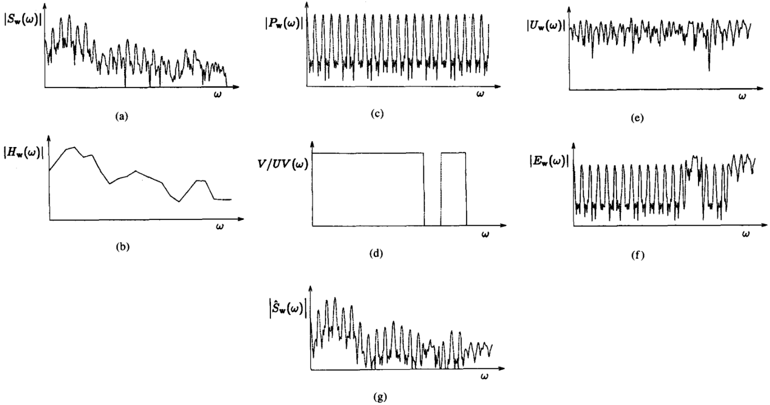


Figure: *(a) Original Spectrum, (b) Spectral Envelope, (c) Periodic Spectrum, (d) V/UV information, (e) Noise Spectrum, (f) Excitation Spectrum, (g) Synthetic Spectrum*

Background

- Sinusoids + band-pass random signals (Abrantes et al.1991 [3])
  - Completely avoids V/UV decision
  - Harmonically related sinusoids model the voiced parts
  - Random band-pass signals model the unvoiced parts
    - White noise filtered by a group of band-pass filters (filterbank) with center frequencies $k\omega_s$
  - $s(t) = \sum_{k=1}^{N_p} a_k(t) \cos(\phi_k(t)) + \sum_{l=1}^{N_r} b_k(t) \epsilon_k(t) \cos(k\omega_s t + \theta_k)$

- Sinusoids + band-pass random signals (Abrantes et al.1991 [3])
  - Completely avoids V/UV decision
  - Harmonically related sinusoids model the voiced parts
  - Random band-pass signals model the unvoiced parts
    - White noise filtered by a group of band-pass filters (filterbank) with center frequencies $k\omega_s$
  - $s(t) = \sum_{k=1}^{N_p} a_k(t)\cos(\phi_k(t)) + \sum_{l=1}^{N_r} b_k(t)\epsilon_k(t)\cos(k\omega_s t + \theta_k)$

Background

- Sinusoids + band-pass random signals (Abrantes et al.1991 [3])
  - Completely avoids V/UV decision
  - Harmonically related sinusoids model the voiced parts
  - Random band-pass signals model the unvoiced parts
    - White noise filtered by a group of band-pass filters (filterbank) with center frequencies $k\omega_s$
  - $s(t) = \sum_{k=1}^{N_p} a_k(t)\cos(\phi_k(t)) + \sum_{l=1}^{N_r} b_k(t)\epsilon_k(t)\cos(k\omega_s t + \theta_k)$

Background

- Sinusoids + band-pass random signals (Abrantes et al.1991 [3])
  - Completely avoids V/UV decision
  - Harmonically related sinusoids model the voiced parts
  - Random band-pass signals model the unvoiced parts
    - White noise filtered by a group of band-pass filters (filterbank) with center frequencies $k\omega_s$
  - $s(t) = \sum_{k=1}^{N_p} a_k(t)\cos(\phi_k(t)) + \sum_{l=1}^{N_r} b_k(t)\epsilon_k(t)\cos(k\omega_s t + \theta_k)$

Background

- Sinusoids + band-pass random signals (Abrantes et al.1991 [3])
  - Completely avoids V/UV decision
  - Harmonically related sinusoids model the voiced parts
  - Random band-pass signals model the unvoiced parts
    - White noise filtered by a group of band-pass filters (filterbank) with center frequencies $k\omega_s$
  - $s(t) = \sum_{k=1}^{N_p} a_k(t) \cos(\phi_k(t)) + \sum_{l=1}^{N_r} b_k(t) \epsilon_k(t) \cos(k\omega_s t + \theta_k)$

Background

- Sinusoids + band-pass random signals (Abrantes et al.1991 [3])
  - Completely avoids V/UV decision
  - Harmonically related sinusoids model the voiced parts
  - Random band-pass signals model the unvoiced parts
    - White noise filtered by a group of band-pass filters (filterbank) with center frequencies $k\omega_s$
  - $s(t) = \sum_{k=1}^{N_p} a_k(t)\cos(\phi_k(t)) + \sum_{l=1}^{N_r} b_k(t)\epsilon_k(t)\cos(k\omega_s t + \theta_k)$

Background

Sinusoids + band-pass random signals (Abrantes et al.1991 [3])

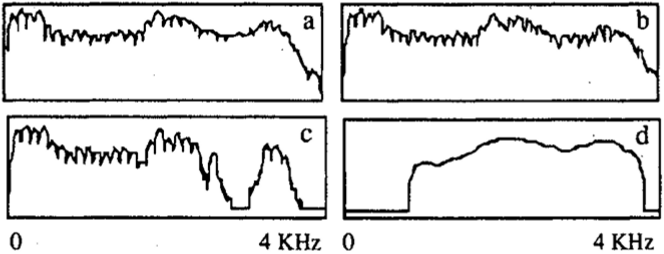


Figure: *(a) Original Spectrum, (b) Hybrid model output, (c) Periodic and (d) random components.*

Background

- Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])
    - The LP residual signal is used as an approximation to the excitation of the vocal tract
    - V/UV analysis is used
    - Frequency regions of harmonic and noise components in the spectral domain are recognized
    - An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions
    - The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal.

- Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])
  - The LP residual signal is used as an approximation to the excitation of the vocal tract
  - V/UV analysis is used
  - Frequency regions of harmonic and noise components in the spectral domain are recognized
  - An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions
  - The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal.

Background

- Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])
  - The LP residual signal is used as an approximation to the excitation of the vocal tract
  - V/UV analysis is used
  - Frequency regions of harmonic and noise components in the spectral domain are recognized
  - An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions
  - The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal.

Background

- Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])
  - The LP residual signal is used as an approximation to the excitation of the vocal tract
  - V/UV analysis is used
  - Frequency regions of harmonic and noise components in the spectral domain are recognized
  - An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions
  - The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal.

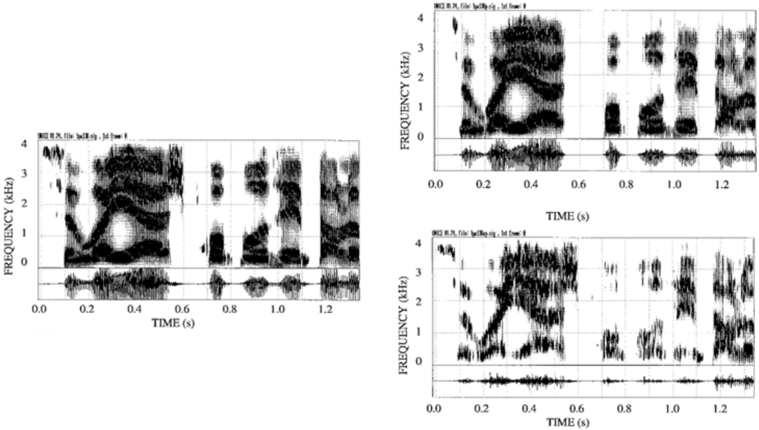
Background

- Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])
  - The LP residual signal is used as an approximation to the excitation of the vocal tract
  - V/UV analysis is used
  - Frequency regions of harmonic and noise components in the spectral domain are recognized
  - An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions
  - The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal.

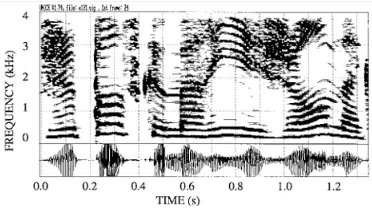
Background

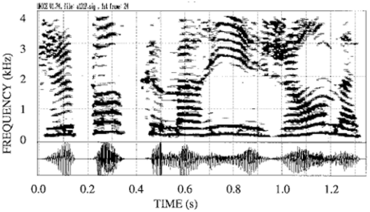
- Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])
  - The LP residual signal is used as an approximation to the excitation of the vocal tract
  - V/UV analysis is used
  - Frequency regions of harmonic and noise components in the spectral domain are recognized
  - An iterative algorithm is proposed which reconstructs the aperiodic component in the harmonic regions
  - The periodic component is obtained by subtracting the reconstructed aperiodic component signal from the residual signal.

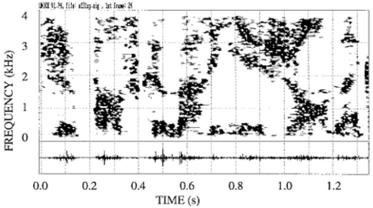
Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])



Periodic + Aperiodic Decomposition (Yegnayarayana et al.1995 [4])







Why decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Why decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Why decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

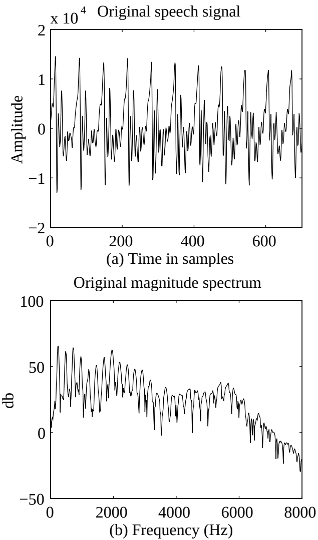
Why decompose?

Decomposing speech into (quasi)periodic and non-periodic part has many applications in:

- Speech modification
- Speech coding
- Pathologic voice detection (i.e., HNR ...)
- Psychoacoustic research

Outline

Motivation for HNM



Brief overview of HNM

- HNM (Stylianou 1995 [5]) 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 *lower* 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.

Brief overview of HNM

- HNM (Stylianou 1995 [5]) 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 *lower* 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.

Brief overview of HNM

- HNM (Stylianou 1995 [5]) 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 *lower* 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.

Brief overview of HNM

- HNM (Stylianou 1995 [5]) 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 *lower* 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.

Brief overview of HNM

- HNM (Stylianou 1995 [5]) 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 *lower* 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:

$$h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j2\pi k f_0(t) t}$$

  where $A_k(t)$ and $f_0(t)$ are the instantaneous complex amplitude and real frequency, respectively

- Noise part:

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

  where $e(t), v(\tau, t), g(t)$ are a time envelope, an estimation of the PSD (filter), and white gaussian noise, respectively

- Speech:

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

HNM in equations

- Harmonic part:

$$h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j2\pi k f_0(t) t}$$

  where $A_k(t)$ and $f_0(t)$ are the instantaneous complex amplitude and real frequency, respectively

- Noise part:

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

  where $e(t), v(\tau, t), g(t)$ are a time envelope, an estimation of the PSD (filter), and white gaussian noise, respectively

- Speech:

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

HNM in equations

- Harmonic part:

$$h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j2\pi k f_0(t) t}$$

  where $A_k(t)$ and $f_0(t)$ are the instantaneous complex amplitude and real frequency, respectively

- Noise part:

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

  where $e(t), v(\tau, t), g(t)$ are a time envelope, an estimation of the PSD (filter), and white gaussian noise, respectively

- Speech:

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

Models for periodic part

- HNM_1: Sum of exponential functions without slope

$$h_1(t) = \sum_{k=-L(t_a^i)}^{L(t_a^i)} a_k(t_a^i) e^{j2\pi k f_0(t_a^i)(t-t_a^i)}$$

- HNM_2: Sum of exponential function with complex slope

$$h_2(t) = \sum_{k=-L(t_a^i)}^{L(t_a^i)} A_k(t) e^{j2\pi k f_0(t_a^i)(t-t_a^i)}$$

where

$$A_k(t) = a_k(t_a^i) + (t - t_a^i) b_k(t_a^i)$$

with $a_k(t_a^i), b_k(t_a^i)$ to be complex numbers (amplitude and slope respectively).

Models for periodic part

- HNM_1: Sum of exponential functions without slope

$$h_1(t) = \sum_{k=-L(t_a^i)}^{L(t_a^i)} a_k(t_a^i) e^{j2\pi k f_0(t_a^i)(t-t_a^i)}$$

- HNM_2: Sum of exponential function with complex slope

$$h_2(t) = \sum_{k=-L(t_a^i)}^{L(t_a^i)} A_k(t) e^{j2\pi k f_0(t_a^i)(t-t_a^i)}$$

where

$$A_k(t) = a_k(t_a^i) + (t - t_a^i) b_k(t_a^i)$$

with $a_k(t_a^i), b_k(t_a^i)$ to be complex numbers (amplitude and slope respectively).

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

$$h_3(t) = \sum_{k=0}^{L(t_a^i)} a_k(t)\cos(\varphi_k(t))$$

where

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

where $a_k(t)$, $\phi_k(t)$ are real functions of discrete time and $p(t)$ 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

$$r(t) = s(t) - h(t)$$

where $h(t)$ is either $h_1(t), h_2(t)$, or $h_3(t)$ (harmonic part of HNM_1, HNM_2, and HNM_3, respectively).

Outline

Initial fundamental frequency

- Get an initial estimation of fundamental frequency f_0 [6]
- Determine the voicing of the frame using normalized error over first four harmonics:

$$E = \frac{\int_{0.7f_0}^{4.3f_0} (|S(f)| - |\tilde{S}(f)|)^2}{\int_{0.7f_0}^{4.3f_0} |S(f)|^2}$$

  where $\tilde{S}(f)$ is a synthetic DFT-based spectrum using the initial f_0 estimation
- If $E < T$, where T an appropriate threshold (e.g. -15 dB), then frame is voiced, else it is labeled as unvoiced

Initial fundamental frequency

- Get an initial estimation of fundamental frequency f_0 [6]
- Determine the voicing of the frame using normalized error over first four harmonics:

$$E = \frac{\displaystyle\int_{0.7f_0}^{4.3f_0} (|S(f)| - |\tilde{S}(f)|)^2}{\displaystyle\int_{0.7f_0}^{4.3f_0} |S(f)|^2}$$

  where $\tilde{S}(f)$ is a synthetic DFT-based spectrum using the initial f_0 estimation
- If $E < T$, where T an appropriate threshold (e.g. -15 dB), then frame is voiced, else it is labeled as unvoiced

Initial fundamental frequency

- Get an initial estimation of fundamental frequency f_0 [6]
- Determine the voicing of the frame using normalized error over first four harmonics:

$$E = \frac{\int_{0.7f_0}^{4.3f_0} (|S(f)| - |\tilde{S}(f)|)^2}{\int_{0.7f_0}^{4.3f_0} |S(f)|^2}$$

  where $\tilde{S}(f)$ is a synthetic DFT-based spectrum using the initial f_0 estimation
- If $E < T$, where T an appropriate threshold (e.g. -15 dB), then frame is voiced, else it is labeled as unvoiced

Maximum Voiced Frequency - MVF

- The MVF F_M is determined frame-wise from the speech spectrum
- Starting from the frequency f_c of the *maximum* spectral peak, $A(f_c)$, in $[f_0/2, 3f_0/2]$, spectral peak values are collected around that maximum peak, along with their frequencies
- The range of collection is $R_{search} = [f_c - f_0/2,\ f_c + f_0/2]$
- Determine peak frequencies f_i in R_{search}, and the corresponding amplitudes, $A(f_i)$ and cumulative amplitudes $A_c(f_i)$
- Cumulative amplitude $A_c(f)$ is the sum of all spectral peak values from previous valley to following valley

Maximum Voiced Frequency - MVF

- The MVF F_M is determined frame-wise from the speech spectrum
- Starting from the frequency f_c of the *maximum* spectral peak, $A(f_c)$, in $[f_0/2, 3f_0/2]$, spectral peak values are collected around that maximum peak, along with their frequencies
- The range of collection is $R_{search} = [f_c - f_0/2,\ f_c + f_0/2]$
- Determine peak frequencies f_i in R_{search}, and the corresponding amplitudes, $A(f_i)$ and cumulative amplitudes $A_c(f_i)$
- Cumulative amplitude $A_c(f)$ is the sum of all spectral peak values from previous valley to following valley

Maximum Voiced Frequency - MVF

- The MVF F_M is determined frame-wise from the speech spectrum
- Starting from the frequency f_c of the *maximum* spectral peak, $A(f_c)$, in $[f_0/2, 3f_0/2]$, spectral peak values are collected around that maximum peak, along with their frequencies
- The range of collection is $R_{search} = [f_c - f_0/2,\ f_c + f_0/2]$
- Determine peak frequencies f_i in R_{search}, and the corresponding amplitudes, $A(f_i)$ and cumulative amplitudes $A_c(f_i)$
- Cumulative amplitude $A_c(f)$ is the sum of all spectral peak values from previous valley to following valley

Maximum Voiced Frequency - MVF

- The MVF F_M is determined frame-wise from the speech spectrum
- Starting from the frequency f_c of the *maximum* spectral peak, $A(f_c)$, in $[f_0/2, 3f_0/2]$, spectral peak values are collected around that maximum peak, along with their frequencies
- The range of collection is $R_{search} = [f_c - f_0/2,\ f_c + f_0/2]$
- Determine peak frequencies f_i in R_{search}, and the corresponding amplitudes, $A(f_i)$ and cumulative amplitudes $A_c(f_i)$
- Cumulative amplitude $A_c(f)$ is the sum of all spectral peak values from previous valley to following valley

Maximum Voiced Frequency - MVF

- The MVF F_M is determined frame-wise from the speech spectrum
- Starting from the frequency f_c of the *maximum* spectral peak, $A(f_c)$, in $[f_0/2, 3f_0/2]$, spectral peak values are collected around that maximum peak, along with their frequencies
- The range of collection is $R_{search} = [f_c - f_0/2,\ f_c + f_0/2]$
- Determine peak frequencies f_i in R_{search}, and the corresponding amplitudes, $A(f_i)$ and cumulative amplitudes $A_c(f_i)$
- Cumulative amplitude $A_c(f)$ is the sum of all spectral peak values from previous valley to following valley

Maximum Voiced Frequency - MVF

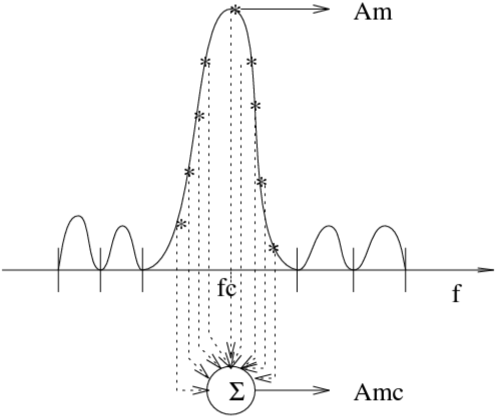


Fig. 1. Cumulative amplitude definition

Maximum Voiced Frequency - MVF

- Compute the average cumulative amplitude for all f_i: $\bar{A}_c(f_i)$
- Pass f_c through the **Voicing Test**:
  - If

$$\frac{A_c(f_c)}{\bar{A}_c(f_i)} > 2$$

    or

$$|A(f_c) - \max\{A(f_i)\}| > 13 \text{ dB}$$

    then

    - if f_c is *really* close to the closest harmonic lf_0, then
    - declare f_c as voiced frequency. Otherwise, declare f_c as unvoiced frequency.

Maximum Voiced Frequency - MVF

- Compute the average cumulative amplitude for all f_i: $\bar{A}_c(f_i)$
- Pass f_c through the **Voicing Test**:
  - If
    $$\frac{A_c(f_c)}{\bar{A}_c(f_i)} > 2$$
    or
    $$|A(f_c) - \max\{A(f_i)\}| > 13 \text{ dB}$$
    then
    - if f_c is *really* close to the closest harmonic lf_0, then
    - declare f_c as voiced frequency. Otherwise, declare f_c as unvoiced frequency.

Maximum Voiced Frequency - MVF

- Compute the average cumulative amplitude for all f_i: $\bar{A}_c(f_i)$
- Pass f_c through the **Voicing Test**:
  - If

$$\frac{A_c(f_c)}{\bar{A}_c(f_i)} > 2$$

    or

$$|A(f_c) - \max\{A(f_i)\}| > 13 \text{ dB}$$

    then

    - if f_c is *really* close to the closest harmonic lf_0, then
    - declare f_c as voiced frequency. Otherwise, declare f_c as unvoiced frequency.

Maximum Voiced Frequency - MVF

- Compute the average cumulative amplitude for all f_i: $\bar{A}_c(f_i)$
- Pass f_c through the **Voicing Test**:
  - If
    $$\frac{A_c(f_c)}{\bar{A}_c(f_i)} > 2$$
    or
    $$|A(f_c) - \max\{A(f_i)\}| > 13 \text{ dB}$$
    then
    - if f_c is *really* close to the closest harmonic lf_0, then
    - declare f_c as voiced frequency. Otherwise, declare f_c as unvoiced frequency.

- Compute the average cumulative amplitude for all f_i: $\bar{A}_c(f_i)$
- Pass f_c through the **Voicing Test**:
    - If

$$\frac{A_c(f_c)}{\bar{A}_c(f_i)} > 2$$

    or

$$|A(f_c) - \max\{A(f_i)\}| > 13 \text{ dB}$$

    then

        - if f_c is *really* close to the closest harmonic lf_0, then
        - declare f_c as voiced frequency. Otherwise, declare f_c as unvoiced frequency.

Maximum Voiced Frequency

Then:

- Search for the maximum spectral peak in $[f_c + f_0/2,\ f_c + 3f_0/2]$, and find new f_c
- Repeat the steps until $f_c \leq f_s/2$.
- Determine voiced and unvoiced spectral areas
- Maximum voiced frequency M_F is the maximum frequency of the last voiced spectral area.

Maximum voiced frequency example

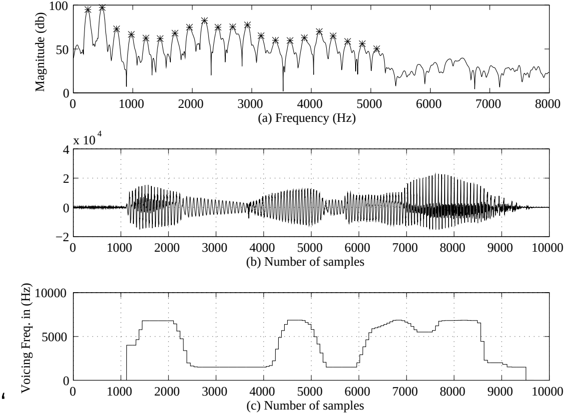


FIGURE: *(a) Maximum voiced frequency estimation for a voiced frame, (b) a voied speech segment, (c) maximum voiced frequency estimation for the voiced speech segment in (b).*

Fundamental frequency refinement

- Using the initial f_0 value and the L detected voiced frequencies f_i, then the refined fundamental frequency, $\hat{f}_0$ is defined as the value that minimizes the error:

$$E(\hat{f}_0) = \sum_{i=1}^{L} |f_i - i \cdot \hat{f}_0|^2$$

- Having the stream of pitch values, we set the *analysis time instants*, t_a^i, as functions of the local pitch period $P(t_a^i)$:
  - voiced frames: $t_a^{i+1} = t_a^i + P(t_a^i)$
  - unvoiced frames: fixed, 10 ms

Fundamental frequency refinement

- Using the initial f_0 value and the L detected voiced frequencies f_i, then the refined fundamental frequency, $\hat{f}_0$ is defined as the value that minimizes the error:

$$E(\hat{f}_0) = \sum_{i=1}^{L} |f_i - i \cdot \hat{f}_0|^2$$

- Having the stream of pitch values, we set the *analysis time instants*, t_a^i, as functions of the local pitch period $P(t_a^i)$:
  - voiced frames: $t_a^{i+1} = t_a^i + P(t_a^i)$
  - unvoiced frames: fixed, 10 ms

Fundamental frequency refinement

- Using the initial f_0 value and the L detected voiced frequencies f_i, then the refined fundamental frequency, $\hat{f}_0$ is defined as the value that minimizes the error:

$$E(\hat{f}_0) = \sum_{i=1}^{L} |f_i - i \cdot \hat{f}_0|^2$$

- Having the stream of pitch values, we set the *analysis time instants*, t_a^i, as functions of the local pitch period $P(t_a^i)$:
  - voiced frames: $t_a^{i+1} = t_a^i + P(t_a^i)$
  - unvoiced frames: fixed, 10 ms

Fundamental frequency refinement

- Using the initial f_0 value and the L detected voiced frequencies f_i, then the refined fundamental frequency, $\hat{f}_0$ is defined as the value that minimizes the error:

$$E(\hat{f}_0) = \sum_{i=1}^{L} |f_i - i \cdot \hat{f}_0|^2$$

- Having the stream of pitch values, we set the *analysis time instants*, t_a^i, as functions of the local pitch period $P(t_a^i)$:
  - voiced frames: $t_a^{i+1} = t_a^i + P(t_a^i)$
  - unvoiced frames: fixed, 10 ms

Refinement frequency example

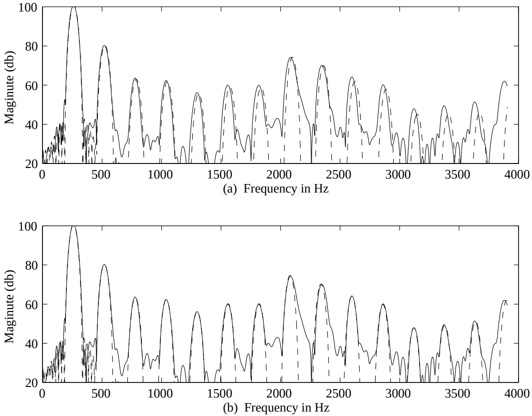


Figure: *(a) Original and synthetic spectrum for the initial pitch estimation, (b) Original and synthetic spectrum for the refined pitch value.*

Pitch Detection Algorithm

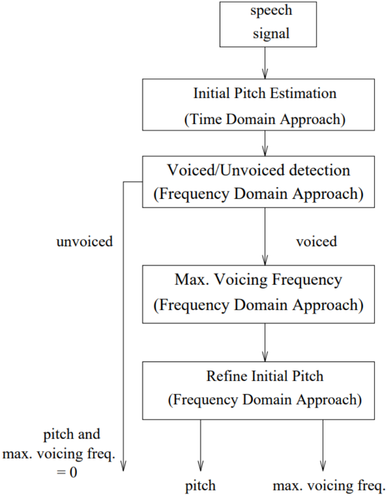


Figure: *The pitch analysis algorithm.*

Amplitudes and phases estimation

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

$$\epsilon = \sum_{n=t_a^i-T}^{t_a^i+T} \Big(w(t)(s(t)-\hat{h}(t))\Big)^2 = \sum_{n=t_a^i-T}^{t_a^i+T} w^2(t)(s(t)-\hat{h}(t))^2$$

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

Amplitudes and phases estimation

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

$$\epsilon = \sum_{n=t_a^i-T}^{t_a^i+T} \Big(w(t)(s(t)-\hat{h}(t))\Big)^2 = \sum_{n=t_a^i-T}^{t_a^i+T} w^2(t)(s(t)-\hat{h}(t))^2$$

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

Reformulate the error function - for HNM_1

Cost function (in discrete time):

$$\epsilon(a_{-L}, ..., a_L, f_0) = \frac{1}{2} \sum_{n=-N}^{N} (e[n])^2 = \frac{1}{2} \mathbf{e}^H \mathbf{e}$$

where

$$e[n] = w[n](s[n] - h[n])$$

or

$$\mathbf{e} = \begin{bmatrix} e[-N], & e[-N+1], & ... & e[N] \end{bmatrix}^T$$

Reformulate the error function - for HNM_1

In matrix form

$$\epsilon(\mathbf{a}) = \frac{1}{2}(\mathbf{s} - \mathbf{E}\mathbf{a})^H \mathbf{W}^2 (\mathbf{s} - \mathbf{E}\mathbf{a})$$

where

$$\mathbf{a} = \begin{bmatrix} a_{-L}, & \dots & a_0, & \dots & a_L \end{bmatrix}^T$$

and

$$\mathbf{E} = \begin{bmatrix} e^{j2\pi(-L)\hat{f}_0(-N)/f_s}, & \dots & e^{j2\pi L\hat{f}_0(-N)/f_s} \\ e^{j2\pi(-L)\hat{f}_0(-N+1)/f_s}, & \dots & e^{j2\pi L\hat{f}_0(-N+1)/f_s} \\ \vdots & \vdots & \vdots \\ e^{j2\pi(-L)\hat{f}_0 N/f_s}, & \dots & e^{j2\pi L\hat{f}_0 N/f_s} \end{bmatrix}^T_{(2L+1 \times 2N+1)}$$

Least Squares - for HNM_1

- Setting:

$$\frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} = 0 \Longrightarrow \mathbf{E}^H\mathbf{W}^2\mathbf{E}\mathbf{a} - \mathbf{E}^H\mathbf{W}^2\mathbf{s} = 0$$

  where ^{H} denotes Hermitian operator
- Solution:

$$\mathbf{a}_{LS} = (\mathbf{E}^H\mathbf{W}^2\mathbf{E})^{-1}\mathbf{E}^H\mathbf{W}^2\mathbf{s}$$

- Properties:
  - Rather fast, $O(L(N+L))$.
  - Assumes no errors in E matrix.

Least Squares - for HNM_1

- Setting:

$$\frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} = 0 \Longrightarrow \mathbf{E}^H\mathbf{W}^2\mathbf{E}\mathbf{a} - \mathbf{E}^H\mathbf{W}^2\mathbf{s} = 0$$

  where ^{H} denotes Hermitian operator
- Solution:

$$\mathbf{a}_{LS} = (\mathbf{E}^H\mathbf{W}^2\mathbf{E})^{-1}\mathbf{E}^H\mathbf{W}^2\mathbf{s}$$

- Properties:
  - Rather fast, $O(L(N+L))$.
  - Assumes no errors in E matrix.

- Setting:

$$\frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} = 0 \Longrightarrow \mathbf{E}^H \mathbf{W}^2 \mathbf{E} \mathbf{a} - \mathbf{E}^H \mathbf{W}^2 \mathbf{s} = 0$$

  where ^{H} denotes Hermitian operator

- Solution:

$$\mathbf{a}_{LS} = (\mathbf{E}^H \mathbf{W}^2 \mathbf{E})^{-1} \mathbf{E}^H \mathbf{W}^2 \mathbf{s}$$

- Properties:
  - Rather fast, $O(L(N+L))$.
  - Assumes no errors in E matrix.

Least Squares - for HNM$_1$

- Setting:

$$\frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} = 0 \Longrightarrow \mathbf{E}^H \mathbf{W}^2 \mathbf{E} \mathbf{a} - \mathbf{E}^H \mathbf{W}^2 \mathbf{s} = 0$$

  where ^{H} denotes Hermitian operator
- Solution:

$$\mathbf{a}_{LS} = (\mathbf{E}^H \mathbf{W}^2 \mathbf{E})^{-1} \mathbf{E}^H \mathbf{W}^2 \mathbf{s}$$

- Properties:
  - Rather fast, $O(L(N+L))$.
  - Assumes no errors in E matrix.

Least Squares - for HNM_1

- Setting:

$$\frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} = 0 \Longrightarrow \mathbf{E}^H \mathbf{W}^2 \mathbf{E} \mathbf{a} - \mathbf{E}^H \mathbf{W}^2 \mathbf{s} = 0$$

  where ^{H} denotes Hermitian operator

- Solution:

$$\mathbf{a}_{LS} = (\mathbf{E}^H \mathbf{W}^2 \mathbf{E})^{-1} \mathbf{E}^H \mathbf{W}^2 \mathbf{s}$$

- Properties:
  - Rather fast, $O(L(N+L))$.
  - Assumes no errors in E matrix.

Avoiding ill-conditioning

- For HNM_1 there is no problem if window length is twice the local pitch period
- Same thing for HNM_2
- For HNM_3 stands the same in case the maximum voiced frequency is less than $3/4$ of the sampling frequency and order of amplitude polynomial is 2

Avoiding ill-conditioning

- For HNM_1 there is no problem if window length is twice the local pitch period
- Same thing for HNM_2
- For HNM_3 stands the same in case the maximum voiced frequency is less than 3/4 of the sampling frequency and order of amplitude polynomial is 2

Avoiding ill-conditioning

- For HNM_1 there is no problem if window length is twice the local pitch period
- Same thing for HNM_2
- For HNM_3 stands the same in case the maximum voiced frequency is less than $3/4$ of the sampling frequency and order of amplitude polynomial is 2

Residual Signal

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

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

TIME DOMAIN CHARACTERISTICS OF $\hat{r}[n]$

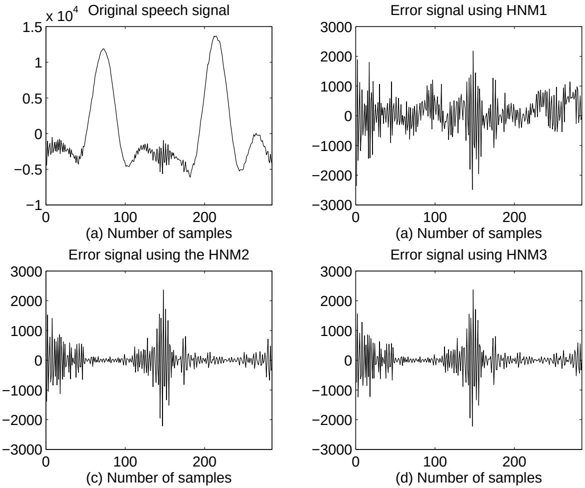


FIGURE: *(a) A fricative voiced of an original recording and the residual error signals from (b) HNM1, (c) HNM2, (d) HNM3.*

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

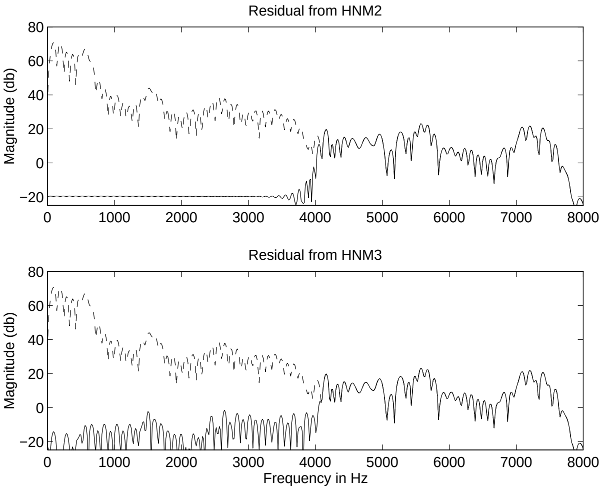


Figure: *(a) FFT-magnitude of the residual signal (solid) from (a) HNM2 and (b) HNM3. The FFT-magnitude of the original signal has been also included (dashed).*

... AND AFTER ADDING NOISE

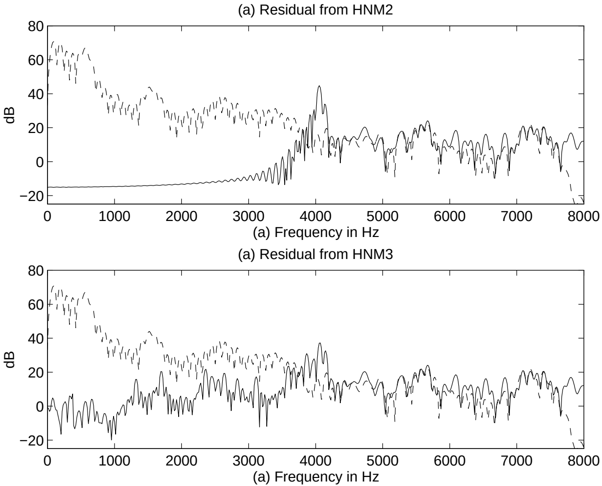


FIGURE: *Same as before, but with additive white noise.*

Modeling error

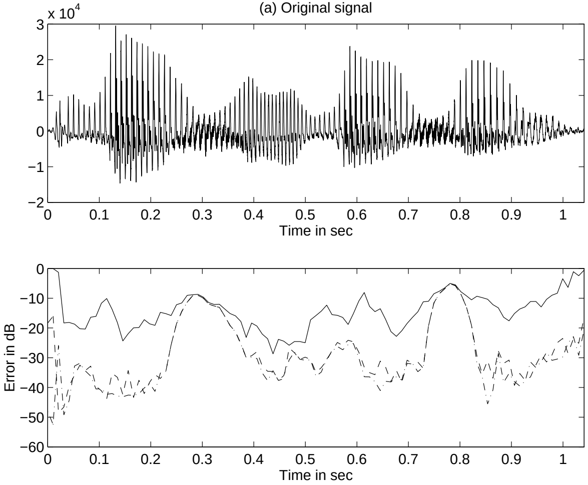


FIGURE: *Modelling error in dB using the three models.*

Modeling the residual signal

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

Outline

For all HNMs

- $t_s^i \longleftrightarrow t_a^i$
- For the periodic part: Overlap-and-Add
- For the stochastic (noise) part:
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

For all HNMs

- $t_s^i \longleftrightarrow t_a^i$
- For the periodic part: Overlap-and-Add
- For the stochastic (noise) part:
    - Instead of AR coefficients we use reflection coefficients
    - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
    - Modulation in time with a deterministic function (i.e., triangular)

For all HNMs

- $t_s^i \longleftrightarrow t_a^i$
- For the periodic part: Overlap-and-Add
- For the stochastic (noise) part:
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

For all HNMs

- $t_s^i \longleftrightarrow t_a^i$
- For the periodic part: Overlap-and-Add
- For the stochastic (noise) part:
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

For all HNMs

- $t_s^i \longleftrightarrow t_a^i$
- For the periodic part: Overlap-and-Add
- For the stochastic (noise) part:
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

For all HNMs

- $t_s^i \longleftrightarrow t_a^i$
- For the periodic part: Overlap-and-Add
- For the stochastic (noise) part:
  - Instead of AR coefficients we use reflection coefficients
  - Sample-by-sample filtering of Gaussian noise using normalized lattice filtering
  - Modulation in time with a deterministic function (i.e., triangular)

For HNM_1 specifically

For Periodic part (as an alternative to OLA)

- Direct frequency matching
- Linear amplitude interpolation
- Linear phase interpolation using average pitch value

For HNM$_1$ specifically

For periodic part (as an alternative to OLA)

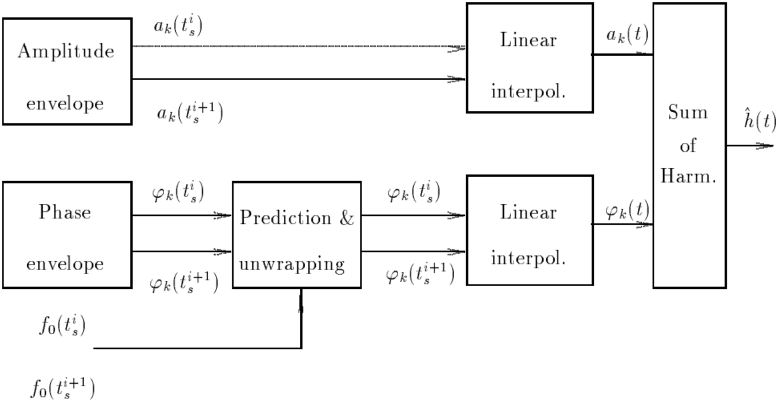


Figure: *Block diagram of the synthesis of the harmonic part for* $t \in [t_s^i, t_s^{i+1}]$.

FOR HNM$_1$ SPECIFICALLY

For the noise part

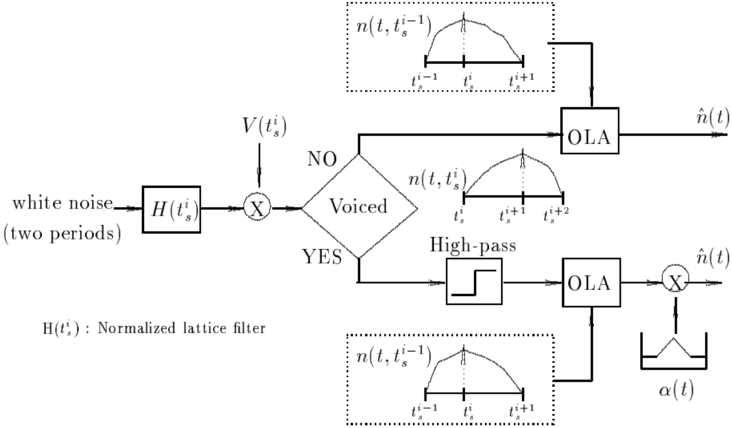


FIGURE: *Block diagram of the synthesis of the noise part for $t \in [t_s^i, t_s^{i+1}]$.*

Outline

Again on the energy modulation

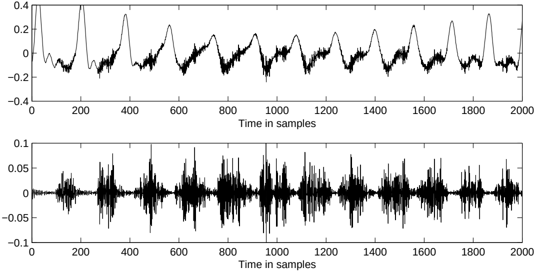
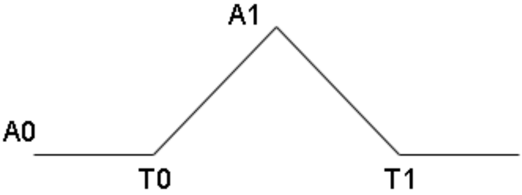


Figure: *Upper plot: 12 pitch periods of voiced fricative phoneme /z/. Lower plot: The same speech signal filtered by a highpass filter at 4 kHz.*

So far, mainly

So far we mainly use the 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.

Hilbert envelope

We may also use the Hilbert envelope, computed as:

$$\tilde{e}_H[n] = \sum_{k=L-M+1}^{L} a_k e^{2\pi k (f_0/f_s) n}$$

Example of energy envelope

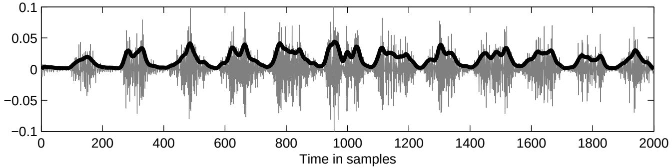


Figure: *Example of Energy Envelope, with* $N = 7$.

Energy envelope

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

Looking at time domain properties

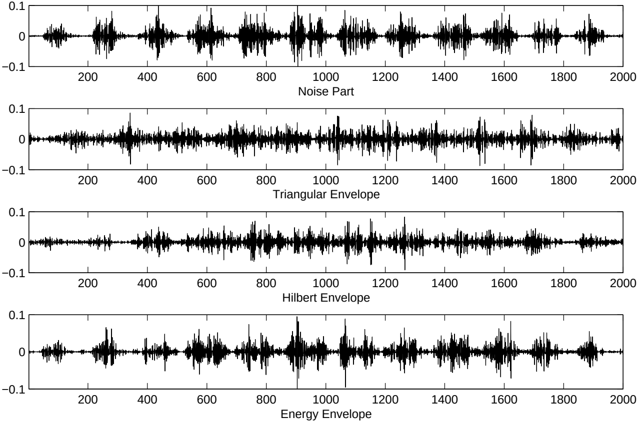


Figure: *A few periods of the noise part of phoneme /z/. First: the original noise part. Second: synthesized noise using triangular envelope. Third: synthesized noise using Hilbert envelope. Fourth: synthesized noise using energy envelope.*

Results from Listening test I

	Triangular	No pref.	Hilbert
Male	8 (8.3%)	43 (44.8%)	45 (46.9%)
Female	40 (41.7%)	47 (48.9%)	9 (9.4%)

	Hilbert	No pref.	Energy
Male	22 (22.9%)	47 (49.0%)	27 (28.1%)
Female	22 (22.9%)	54 (56.3%)	20 (20.8%)

	Energy	No pref.	Triangular
Male	43 (44.8%)	50 (52.0%)	3 (3.2%)
Female	16 (16.7%)	67 (69.8%)	13 (13.5%)

Table: Results from the listening test for the English sentences.

Results from Listening test II

	Triangular	No pref.	Hilbert
Male	10 (10.4%)	47 (49.0%)	39 (40.6%)
Female	8 (8.3%)	71 (74.0%)	17 (17.7%)

	Hilbert	No pref.	Energy
Male	11 (11.5%)	58 (60.4%)	27 (28.1%)
Female	13 (13.5%)	58 (60.4%)	25 (26.1%)

	Energy	No pref.	Triangular
Male	42 (43.7%)	48 (50.0%)	6 (6.3%)
Female	16 (16.7%)	68 (70.8%)	12 (12.5%)

Table: Results from the listening test for the French sentences.

Outline

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Estimating Sinusoidal parameters

- Sinusoidal representation for a speech/signal frame:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t)$$

- Methods:
  - FFT-based methods (i.e., QIFFT (Abe et al., 2004-05 [7, 8]))
  - Subspace methods
  - Least Squares (LS) method
- Frequency mismatch:

$$\hat{f}_k = f_k + \eta_k$$

- How to deal with that?
- You will discuss more advanced sinusoidal models in the following lecture! ☺

Outline

THANK YOU

for your attention

Outline

References I

- R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug 1986.
- D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, Fev 1988.
- A. Abrantes, J. Marques, and I. Transcoso, "Hybrid sinusoidal modeling of speech without voicing decision," *Eurospeech-91*, pp. 231–234, 1991.
- B.Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, 1998.

References II

Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.*
PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Jan 1996.

W. Hess, *Pitch determination of Speech Signals: Algorithmes and Devices.*
Berlin: Springer, 1983.

M. Abe and J. S. III, "CQIFFT: Correcting Bias in a Sinusoidal Parameter Estimator based on Quadratic Interpolation of FFT Magnitude Peaks," Tech. Rep. STAN-M-117, Stanford University, California, Oct 2004.

M. Abe and J. S. III, "AM/FM Estimation for Time-varying Sinusoidal Modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Philadelphia), pp. III 201–204, 2005.