Sinusoidal Models for Text-to-Speech Synthesis

… or …

Development of Ahocoder, an HNM-based vocoder

Daniel Erro - derro@aholab.ehu.es

ikerbasque
Basque Foundation for Science

Bilbao, Spain

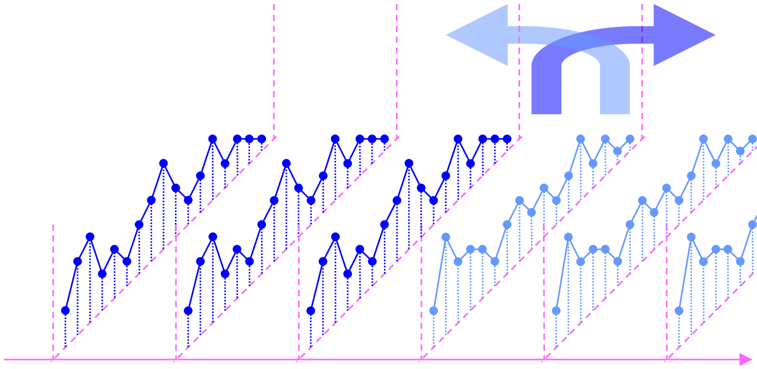
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

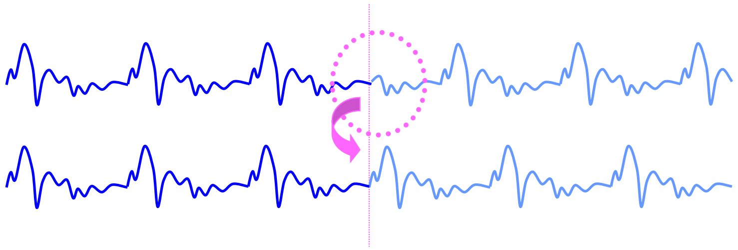
Outline

- Introduction
- Ahocoder, an HNM-based vocoder
  - F0 estimation
  - MVF estimation
  - Spectral analysis
  - Speech waveform reconstruction
  - Evaluation
- Conclusions

Introduction

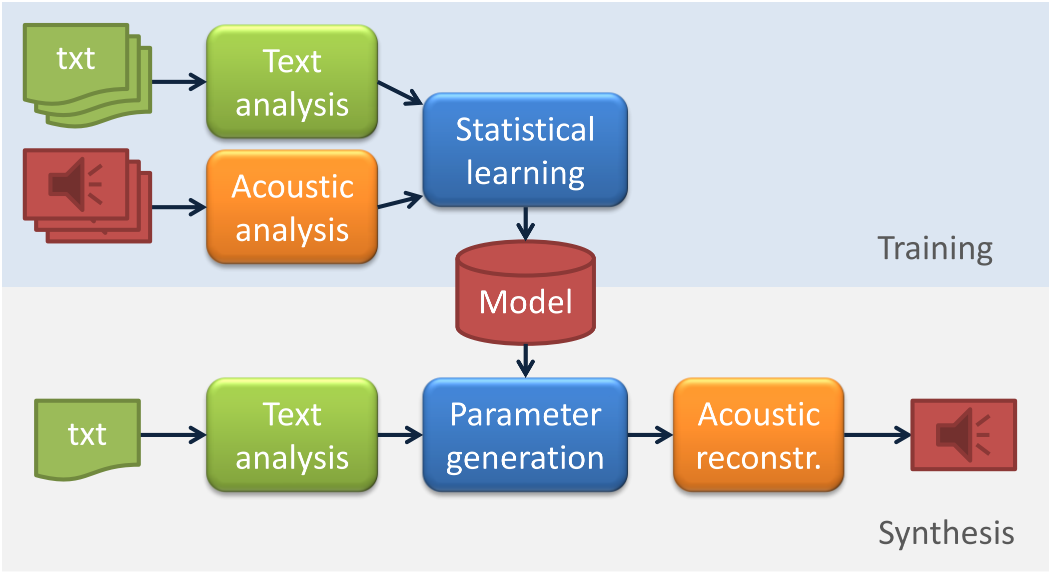
- Role of sinusoidal models in TTS:
  - Concatenative speech synthesis: prosodic modification and smoothing of boundary effects (Y. Stylianou, 2001)

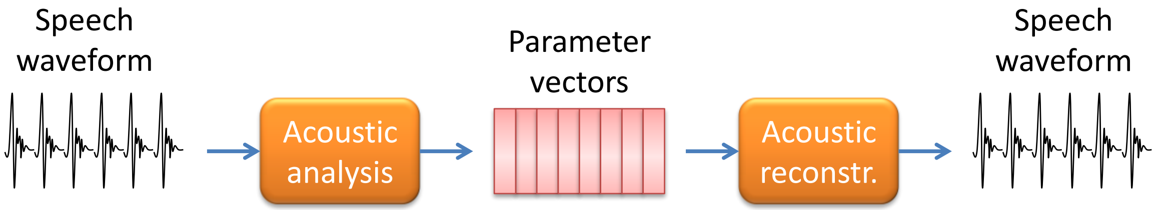




Introduction

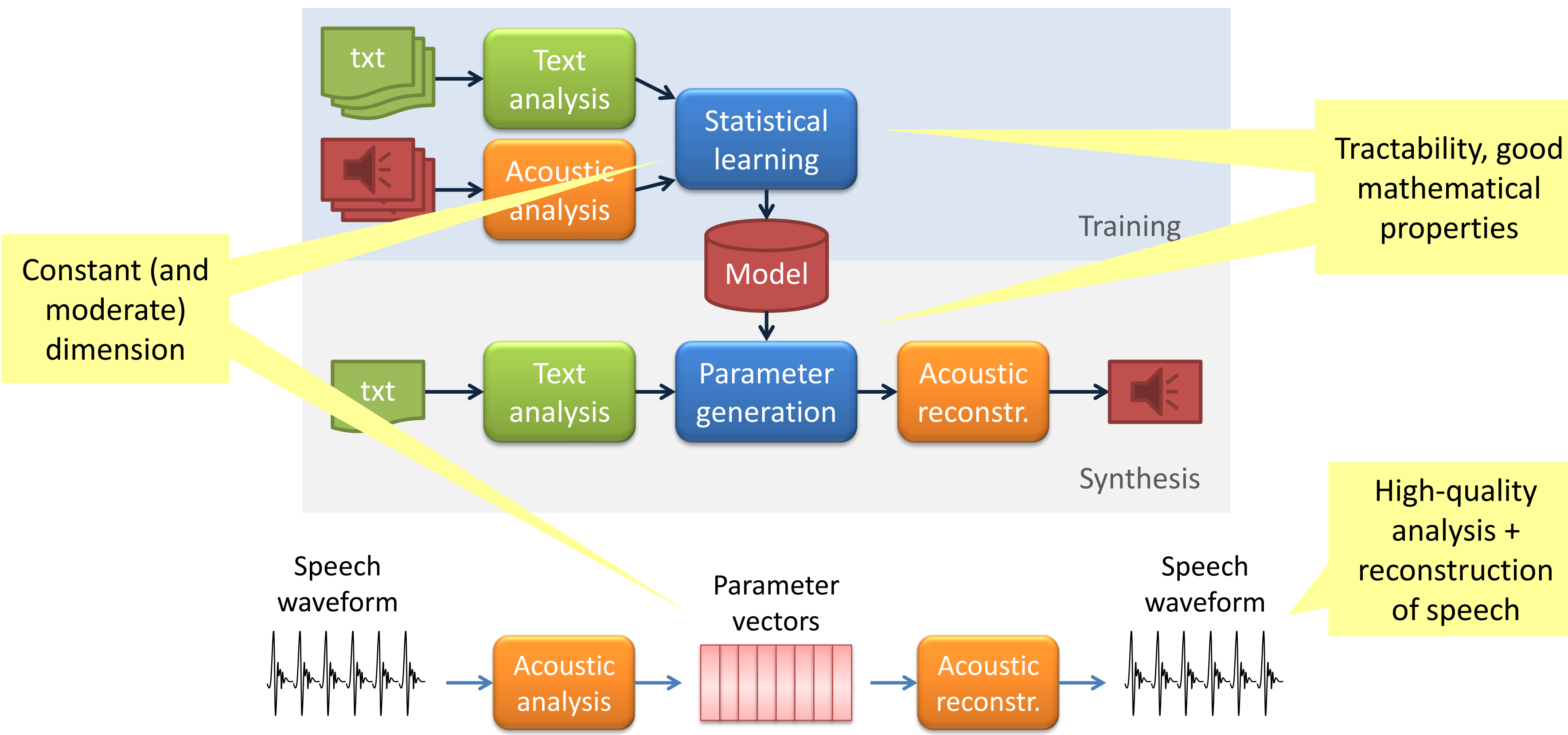
- Role of sinusoidal models in TTS:
  - Statistical parametric speech synthesis: **vocoders**





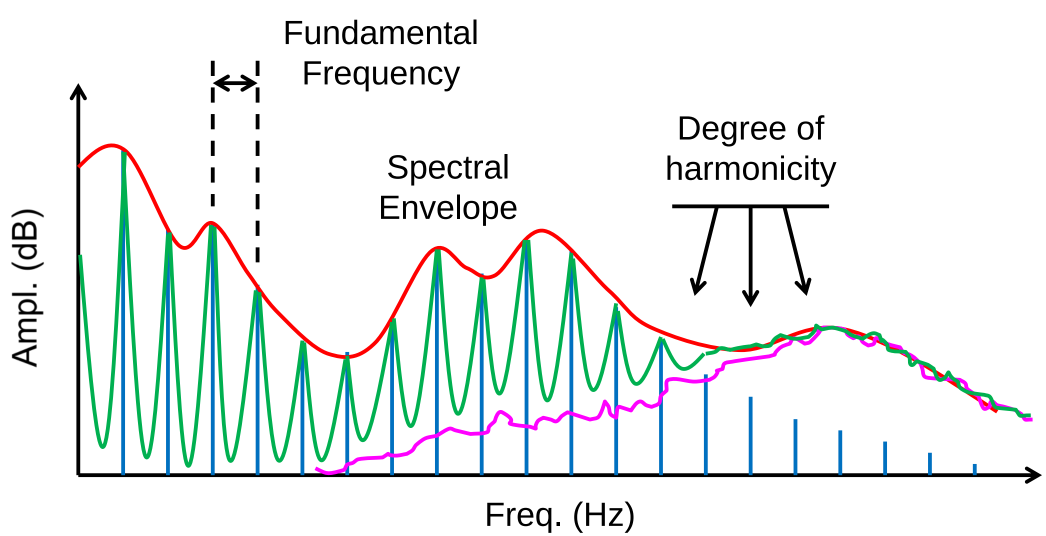
Introduction

- Role of sinusoidal models in TTS:
  - Statistical parametric speech synthesis: **vocoders**



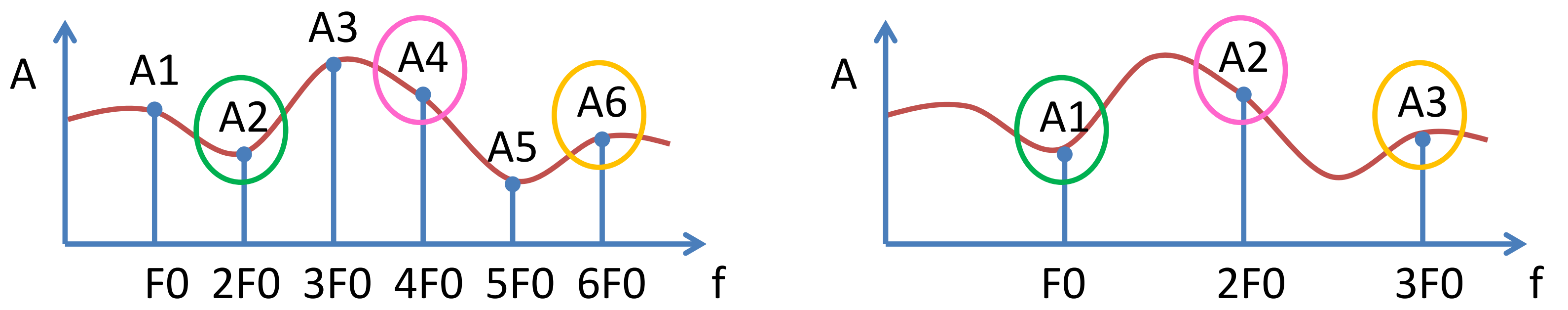
Introduction

- A typical vocoder extracts info at 3 levels
  - F0 (log): get_f0, RAPT, YIN, Tempo, PRAAT, SRH…
  - Spectral envelope: SPTK, STRAIGHT+MCEP (Kawahara, 1999), GlottHMM+LSF (Raitio et al., 2011)…
  - Degree of harmonicity: BAP, HNR, MVF…



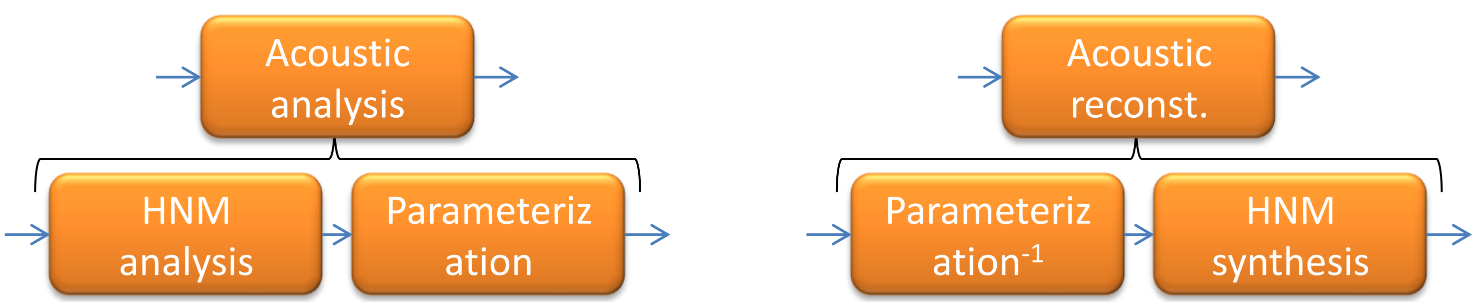
Introduction

- Sinusoids(+noise) based vocoder?
  - HQ resynthesis and modification, but…
  - Variable dimension
  - Not very tractable, complicated dependencies



Introduction

- Sinusoids(+noise) based vocoder?
  - Use them as an intermediate stage between waveforms and parameters
    - Sinusoidal frequencies → logF0
    - Sinusoidal amplitudes → MCEP, MGC, LSF…
    - Sinusoidal phases → RPS, PD… or nothing!
    - Noise → HNR, MVF…



Introduction

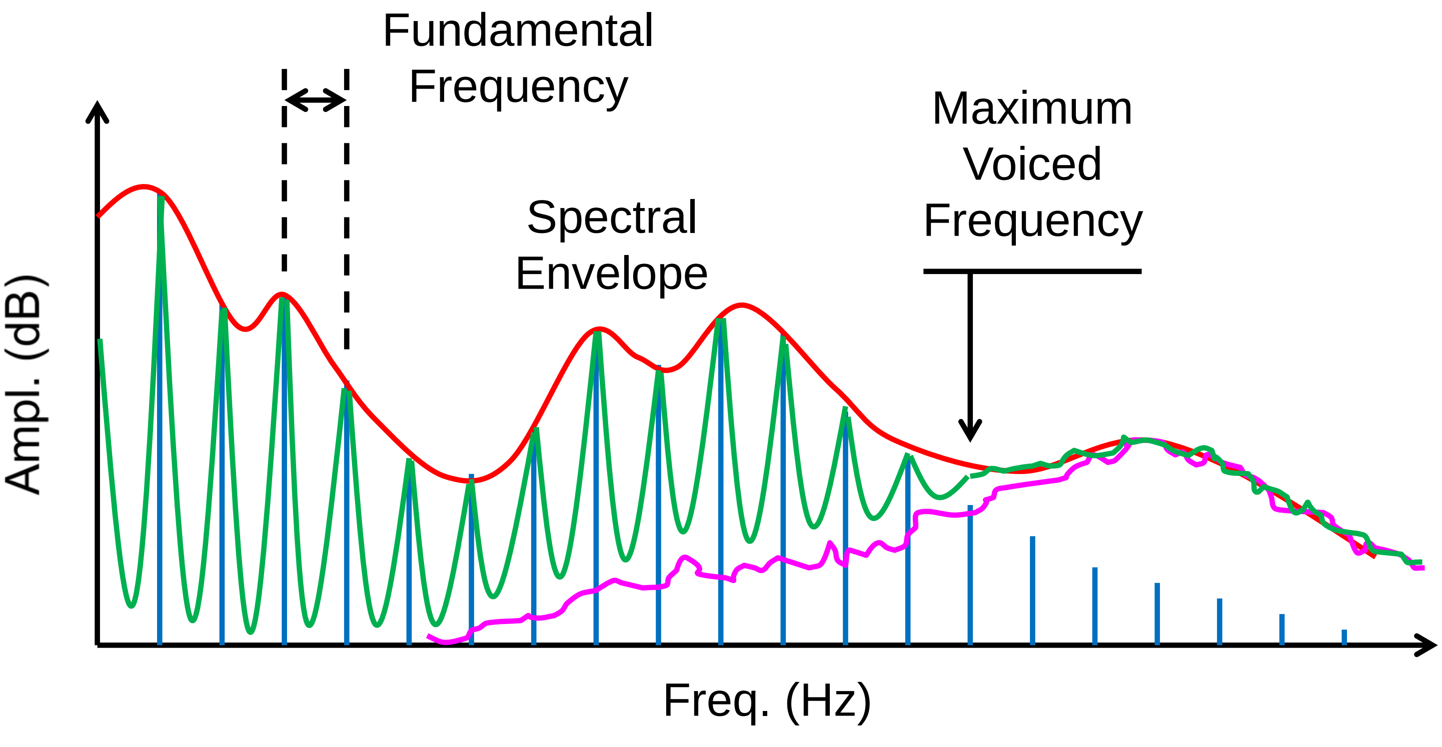
- Sinusoids(+noise) based vocoder?
  - Use them as an intermediate stage between waveforms and parameters
    - Sinusoidal frequencies → logF0
    - Sinusoidal amplitudes → MCEP, MGC, LSF…
    - Sinusoidal phases → RPS, PD… or nothing!
    - Noise → HNR, MVF…
  - Enables intermediate modifications (Thursday)

Outline

- ~~Introduction~~
- Ahocoder, an HNM-based vocoder
  - F0 estimation
  - MVF estimation
  - Spectral analysis
  - Speech waveform reconstruction
  - Evaluation
- Conclusions

Ahocoder, an HNM-based vocoder

- Every 5ms, fs=16kHz



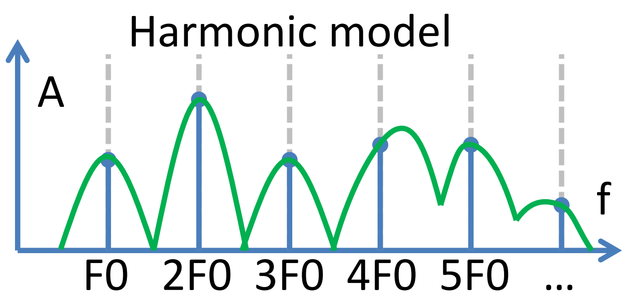
Ahocoder, an HNM-based vocoder

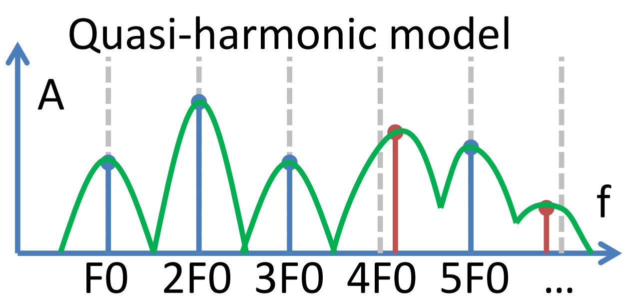
- F0 estimation
  - Praat / external, every 5ms starting at n=0
  - QHM refinement

$$s(t) = \sum_{i=1}^{I} (a_i + t b_i) \exp(j 2\pi f_i t)$$

$$\Delta f_i = \frac{\mathrm{Re}\{a_i\}\mathrm{Im}\{b_i\} - \mathrm{Im}\{a_i\}\mathrm{Re}\{b_i\}}{2\pi |a_i|^2}$$

$$\Delta f_0 = \frac{\sum_{i=1}^{I} w_i \cdot \Delta f_i / i}{\sum_{i=1}^{I} w_i}$$





Ahocoder, an HNM-based vocoder

- F0 estimation
  - Praat / external, every 5ms starting at n=0
  - QHM refinement

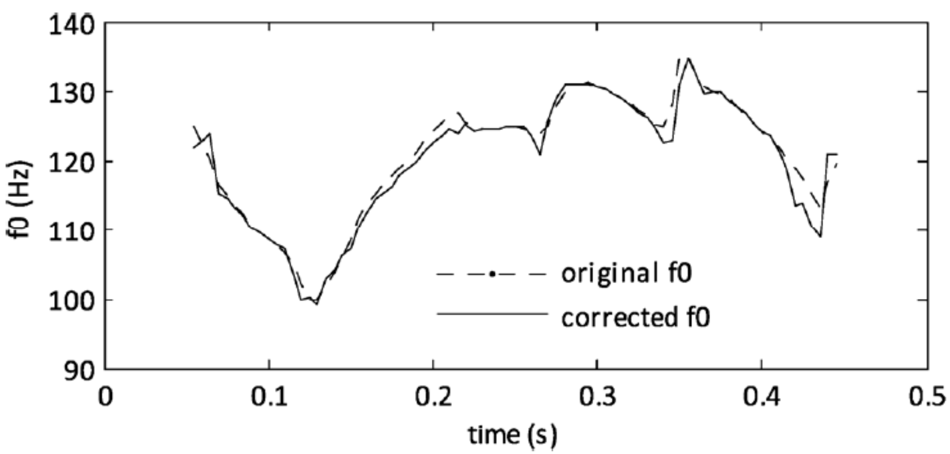
$$s(t) = \sum_{i=1}^{I} (a_i + tb_i) \exp(j2\pi f_i t)$$

$$\Delta f_i = \frac{\mathrm{Re}\{a_i\}\mathrm{Im}\{b_i\} - \mathrm{Im}\{a_i\}\mathrm{Re}\{b_i\}}{2\pi |a_i|^2}$$

$$\Delta f_0 = \frac{\sum_{i=1}^{I} w_i \cdot \Delta f_i / i}{\sum_{i=1}^{I} w_i}$$

Analysis band? 1kHz? 2? 4? 8?

Constant weights? Amplitude-related weights?



Ahocoder, an HNM-based vocoder

- F0 estimation
  - Experiment: more accurate F0 → more accurate harmonic reconstruction of signals

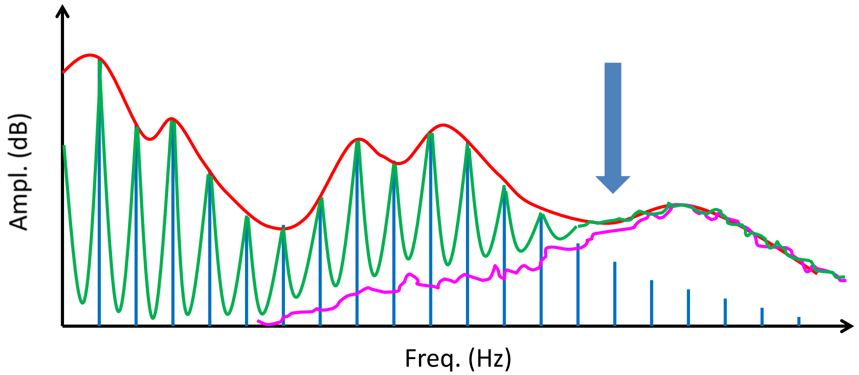
Constant: wi=1
By ampl.: wi=sqrt(Ai)

Weighting, Band	0-1kHz	1-2kHz	2-4kHz	4-8kHz	Total
Constant, MVF	2.1%	-14.2%	-28.0%	-14.4%	-4.5%
By ampl., MVF	-8.1%	-15.5%	-21.5%	-8.5%	-10.9%

$$\Delta f_0 = \frac{\sum_{i=1}^{I} w_i \cdot \Delta f_i / i}{\sum_{i=1}^{I} w_i}$$

Analysis band?
1kHz? 2? 4? 8?

Constant weights?
Amplitude-related weights?



Ahocoder, an HNM-based vocoder

- F0 estimation
  – Experiment: more accurate F0 → more accurate harmonic reconstruction of signals

Constant: wi=1
By ampl.: wi=sqrt(Ai)

Weighting, Band	0-1kHz	1-2kHz	2-4kHz	4-8kHz	Total
Constant, MVF	2.1%	-14.2%	-28.0%	-14.4%	-4.5%
By ampl., MVF	-8.1%	-15.5%	-21.5%	-8.5%	-10.9%
By ampl., 4kHz	-6.7%	-16.1%	-23.5%	-12.6%	-10.4%
By ampl., 2kHz	-10.1%	-15.1%	-3.2%	4.2%	-9.8%
By ampl., 1kHz	-14.8%	2.1%	17.5%	12.7%	-7.7%

$$\Delta f_0 = \frac{\sum_{i=1}^{I} w_i \cdot \Delta f_i / i}{\sum_{i=1}^{I} w_i}$$

Analysis band?
1kHz? 2? 4? 8?

Constant weights?
Amplitude-related weights?

Ahocoder, an HNM-based vocoder

- F0 estimation
  - Experiment: more accurate F0 → more accurate harmonic reconstruction of signals

Constant: wi=1
By ampl.: wi=sqrt(Ai)

Weighting, Band	0-1kHz	1-2kHz	2-4kHz	4-8kHz	Total
Constant, MVF	2.1%	-14.2%	-28.0%	-14.4%	-4.5%
By ampl., MVF	-8.1%	-15.5%	-21.5%	-8.5%	-10.9%
By ampl., 4kHz	-6.7%	-16.1%	-23.5%	-12.6%	-10.4%
By ampl., 2kHz	-10.1%	-15.1%	-3.2%	4.2%	-9.8%
By ampl., 1kHz	-14.8%	2.1%	17.5%	12.7%	-7.7%

  - What about quasi-harmonic reconstruction?

Apparently QHM much better than HM, with or without F0 refinement

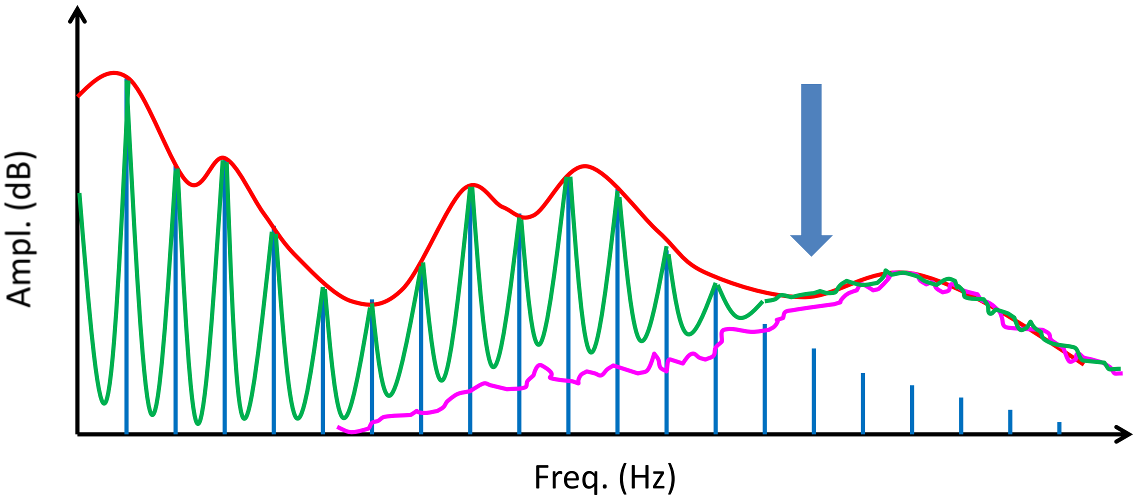
Weighting, Band	0-1kHz	1-2kHz	2-4kHz	4-8kHz	Total
No refinement	-45.8%	-59.1%	-63.2%	-55.5%	-50.3%
By ampl., MVF	-47.1%	-61.9%	-66.3%	-57.9%	-52.0%

Outline

- ~~Introduction~~
- Ahocoder, an HNM-based vocoder
  - ~~F0 estimation~~
  - MVF estimation
  - Spectral analysis
  - Speech waveform reconstruction
  - Evaluation
- Conclusions

Ahocoder, an HNM-based vocoder

- MVF estimation



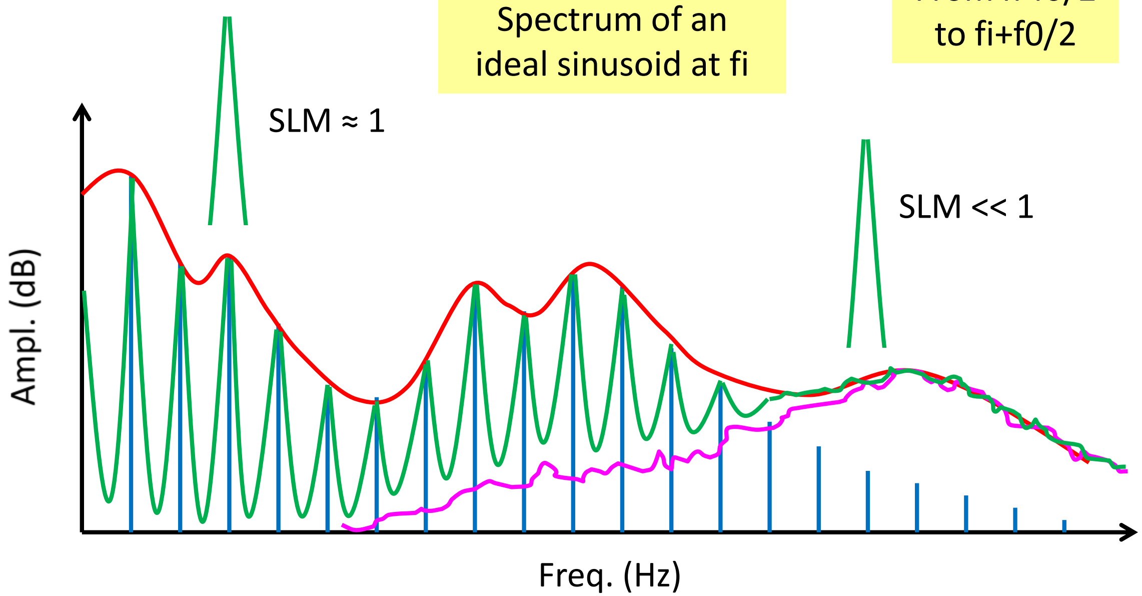
Ahocoder, an HNM-based vocoder

- MVF estimation
  - Sinusoidal likeness measure (SLM) (Rodet, 1997)

$$\lambda_i = \frac{\left|\sum X[m] \cdot W_i^*[m]\right|}{\sqrt{\sum |X[m]|^2 \cdot \sum |W_i[m]|^2}} \qquad \forall m, \left| m\frac{f_s}{N} - f_i \right| < \frac{f_0}{2}$$

Spectrum of an ideal sinusoid at fi

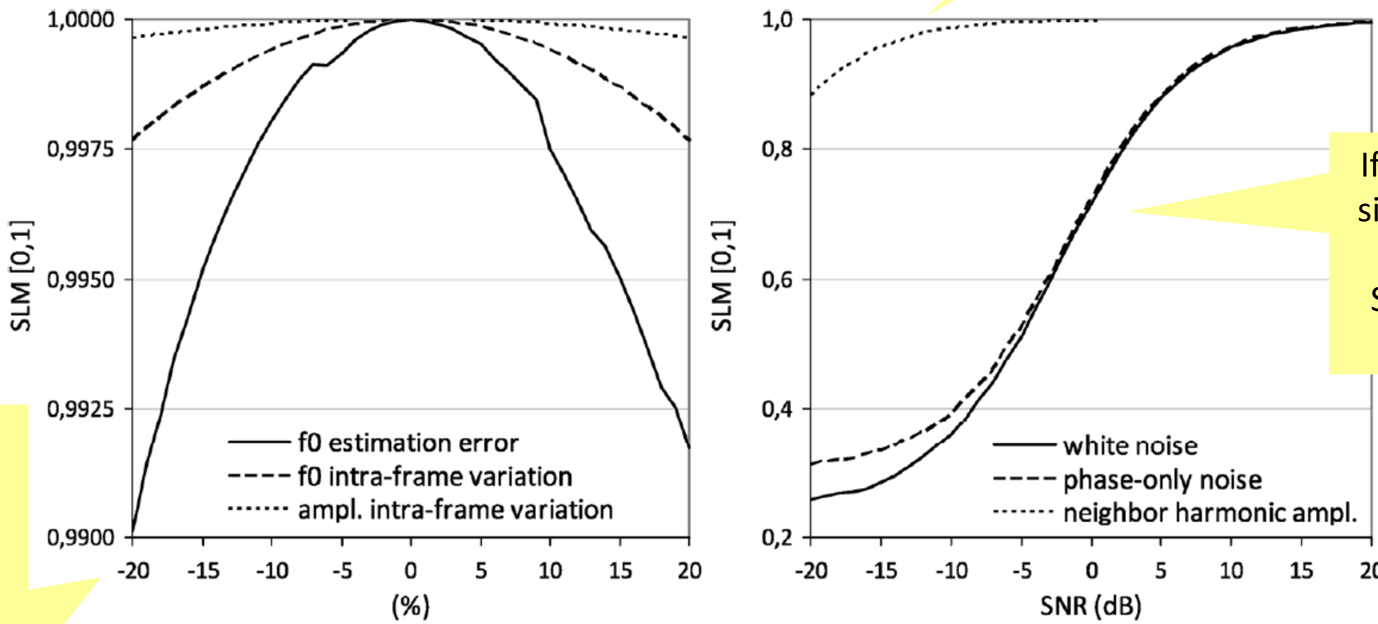
From fi-f0/2 to fi+f0/2



Ahocoder, an HNM-based vocoder

- MVF estimation
  - Sinusoidal likeness measure (SLM)

Be careful with high-A adjacent sinusoids

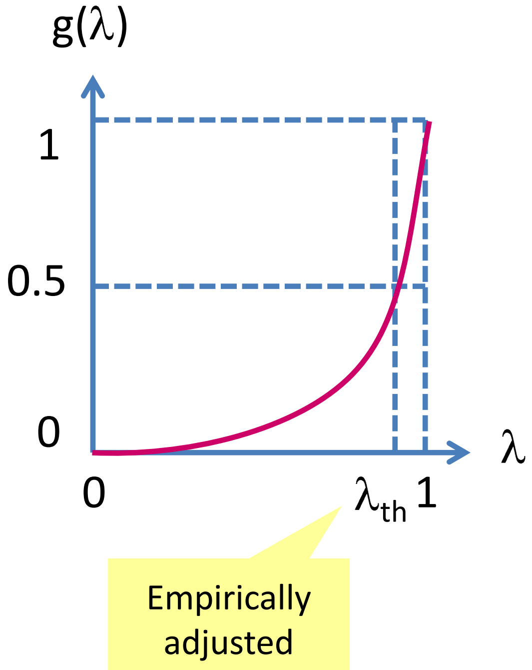


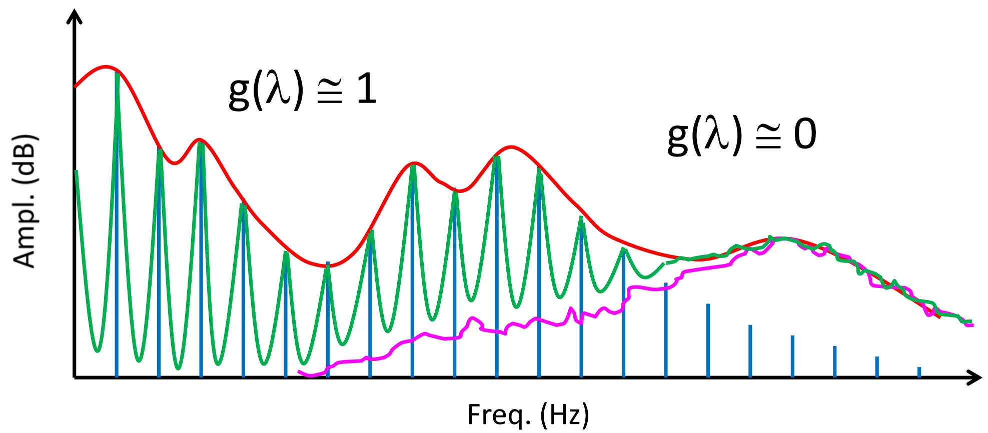
If we mix the sinusoid with noise, the SLM decays rapidly

If we calculate the SLM of a time-varying sinusoid (f, A…), it remains very close to 1

Ahocoder, an HNM-based vocoder

- MVF estimation
  - Sinusoidal likeness measure (SLM)
  - Probability of voicing of each “peak”

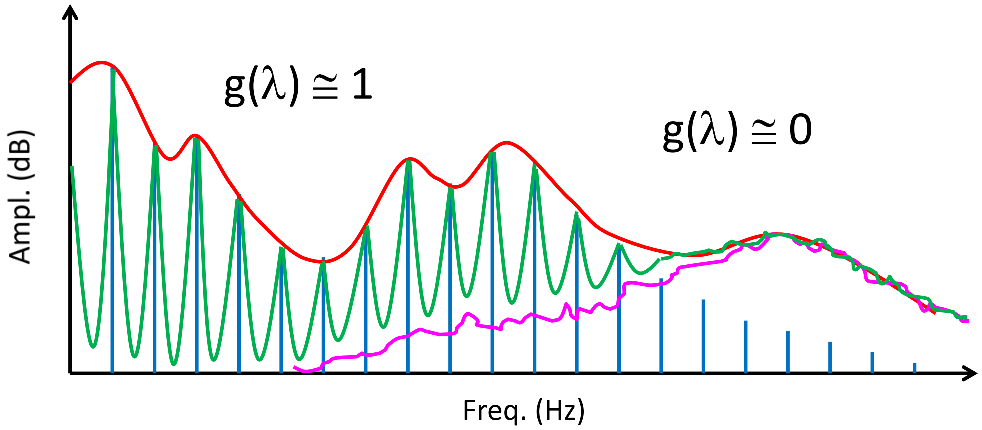




Ahocoder, an HNM-based vocoder

- MVF estimation
  - Sinusoidal likeness measure (SLM)
  - Probability of voicing of each "peak"
  - Local decision
  - Median filter over t

$$i = \underset{i}{\operatorname{argmax}} \underbrace{\prod_{j=1}^{i-1} g(\lambda_j)}_{\text{Voiced below i}} \underbrace{\prod_{j=i}^{I} \left(1 - g(\lambda_j)\right)}_{\text{Unvoiced above i}}$$



Ahocoder, an HNM-based vocoder

- MVF estimation
  - Experiments:
    - Baseline method: prediction from c_0

$$v^{(k)} = v^{(\max)} \cdot \frac{c_0^{(k)} - c_0^{(\min)}}{c_0^{(\max)} - c_0^{(\min)}}$$

~4.5kHz in "normal" voices

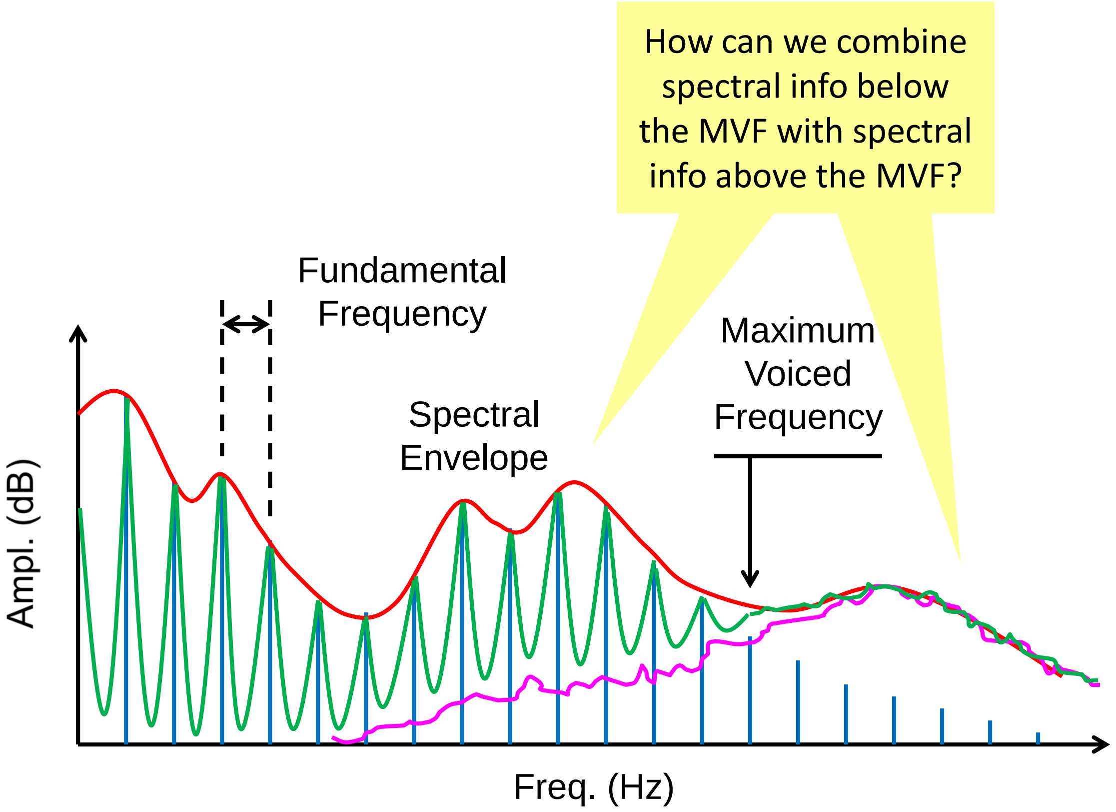
    - Subjective preference in resynthesis: 38% vs 17%
    - Subjective preference in synthesis: 29-30% vs 17-22%

Outline

- ~~Introduction~~
- Ahocoder, an HNM-based vocoder
  - ~~F0 estimation~~
  - ~~MVF estimation~~
  - Spectral analysis
  - Speech waveform reconstruction
  - Evaluation
- Conclusions

Ahocoder, an HNM-based vocoder

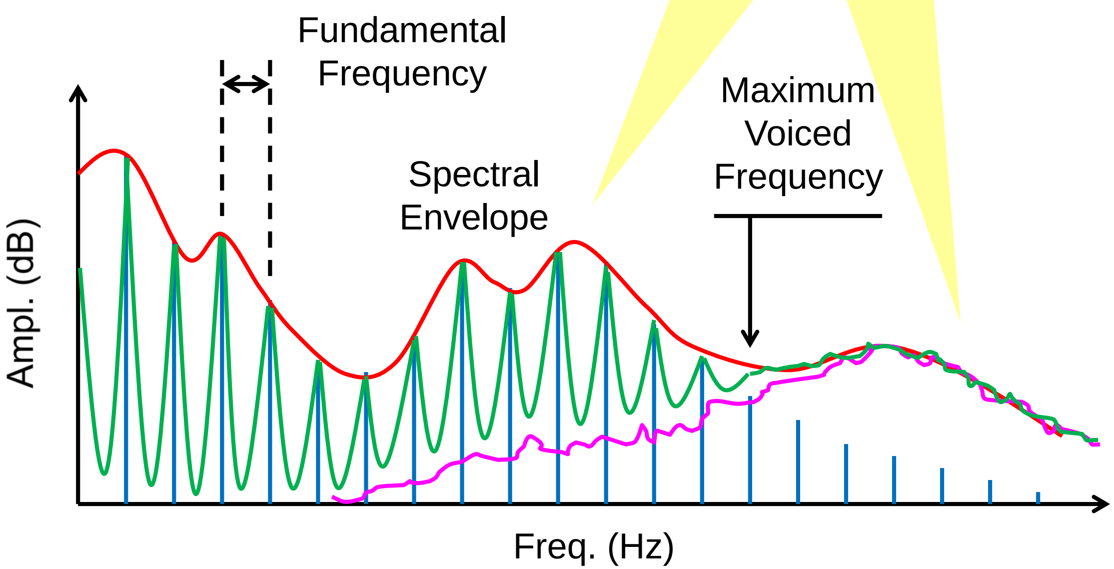
- Spectral analysis



Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM

How can we combine spectral info below the MVF with spectral info above the MVF?



Ahocoder, an HNM-based vocoder

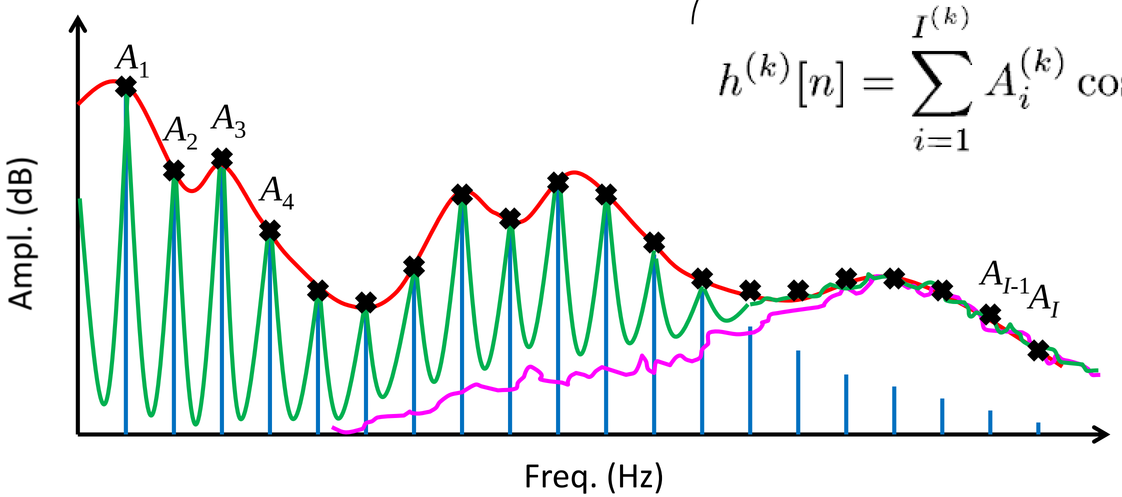
- Spectral analysis
  - HM instead of HNM

Its harmonic reconstruction

Original frame k

$$\{A_i^{(k)}, \varphi_i^{(k)}\} = \arg\ \min \sum_n w^2[n]\Big(s[n+kN] - h^{(k)}[n]\Big)^2$$

$$h^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right)$$

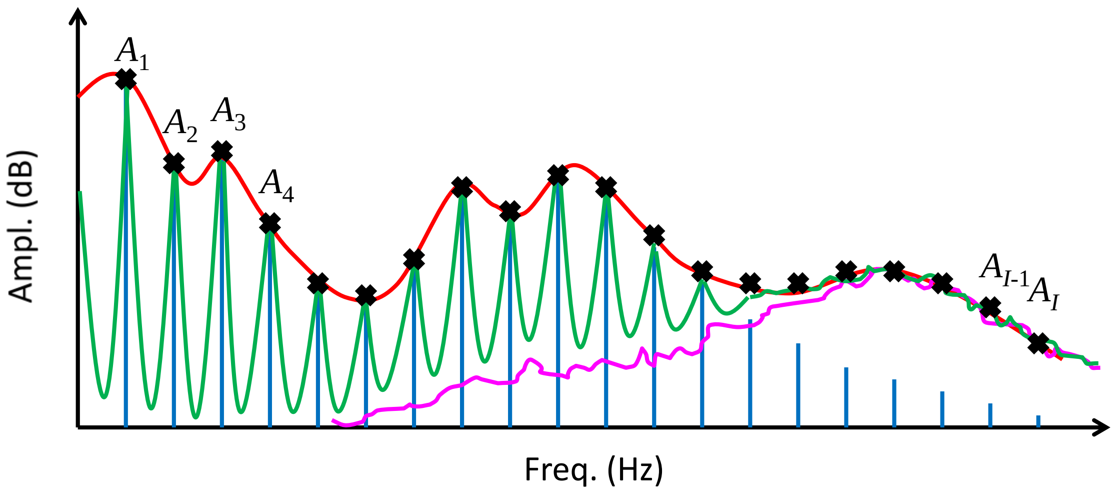


Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM

DFT $\qquad$ DFT^{-1}

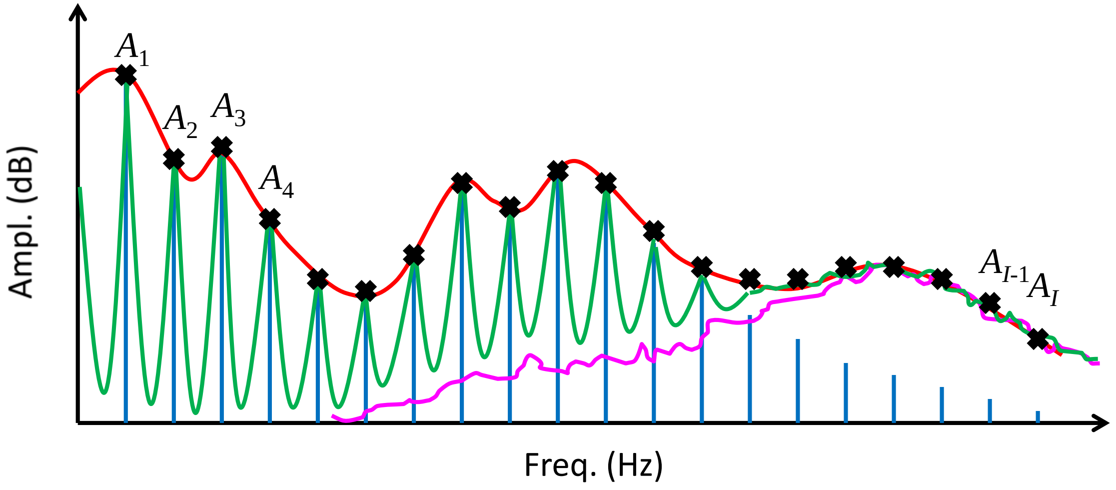
$$X[m] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi mn}{N}} \Leftrightarrow x[n] = \frac{1}{N}\sum_{m=0}^{N-1} X[m] e^{j\frac{2\pi mn}{N}}$$



Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM

$$X[m] = \frac{1}{N}\sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi mn}{N}} \Leftrightarrow x[n] = \sum_{m=0}^{N-1} X[m] e^{j\frac{2\pi mn}{N}}$$



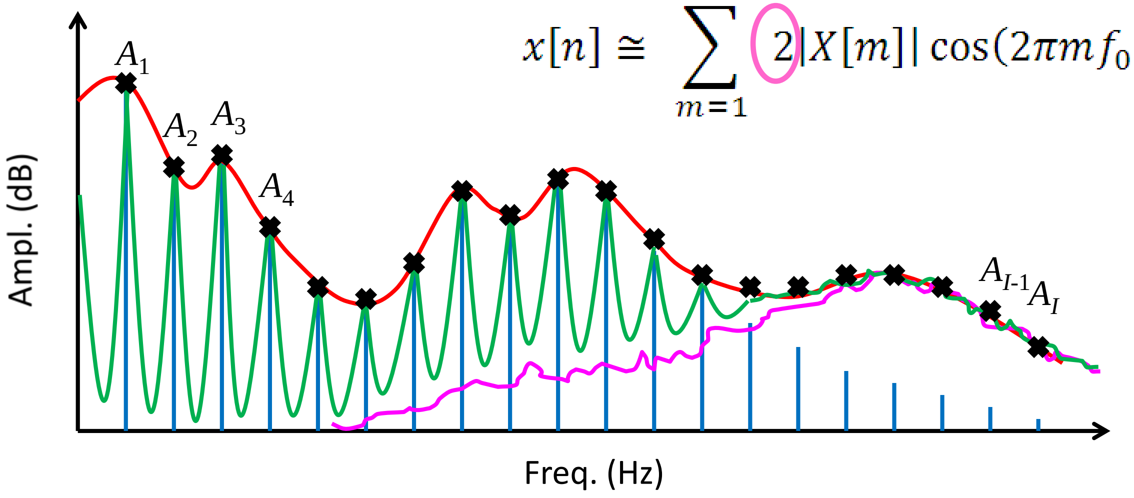
Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM

$$X[m]=\frac{1}{N}\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi mn}{N}} \Leftrightarrow x[n]=\sum_{m=0}^{N-1}X[m]e^{j\frac{2\pi mn}{N}}$$

$$x[n]\cong\sum_{m=1}^{N/2-1}2|X[m]|\cos(2\pi m f_0 n/f_s+\measuredangle X[m])$$

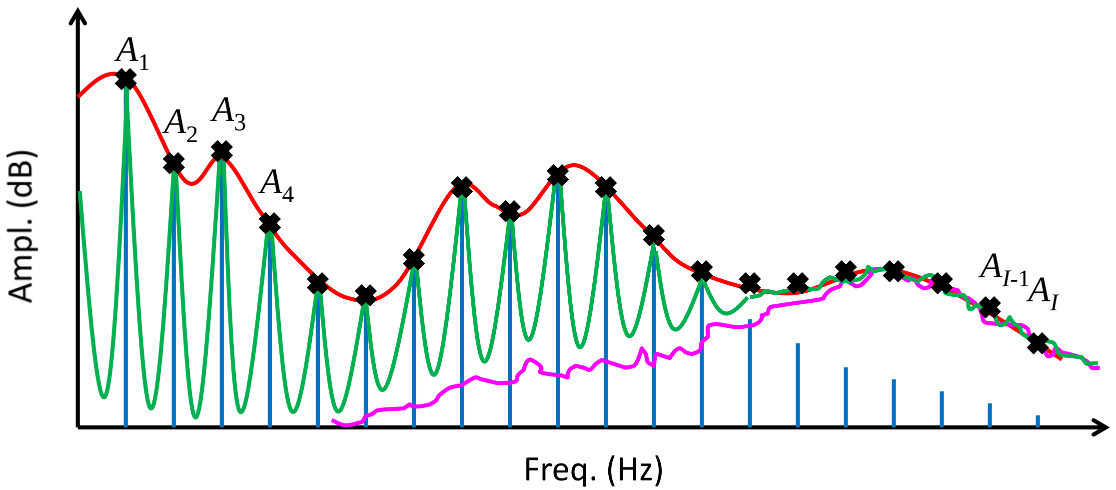
$$f_0=f_s/N$$



Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM

$$X[m] = \frac{1}{N}\,\mathrm{FFT}_N\{x[n]\} \quad \Leftrightarrow \quad x[n] = N\,\mathrm{FFT}_N^{-1}\{X[m]\}$$

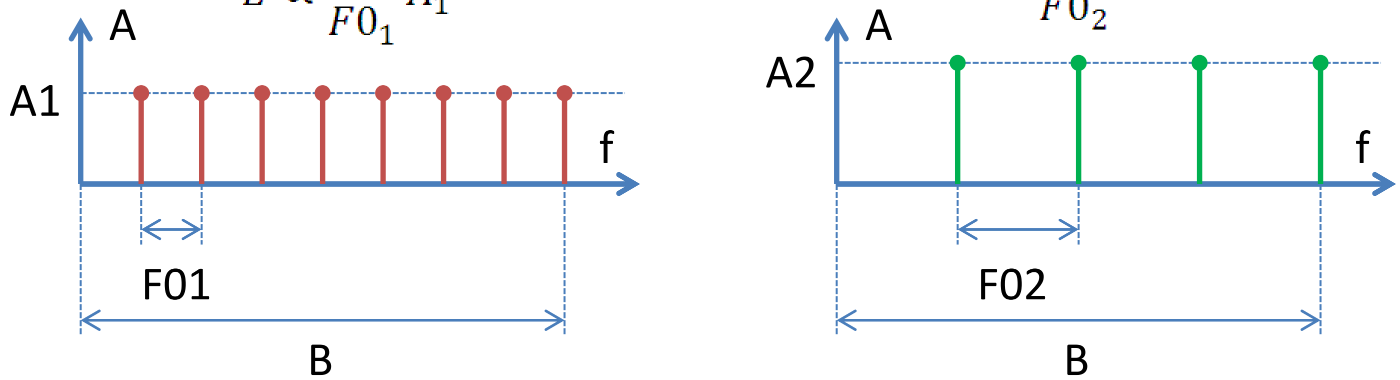


Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0

$$\frac{A_1}{\sqrt{F0_1}} = \frac{A_2}{\sqrt{F0_2}}$$

$$E \propto \frac{B}{F0_1} A_1^2 \qquad \frac{B}{F0_2} A_2^2$$



Ahocoder, an HNM-based vocoder

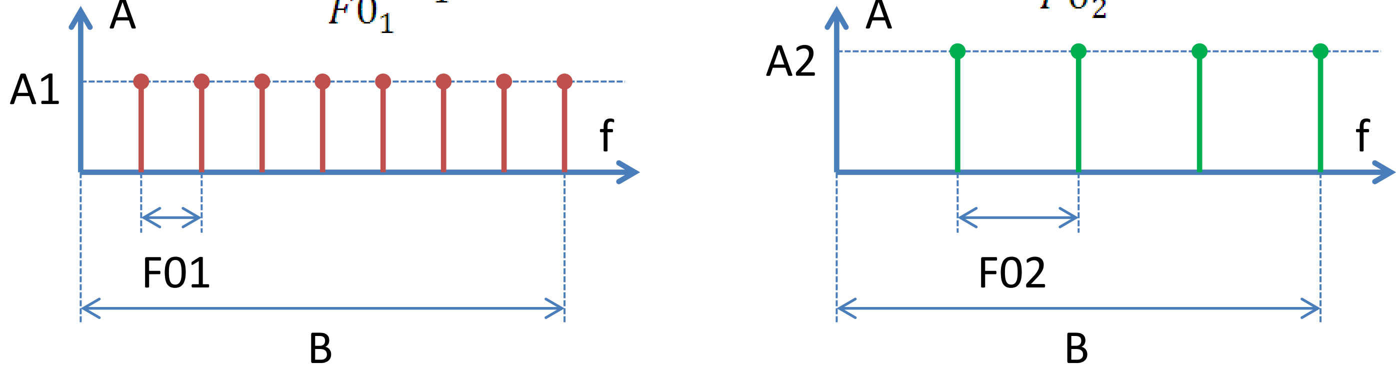
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0

$$\hat{A}_i = \frac{A_i}{2\sqrt{f_0}}$$

Crucial to separate F0 from spectrum and allow signal reconstruction at different F0 and voicing contours

$$\frac{A_1}{\sqrt{F0_1}} = \frac{A_2}{\sqrt{F0_2}}$$

$$E \propto \frac{B}{F0_1} A_1^2 \qquad \frac{B}{F0_2} A_2^2$$



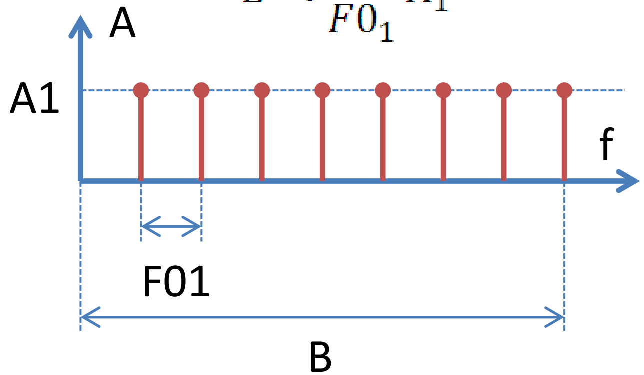
Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0

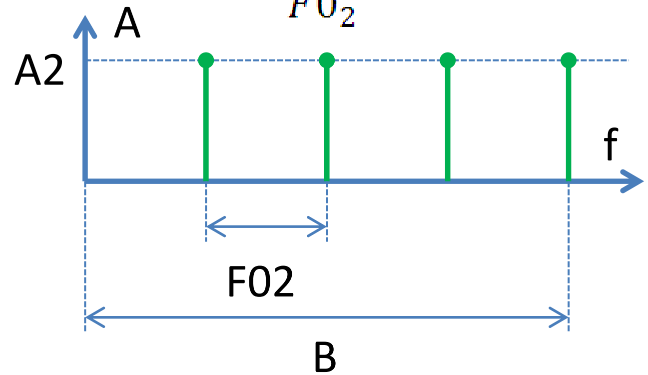
$$X[m] = \frac{1}{N}\mathrm{FFT}_N\{x[n]\} \quad \Leftrightarrow \quad x[n] = N\ \mathrm{FFT}_N^{-1}\{X[m]\}$$

$$\frac{A_1}{\sqrt{F0_1}} = \frac{A_2}{\sqrt{F0_2}}$$

$$E \propto \frac{B}{F0_1}A_1^2$$



$$\frac{B}{F0_2}A_2^2$$



Ahocoder, an HNM-based vocoder

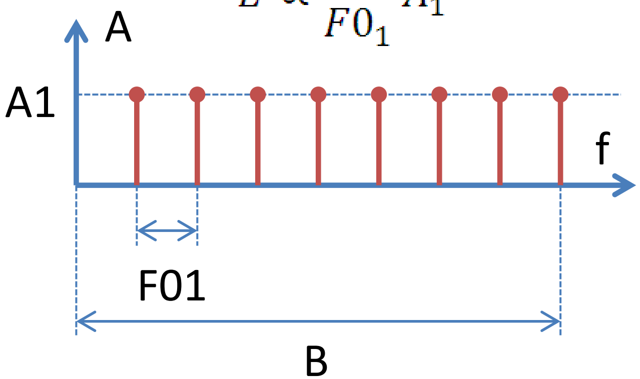
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0

$$X[m] = \frac{1}{\sqrt{N f_s}} \mathrm{FFT}_N\{x[n]\} \iff x[n] = \sqrt{N f_s}\, \mathrm{FFT}_N^{-1}\{X[m]\}$$

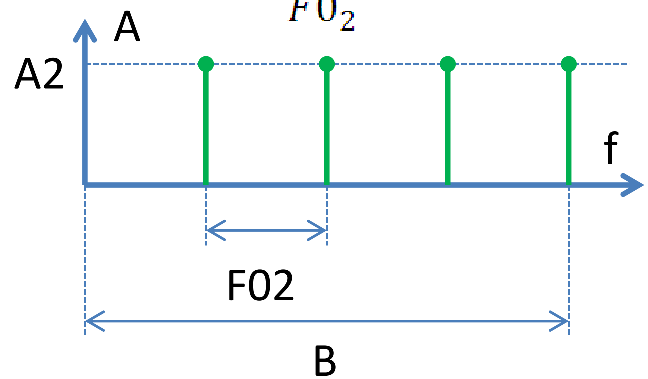
F0 norm $f_0 = f_s / N$

$$\frac{A_1}{\sqrt{F0_1}} = \frac{A_2}{\sqrt{F0_2}}$$

$$E \propto \frac{B}{F0_1} A_1^2$$



$$\frac{B}{F0_2} A_2^2$$



Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

Cepstral order

Scaled freq (Mel)

$$S(\omega) = c_0 + 2\sum_{q=1}^{p} c_q \cos q\tilde{\omega}$$

F0-normalized log-amplitude envelope

Not always used (not in HTS, for instance)

Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

F0-normalized amplitudes → spectral envelope → MCEP coeffs

F0-normalized amplitudes → MCEP coeffs

Ahocoder, an HNM-based vocoder

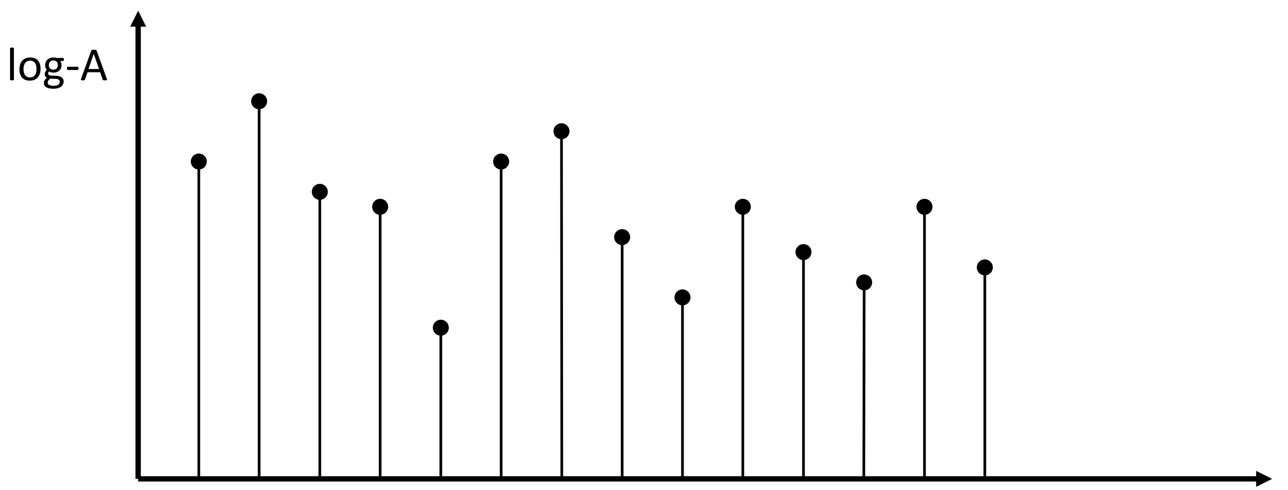
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

Unvoiced: $$S^{(k)}[m] = \log\left(\frac{1}{\sqrt{Lf_s}}\left|FFT_N\left\{s[n]\cdot w[n-n_k]\right\}\right|\right)$$

Ahocoder, an HNM-based vocoder

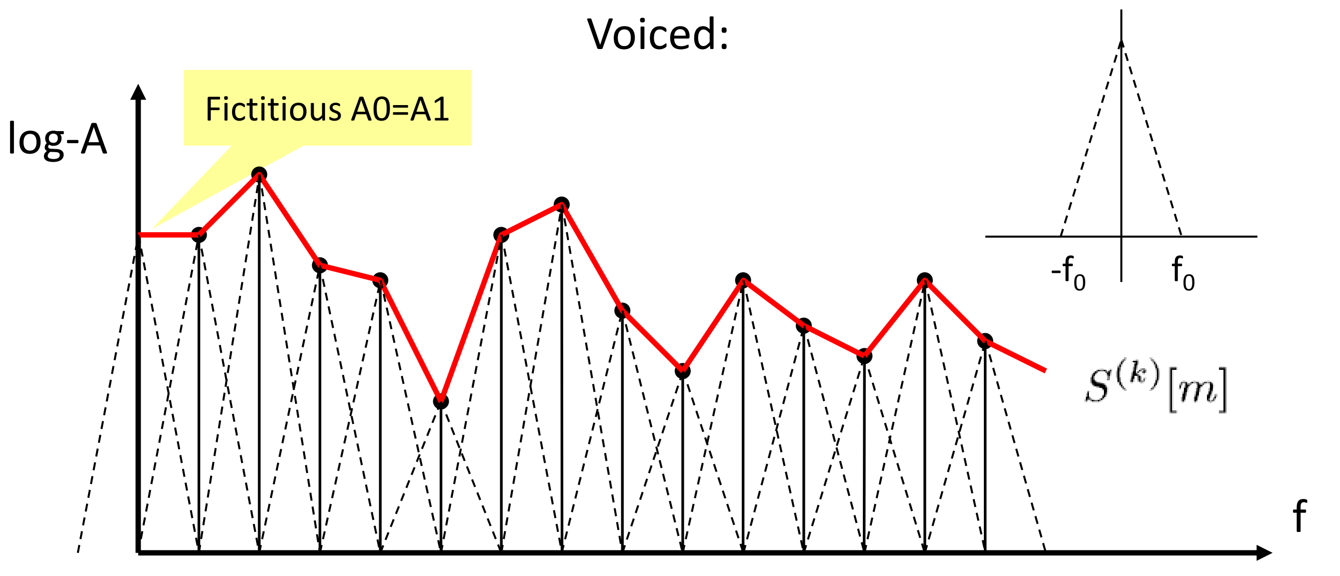
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

Voiced:



Ahocoder, an HNM-based vocoder

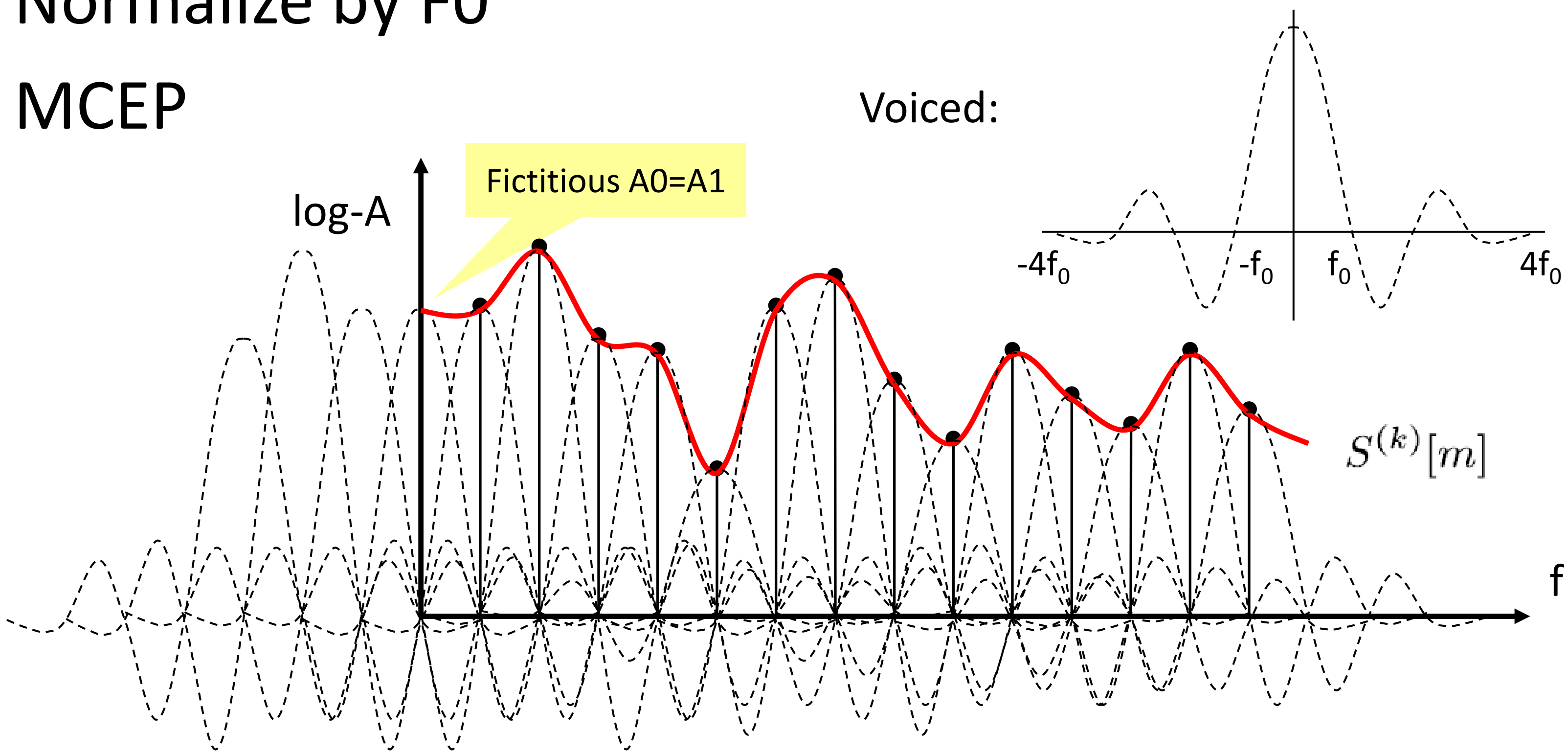
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP



Ahocoder, an HNM-based vocoder

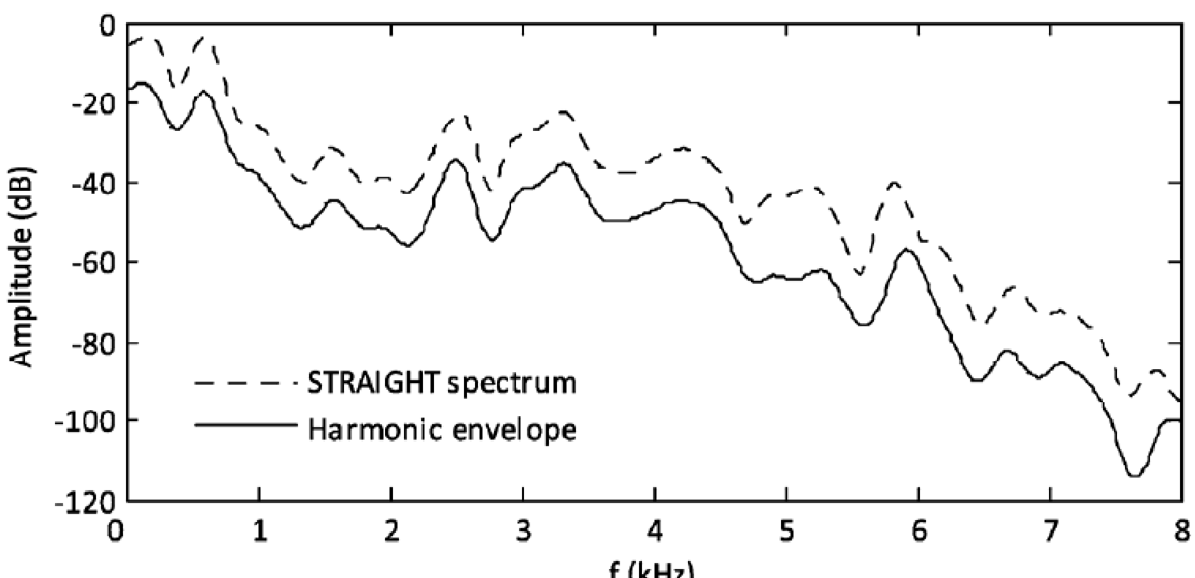
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

Voiced:



Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

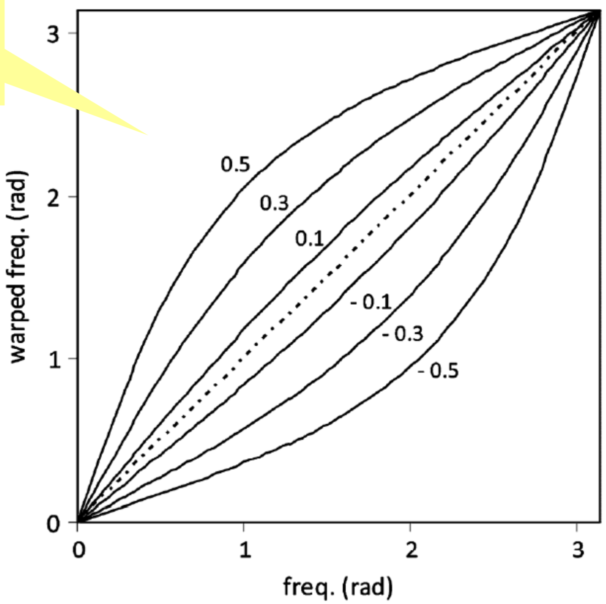


Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

Mel scale: alpha=0.42 for fs=16kHz

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}$$



$$\boldsymbol{c} \stackrel{\text{def}}{=} \text{FFT}^{-1}\{S[m]\}$$

Without 2

$$c_\alpha^{(i)}(m) = \left\{ \begin{array}{ll} c(-i) + \alpha\, c_\alpha^{(i-1)}(0), & m = 0 \\ (1-\alpha^2)\, c_\alpha^{(i-1)}(0) + \alpha\, c_\alpha^{(i-1)}(1), & m = 1 \\ c_\alpha^{(i-1)}(m-1) + \alpha\left(c_\alpha^{(i-1)}(m) - c_\alpha^{(i)}(m-1)\right), & m = 2, 3, \ldots, M_2 \end{array} \right\}, \; i = -M_1, \ldots, -1, 0$$

(Tokuda et al., 1994)

Ahocoder, an HNM-based vocoder

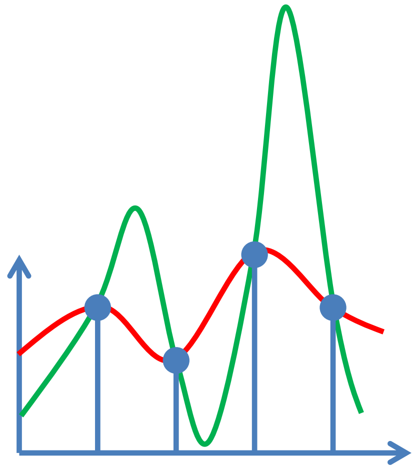
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

$$S(\omega) = c_0 + 2\sum_{q=1}^{p} c_q \cos q\tilde{\omega}$$

$$\begin{bmatrix} 1 & 2\cos\tilde{\omega}_1 & \cdots & 2\cos p\tilde{\omega}_1 \\ 1 & 2\cos\tilde{\omega}_2 & \cdots & 2\cos p\tilde{\omega}_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos\tilde{\omega}_I & \cdots & 2\cos p\tilde{\omega}_I \end{bmatrix} \mathbf{c} = \begin{bmatrix} \log \hat{A}_1 \\ \log \hat{A}_2 \\ \vdots \\ \log \hat{A}_I \end{bmatrix}$$

Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP



$$S(\omega) = c_0 + 2\sum_{q=1}^{p} c_q \cos q\tilde{\omega}$$

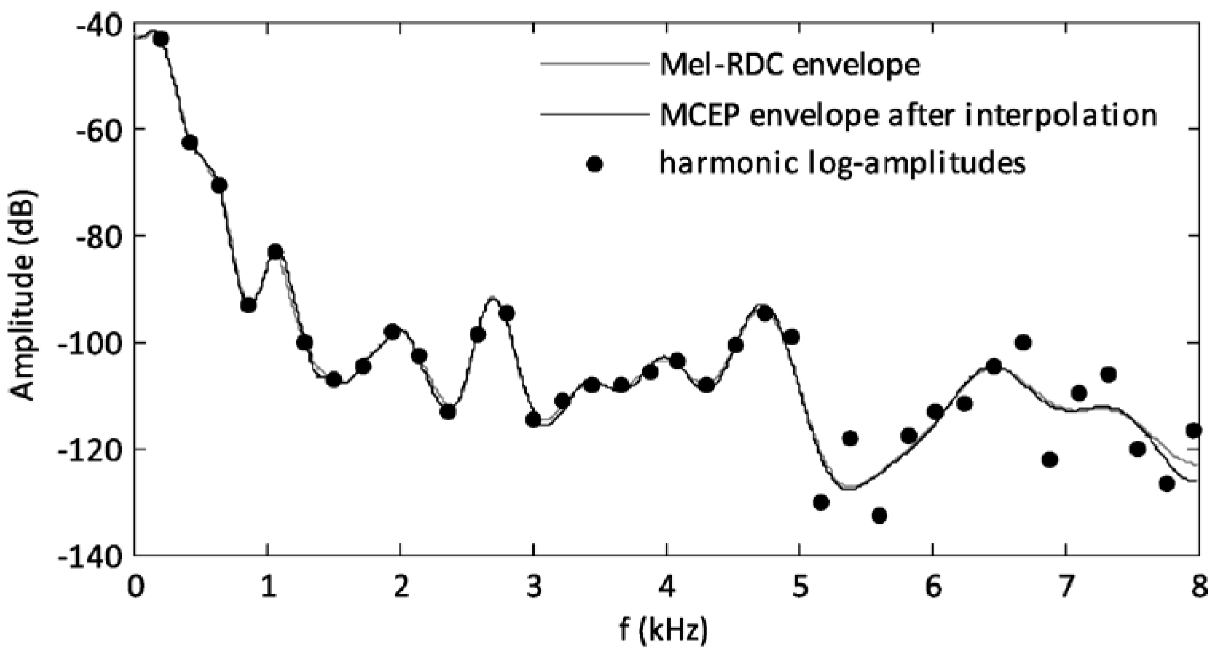
$$\begin{bmatrix} 1 & 2\cos\tilde{\omega}_1 & \cdots & 2\cos p\tilde{\omega}_1 \\ 1 & 2\cos\tilde{\omega}_2 & \cdots & 2\cos p\tilde{\omega}_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos\tilde{\omega}_I & \cdots & 2\cos p\tilde{\omega}_I \end{bmatrix} \mathbf{c} + Reg.Term. = \begin{bmatrix} \log\hat{A}_1 \\ \log\hat{A}_2 \\ \vdots \\ \log\hat{A}_I \end{bmatrix}$$

Mel Regularized Discrete Cepstrum

(Cappé et al., 1995)

Ahocoder, an HNM-based vocoder

- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP



Ahocoder, an HNM-based vocoder

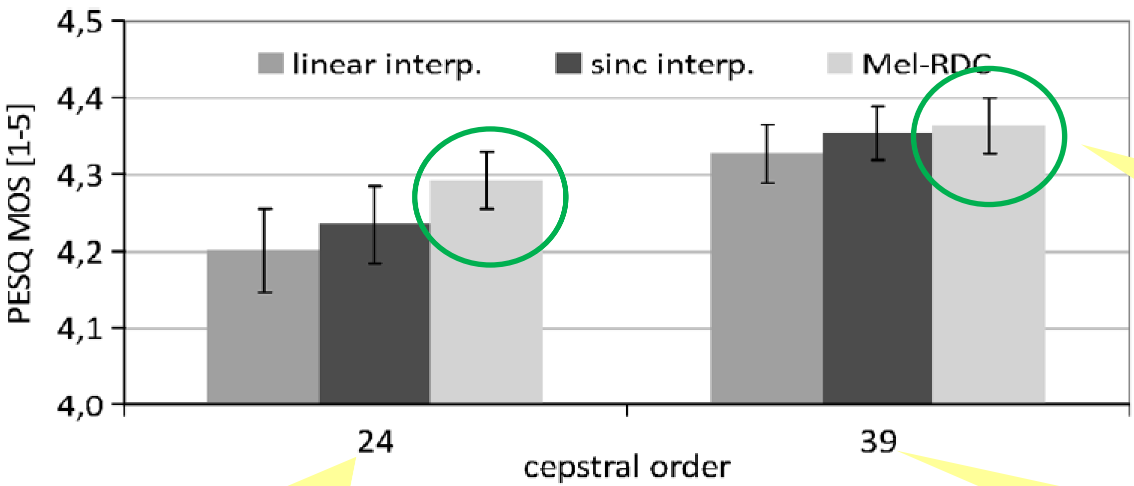
- Spectral analysis
  - HM instead of HNM
  - Normalize by F0
  - MCEP

$$\{A_i^{(k)}, \varphi_i^{(k)}\} = \arg\ \min \sum_n w^2[n]\Big(s[n+kN] - \underbrace{h^{(k)}[n]}_{h^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right)}\Big)^2$$

  - FFT-based harmonic analysis
    - Much faster and less accurate, BUT when MCEP coeffs are involved the difference is hard to perceive!!

Ahocoder, an HNM-based vocoder

- Spectral analysis
  - Experiments:
    - Resynthesis quality MOS predicted by PESQ, ITU-T/P.862 (keeping uv frames and measured phases)



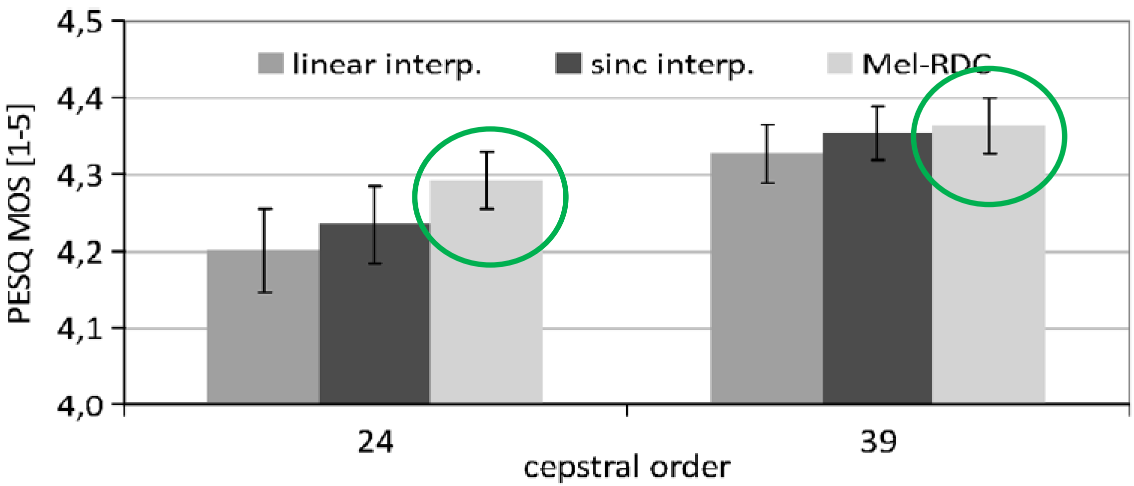
Mel-RDC slightly superior to interpolation of envelope + MCEP

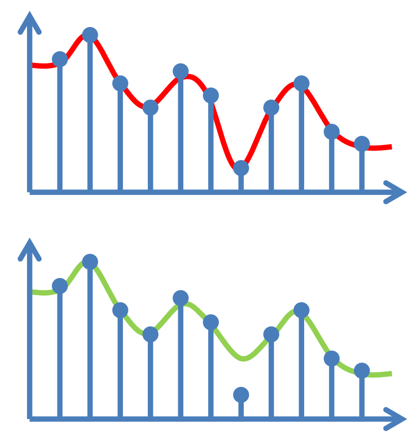
Typical order in voice conversion for fs=16kHz

Typical order in HMM-based speech synthesis for fs=16kHz

Ahocoder, an HNM-based vocoder

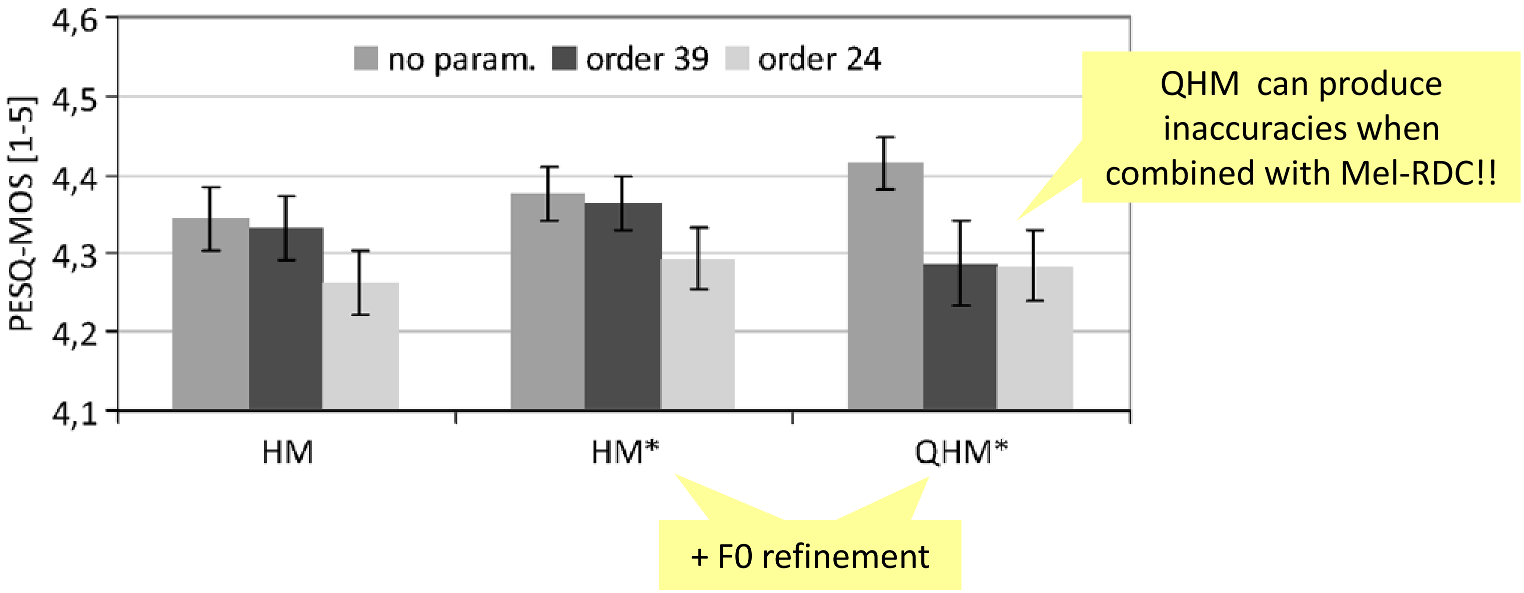
- Spectral analysis
  – Experiments:
    - Resynthesis quality MOS predicted by PESQ, ITU-T/P.862 (keeping uv frames and measured phases)





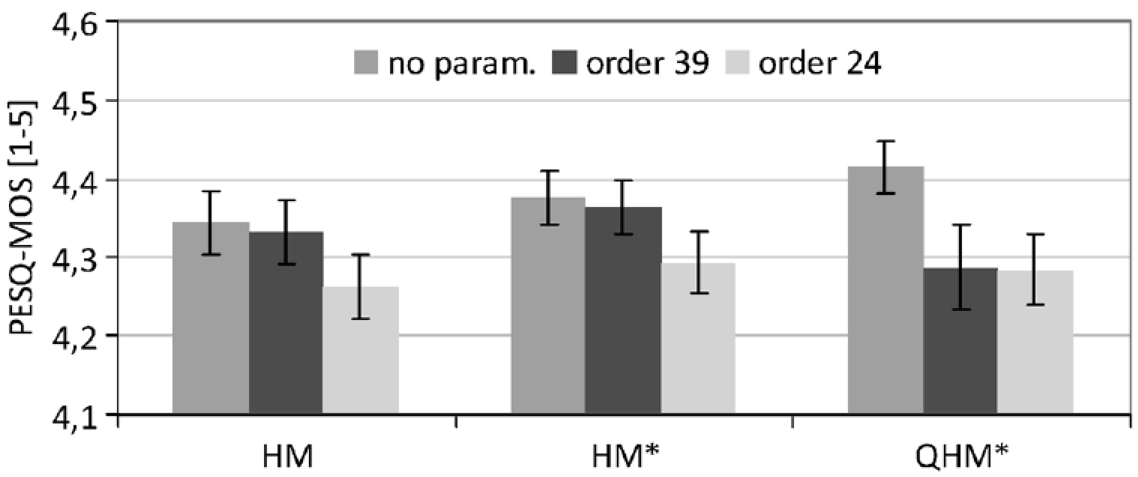
Ahocoder, an HNM-based vocoder

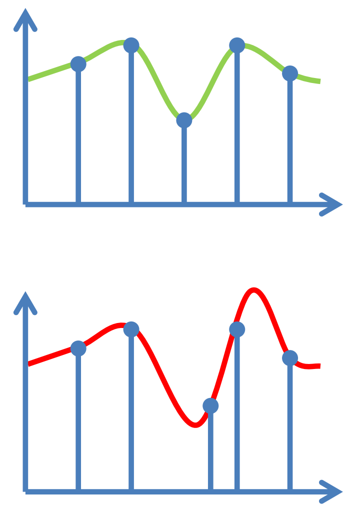
- Spectral analysis
  - Experiments:
    - Resynthesis quality MOS predicted by PESQ, ITU-T/P.862 (keeping uv frames and measured phases)



Ahocoder, an HNM-based vocoder

- Spectral analysis
  - Experiments:
    - Resynthesis quality MOS predicted by PESQ, ITU-T/P.862 (keeping uv frames and measured phases)





Ahocoder, an HNM-based vocoder

- Spectral analysis
  - Experiments:
    - Resynthesis quality MOS predicted by PESQ, ITU-T/P.862 (keeping uv frames and measured phases)
    - Accuracy of statistical modeling: average log-probability per frame given by HTS (v2.1.1)

Method \ Voice	Female	Male
Sinc interp. + MCEP	$1.0095 \cdot 10^2$	$1.1034 \cdot 10^2$
Mel-RDC	$1.0339 \cdot 10^2$	$1.1176 \cdot 10^2$
f_0 ref. + Mel-RDC	$1.0446 \cdot 10^2$	$1.1519 \cdot 10^2$

Again, Mel-RDC slightly superior to interpolation of envelope + MCEP, and F0 refinement helps

Outline

- ~~Introduction~~
- Ahocoder, an HNM-based vocoder
  - ~~F0 estimation~~
  - ~~MVF estimation~~
  - ~~Spectral analysis~~
  - Speech waveform reconstruction
  - Evaluation
- Conclusions

Ahocoder, an HNM-based vocoder

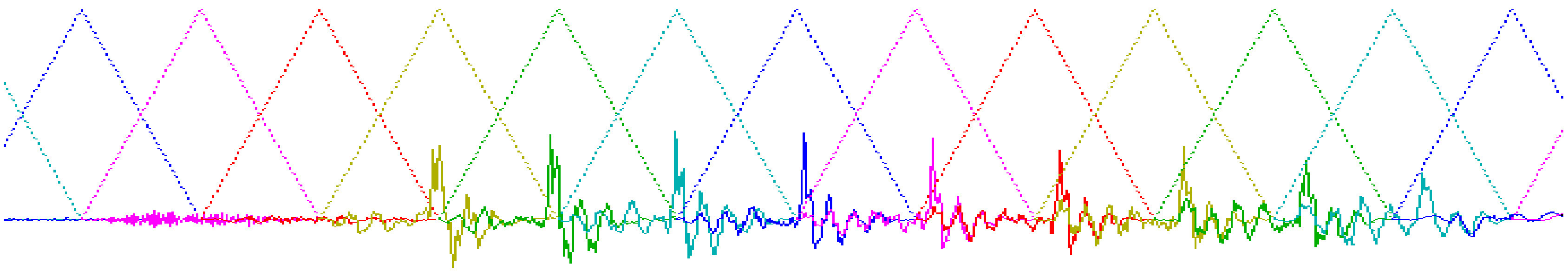
- Speech waveform reconstruction

Only part that will take part in the TTS process!!

Triang win

Frame k

$$s[n] = \sum_{k=1}^{K} t[n - n_k] \cdot s^{(k)}[n - n_k]$$



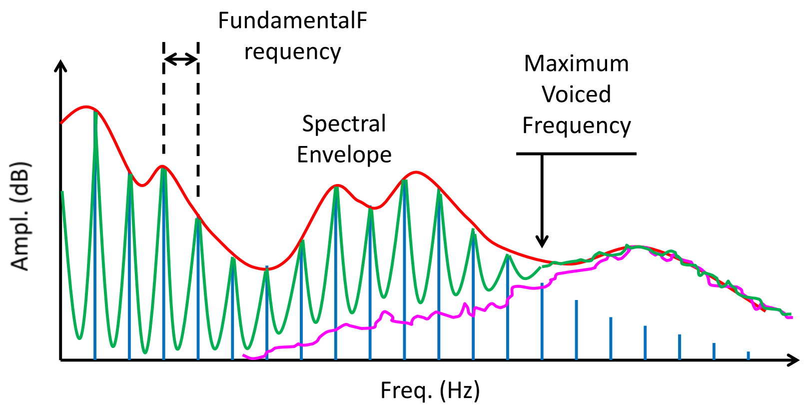
Not pitch-synchronous but constant frame length of 10ms (from nk-5ms to nk+5ms)

Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s[n] = \sum_{k=1}^{K} \overbrace{t[n-n_k]}^{\text{Triang win}} \cdot \overbrace{s^{(k)}}^{\text{Frame k}}[n-n_k]$$

$$s^{(k)}[n] = \underbrace{\sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right)}_{\text{Harmonic component}} + \underbrace{\rho \cdot r^{(k)}[n] \cdot e^{(k)}[n]}_{\text{Noisy component}}$$



Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot e^{(k)}[n]$$

Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot \boxed{e^{(k)}[n]}$$

F0 denorm

$$e^{(k)}[n] = \sqrt{N f_s}\, \mathrm{FFT}_N^{-1}\{E^{(k)}[m]\}$$

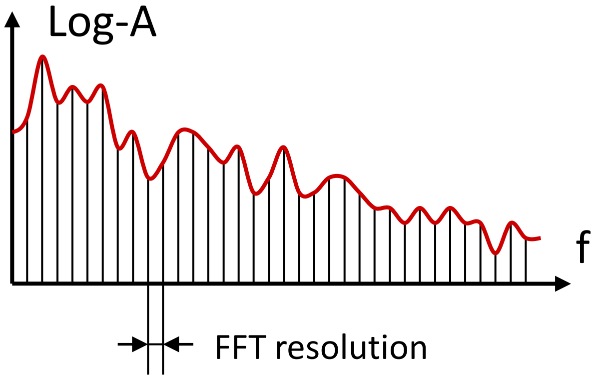
Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot \boxed{e^{(k)}[n]}$$

F0 denorm

$$e^{(k)}[n] = \sqrt{N f_s}\ \mathrm{FFT}_N^{-1}\{E^{(k)}[m]\}$$



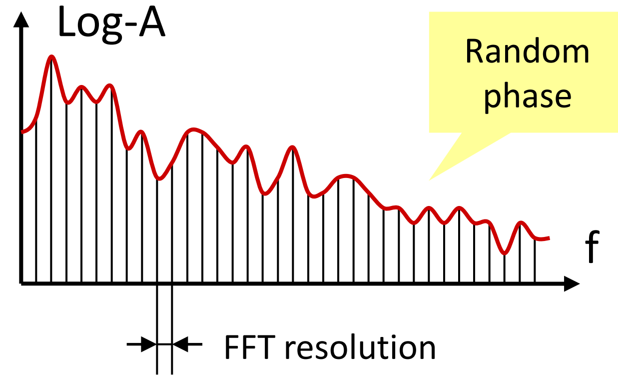
Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot \boxed{e^{(k)}[n]}$$

F0 denorm

$$e^{(k)}[n] = \sqrt{N f_s}\, \mathrm{FFT}_N^{-1}\{E^{(k)}[m]\}$$



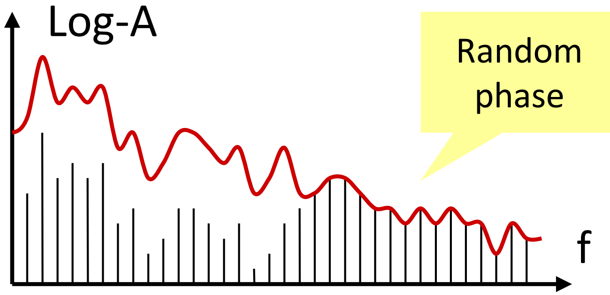
Ahocoder, an HNM-based vocoder

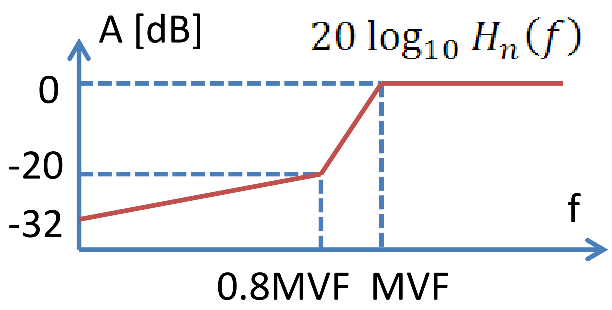
- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot \boxed{e^{(k)}[n]}$$

F0 denorm

$$e^{(k)}[n] = \sqrt{N f_s}\, \mathrm{FFT}_N^{-1}\{E^{(k)}[m]\}$$





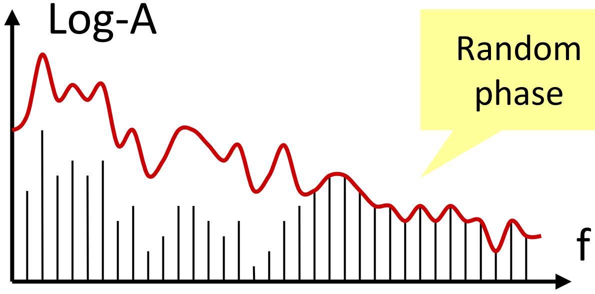
Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

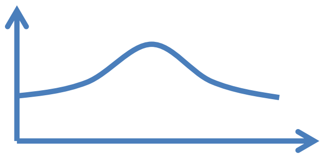
$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot \boxed{r^{(k)}[n]} \cdot e^{(k)}[n]$$

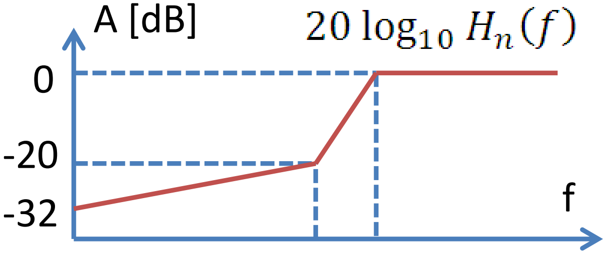
F0 denorm

$$e^{(k)}[n] = \sqrt{N f_s}\, \mathrm{FFT}_N^{-1}\{E^{(k)}[m]\}$$



Random phase





Ahocoder, an HNM-based vocoder

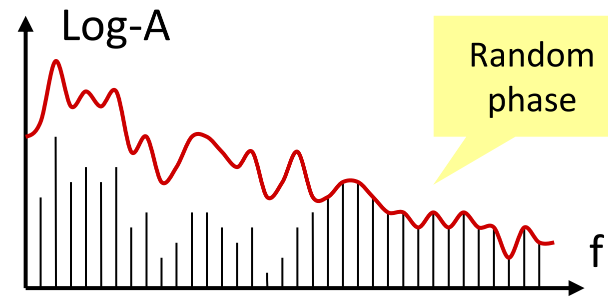
- Speech waveform reconstruction

Compensate interference: 1.21

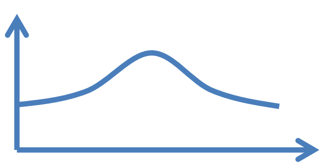
$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \boxed{\rho} \cdot r^{(k)}[n] \cdot e^{(k)}[n]$$

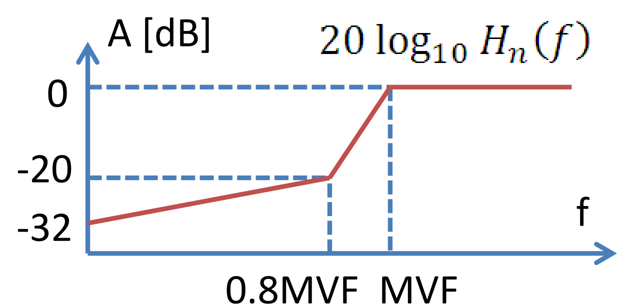
F0 denorm

$$e^{(k)}[n] = \sqrt{N f_s}\, \mathrm{FFT}_N^{-1}\{E^{(k)}[m]\}$$



Random phase





Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} \boxed{A_i^{(k)}} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot e^{(k)}[n]$$

F0 denorm

By definition

$$A_i^{(k)} = 2\sqrt{f_0^{(k)}} \cdot H_h^{(k)}\left(i f_0^{(k)}\right) \cdot \exp\left(c_0^{(k)} + 2\sum_{q=1}^{p} c_q^{(k)} \cos q i \widetilde{\omega}_0^{(k)}\right)$$

Low-pass filter (MVF)

Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} \boxed{A_i^{(k)}} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot e^{(k)}[n]$$

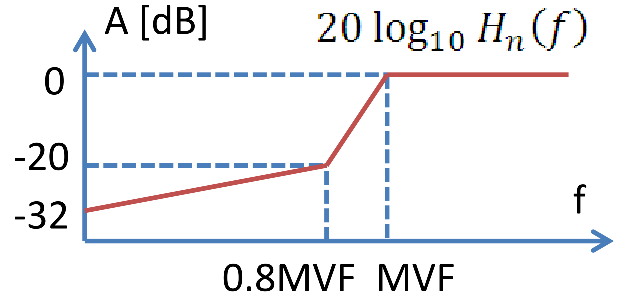
F0 denorm

By definition

$$A_i^{(k)} = 2\sqrt{f_0^{(k)}} \cdot H_h^{(k)}\left(i f_0^{(k)}\right) \cdot \exp\left(c_0^{(k)} + 2\sum_{q=1}^{p} c_q^{(k)} \cos q \widetilde{i\omega}_0^{(k)}\right)$$

Low-pass filter (MVF)

$$H_h^{(k)}(f) = \sqrt{1 - {H_n^{(k)}}^2(f)}$$



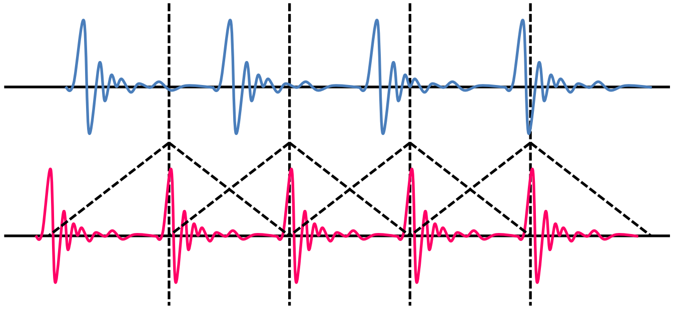
Ahocoder, an HNM-based vocoder

- Speech waveform reconstruction

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \boxed{\varphi_i^{(k)}}\right) + \rho \cdot r^{(k)}[n] \cdot e^{(k)}[n]$$

$$\varphi_i^{(k)} = \overbrace{-2 \sum_{q=1}^{p} c_q^{(k)} \sin q i \widetilde{\omega}_0^{(k)}}^{\text{Min phase}} \overbrace{+\, i\phi^{(k)}}^{\text{Lin term}}$$

$$\phi^{(k)} = \phi^{(k-1)} + \frac{1}{2}\left(\omega_0^{(k)} + \omega_0^{(k-1)}\right)\left(n_k - n_{k-1}\right)$$



(McAulay & Quatieri, 1995)

Ahocoder, an HNM-based vocoder

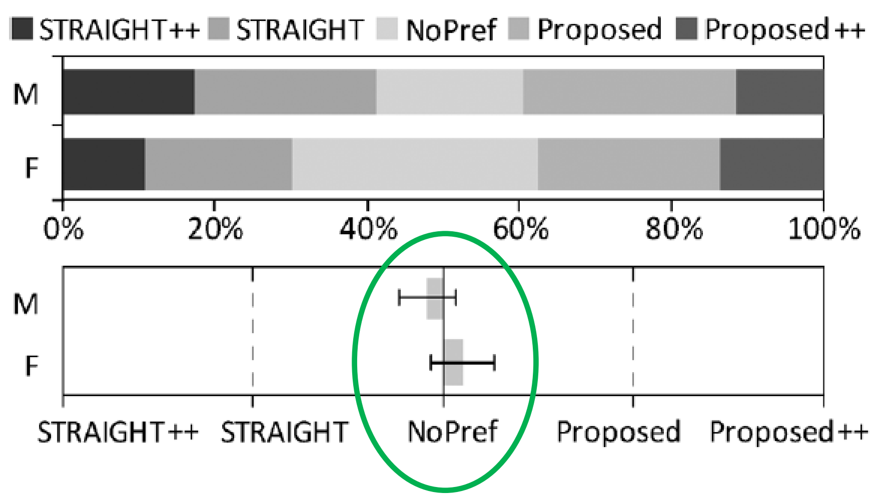
- Speech waveform reconstruction
  - Phase info is discarded by Ahocoder
  - There are attempts to model the non-linear non-minimum part of phase (Degottex & Erro, 2014)

Outline

- ~~Introduction~~
- Ahocoder, an HNM-based vocoder
  - ~~F0 estimation~~
  - ~~MVF estimation~~
  - ~~Spectral analysis~~
  - ~~Speech waveform reconstruction~~
  - Evaluation
- Conclusions

Ahocoder, an HNM-based vocoder

- Evaluation
  - Comparison with STRAIGHT, 30 listeners



Not equivalent but equally suitable for synthesis

Outline

- ~~Introduction~~
- ~~Ahocoder, an HNM-based vocoder~~
  - ~~F0 estimation~~
  - ~~MVF estimation~~
  - ~~Spectral analysis~~
  - ~~Speech waveform reconstruction~~
  - ~~Evaluation~~
- Conclusions

Conclusions

- HM/HNM good framework for vocoder development
- QHM-based pitch refinement (0-MVF band, amplitude-related weights) helps
- Explicit MVF analysis and modeling helps
- Mel-RDC slightly better than log-amplitude envelope intepolation + MCEP
- Don’t discard the FFT-based harmonic analysis approach
- Don’t forget to normalize amplitudes by F0!
- Be careful with QHM + Mel-RDC!!
- Overall, the result is comparable with STRAIGHT

Outline

- ~~Introduction~~
- ~~Ahocoder, an HNM-based vocoder~~
  - ~~F0 estimation~~
  - ~~MVF estimation~~
  - ~~Spectral analysis~~
  - ~~Speech waveform reconstruction~~
  - ~~Evaluation~~
- ~~Conclusions~~

Acknowledgements

Iñaki Sainz

Eva Navas

Inma Hernáez

GOBIERNO DE ESPAÑA

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

UNIÓN EUROPEA

FONDO EUROPEO DE DESARROLLO REGIONAL

"Una manera de hacer Europa"

EUSKO JAURLARITZA

GOBIERNO VASCO

References

- D. Erro et al., “Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis”, IEEE J STSP, 2014
- Y. Stylianou, “Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis”, IEEE T SAP, 2001
- H. Kawahara et al., “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction (…)”, Speech Commun., 1999
- T. Raitio et al., “HMM-based speech synthesis utilizing glottal inverse filtering”, IEEE T ASLP, 2011
- Y. Pantazis et al., “Iterative Estimation of Sinusoidal Signal Parameters”, IEEE SPL, 2010
- X. Rodet, “Musical sound signals analysis/synthesis: Sinusoidal + residual and elementary waveform models”, Appl. Signal Process., 1997
- K. Tokuda et al., “Mel-generalized cepstral analysis - a unified approach to speech spectral estimation”, Proc. ICSLP, 1994
- O. Cappé et al., “Regularized estimation of cepstrum envelope from discrete frequency points,” Proc. IEEE WASPAA, 1995
- G. Degottex, D. Erro, “A uniform phase representation for the harmonic model in speech synthesis applications”, EURASIP J ASMP, 2014
- R. McAulay, T. Quatieri, “Sinusoidal Coding”, chapter in “Speech Coding and Synthesis”, Elsevier, 1995

Sinusoidal Models for Text-to-Speech Synthesis

… or …

Development of Ahocoder, an HNM-based vocoder

Daniel Erro - derro@aholab.ehu.es

ikerbasque
Basque Foundation for Science

Bilbao, Spain

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea