KATHOLIEKE UNIVERSITEIT
LEUVEN



Linear Predictive Modeling and Sparse Approximation of Audio Signals

Toon van Waterschoot
Dept. ESAT-SCD, KU Leuven, Belgium

Statistics Seminar
Nov. 30, 2012, Lund University

Outline

- Introduction
- Recap: Linear prediction of speech signals
- Linear prediction of audio signals [modeling]
  - fundamental issues
  - alternative models
  - performance evaluation
- Sparse linear prediction of audio signals [estimation]
- Conclusion



Outline

- Introduction
- Recap: Linear prediction of speech signals
- Linear prediction of audio signals
  - fundamental issues
  - alternative models
  - performance evaluation
- Sparse linear prediction of audio signals
- Conclusion



Introduction (1): Motivation

- Linear prediction (LP) provides a simple yet powerful and widely used model for *speech signals*
- Musical *audio signals* are typically represented using
  - non-parametric (transform-domain) models (MP3, AAC, ...)
  - parametric (transients + sinusoids + noise) models

We claim that LP modeling of audio signals has high potential, provided that some fundamental issues are dealt with

- Envisaged applications:
  - **coding:** represent, store, and transmit signals using relatively small number of parameters
  - **analysis:** summarize characteristic signal behavior in low-dimensional parameter space
  - **synthesis:** generate synthetic signals from limited number of parameters
  - **whitening:** invertible parametric signal models can be useful for signal whitening

Introduction (2): LP models

- All-pole / Autoregressive (AR) models & Pole-Zero / Autoregressive Moving Average (ARMA) models:
  - **linear prediction interpretation:** prediction of current signal sample based on past signal samples and (moving average of) excitation signal

AR: $x(t) = -a_1 x(t-1) - a_2 x(t-2) - \ldots - a_P x(t-P) + e(t)$

ARMA: $x(t) = -a_1 x(t-1) - a_2 x(t-2) - \ldots - a_P x(t-P) + b_0 e(t) + b_1 e(t-1) + \ldots b_Q e(t-Q)$

  - **source-filter interpretation:** modeling signal as output of linear all-pole/pole-zero filter driven by excitation (source) signal

AR: $x(t) = \frac{1}{A(z)} e(t) \qquad A(z) = 1 + a_1 z^{-1} + \ldots + a_P z^{-P}$

ARMA: $x(t) = \frac{B(z)}{A(z)} e(t) \qquad B(z) = b_0 + b_1 z^{-1} + \ldots + b_Q z^{-Q}$

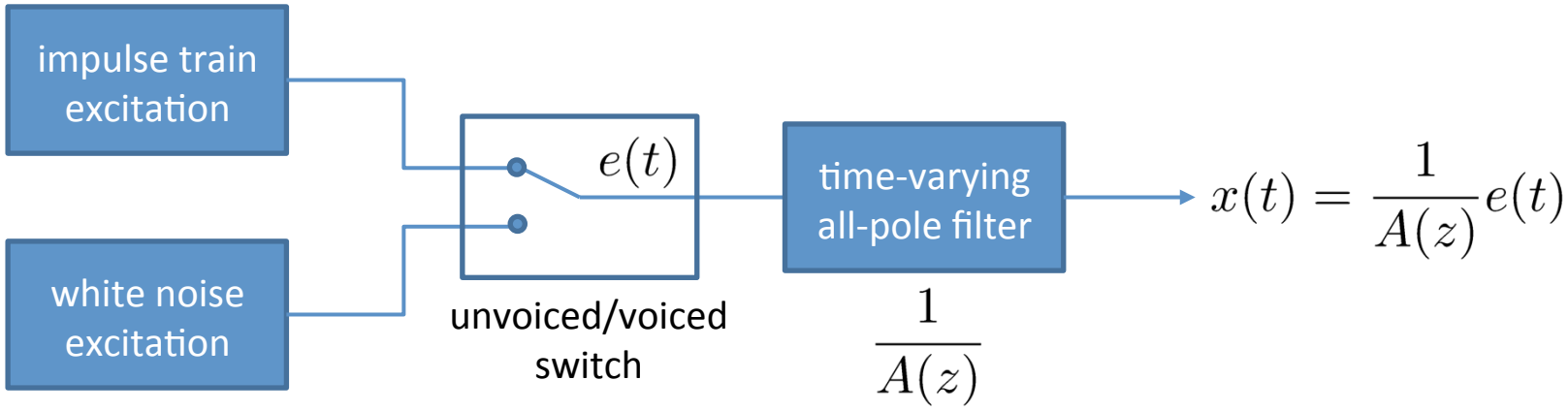
Outline

- Introduction
- Recap: Linear prediction of speech signals
- Linear prediction of audio signals
  - fundamental issues
  - alternative models
  - performance evaluation
- Sparse linear prediction of audio signals
- Conclusion



Linear prediction of speech signals (1): Source-filter model of speech production

- Correspondence between all-pole source-filter model and human speech production system:



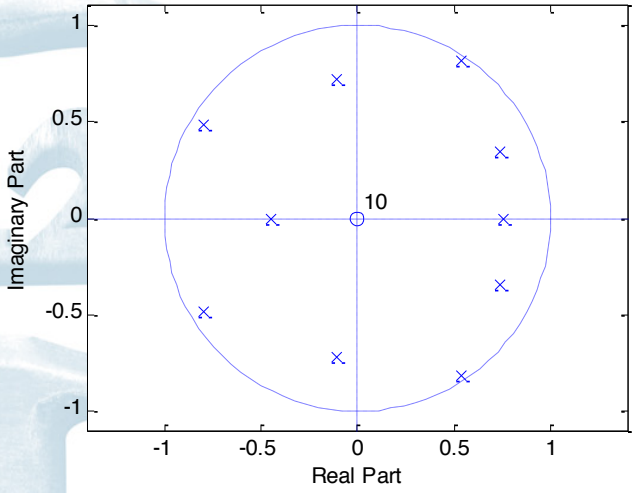
  - glottal air flow represented as broadband noise signal (white noise excitation for unvoiced speech)
  - vocal cords shape glottal air flow into periodic signal (impulse train excitation for voiced speech)
  - vocal tract behaves as time-varying all-pole filter (spectral shaping filter applied to speech excitation signal)

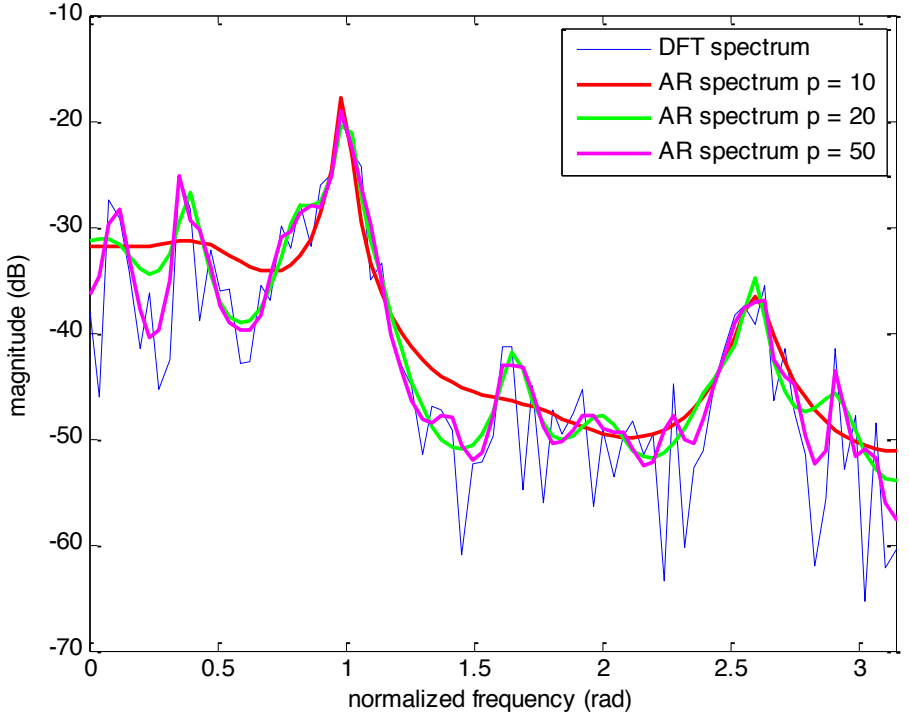
Linear prediction of speech signals (2): Practical issues

- User choices:
  - **length of speech signal segment** (N) = compromise between
    - accurate estimation of autocorrelation function: $N \nearrow$
    - speech stationarity throughout signal segment: $N \searrow$

    rule of thumb: N ~ 20 ms (e.g., $N = 160$ at $f_s = 8$ kHz)
  - **model order** (P) = compromise between
    - model accuracy: $P \nearrow$
    - model complexity: $P \searrow$

    rule of thumb: $P = 10-20$





Linear prediction of speech signals (3): Pitch prediction

- Voiced speech:
  - periodicity of voiced speech (originating at vocal cords) cannot be modeled using (low-order) all-pole model

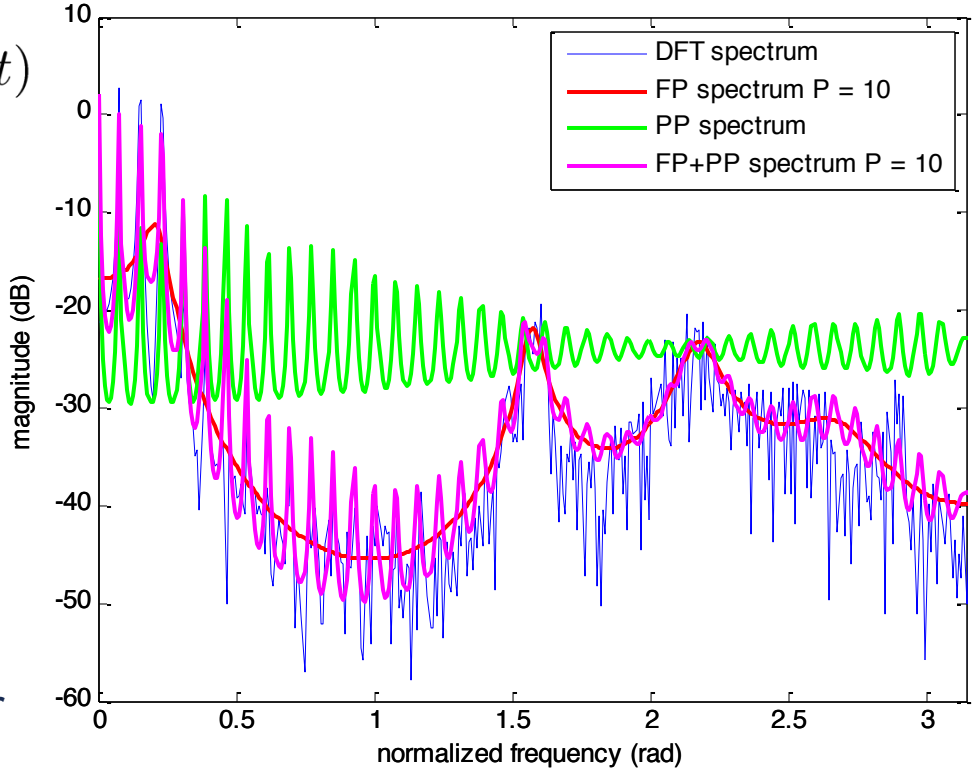
    (all-pole model only represents signal autocorrelation up to lag P)
  - cascade of two all-pole models (formant predictor + pitch predictor)

$$x(t) = \frac{1}{A_{\mathrm{FP}}(z)A_{\mathrm{PP}}(z)}e(t)$$

- Pitch predictor:

$$A_{\mathrm{PP}}(z) = 1 - z^{-K}$$

  - pitch lag: $20 \leq K \leq 160$
  - comb filter behavior
  - improvements:
    - interpolation filter
    - multi-tap pitch predictor



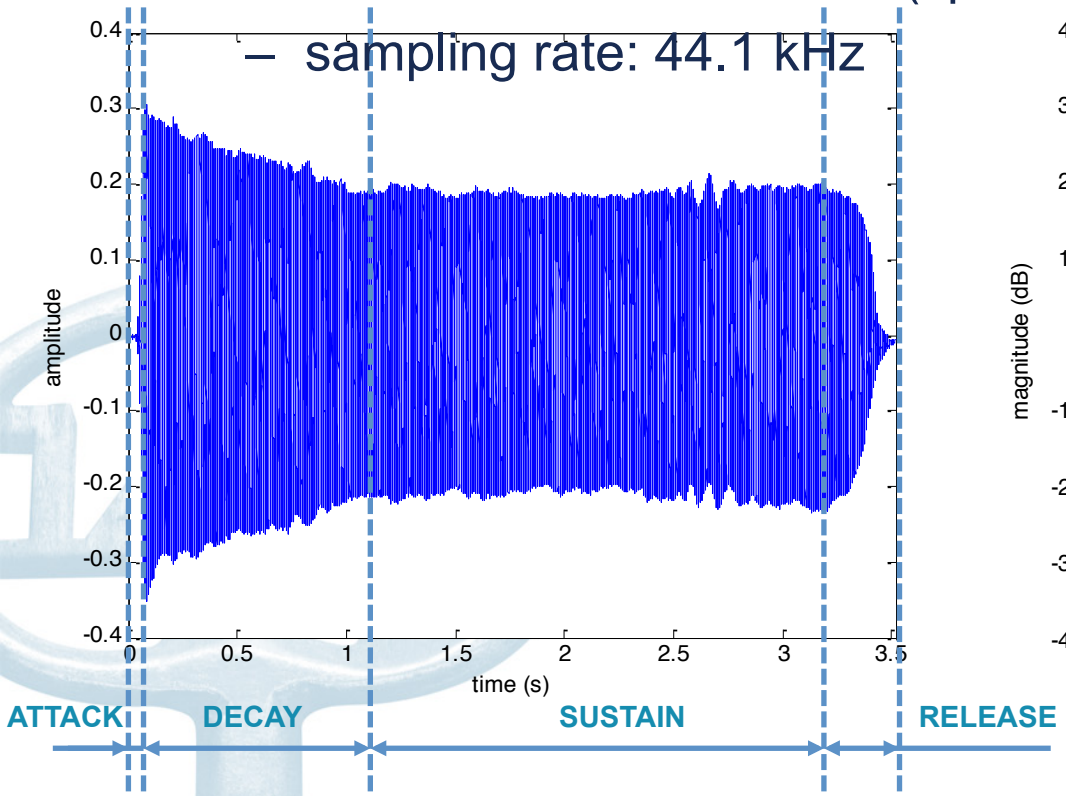
Outline

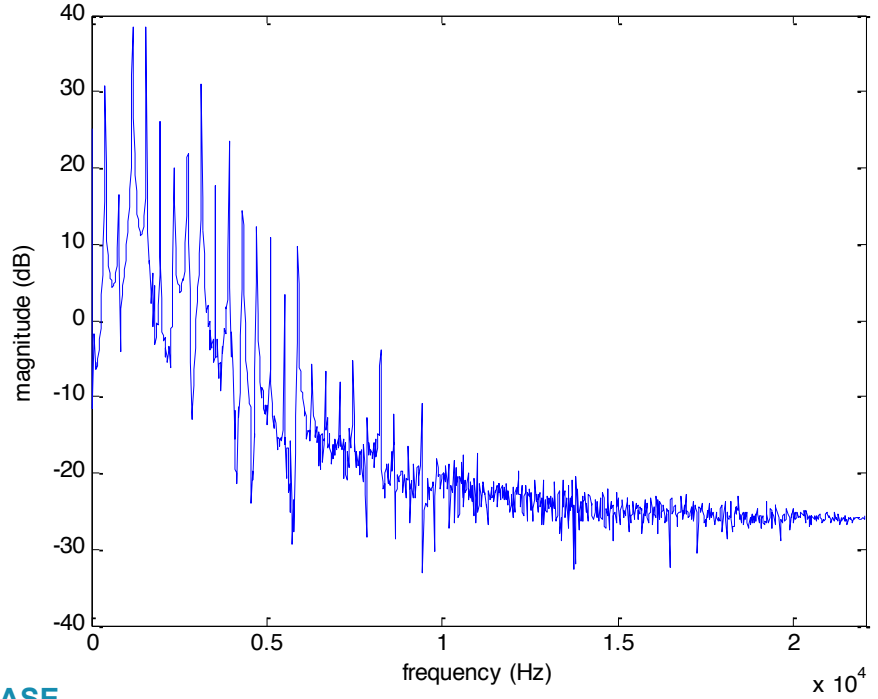
- Introduction
- Recap: Linear prediction of speech signals
- Linear prediction of audio signals*
  - fundamental issues
  - alternative models
  - performance evaluation
- Sparse linear prediction of audio signals
- Conclusion

* Toon van Waterschoot and Marc Moonen, "Comparison of linear prediction models for audio signals", *EURASIP J. Audio, Speech, Music Process.*, vol. 2008, Article ID 706935, 24 p., 2008.

Linear prediction of audio signals (1): Fundamental issues (1)

- Difference between speech and audio signals:
  - temporal structure: *attack-decay-sustain-release*
    (stationarity interval ranges from a few millisec to a few sec)
  - spectral structure: *tonality* (damped sinusoidal components)
    timbre (spectral envelope)
  - sampling rate: 44.1 kHz





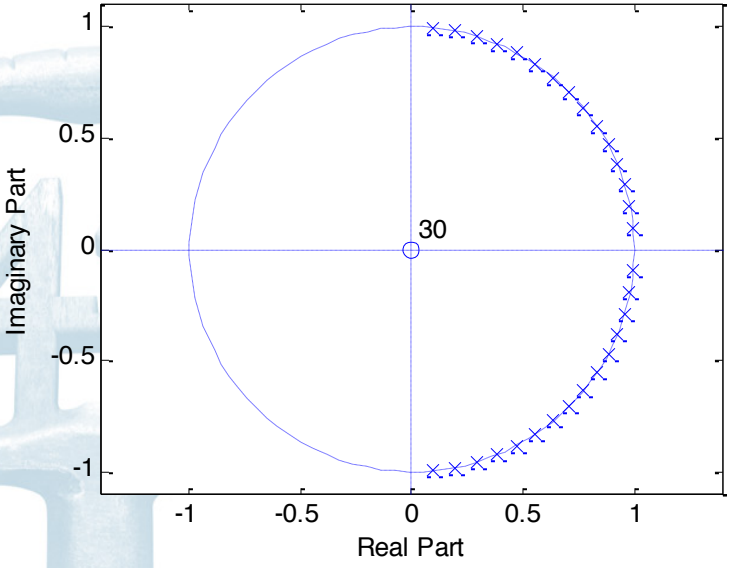
Linear prediction of audio signals (2): Fundamental issues (2)

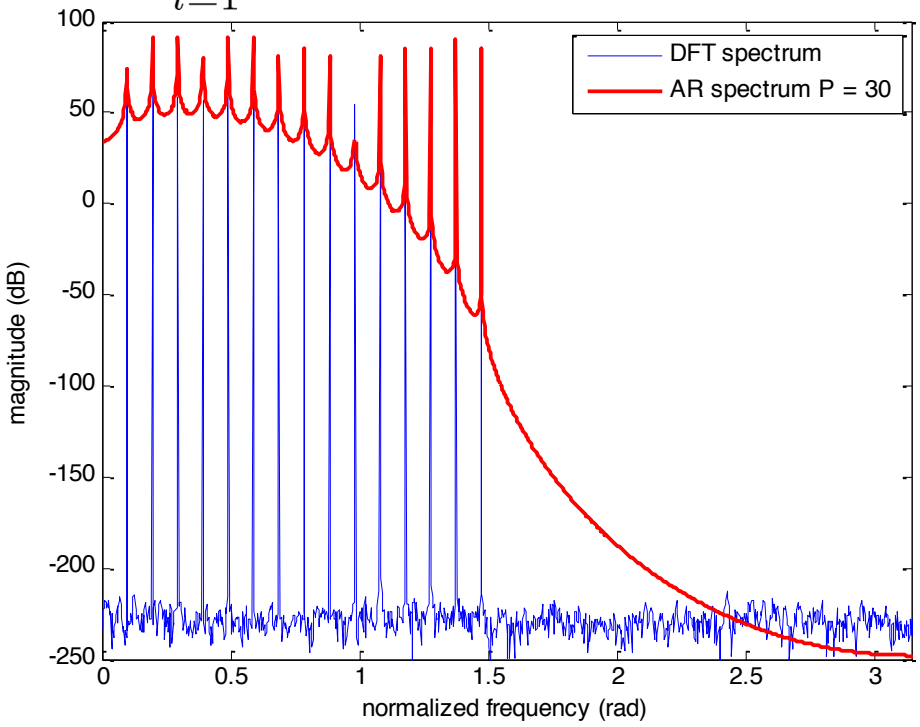
- Sinusoidal signals:
  - sum of M sinusoids admits AR($2M$) representation:

$$x(t) = \sum_{m=1}^{M} \alpha_m \cos(\omega_m t + \phi_m) \Leftrightarrow x(t) = -a_1 x(t-1) - \ldots - a_{2M} x(t-2M) + b_0 \delta(t)$$

  - prediction error filter (PEF) $A(z) = \prod_{l=1}^{P} \left(1 - 2\nu_l \cos\theta_l z^{-1} + \nu_l^2 z^{-2}\right)$

$$P = 2M \Rightarrow \begin{cases} \theta_l = \omega_m \\ \nu_l = 1 \end{cases}$$

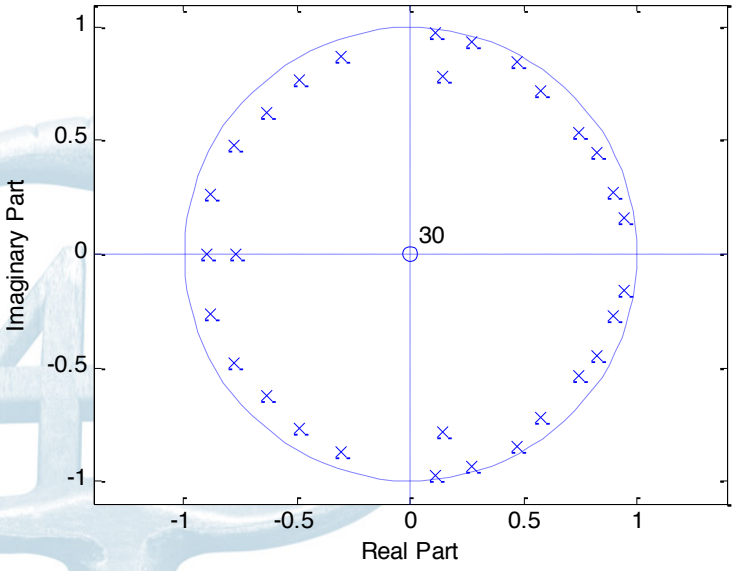


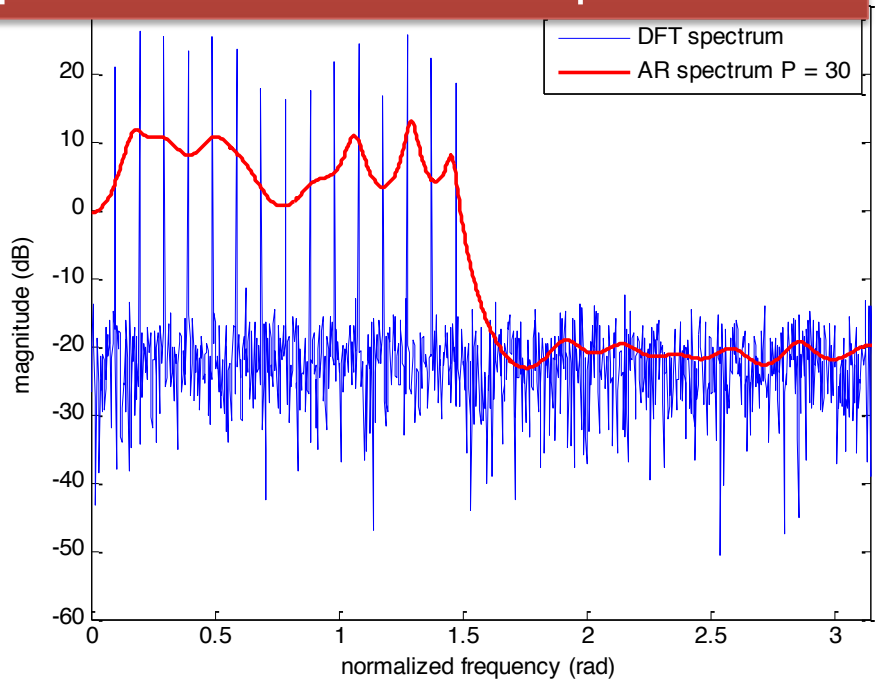


Linear prediction of audio signals (3): Fundamental issues (3)

- Sinusoidal signals with additive broadband noise:
  - *pole radii decrease:* poles are drawn to the origin
  - *pole angles increase:* poles tend to uniform angular distribution

audio signals are generally sampled at very high rate
⇨ distribution of dominating tonal components ≠ uniform
⇨ LP models **spectral envelope** rather than tonal components

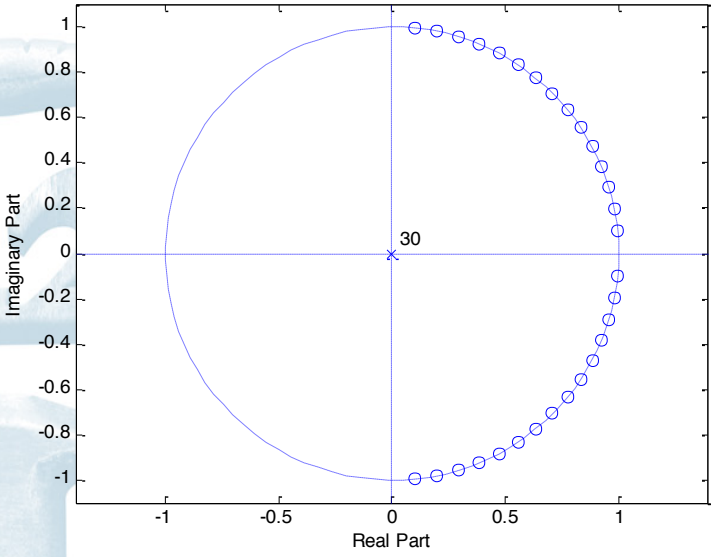


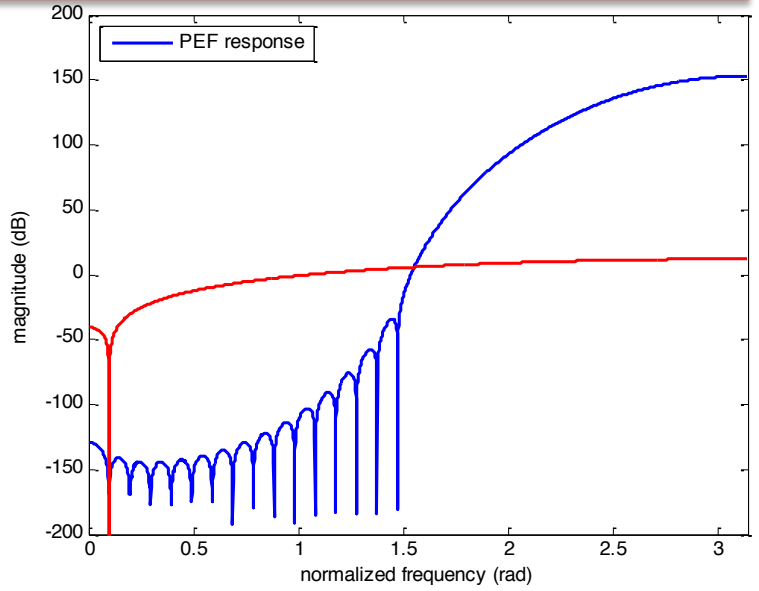


Linear prediction of audio signals (4): Fundamental issues (4)

- Poles tend to uniform angle distribution: **Intuition**
  - in noiseless case, $A(z)$ behaves as cascade of P notch filters
  - two-zero notch filter at $\theta_l < \pi/4$ exhibits high-frequency boost
  - cascade of P two-zero notch filters exhibits very high boost
  - additive broadband noise component in PEF output is amplified in $[\pi/4, \pi/2)$ if majority of notch filters are centered in $[0, \pi/4]$

→ contradicts with the PEF output whitening property of LP





Linear prediction of audio signals (5): Fundamental issues (5)

- Poles tend to uniform angle distribution: **Proof**
  - LP model parameter estimation criterion:

$$\min_{\mathbf{a}} \sum_{t=-\infty}^{\infty} e^2(t,\mathbf{a}) = \min_{\mathbf{a}} \frac{1}{2\pi}\int_0^{2\pi} \left|E\left(e^{j\omega},\mathbf{a}\right)\right|^2 d\omega$$

$$= \min_{\mathbf{a}} \frac{1}{2\pi}\int_0^{2\pi} \left|A\left(e^{j\omega}\right)\right|^2 \left|X\left(e^{j\omega}\right)\right|^2 d\omega$$

$$= \min_{\mathbf{a}} \left[\sum_{m=1}^{M} \frac{\alpha_m^2}{2} \left|A\left(e^{j\omega_m}\right)\right|^2 + \frac{\sigma_r^2}{2\pi}\int_0^{2\pi} \left|A\left(e^{j\omega}\right)\right|^2 d\omega \right]$$

$\theta_l \to \omega_m$ $\qquad$ $\theta_l \to \pi/P$

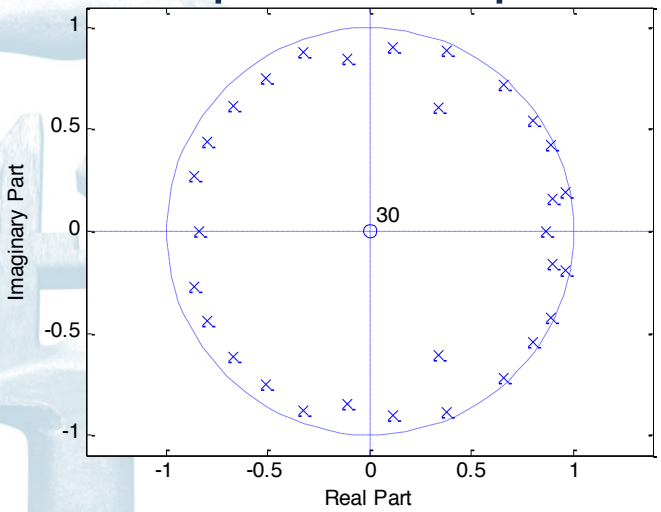
  - $\theta_l \to \pi/P$ can be shown by considering

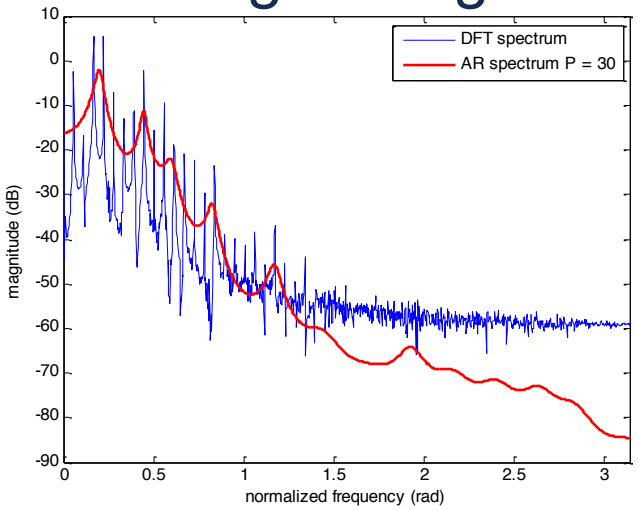
$$\min_{\mathbf{a}} \sigma_r^2 \left(1 + \sum_{i=1}^{2P} a_i^2\right) \quad \text{s.t.} \quad \nu_l > 0,\ l = 1,\ldots,P$$

which leads to a comb filter solution $a_1 = \ldots = a_{2P-1} = 0, a_{2P} = \beta$

Linear prediction of audio signals (6): Alternative models (1)

- LP models different from low-order all-pole model
  - pole-zero model (PZLP)
  - high-order all-pole model (HOLP)
  - pitch prediction model (PLP)
- Low-order all-pole model of transformed audio signal
  - warped all-pole model (WLP)
  - selective all-pole model (SLP)
- Example: monophonic audio signal segment





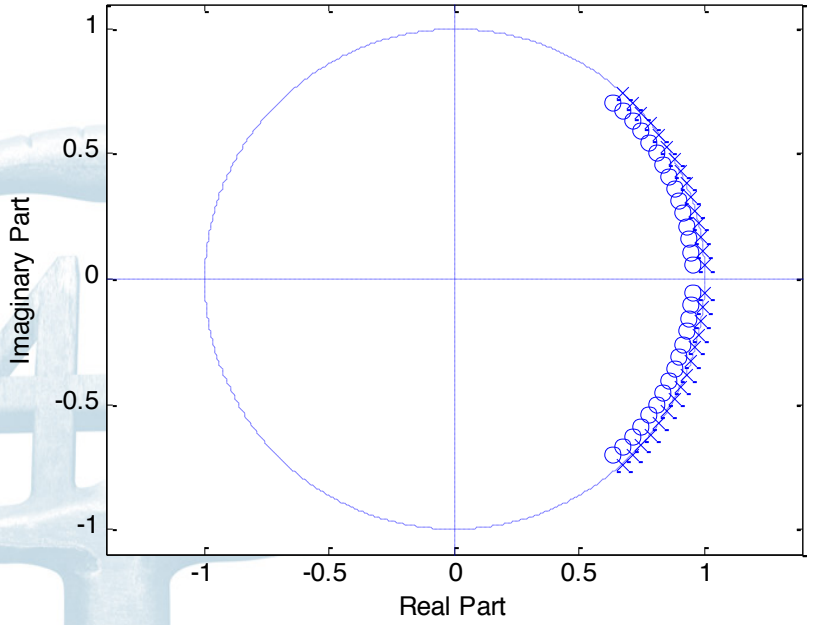
$$N = 2048$$
$$M = 15$$
$$f_0 = 387.6 \text{ Hz}$$
$$f_s = 44.1 \text{ kHz}$$

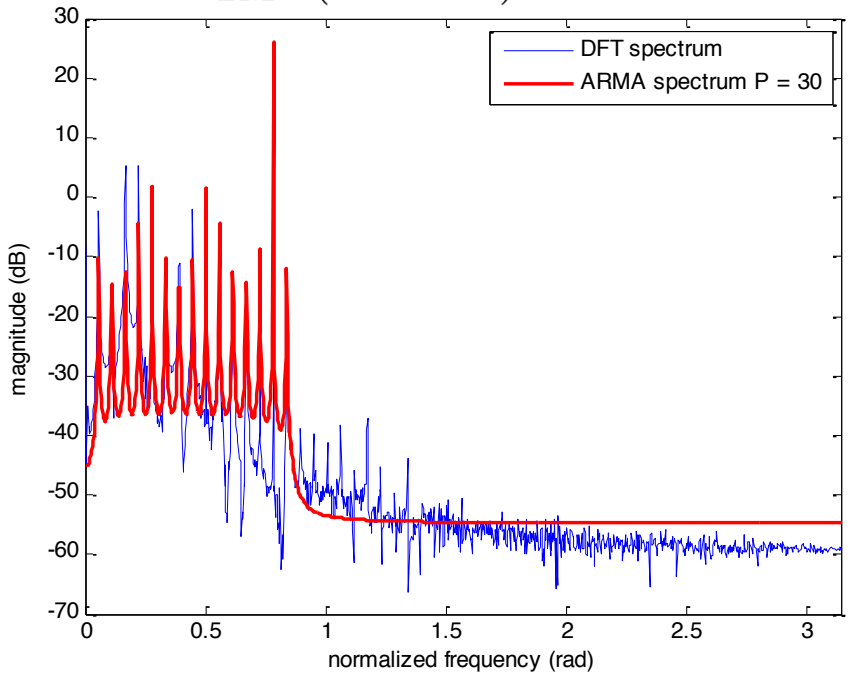
Linear prediction of audio signals (7): Alternative models (2)

1. Pole-zero model (PZLP):
   - sum M sinusoids + white noise admits ARMA($2M$) representation

$$x(t) = \sum_{m=1}^{M} \alpha_m \cos(\omega_m t + \phi_m) + w(t)$$

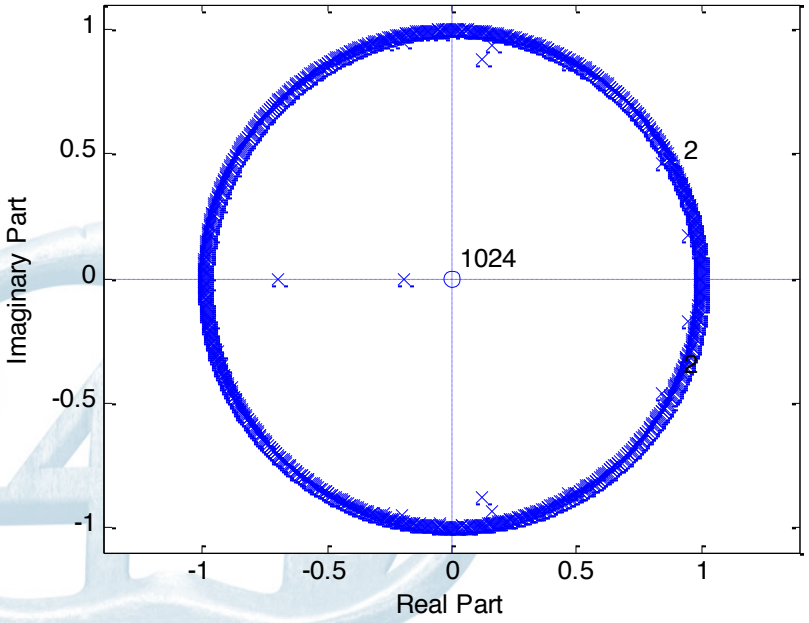
$$\Leftrightarrow \quad x(t) = -a_1 x(t-1) - \ldots - a_{2M} x(t-2M) + b_0 \delta(t) + b_1 \delta(t-1) + \ldots + b_{2M} \delta(t-2M)$$

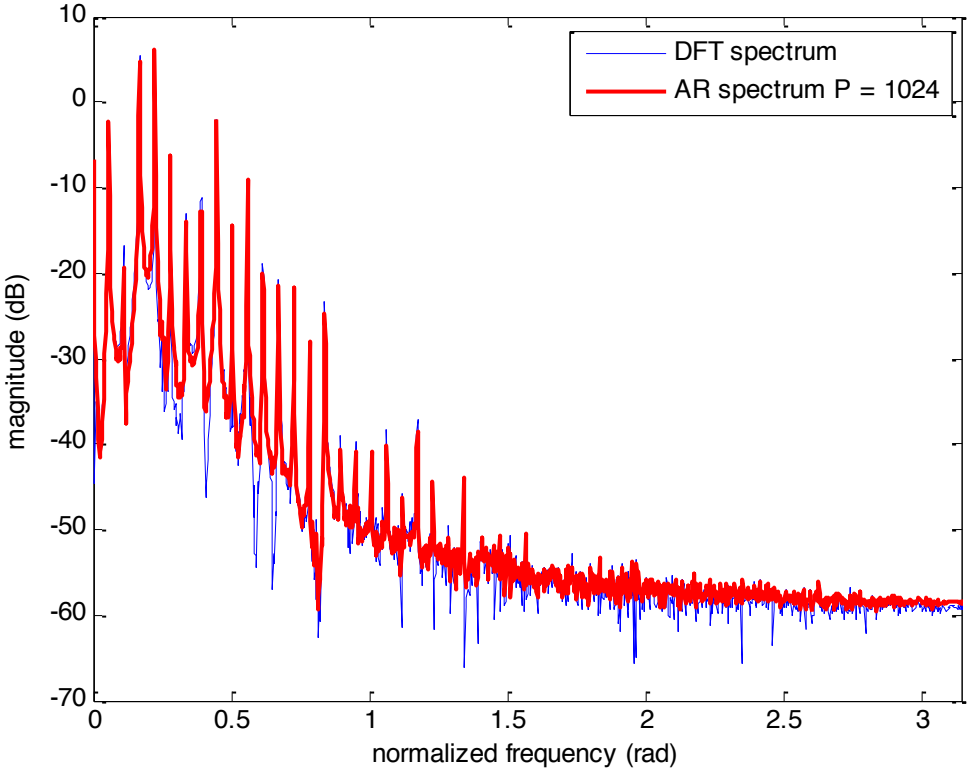




Linear prediction of audio signals (8): Alternative models (3)

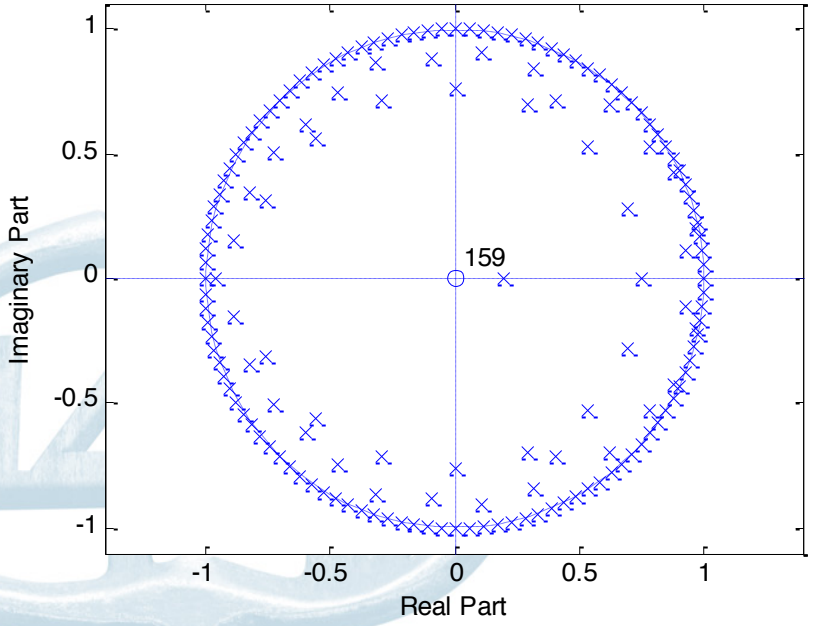
2. High-order all-pole models (HOLP):
   - high-order all-pole model can closely approximate pole-zero model
   - compression still feasible by exploiting sparsity of all-pole model parameter vector (*sparse linear prediction*, see later)

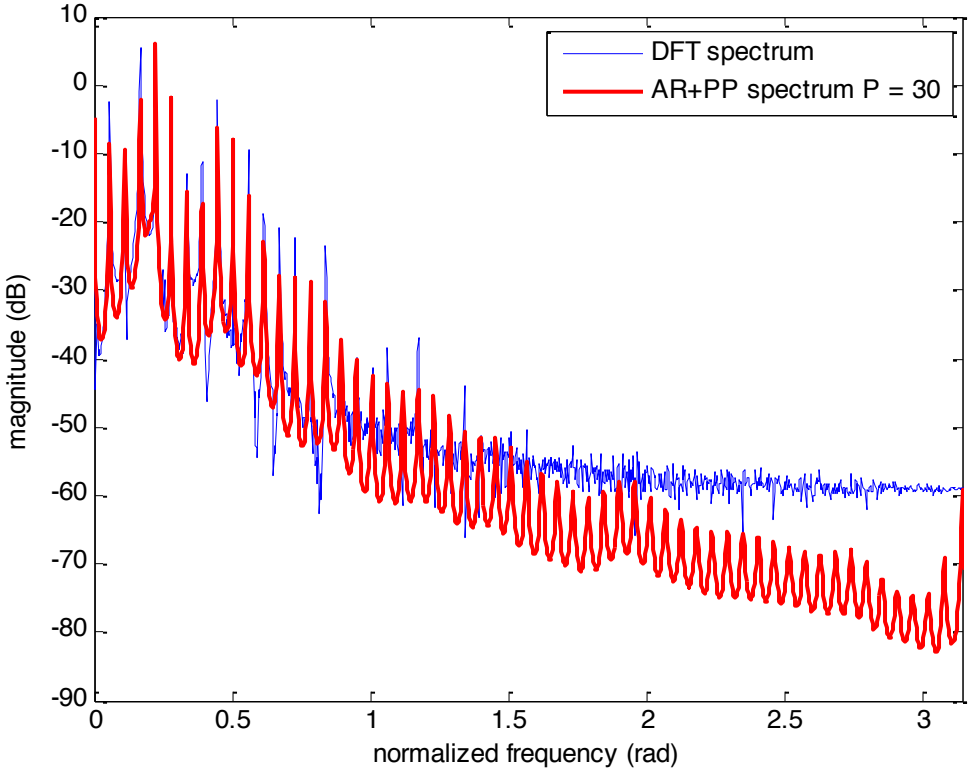




Linear prediction of audio signals (9): Alternative models (4)

3. Pitch prediction model (PLP):
   - comparable to cascade formant + pitch predictor for voiced speech
   - 1st all-pole model=periodicity, 2nd all-pole model=spectral envelope
   - only appropriate for modeling monophonic signals (harmonicity)

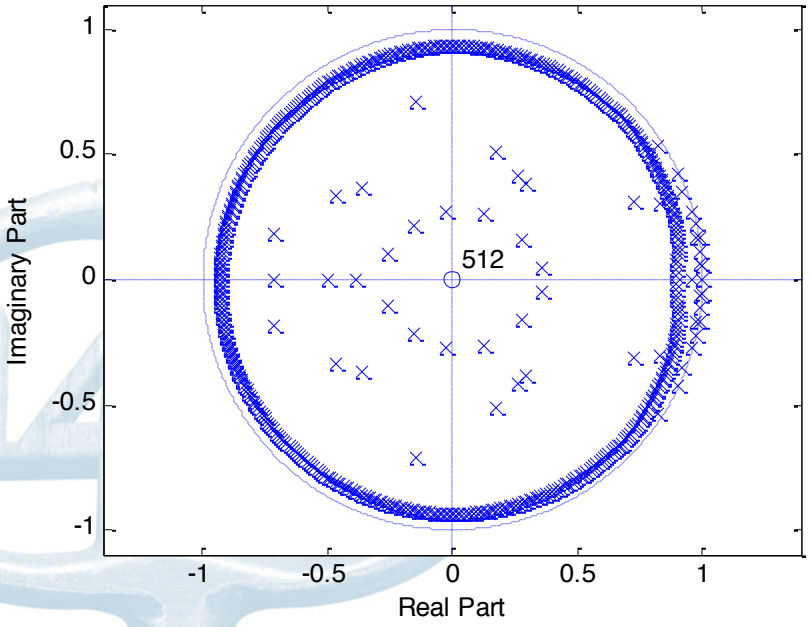


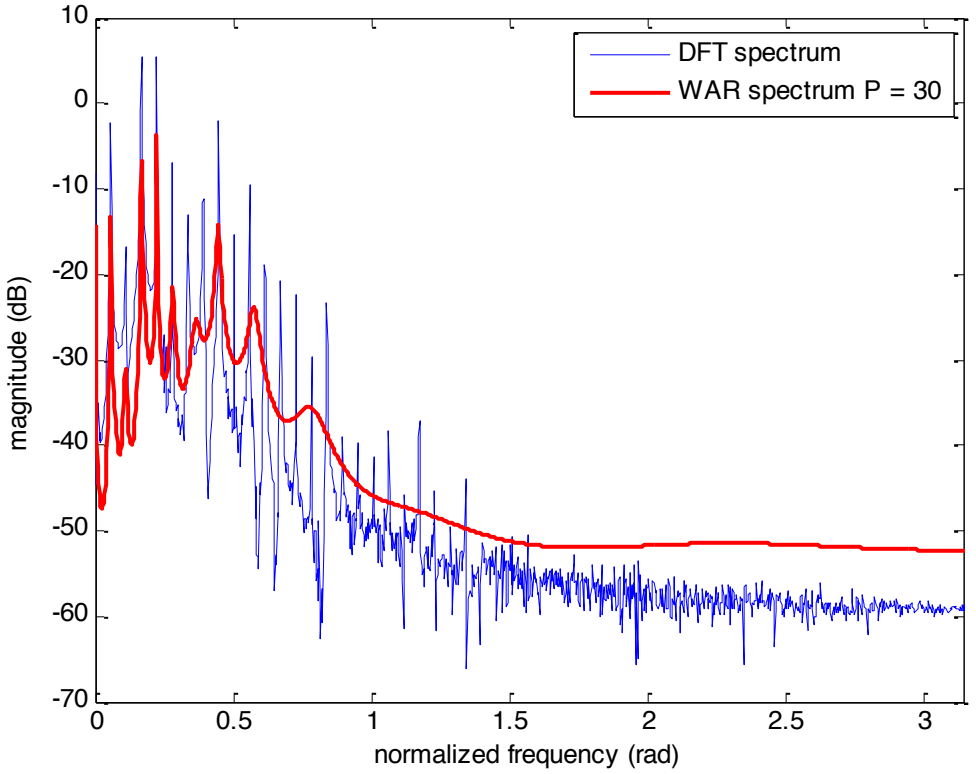


Linear prediction of audio signals (10): Alternative models (5)

4. Warped all-pole model (WLP):
   - frequency warping = non-uniform transformation of frequency axis
   - pole angle distribution: uniform (warped f) $\rightarrow$ logarithmic (linear f)
   - model order: finite (warped f) $\rightarrow$ infinite (linear f)

$$z^{-1} \mapsto \tilde{z}^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}$$



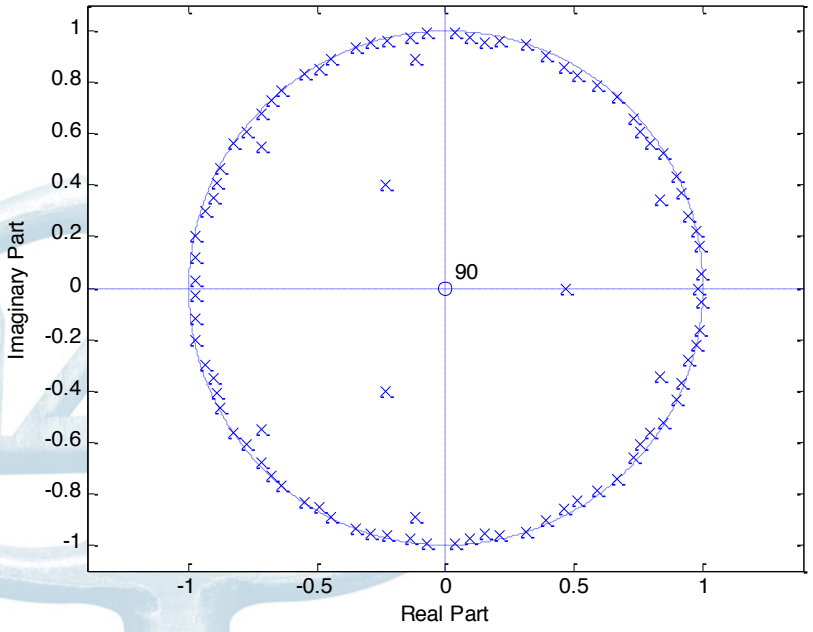


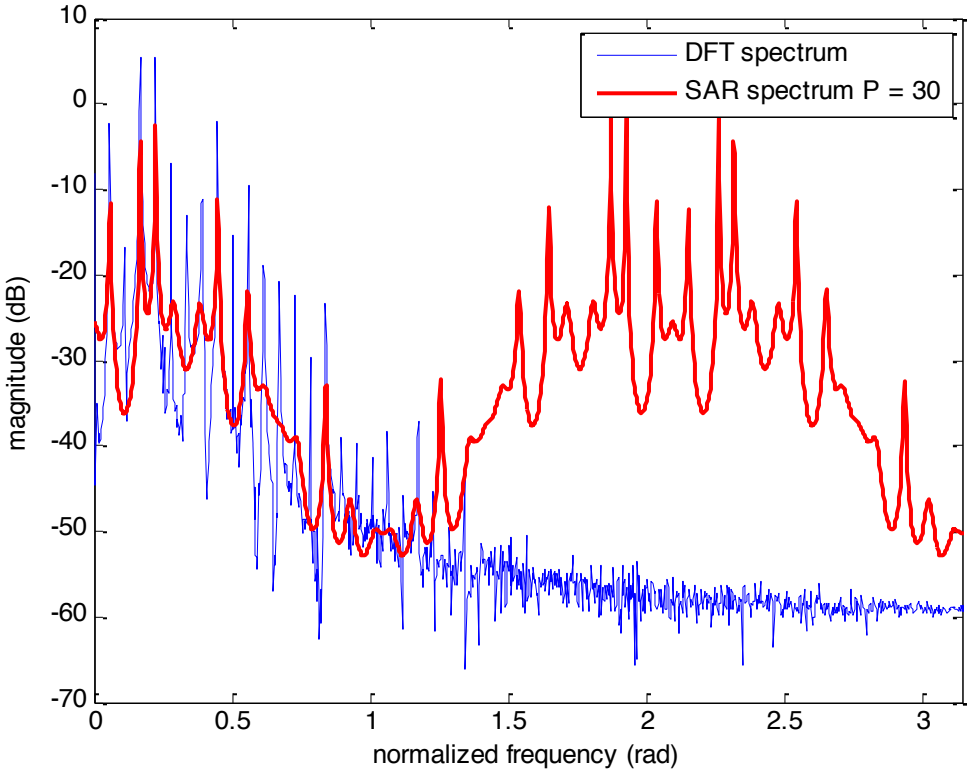
Linear prediction of audio signals (11): Alternative models (6)

5. Selective all-pole model (SLP):
   - frequency zooming = uniform transformation of frequency axis
   - downsampling → conventional LP → upsampling

$$z^{-1} \mapsto \tilde{z}^{-1} = z^{-\Gamma}$$

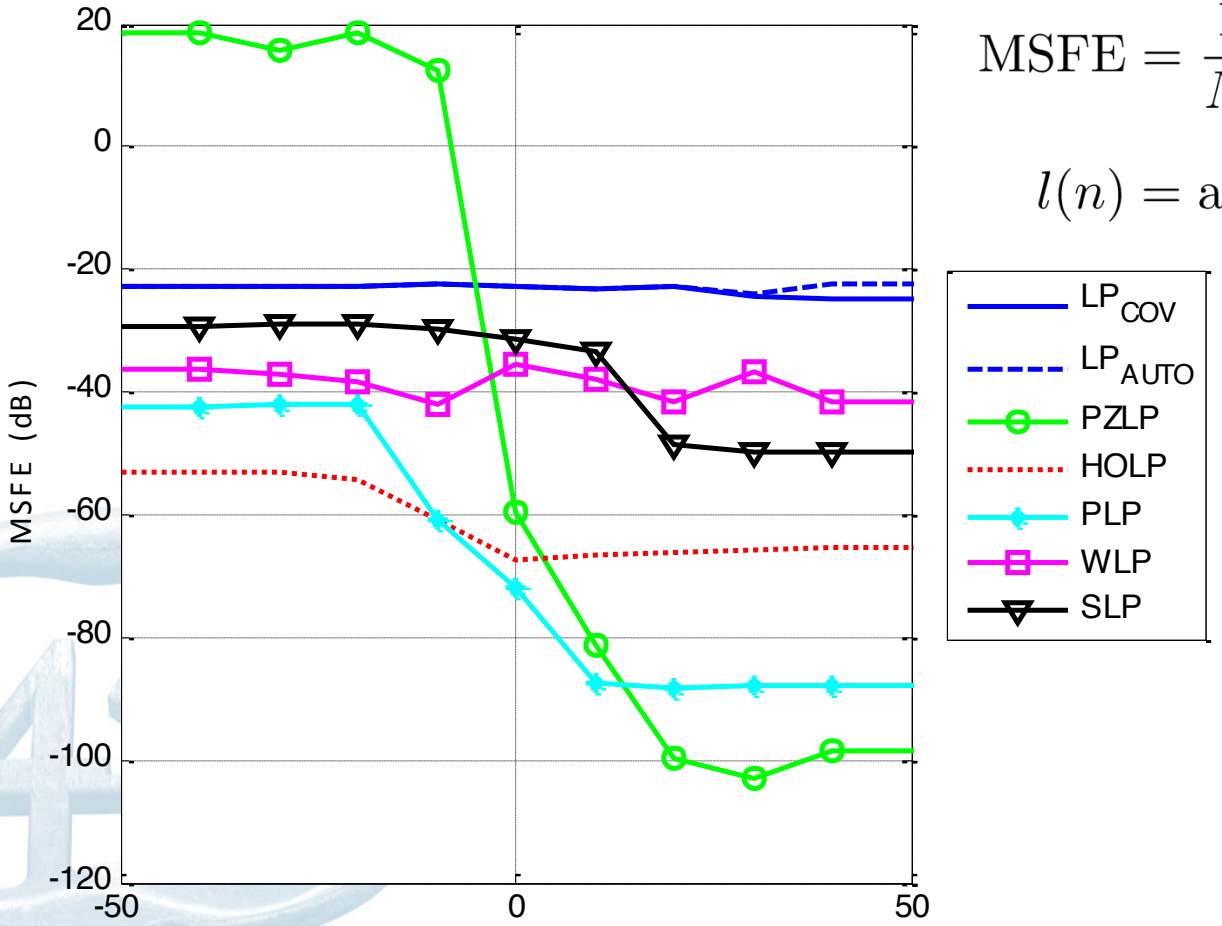
$$\Gamma = \frac{\pi}{\omega_1 + \omega_M}$$





Linear prediction of audio signals (12): Performance evaluation (1)

- Mean square frequency estimation error *vs.* SNR



$$\mathrm{MSFE} = \frac{1}{N}\sum_{n=1}^{N}(\theta_{l(n)} - \omega_n)^2$$

$$l(n) = \arg\min_{l} \|\nu_l e^{j\theta_l} - e^{j\omega_n}\|^2$$

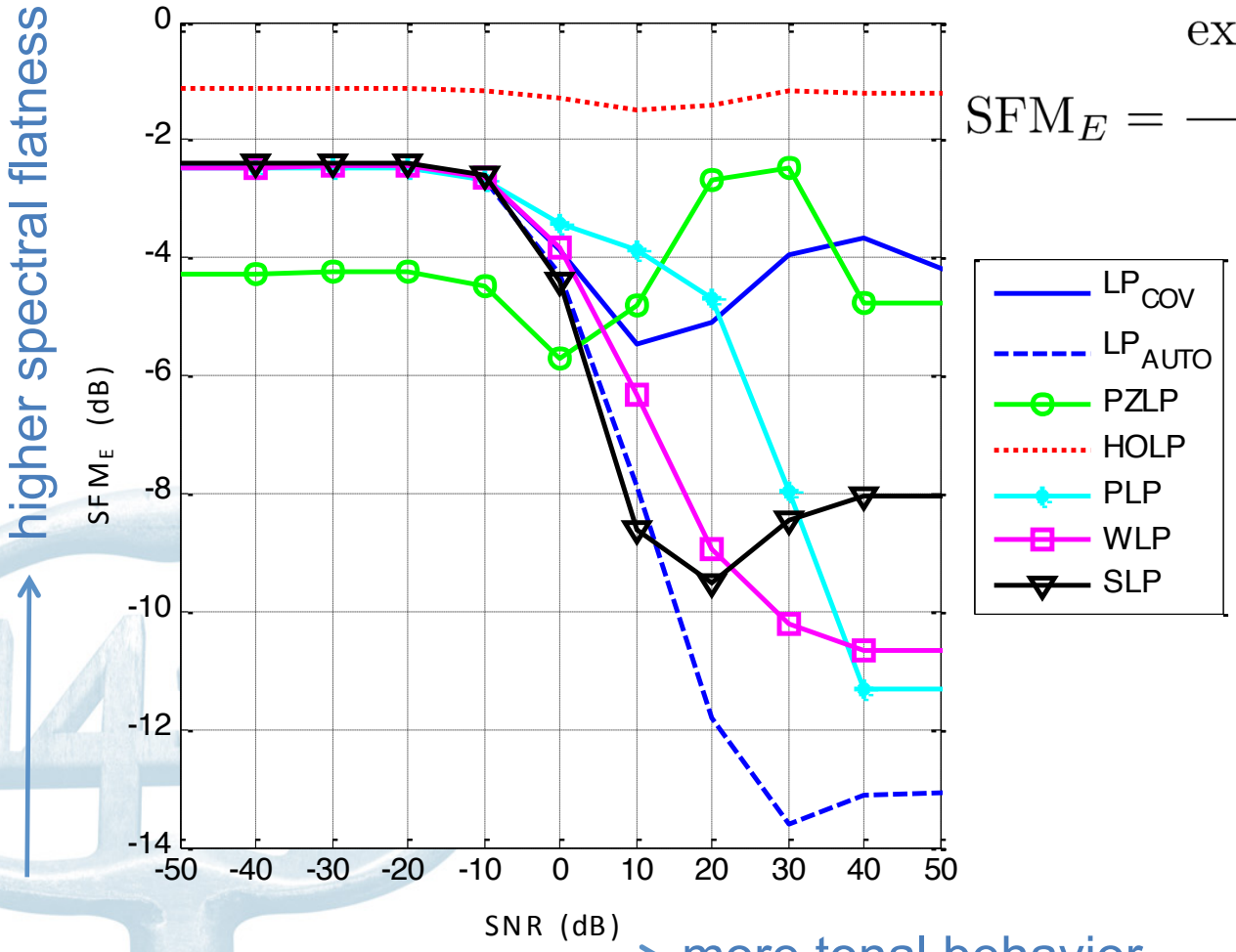
higher frequency accuracy

more tonal behavior

Linear prediction of audio signals (13): Performance evaluation (2)

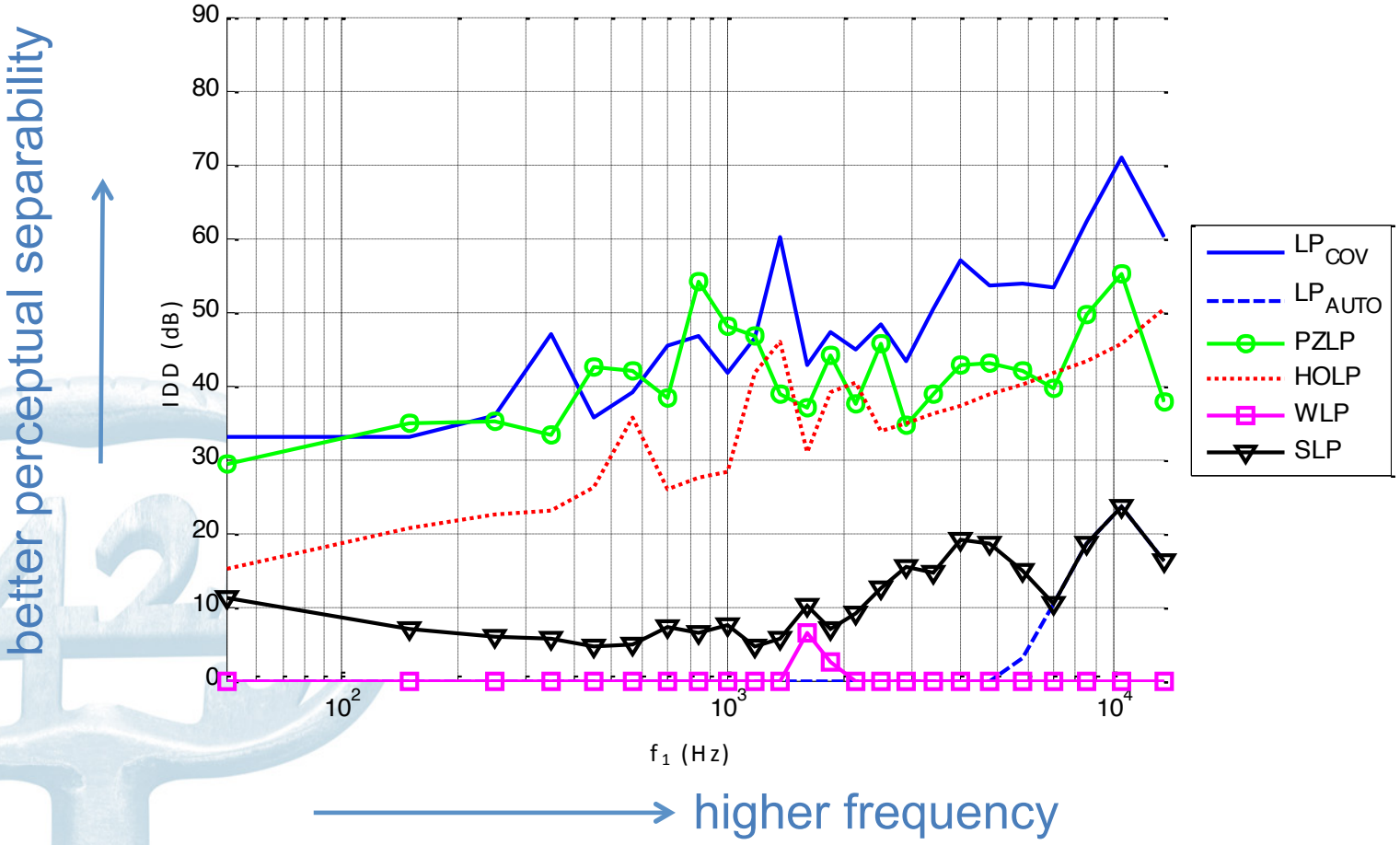
- Spectral flatness measure (SFM) of LP residual

$$\mathrm{SFM}_E = \frac{\exp\left[\frac{1}{L}\sum_{k=0}^{L-1} \ln\left|E\left(e^{j\frac{2\pi k}{L}}, \mathbf{a}\right)\right|\right]}{\frac{1}{L}\sum_{k=0}^{L-1}\left|E\left(e^{j\frac{2\pi k}{L}}, \mathbf{a}\right)\right|}$$



Linear prediction of audio signals (14): Performance evaluation (3)

- Interpeak dip depth (IDD):
  - quantifies perceptual separability of spectral peaks in LP model representing sum of two sinusoids separated by twice ERB



Outline

- Introduction
- Recap: Linear prediction of speech signals
- Linear prediction of audio signals
  - fundamental issues
  - alternative models
  - performance evaluation
- Sparse linear prediction of audio signals*
- Conclusion

* Daniele Giacobello, Toon van Waterschoot, Mads Græsbøll Christensen, Søren Holdt Jensen, and Marc Moonen, "High-order sparse linear predictors for audio processing", in *Proc. 18th European Signal Process. Conf. (EUSIPCO '10)*, Aalborg, Denmark, Aug. 2010, pp. 234-238.

Sparse linear prediction of audio signals (1)

- Motivation:
  - the high-order all-pole model (HOLP) performs very well at representing audio signals with varying degrees of tonality
  - harmonic relations between dominating tonal components induce sparsity and structure in HOLP parameter vector
- High-order sparse linear prediction (HOSpLP):
  - l1-regularized estimation problem

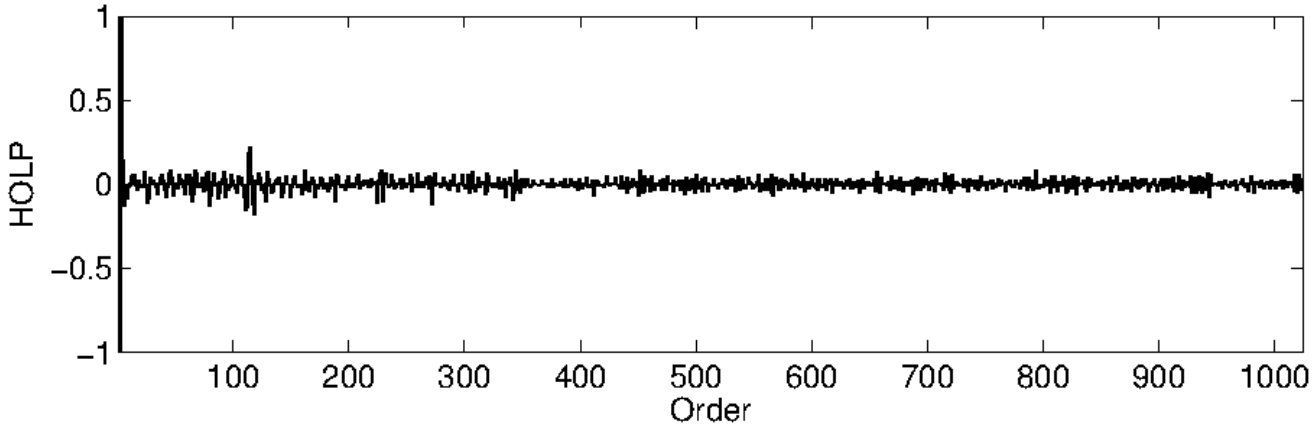
$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 + \gamma\|\mathbf{a}\|_1$$

with high-order parameter vector: $P = N/2$

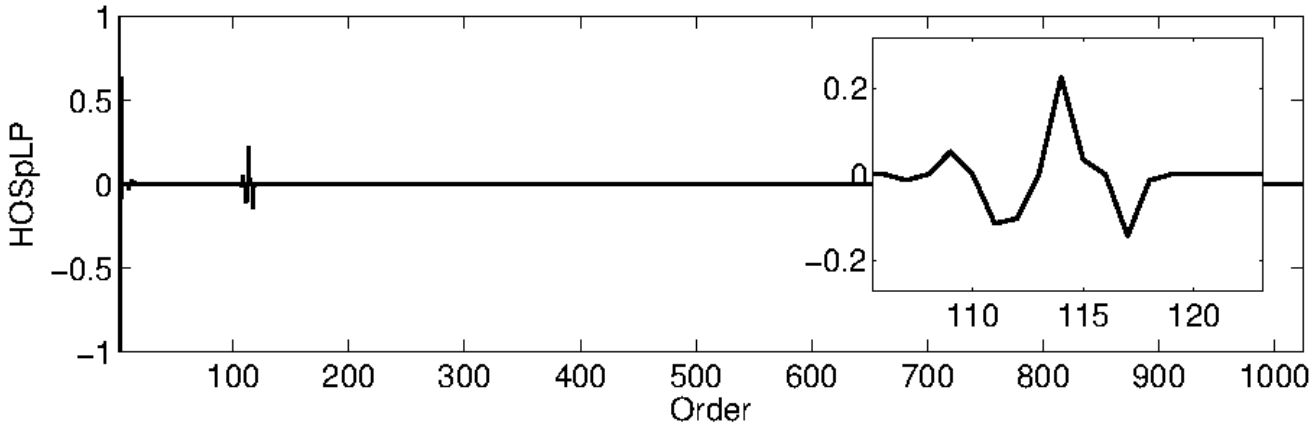
  - influence of γ:
    - related to prior knowledge on parameter vector $\mathbf{a}$ (MAP approach) or how sparse the predictor should be
    - similar to Tikhonov regularization ($\mathbf{X}$ has highly correlated rows, so for $\gamma = 0$ the HOLP problem is ill-posed)

Sparse linear prediction of audio signals (2)

- Example 1: Monophonic audio signal
  - comparison of HOLP and HOSpLP parameter vector estimates



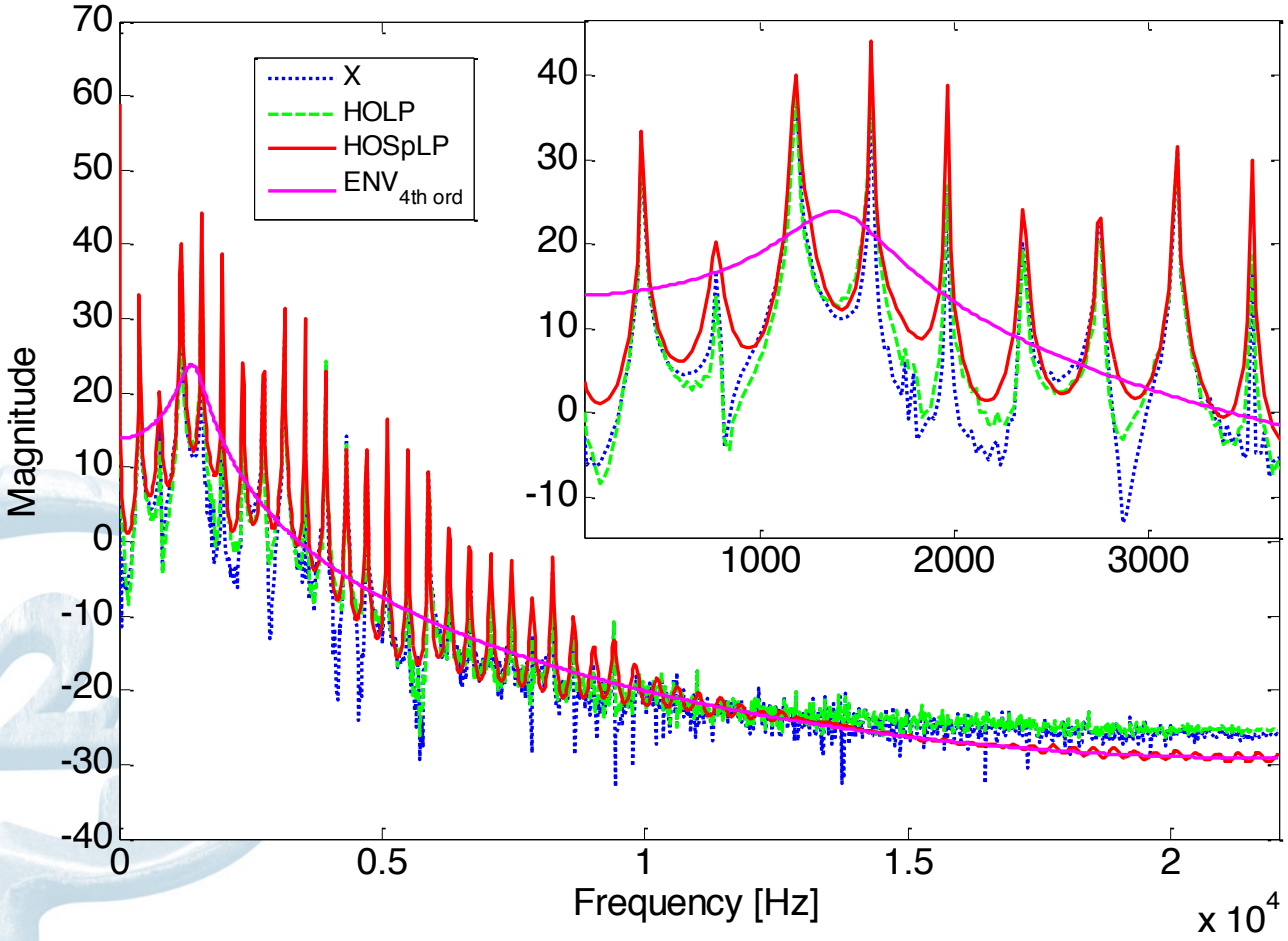
$$\|\hat{\mathbf{a}}\|_0 = 1024$$



$$\|\hat{\mathbf{a}}\|_0 = 25$$

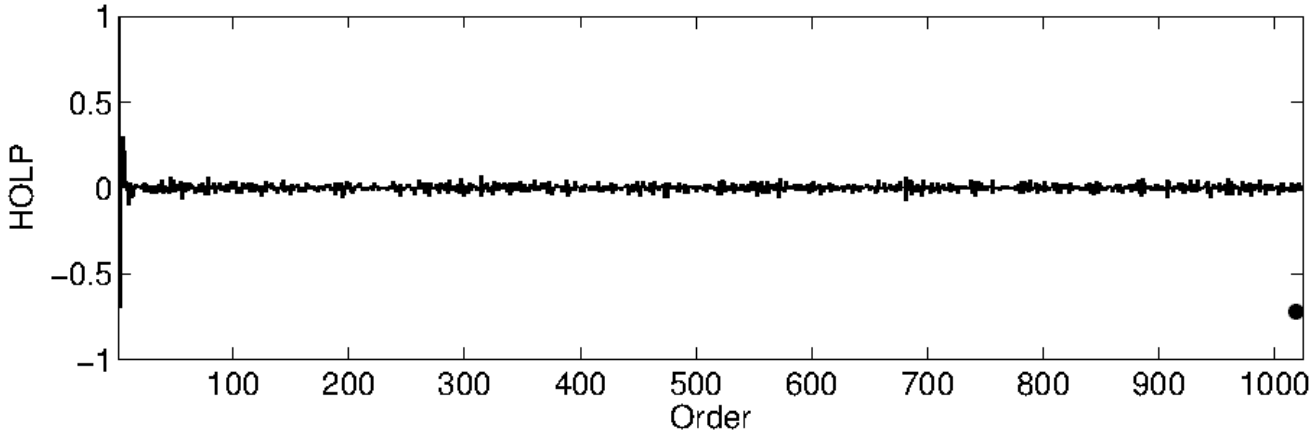
Sparse linear prediction of audio signals (3)

- Example 1: Monophonic audio signal
  - comparison of HOLP and HOSpLP signal spectrum estimates

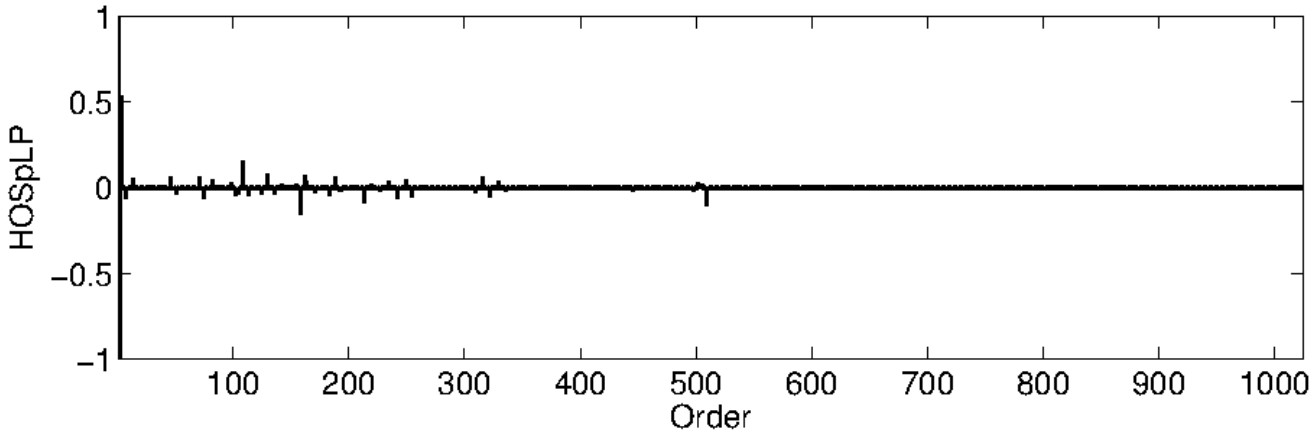


Sparse linear prediction of audio signals (4)

- Example 2: Polyphonic audio signal
  - comparison of HOLP and HOSpLP parameter vector estimates



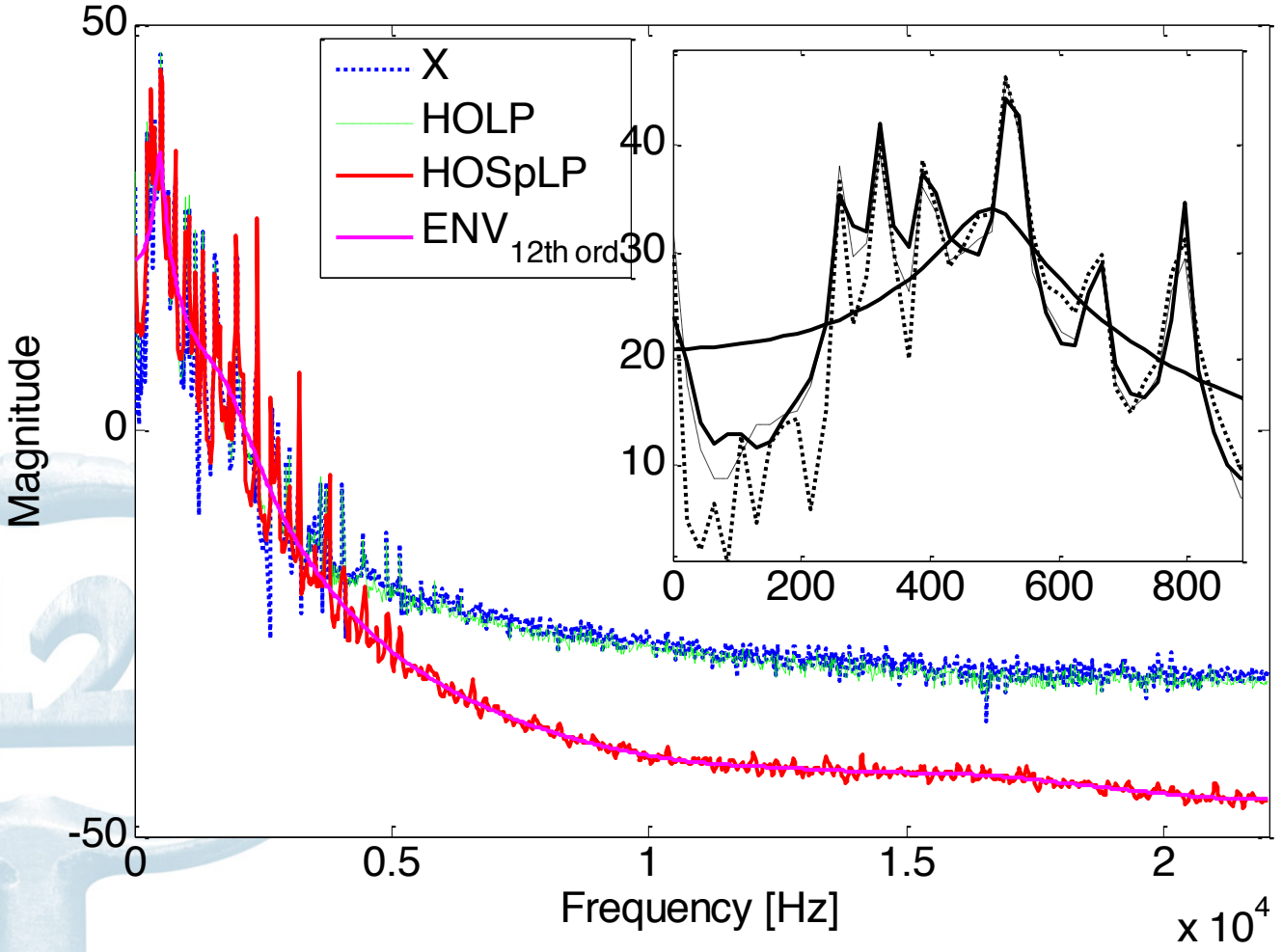
$\|\hat{\mathbf{a}}\|_0 = 1024$



$\|\hat{\mathbf{a}}\|_0 = 53$

Sparse linear prediction of audio signals (5)

- Example 2: Polyphonic audio signal
  - comparison of HOLP and HOSpLP signal spectrum estimates



Conclusion

- Why has LP not been as popular for audio as for speech?
  - audio contains strong tonal (near-sinusoidal) components
  - dominating tonal components located in low-frequency region
  - PEF zeros tend to uniform angular distribution
- Alternative LP models do provide accurate representation
  - LP models different from low-order all-pole model
  - low-order all-pole model of transformed audio signal
- Sparse approximation leads to efficient representation
  - HOLP model provides most accurate audio signal representation
  - harmonic relations induce sparsity and structure in HOLP model
  - HOSpLP model results from solving l1-regularized estimation
- Ongoing work:
  - efficient first-order numerical optimization algorithms for HOSpLP
  - applications of sparse LP-based audio signal models