# DyNaVoiceR: Second annual project meeting

# High-quality parametric synthesis of voiced sounds

Aníbal Ferreira

Department of Electrical and Computer Engineering
University of Porto, Portugal

**ajf@fe.up.pt**

**FEUP, Porto, November 09, 2019**

Cofinanciado por:

COMPETE 2020
PROGRAMA OPERACIONAL COMPETITIVIDADE E INTERNACIONALIZAÇÃO

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Europeu
de Desenvolvimento Regional

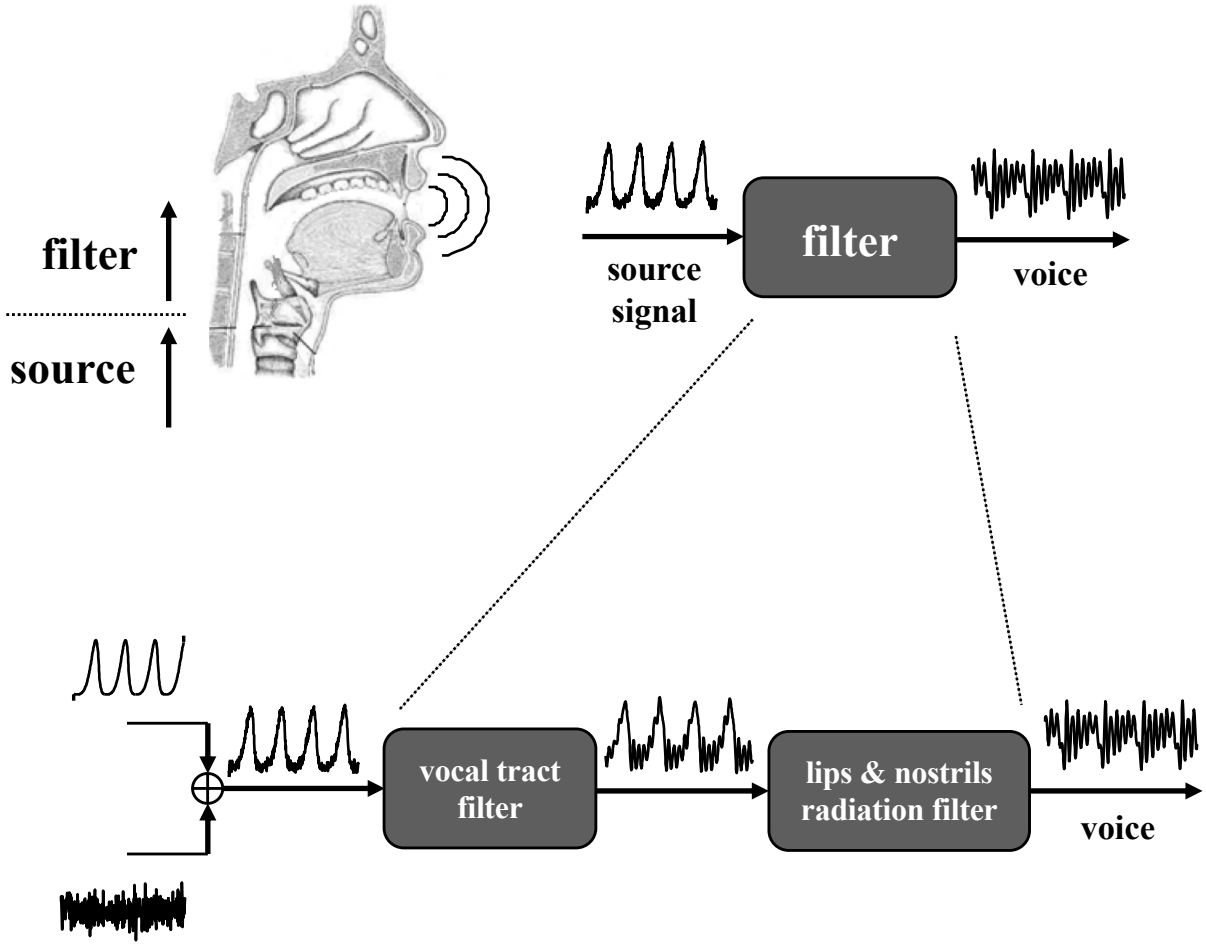
FCT Fundação para a Ciência e a Tecnologia

# Outline

- The source-filter model of voice-production
- What is the voiced part of a speech sound ?
- NRD: a shift-invariant phase-related feature
  - spectrogram and phasegram
  - can phase be shift-invariant ?
  - NRD computation
- Reverse engineering of a voiced signal (version 1)
- Reverse engineering of a voiced signal (version 2)
- Vowel morphing
- Conclusions

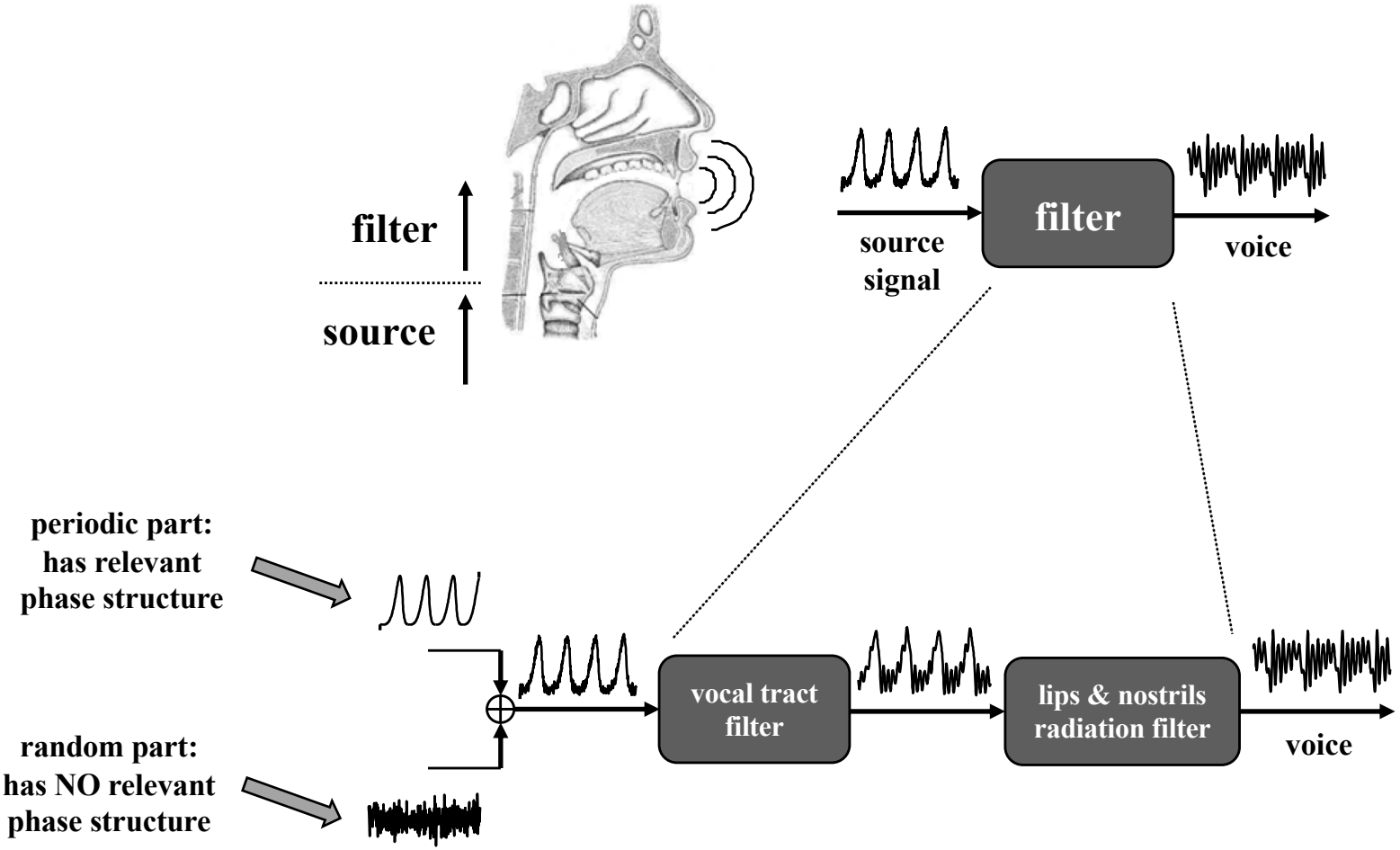
# The source-filter model of voice production

- (Fant, 1960)



# The source-filter model of voice production

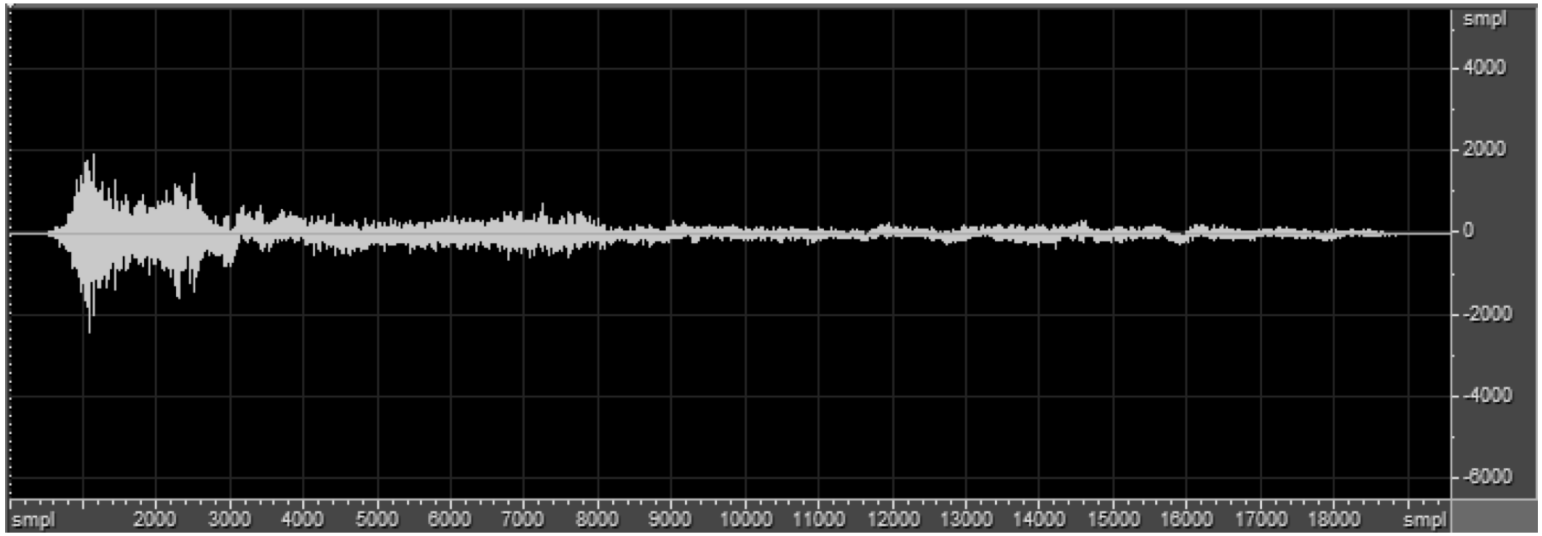
- (Fant, 1960)



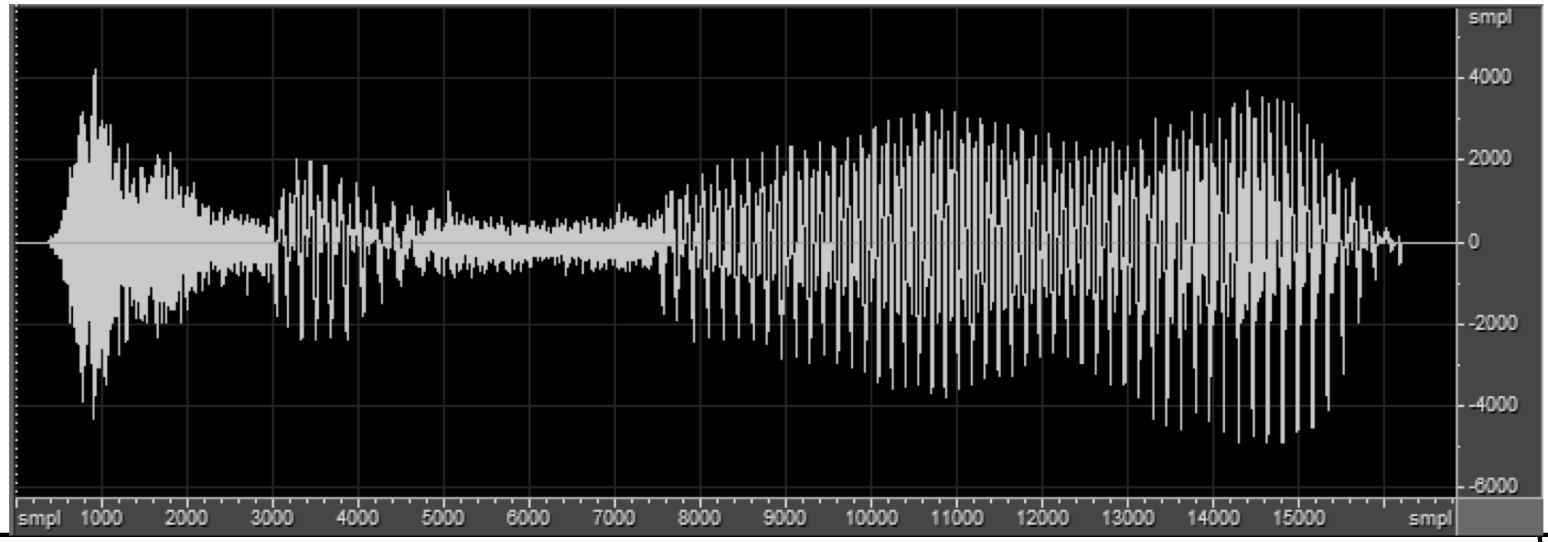
# What is the voiced part of a speech sound ?

- example: whispered and voiced version of same word

**whispered speech**

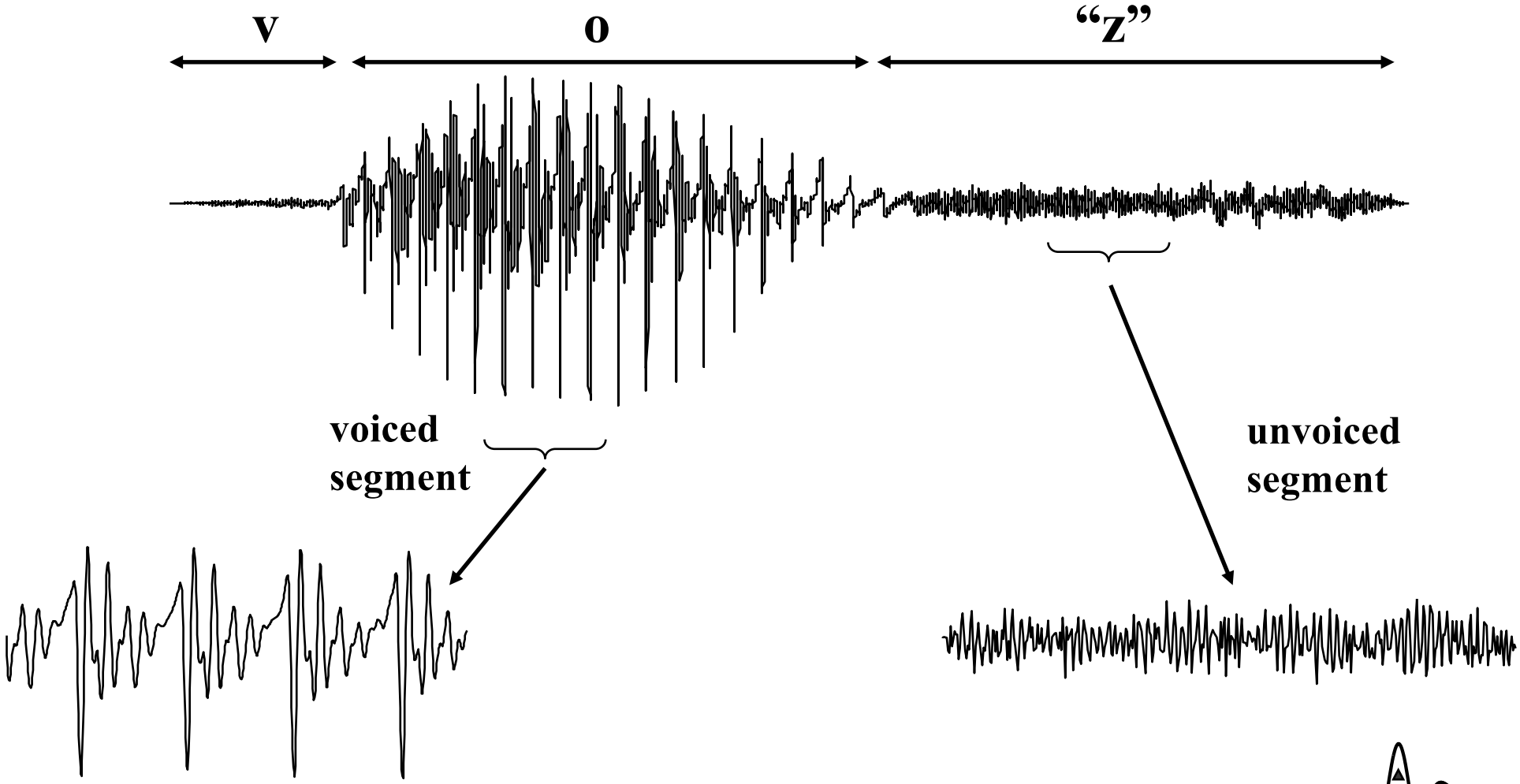



**voiced speech**



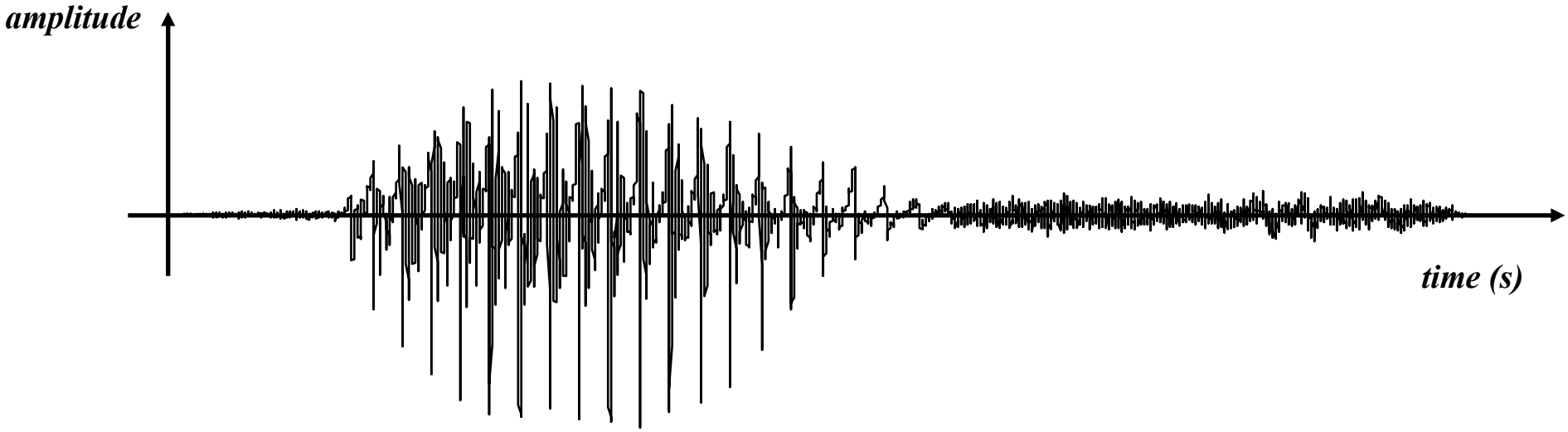
# What is the voiced part of a speech sound ?

- a simple speech sound



# What is the voiced part of a speech sound ?

- time and spectral representation of signals
  - time representation
    - amplitude of the wave as a function of time



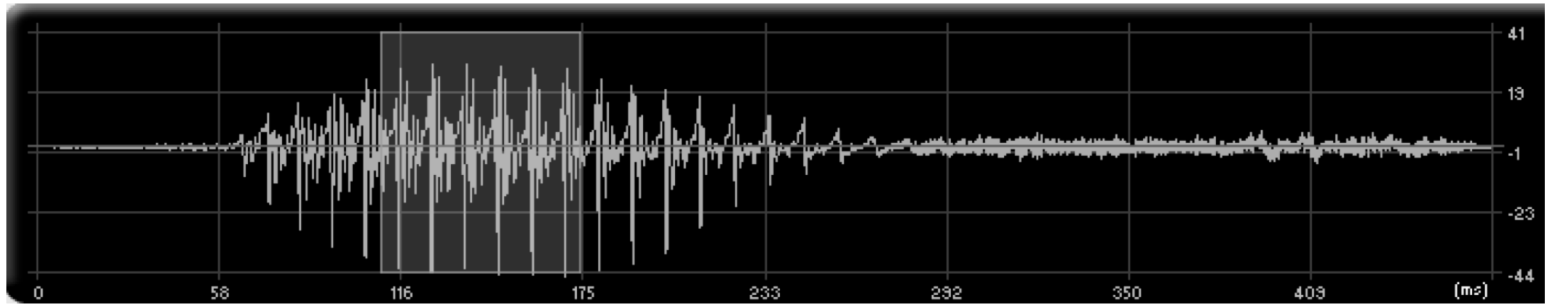

  - frequency representation (magnitude spectrum, spectrogram)
    - magnitude of the different spectral components of the signal

# What is the voiced part of a speech sound ?

- time representation, magnitude spectrum and spectrogram

**time representation**

*Y: amplitude*
*X: time (s)*

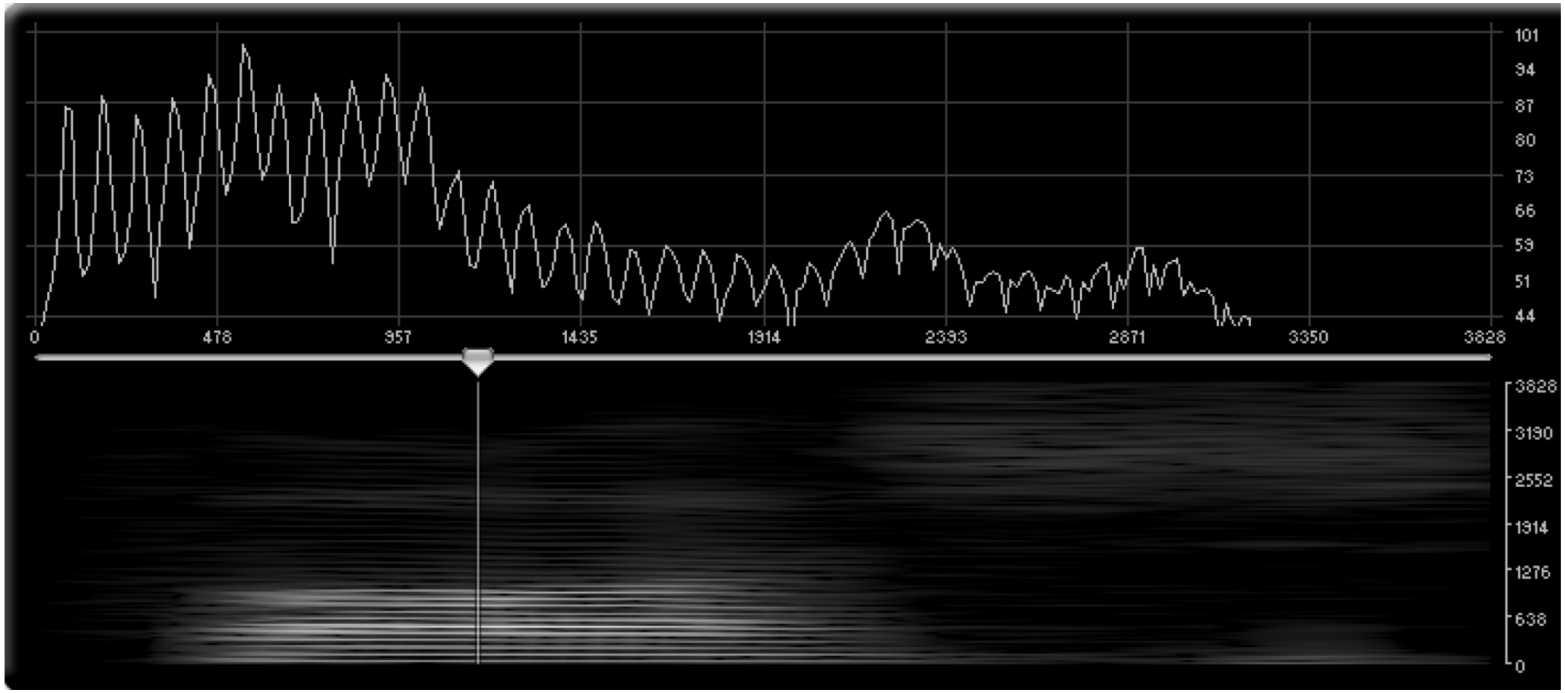



**magnitude spectrum**

*Y: magnitude*
*X: frequency (Hz)*

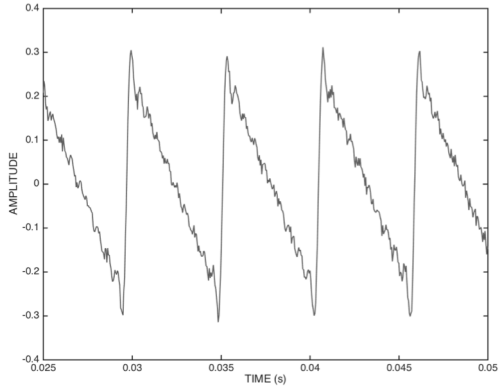
**spectrogram**

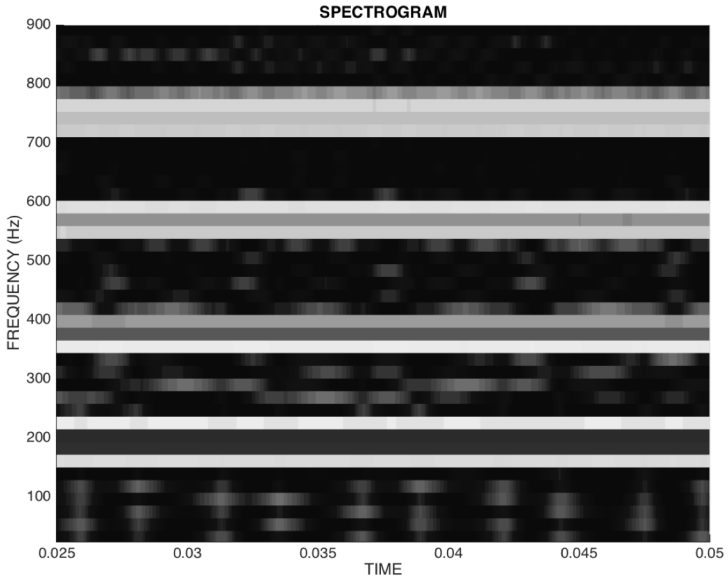
*Z: magnitude*
*Y: frequency (Hz)*
*X: time (s)*

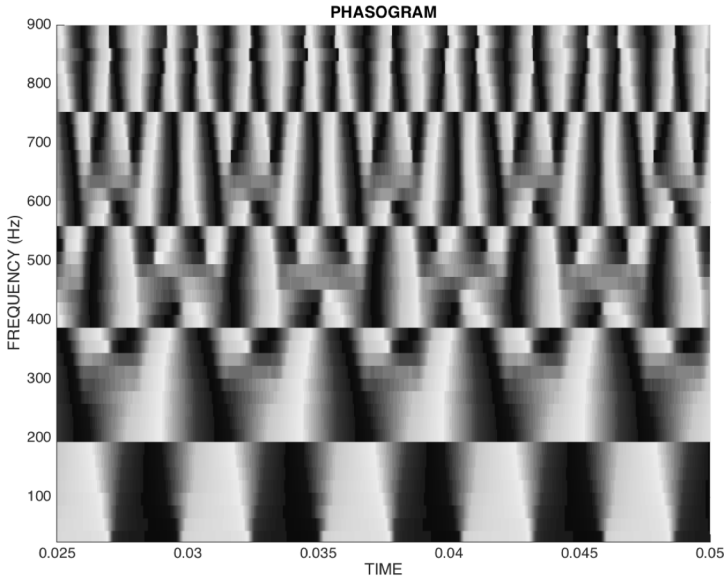


# NRD: a shift-invariant phase-related feature

- spectrogram and phasegram

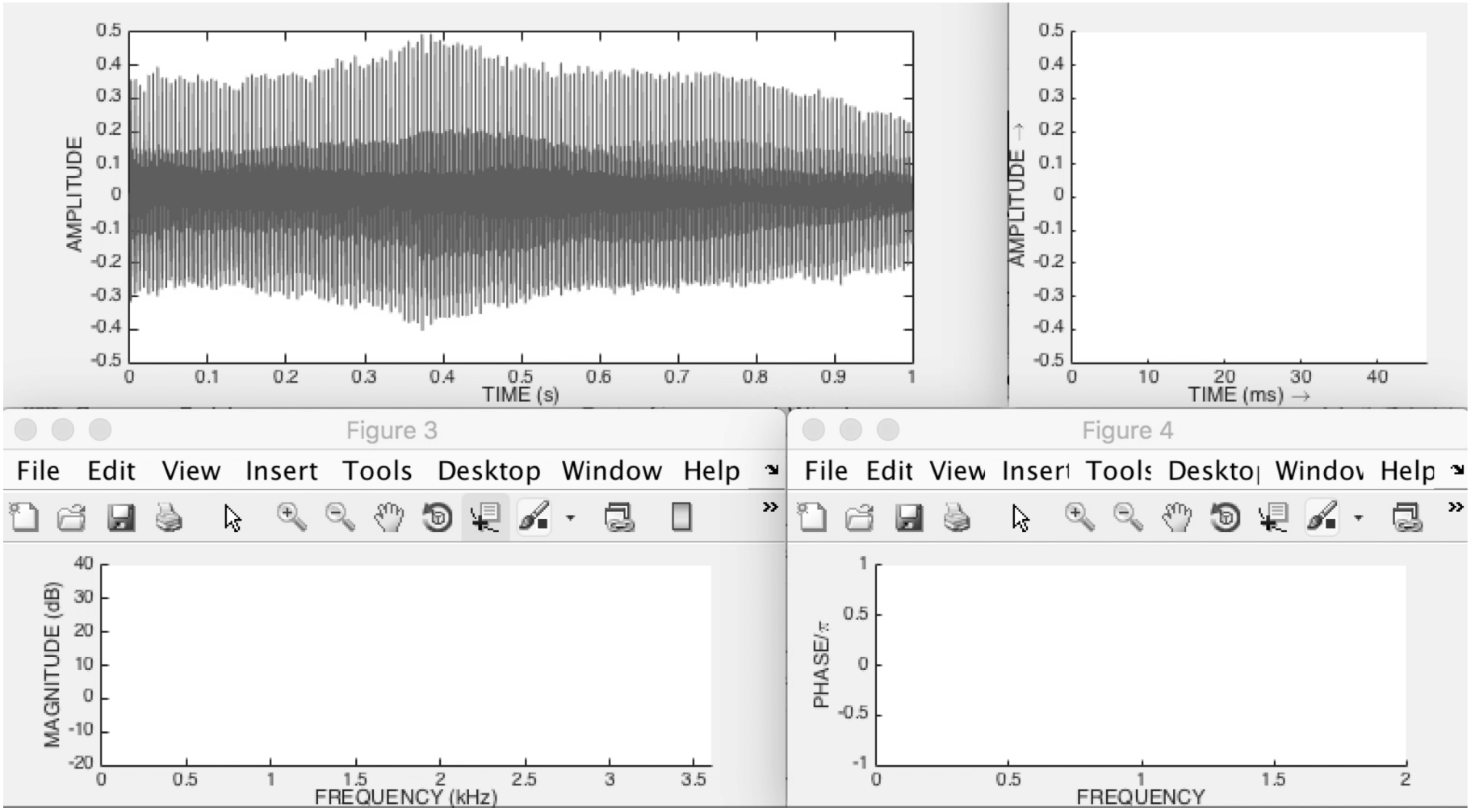


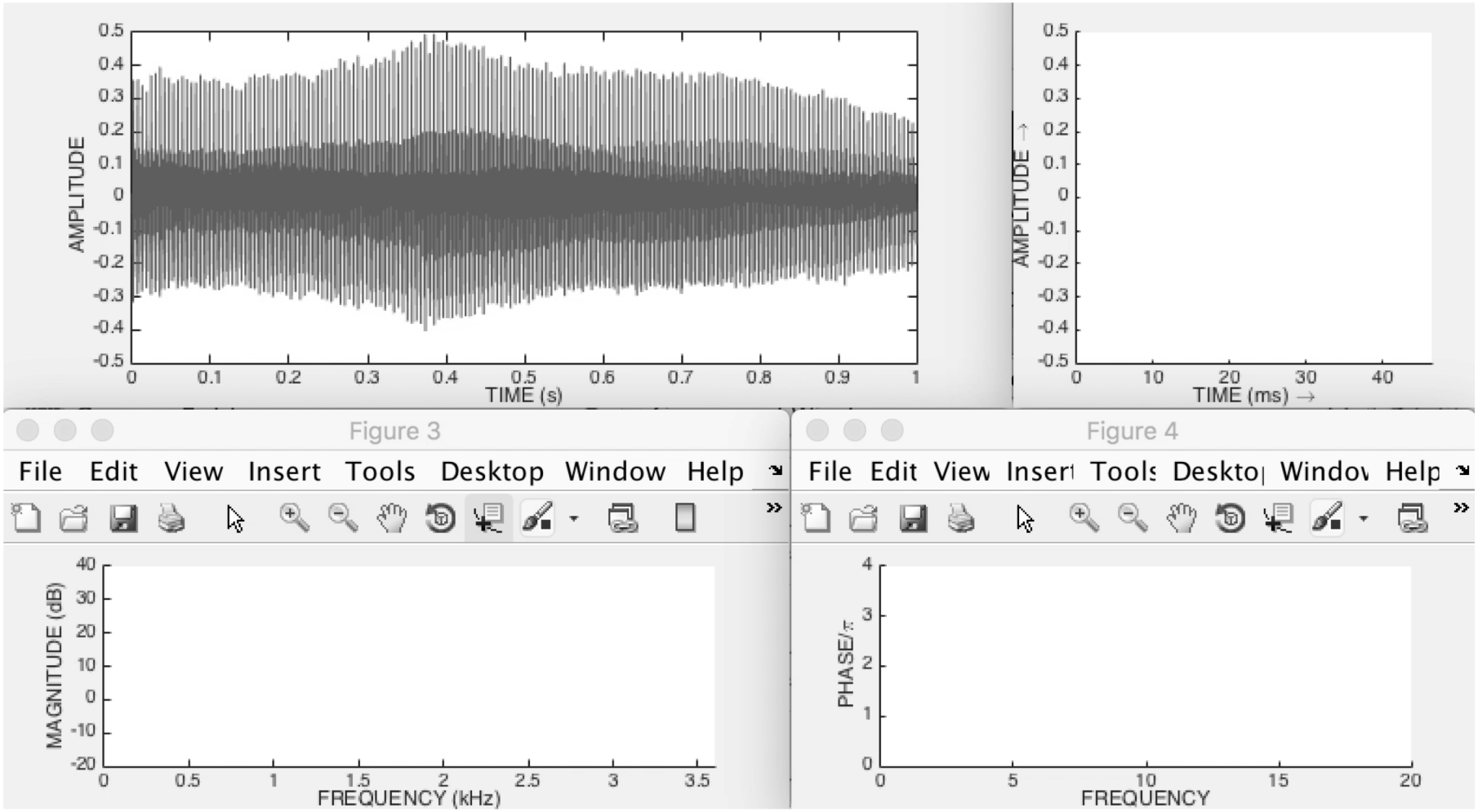
# NRD: a shift-invariant phase-related feature

- can phase be shift-invariant ? (natural sustained vowel, raw phase)



# NRD: a shift-invariant phase-related feature

- can phase be shift-invariant ? (natural sustained vowel, NRD)



# NRD: a shift-invariant phase-related feature

- Normalized Relative Delay (NRD)
  - is a phase-related feature that is relative to the phase of the fundamental frequency and that is further normalized by the period of the harmonic it is associated with

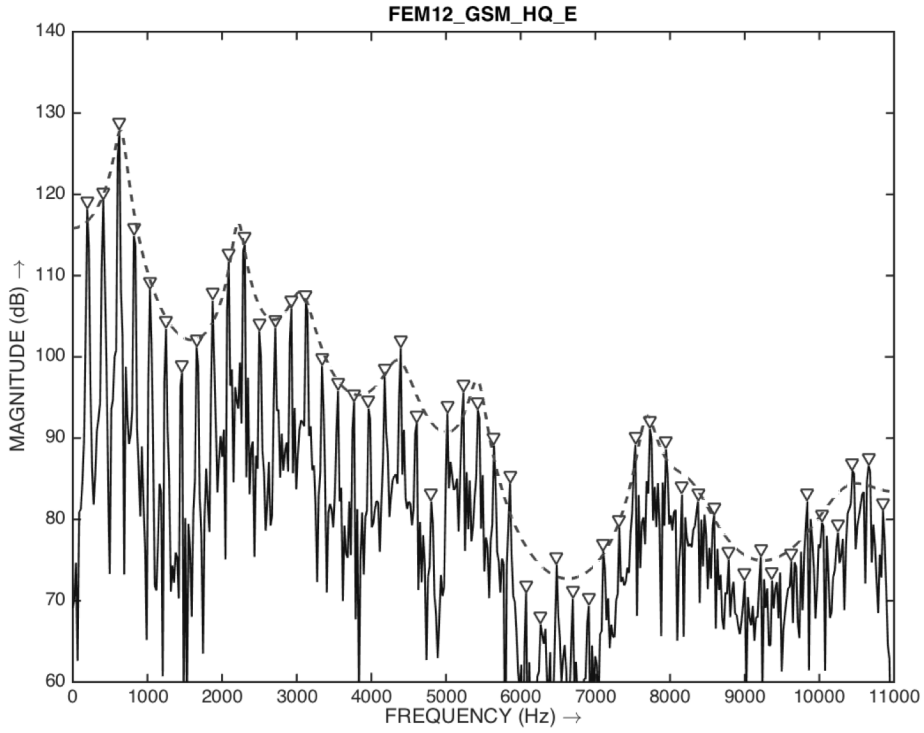
$$\boldsymbol{NRD}_{\ell} = \frac{t_{\ell} - t_0}{(T/\ell)} = \frac{\phi_{\ell} - \ell\phi_0}{2\pi}$$

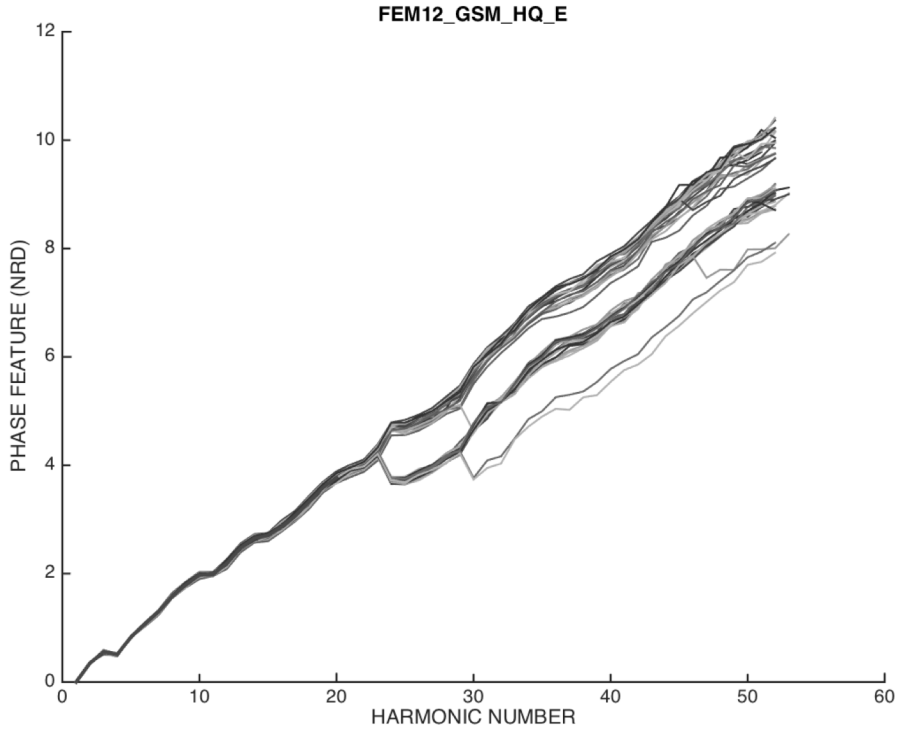
$\ell$ is harmonic index

  - properties: by definition NRDs are time-shift invariant and pitch (i.e. $F_0$) independent
  - as a phase-related feature, the wrapping and unwrapping operations also apply
  - in most cases of natural sustained vowel signals, the first 20 harmonics generate very stable and consistent NRD patterns

# NRD: a shift-invariant phase-related feature

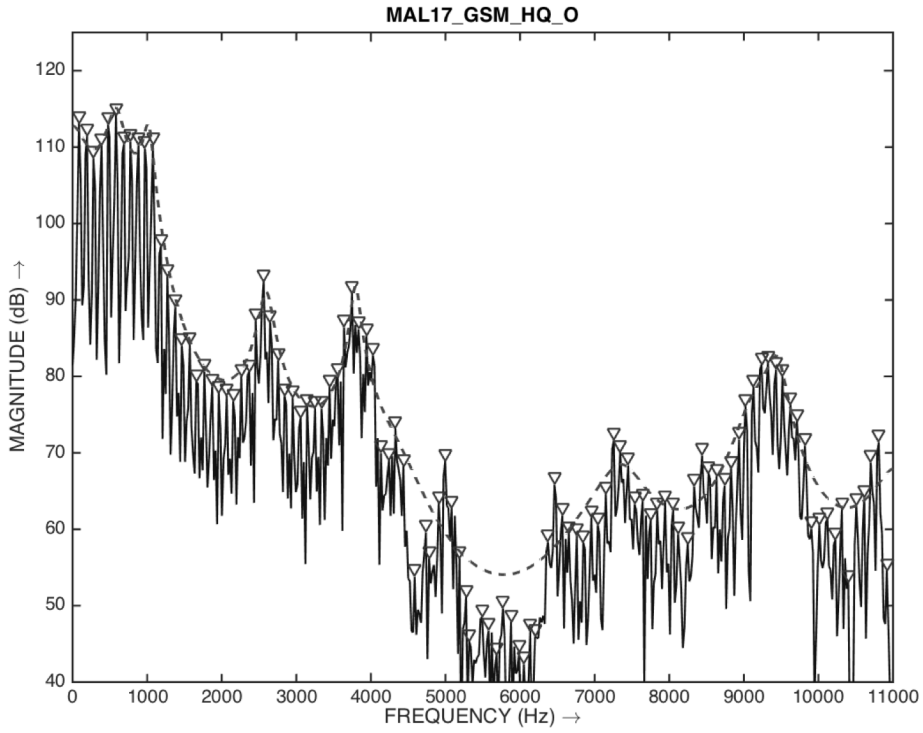
- sustained /e/ vowel by a female (subject FEM12)

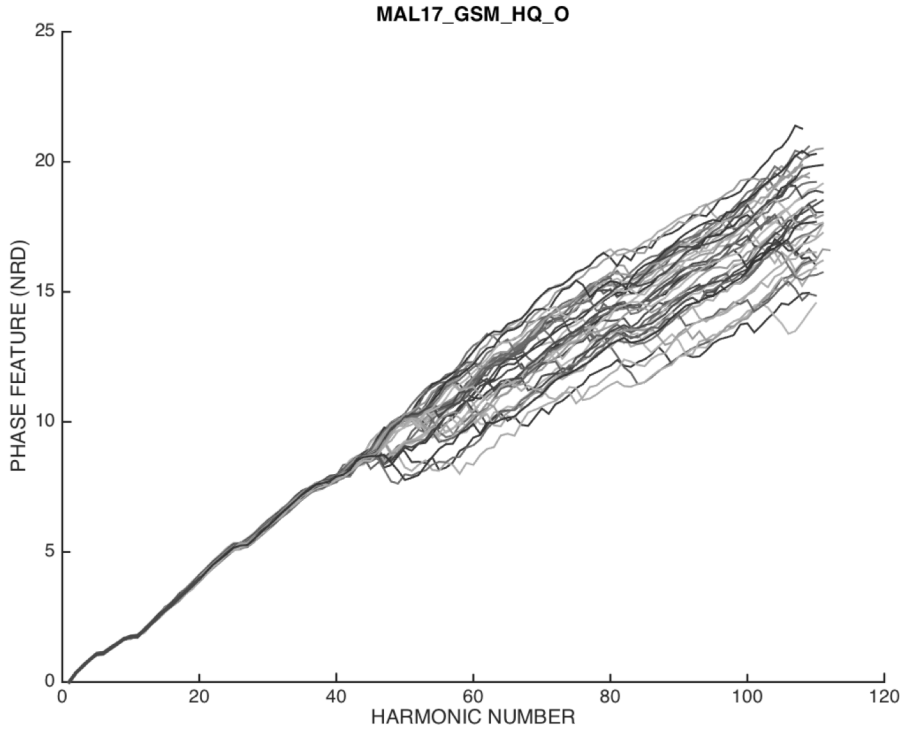


# NRD: a shift-invariant phase-related feature

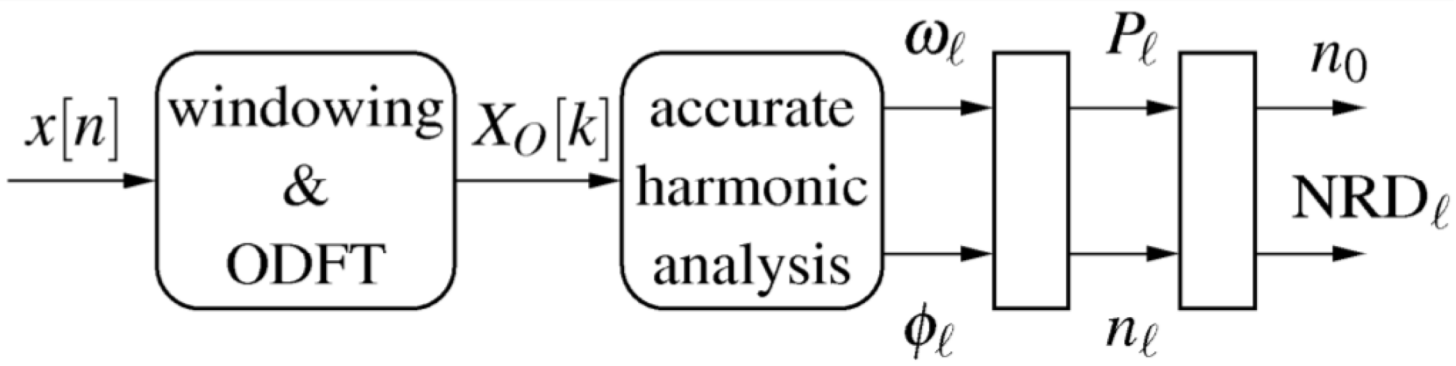
- sustained /a/ vowel by a male



# NRD: a shift-invariant phase-related feature

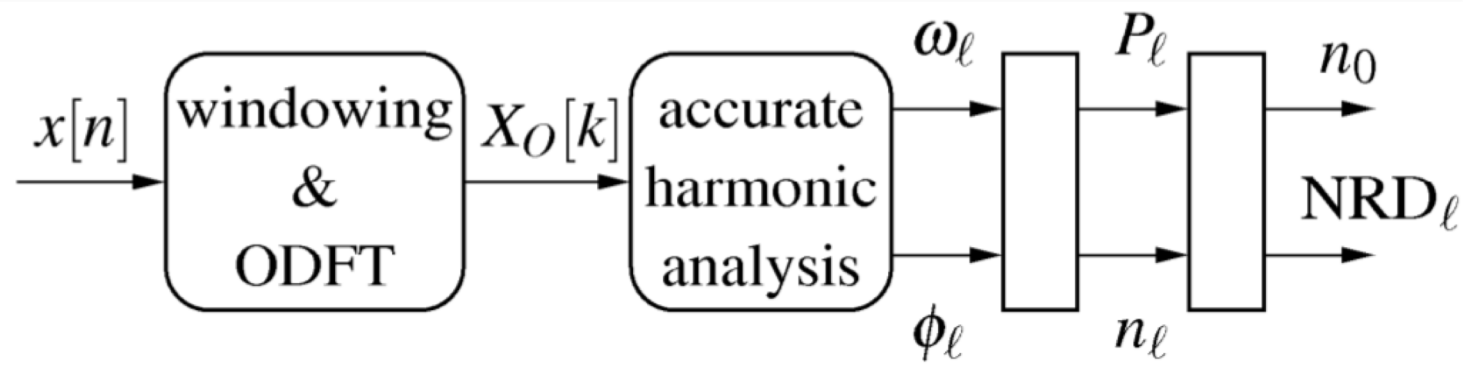
- NRD estimation algorithm




  - we have used the NRD concept in such applications as glottal source modelling, speaker identification, parametric audio coding, and dysphonic voice reconstruction

# NRD: a shift-invariant phase-related feature

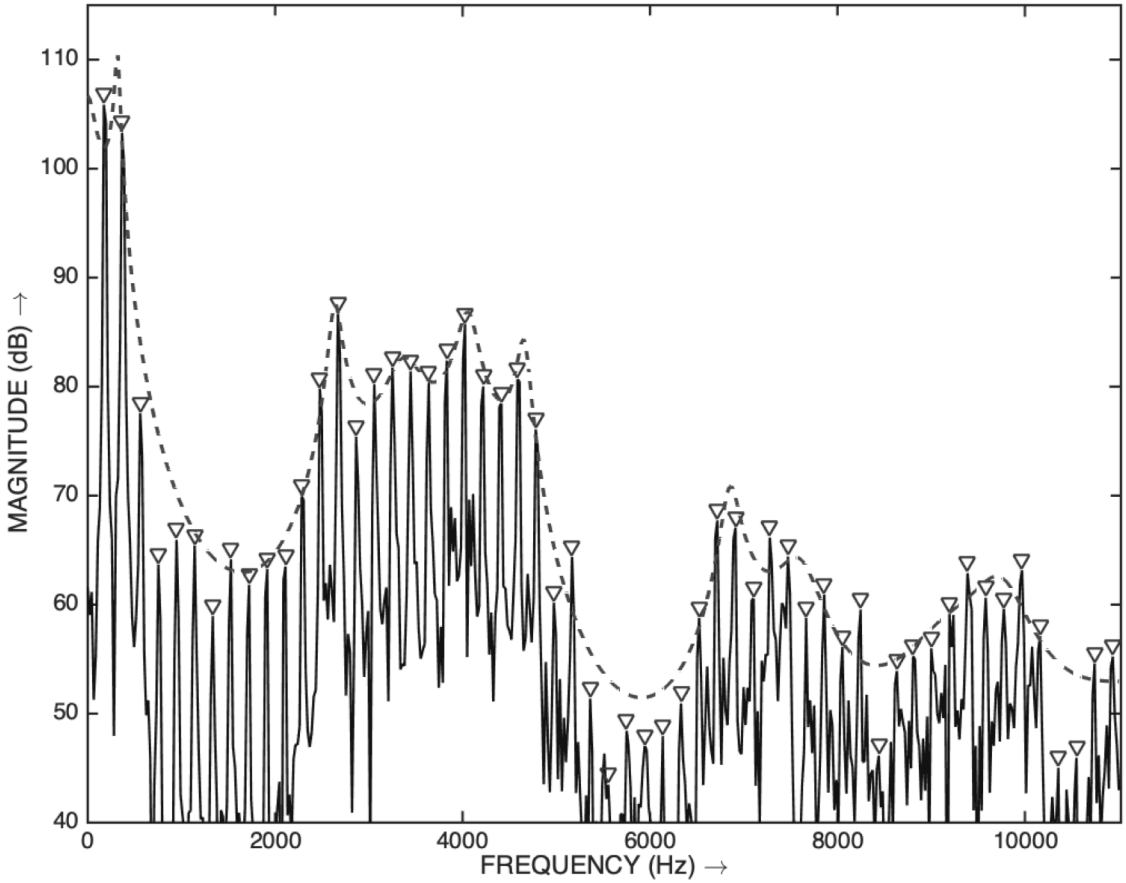
- NRD estimation algorithm




  - Lessons from Nature

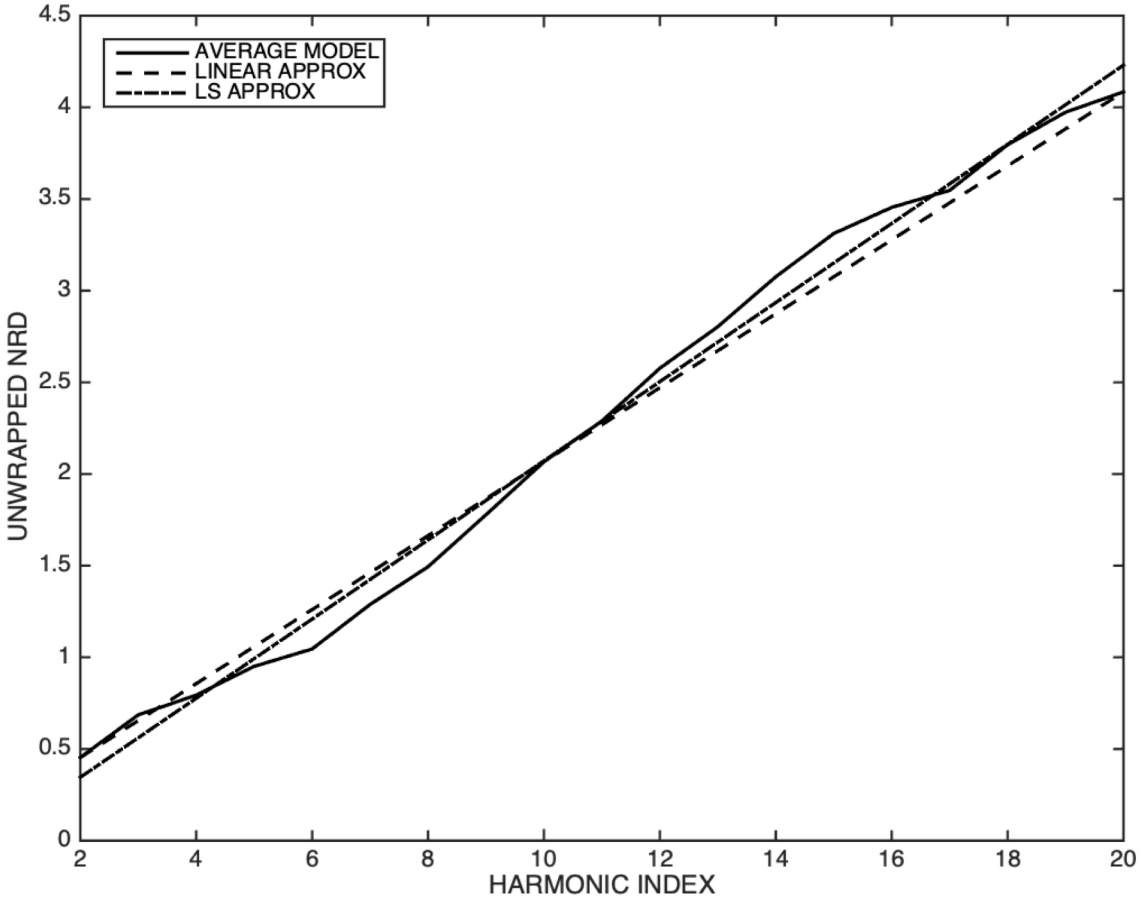
# Reverse engineering of a voiced signal (v1)

- Using LPC spectral envelope model to represent harmonic magnitudes



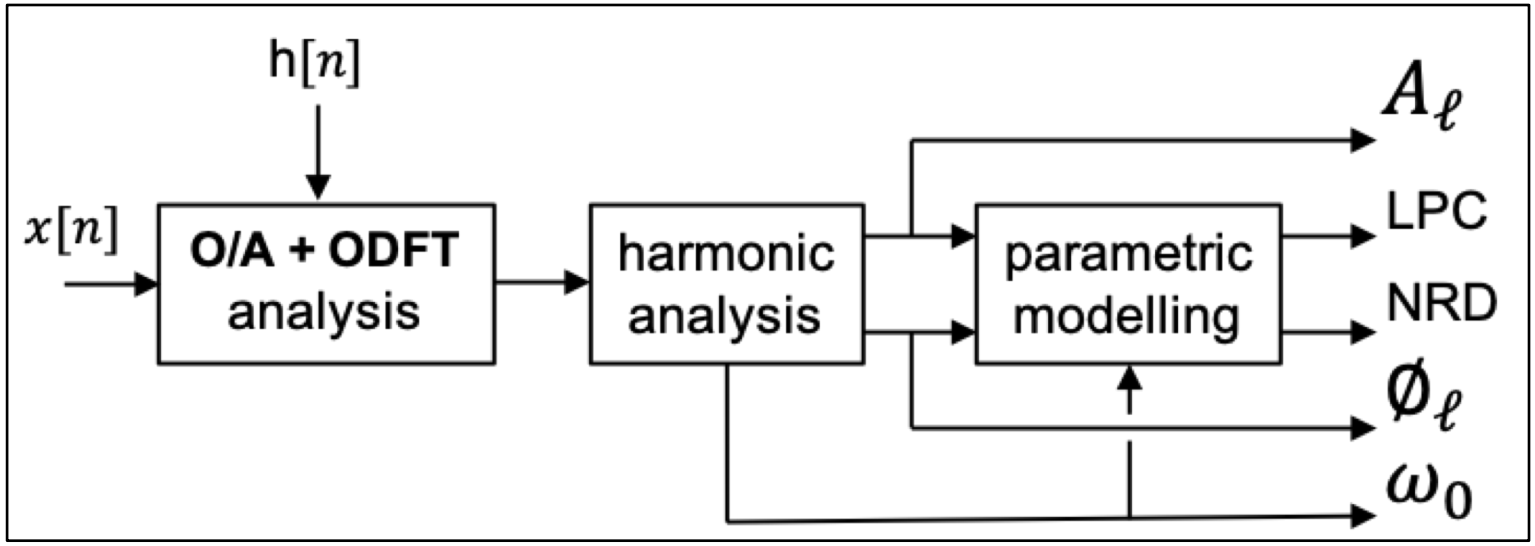
# Reverse engineering of a voiced signal (v1)

- Using a first-order approximation to the stable part of NRD as a model of the phase structure of the periodic signal



# Reverse engineering of a voiced signal (v1)

- Subjective test evaluating the perceptual impact in the synthesis of the signal when the exact harmonic magnitudes ($A_\ell$) are replaced by the LPC model, and when the exact harmonic phases ($\phi_\ell$) are replaced by the NRD model



# Distance metric and performance criteria

- Subjective tests

**/a/ spoken (REFERENCE)** 

 **A version (FREQ)**

 **D version (TIME)**

 **B version (FREQ, fixNRD)**

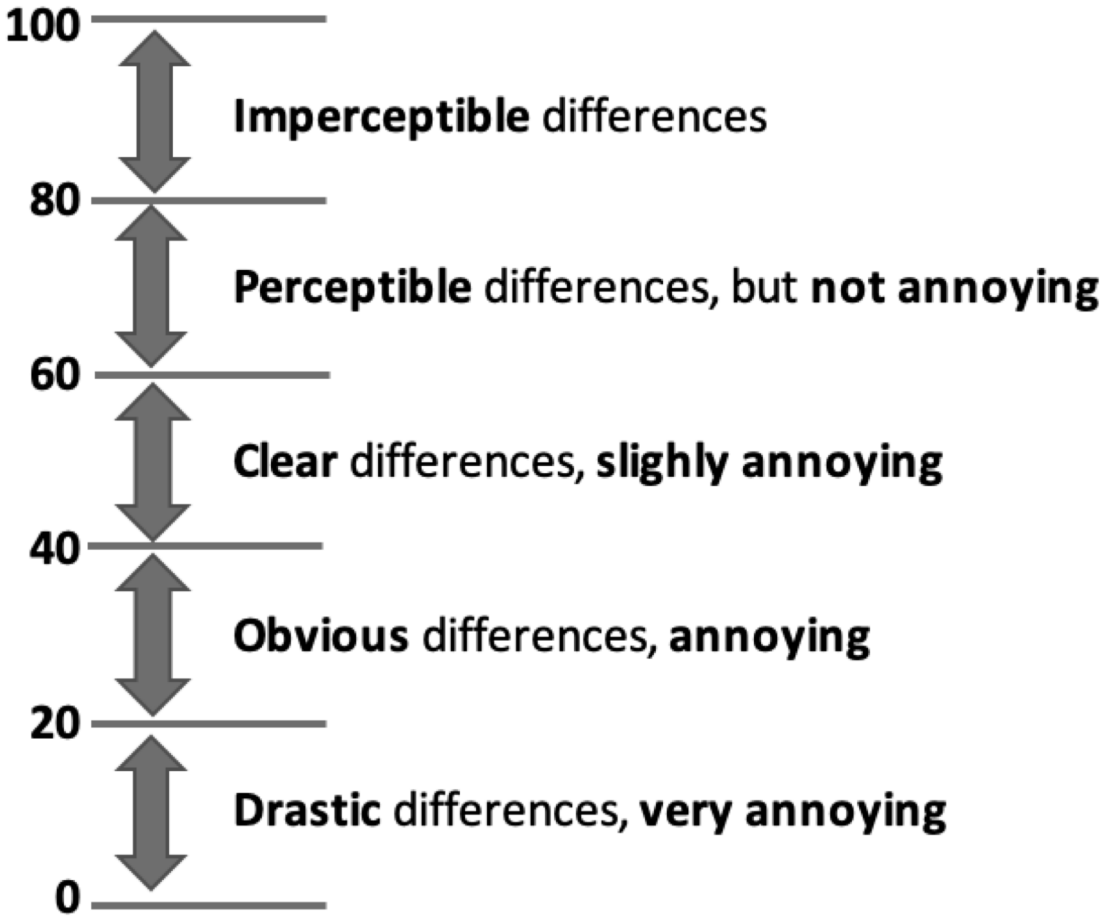
 **E version (TIME, fixNRD)**

 **C version (FREQ, modLPC)**

 **F version (TIME, modLPC)**

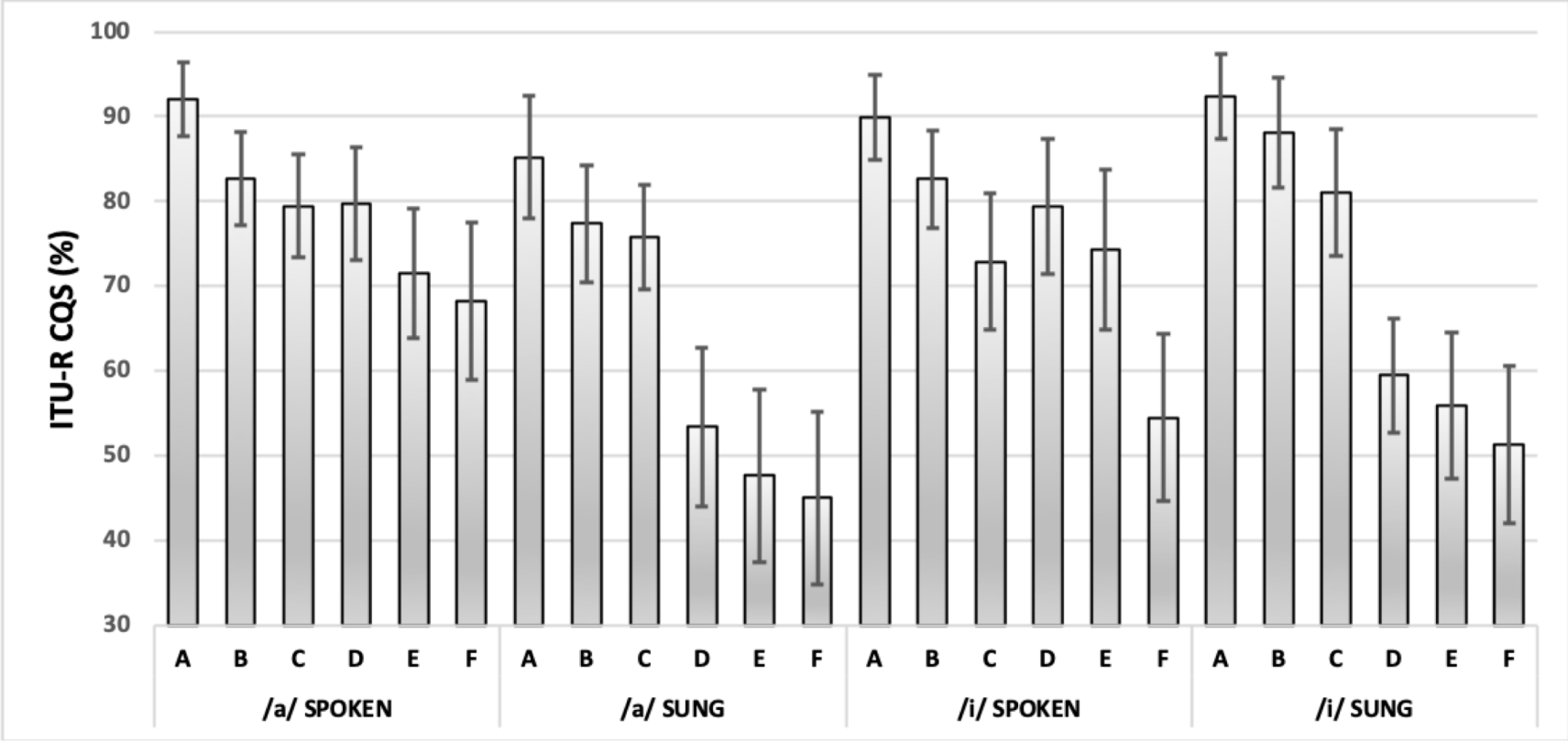
# Distance metric and performance criteria

- Grading scale



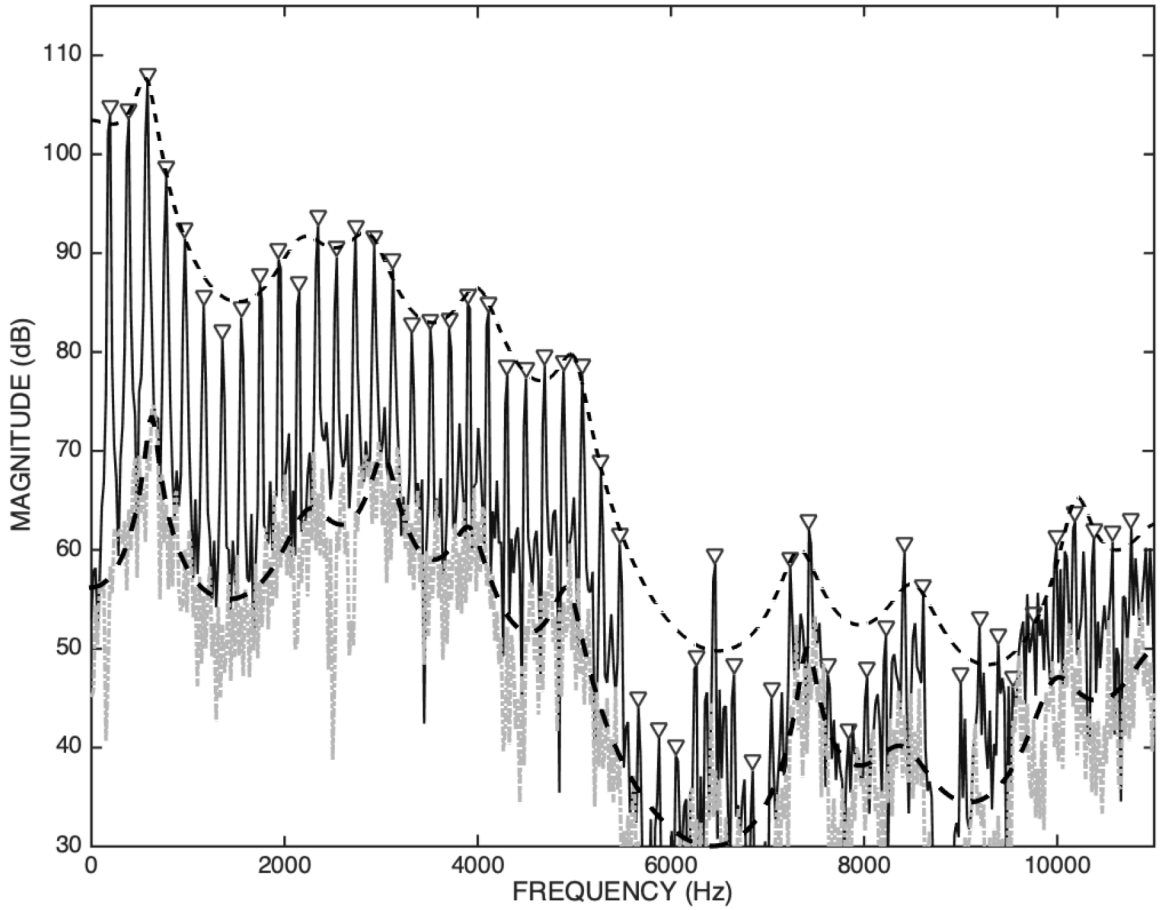
# Distance metric and performance criteria

- Test results



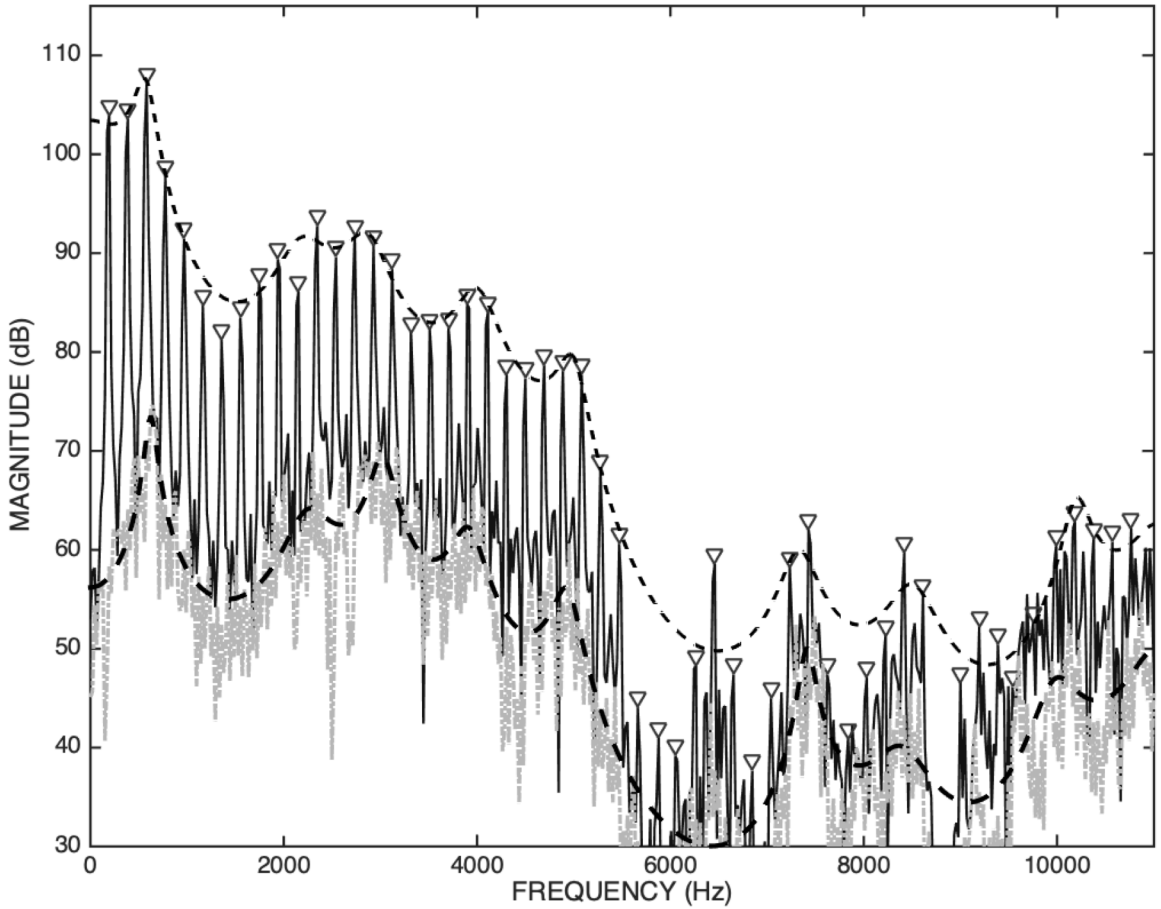
# Reverse engineering of a voiced signal (v2)

- Using one LPC spectral envelope model to represent harmonic magnitudes, and another LPC to represent the spectral residual



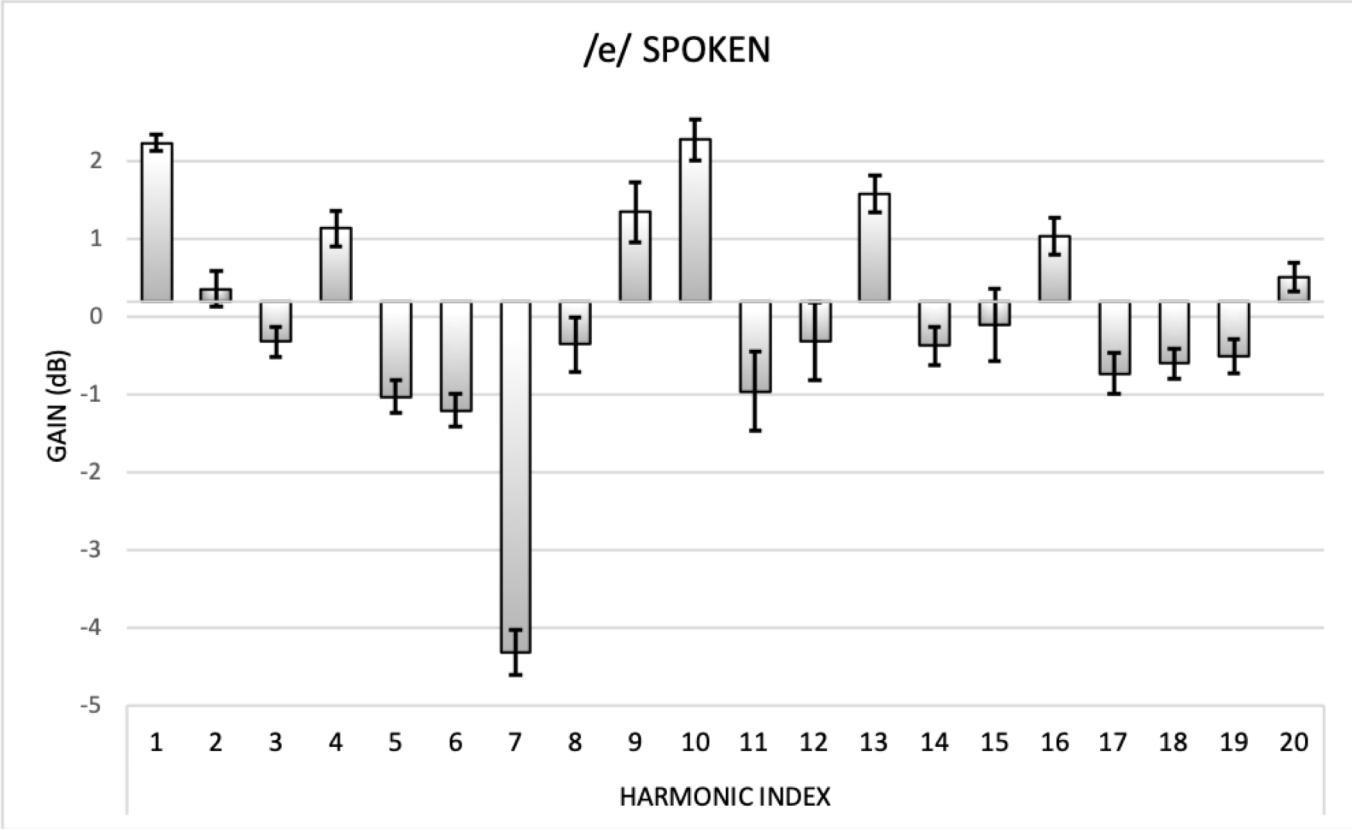
# Reverse engineering of a voiced signal (v2)

- Using the average magnitude difference ($D_\ell$) between each harmonic magnitude and the harmonic LPC magnitude model



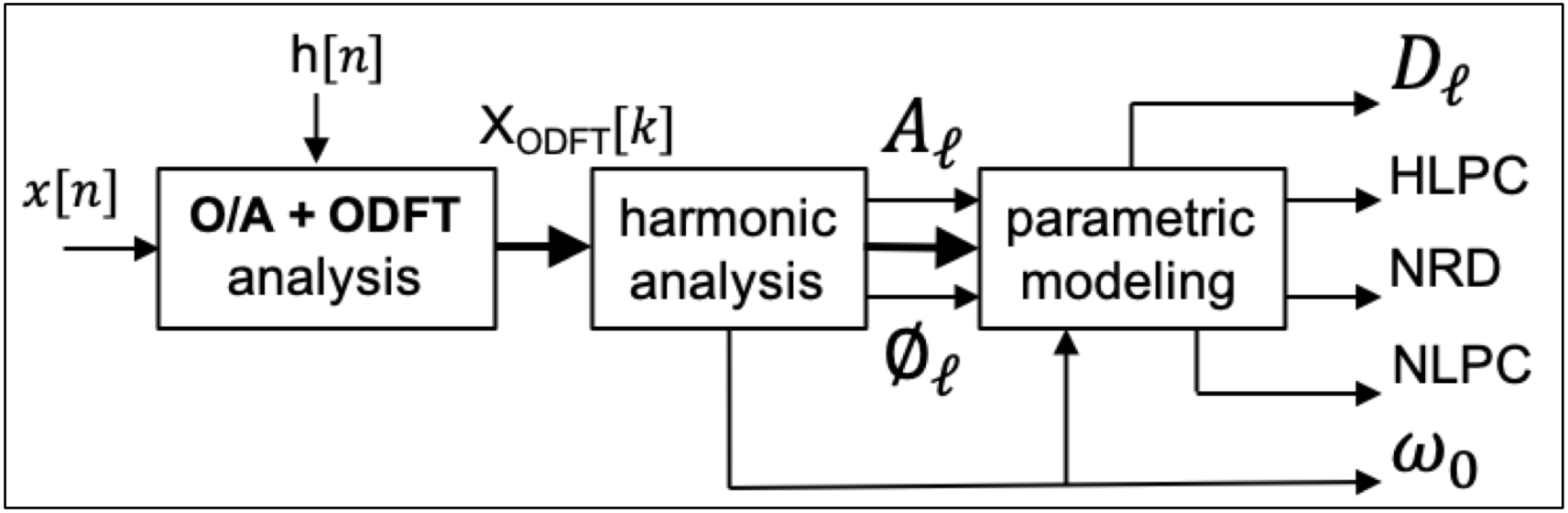
# Reverse engineering of a voiced signal (v2)

- Using the average magnitude difference ($D_\ell$) between each harmonic magnitude and the harmonic LPC magnitude model



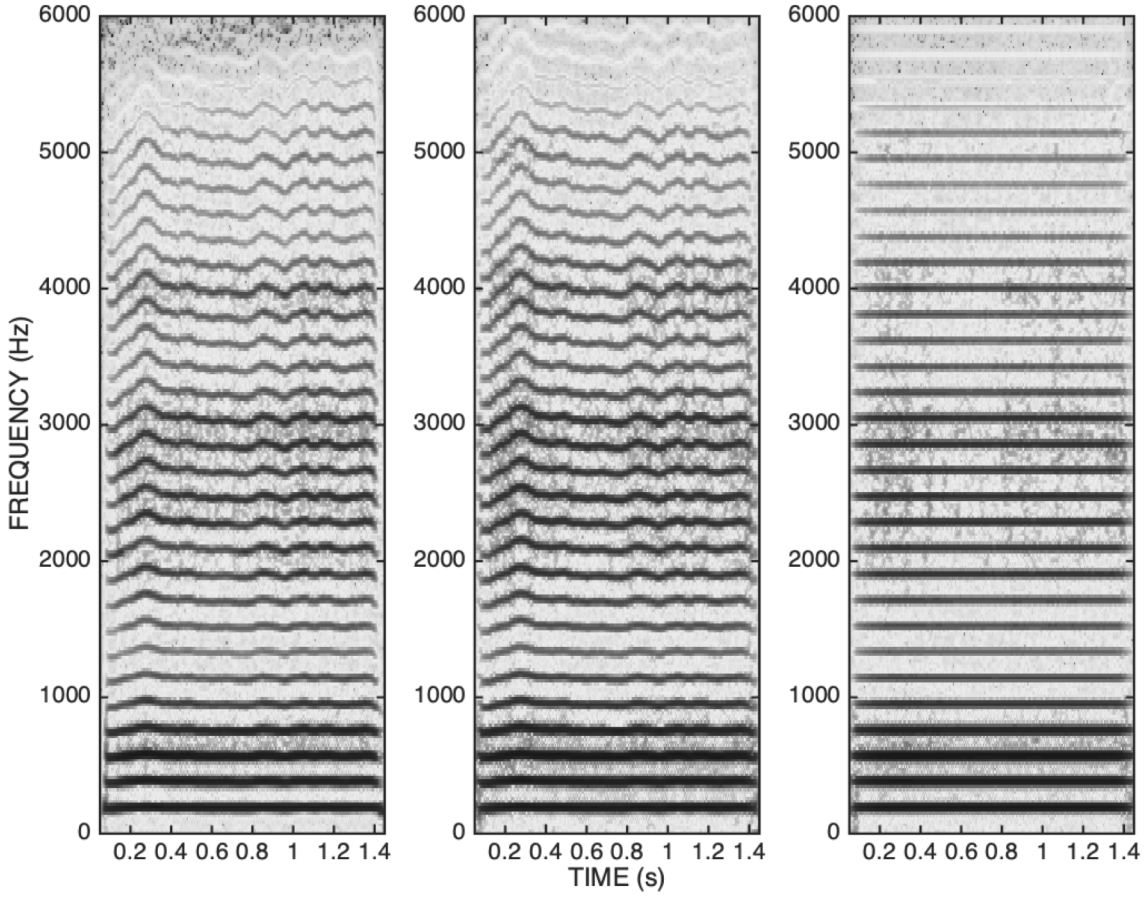
# Reverse engineering of a voiced signal (v2)

- Subjective test evaluating the perceptual impact in the synthesis of the signal when the exact harmonic magnitudes are replaced by the HLPC model and the average harmonic magnitude differences ($D_\ell$), when the spectral residual is replaced by the NLPC model and when the exact harmonic phases ($\phi_\ell$) are replaced by the average NRD model



# Reverse engineering of a voiced signal (v2)

- Anchor signal: $\omega_0$ is forced to be constant and equal to the average $\omega_0$ value



# Reverse engineering of a voiced signal (v2)

- Subjective tests

/a/ spoken (REFERENCE)

A version (spoken /a/)

B version (spoken /a/)

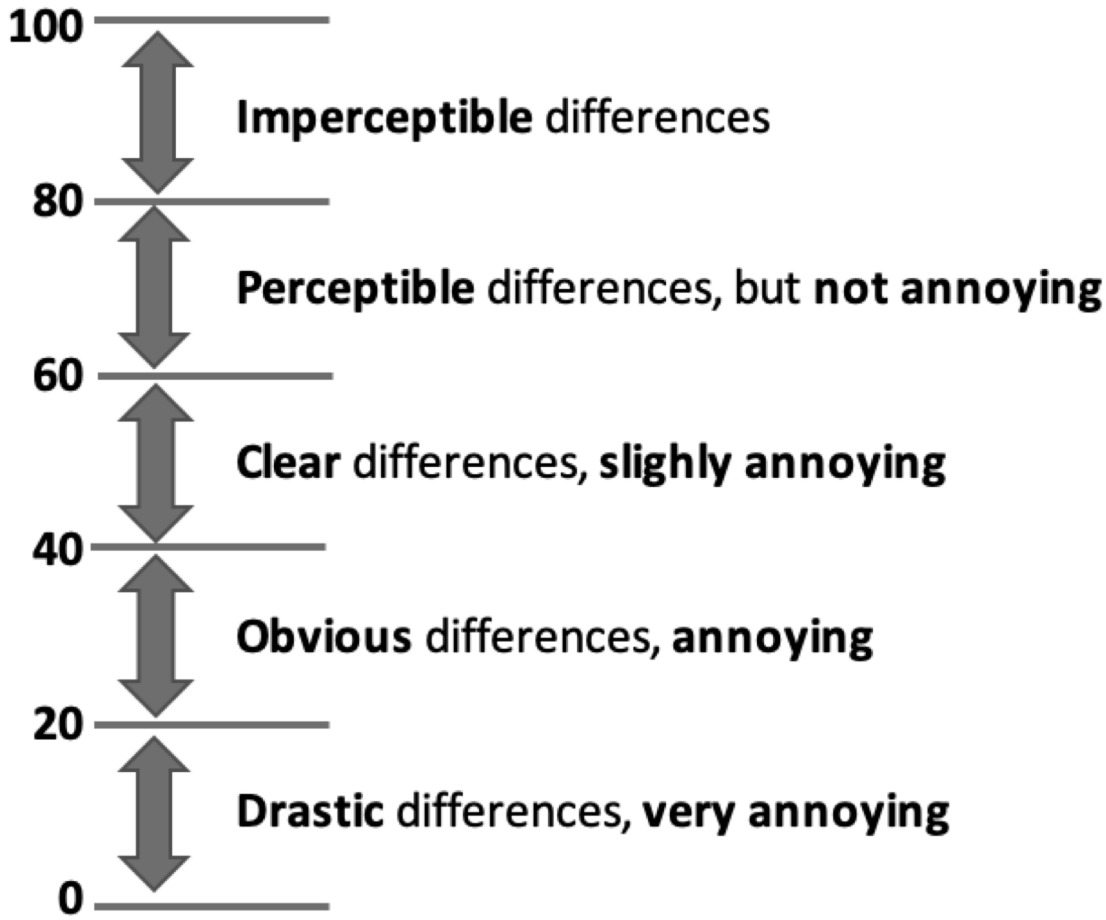
/a/ sung (REFERENCE)

A version (sung /a/)

B version (sung /a/)

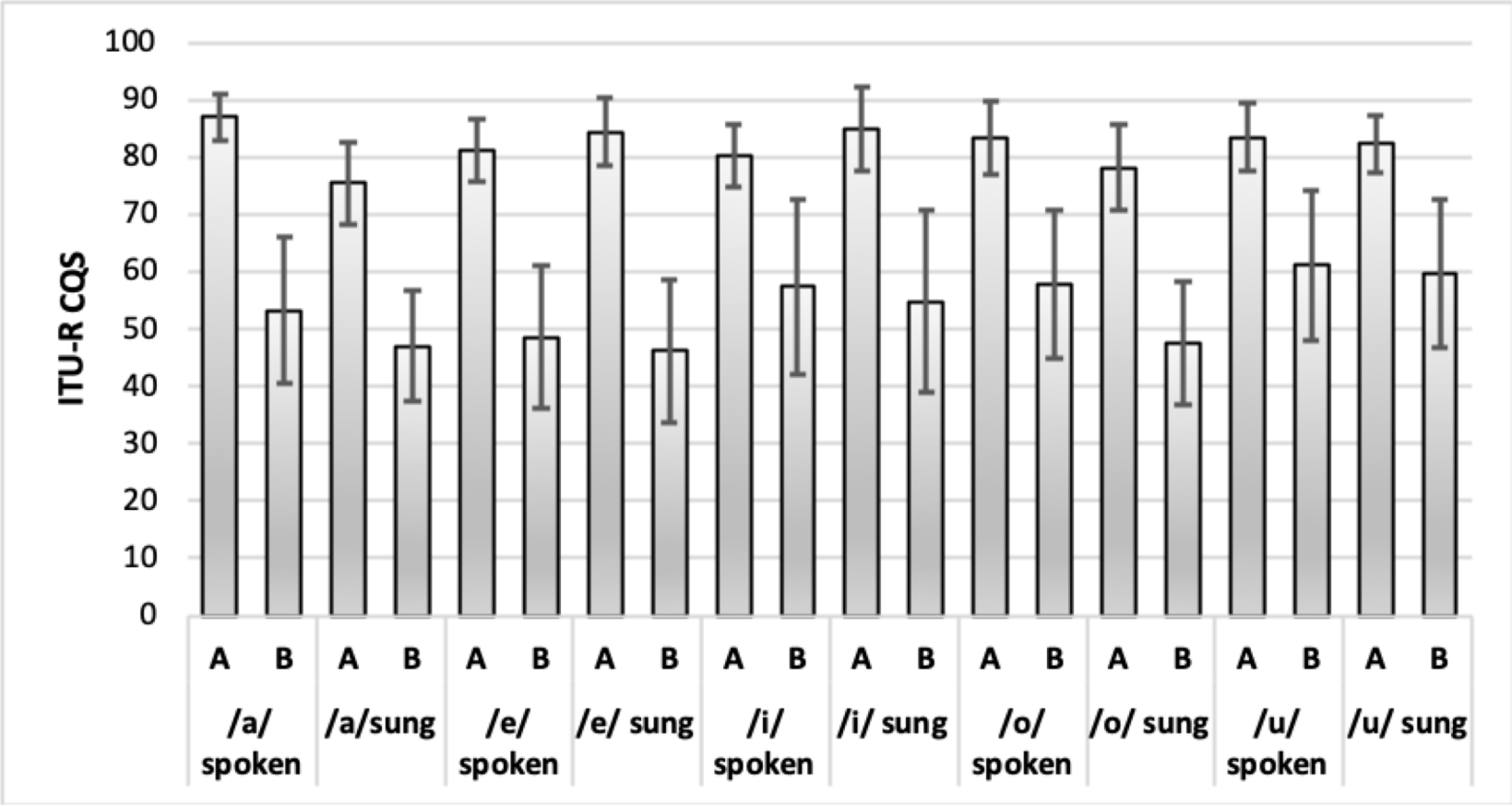
# Reverse engineering of a voiced signal (v2)

- Grading scale



# Reverse engineering of a voiced signal (v2)

- Test results



# Vowel morphing

**sound 1**
**(spoken vowels)**

**sound 2**
**(sung vowels)**

# Conclusions

- Parametric modeling and synthesis of a voiced sound can be achieved using shift-invariant spectral magnitude and shift-invariant phase-related features

- 5 dimensions are sufficient for a (perceptually) complete reverse engineering of a voiced sound (while preserving linguistics, naturalness and idiosyncrasy)

- Each one of those 5 dimensions (average harmonic magnitude, fine harmonic magnitude, average spectral noise magnitude, average shift-invariant harmonic phase structure and F0 contour) can be *controlled independently*

- Synthetic voicing is ready to be used and depends critically on correct phonetic-oriented segmentation of whispered speech which is a project task currently in progress