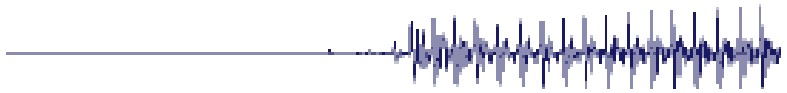
Pretext tasks
4. CPC
|| AR / LPC / VAR / RNN / APC / CPC ||

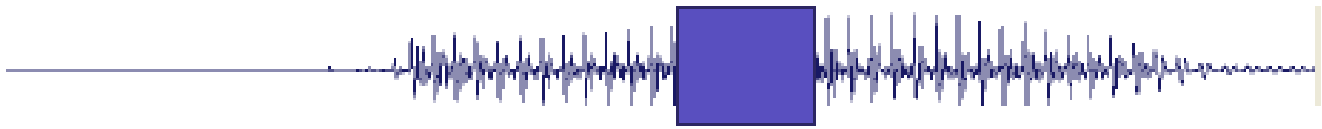
1, 2	SELF-PREDICTION	INNATE RELATIONSHIP (Context-based)	1. ROTATION 2. RELATIVE POSITION	IMAGE
3	CONTRASTIVE LEARNING	INTER-SAMPLE CLASSIFICATION	1. Instance Discrimination 2. SimCLR [Contrastive Loss] 3. Theory – Guarantees / Bounds	IMAGE
4	CONTRASTIVE LEARNING	INTER-SAMPLE CLASSIFICATION	Contrastive Predictive Coding (CPC), [NCE, InfoNCE Loss]	AUDIO/ SPEECH
5	SELF-PREDICTION	GENERATIVE (VAE)	1. AE – Variational Bayes 2. VQ-VAE + AR	IMAGE AUDIO/ SPEECH
6	SELF-PREDICTION	GENERATIVE (AR)	1. AR-LM – GPT 2. Masked-LM – BERT	LANGUAGE
7	SELF-PREDICTION	MASKED-GEN (Masked LM for ASR)	1. Wav2Vec / 2.0 2. HuBERT	AUDIO/ SPEECH

Learning with or without supervision – speech and audio

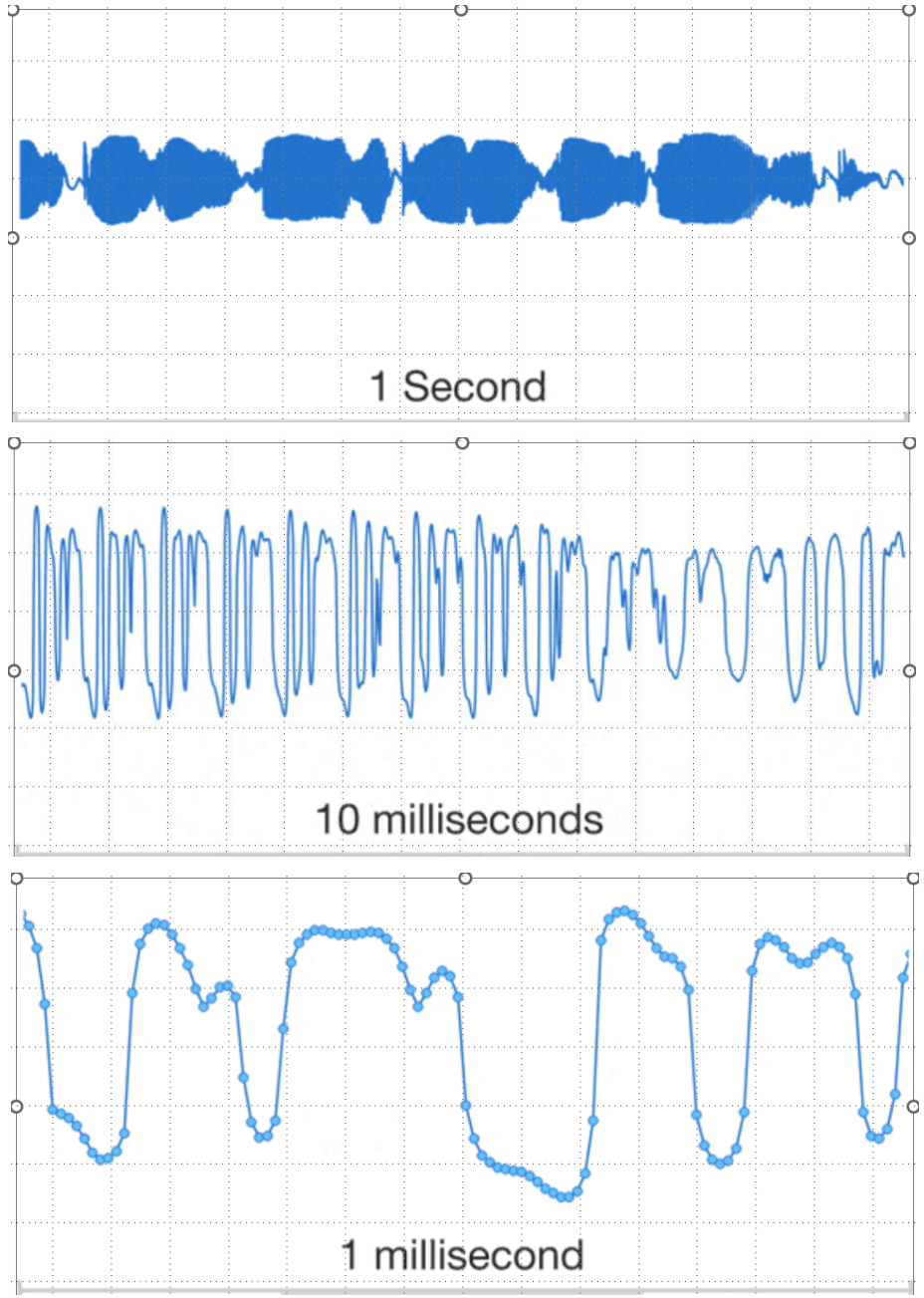
- Next frame prediction



- Masked prediction

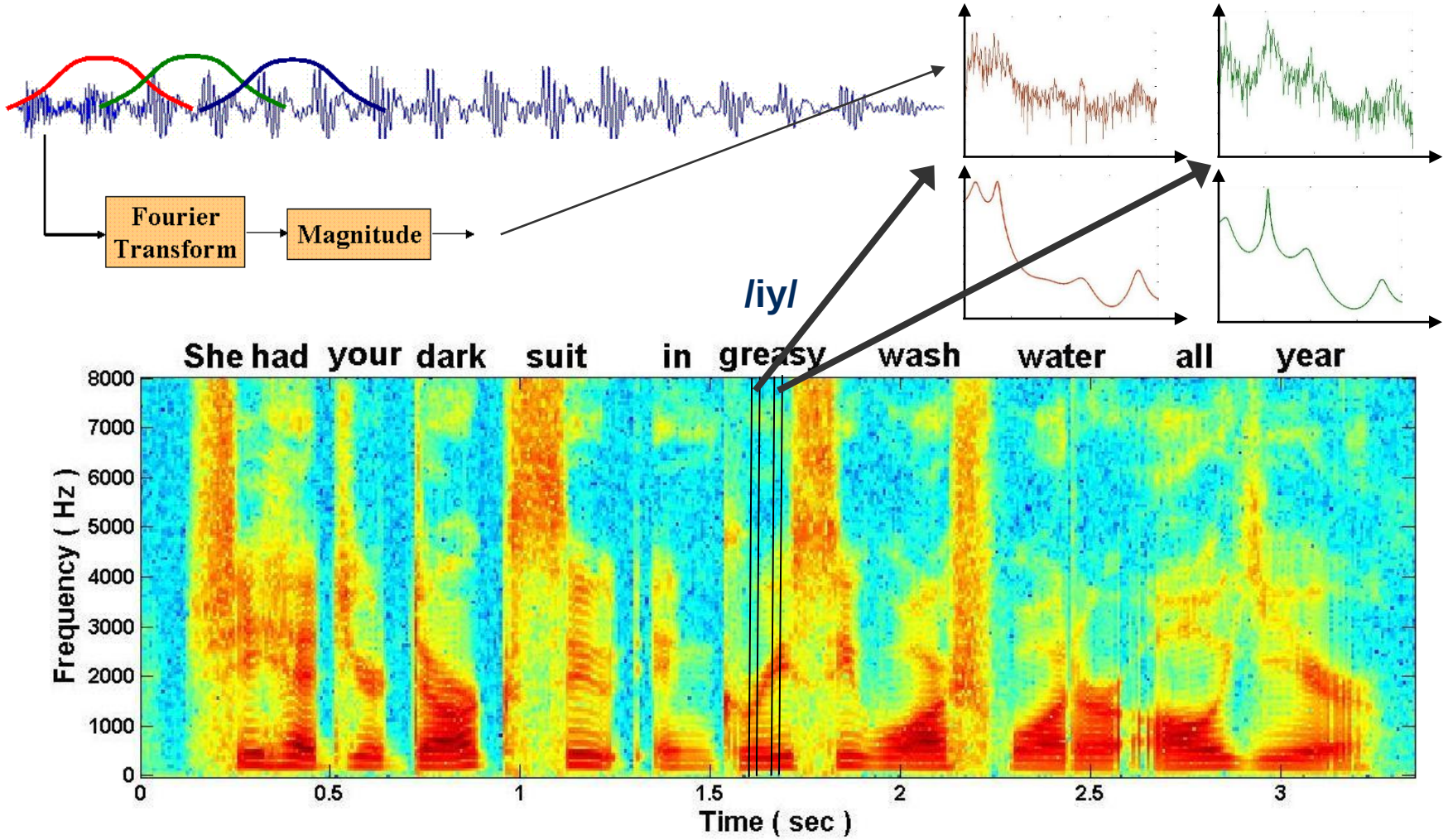


Auto-regressive Model and Speech Spectrum

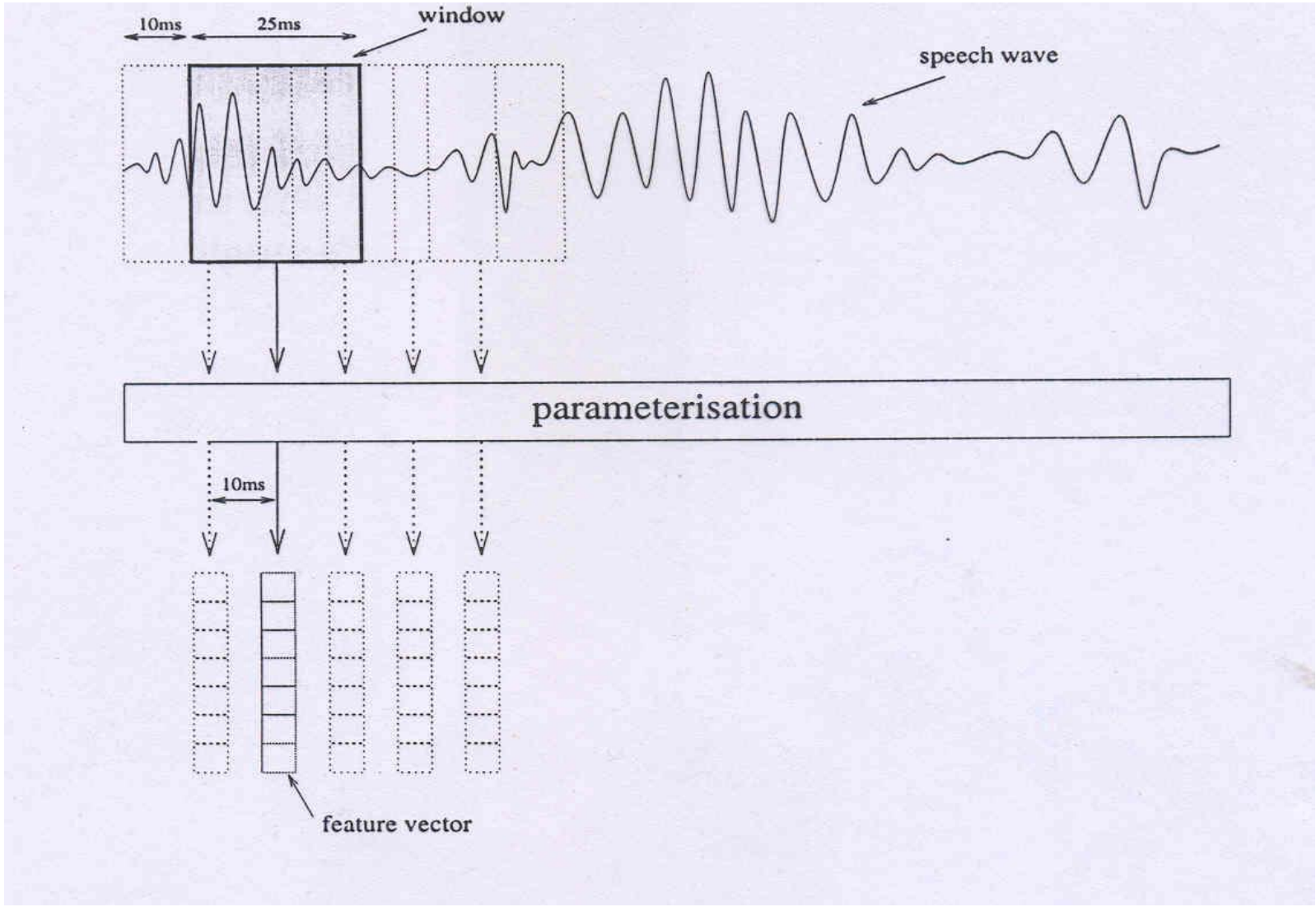

1 Second
10 milliseconds
1 millisecond

Spectrogram

- Speech is a continuous evolution of the vocal tract
- Spectrogram shows time-frequency evolution
- Represented as a time-series of short-time spectra



Short-time Analysis and Parameterization



AR/LPC/VAR/RNN/APC/CPC

AR — Auto-Regressive

LPC — Linear Predictive Coding (Analysis)

VAR — Vector Auto-Regression

RNN — Recurrent Neural Networks

APC — Auto-regressive Predictive Coding

CPC — Contrastive Predictive Coding

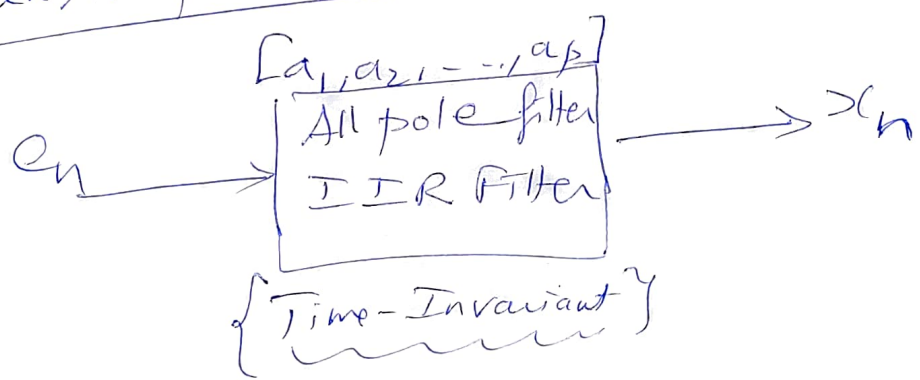
AR(p) : Auto-regressive model of order 'p'

$$x_n = \sum_{i=1}^{p} a_i \, x_{n-i} + e_n \qquad -(1)$$

$\{a_1, a_2, \ldots a_i, \ldots a_p\}$: Parameters of the model

e_n : white noise

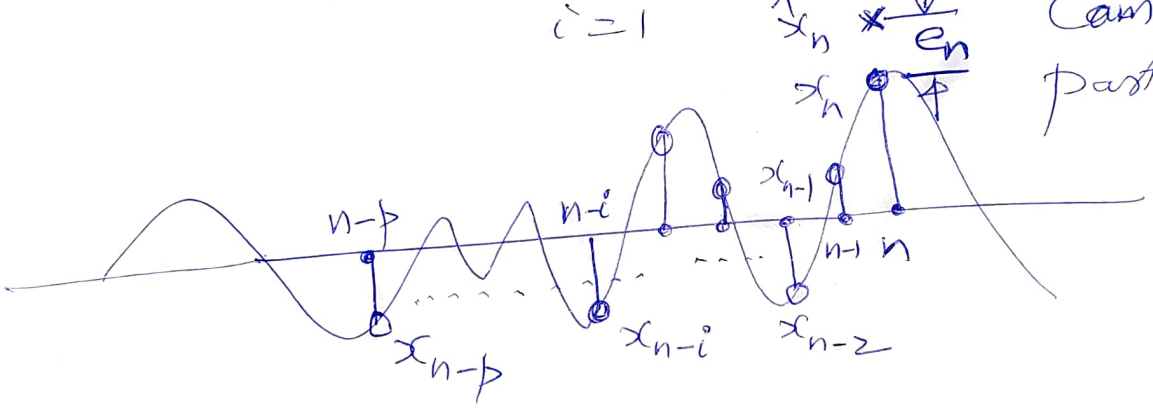
x_n as an AR process



$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + e_n \quad \text{— (1)}$$

$$\hat{x}_n = \sum_{i=1}^{p} a_i x_{n-i}$$

Predict x_n as $\hat{x}_n$ as a linear combination of past 'p' samples.



$$\hat{x}_n = a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_i x_{n-i} + \cdots + a_p x_{n-p}$$

$$x_n = \hat{x}_n + e_n$$

$$\hat{x}_n = \sum_{i=1}^{p} a_i x_{n-i}$$

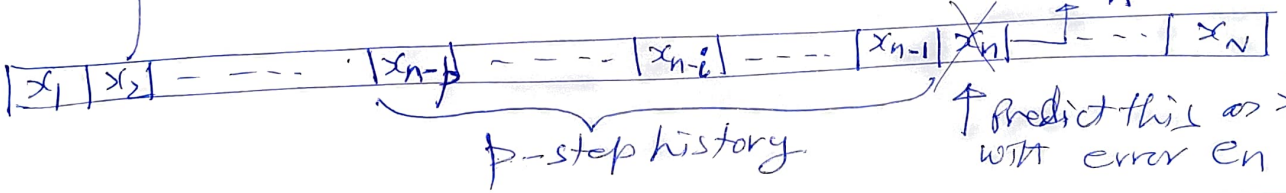
or

$$e_n = x_n - \hat{x}_n$$

x_n: True Signal Value at n^{th} time instant

$\hat{x}_n$: Predicted Signal Value at n^{th} time instant

Frame/Record of N-samples e.g. 10 ms of speech



Optimization / Solution for $\{a_i\}_{i=1}^{p}$ $\left\{\begin{matrix}\text{PREDICTION}\\ \text{CO-EFFICIENTS}\end{matrix}\right\}$

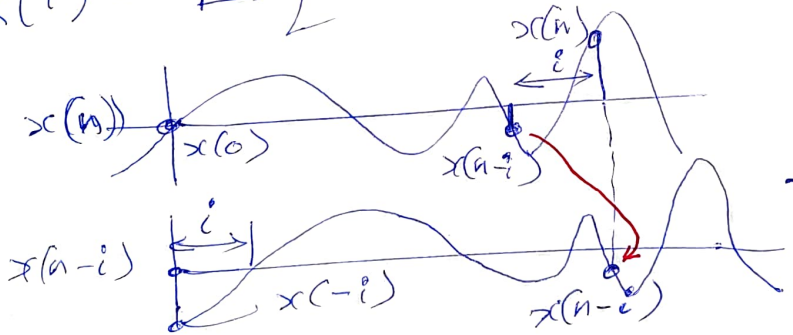
Minimize $E\left[e^2(n)\right]$ over a range of values

$$\Rightarrow \sum_{i=1}^{p} a_i R(j-i) = R(j) \qquad j = 1, \ldots, p$$

Normal Equations or Yule-Walker Equations

$R(i)$: Auto Correlation Coeff of $x(n)$ for lag 'i'

$$R(i) = E\left[x(n)\, x(n-i)\right]$$



→ Right shifted by 'i' samples

Solution for $\{a_i\}_{i=1}^{p}$

$$\sum_{i=1}^{p} a_i R(j-i) = R(j), \quad 1 \le j \le p$$

Matrix form

$$[a_1\ a_2\ \cdots\ a_{p-1}\ a_p]^T \rightarrow [p \times 1]$$

Symmetric, Toeplitz

$$R\ A = r$$

$[p \times p]$ $[p \times 1]$ $[p \times 1]$

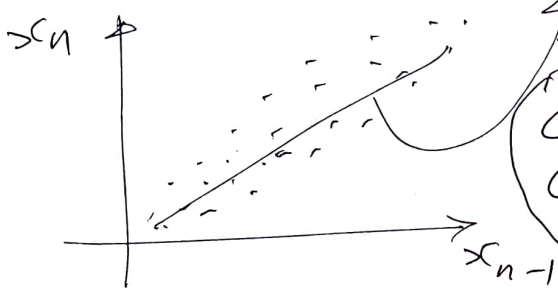
$$r = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(j) \\ \vdots \\ R(p) \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ j \\ \vdots \\ p \end{matrix}$$

R: $p \times p$ matrix, columns $1, 2, \ldots, j, \ldots, p$; rows $1, 2, \ldots, i, \ldots, p$

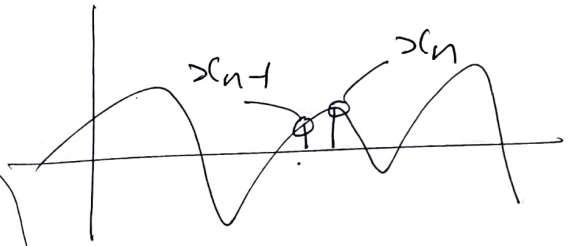
$\gamma_{ij} = R(i-j)$

Order 1 AR: AR(1)

$$x_n = a_1 x_{n-1}$$



Correlation Coefficient of lag 1



AR(0): Noise : $x_n = \cancel{p=0} \; \cancel{\sum} \; a_i x_{n-i} \; \cancel{i=1} + e_n$

AR(1): $x_n = a_1 x_{n-1} + e_n$

For small $a_1 = .3$

$\Rightarrow x_n = .3 x_{n-1} + e_n$ ‖ x_n looks like e_n

$a_1 \to 1$ eg $a_1 = .9$

$x_n = .9 x_{n-1} + e_n$ ‖ "smoothing" of output (Low Pass filter)

$\Rightarrow$ AR(1) $\Rightarrow$ a_1 $\Rightarrow$ Spectral Property

e.g $S(f) = \dfrac{\sigma^2}{1 + a_1^2 + 2a_1 \cos 2\pi f}$

In general for p^{th} order AR(p)

$[a_1\ a_2 \cdots a_p] \Longleftrightarrow [\rho_1\ \rho_2 \cdots \rho_p]$ Auto Correlation Coefficients

Fourier Transform Pair

Power Spectral Density

of $[x_1\ x_2 \cdots x_N]$

e.g 10ms or short time frame/window of speech/audio.

More precisely in z-transform

$$E(z)\left[\frac{1}{1-a_1 z^{-1} - \cdots - a_p z^{-p}}\right] = x(z)$$

$$\Rightarrow \quad e(n) * h(n) = x(n)$$

↳ impulse response of filter.

$$E(z) \cdot H(z) = X(z)$$

$$H(z) = \frac{1}{1-a_1 z^{-1} - \cdots - a_p z^{-p}} = \frac{1}{\prod_{i=1}^{p}(z - z_i)}$$

z_i : Roots of $A(z) = 0$

Spectral Envelope

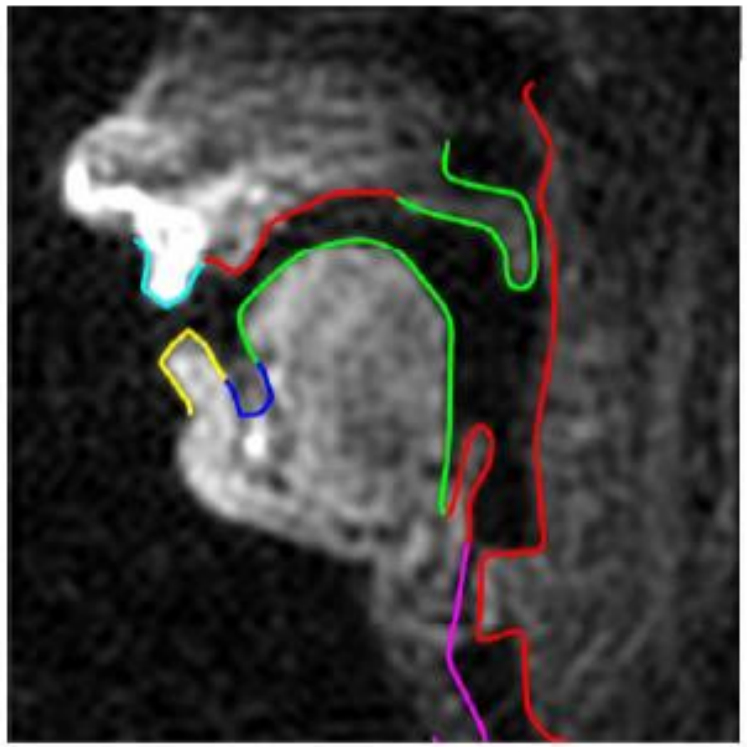
$p = 10$ for speech $\Rightarrow$ 5 Resonances

$$|S(f)| = \frac{1}{|A(z)|}$$

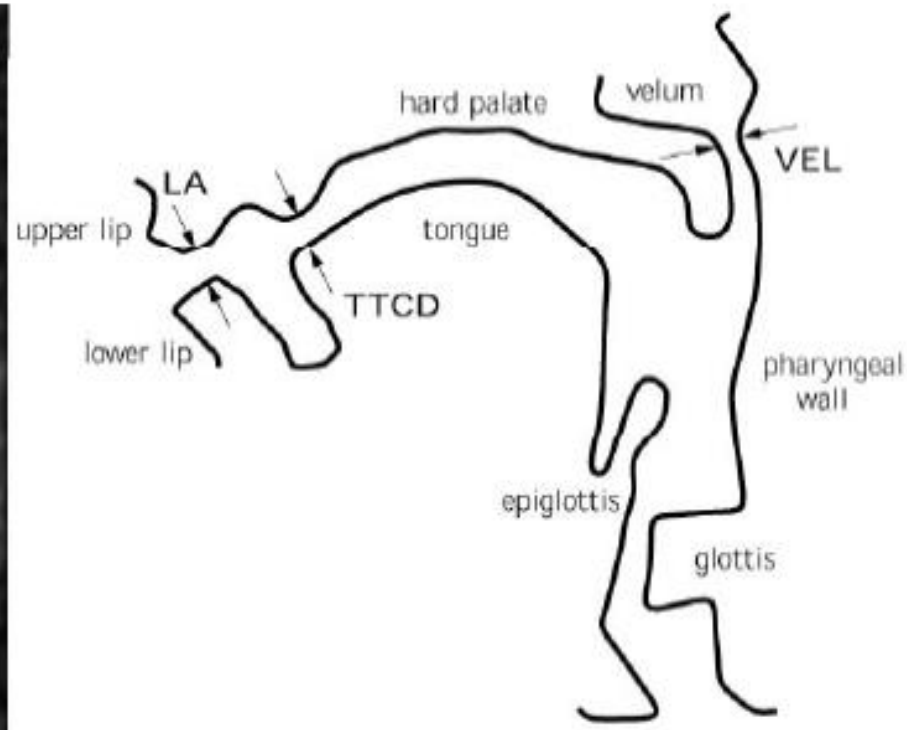
$[a_1\ a_2 \cdots\ a_p] \Rightarrow A(z) \Rightarrow S(f)$

$\longrightarrow$ Feature Vector of $[x_1\ x_2 - x_4 \cdots\ x_N]$

MRI of Speech (Prof. Shri Narayanan, USC)



(a)

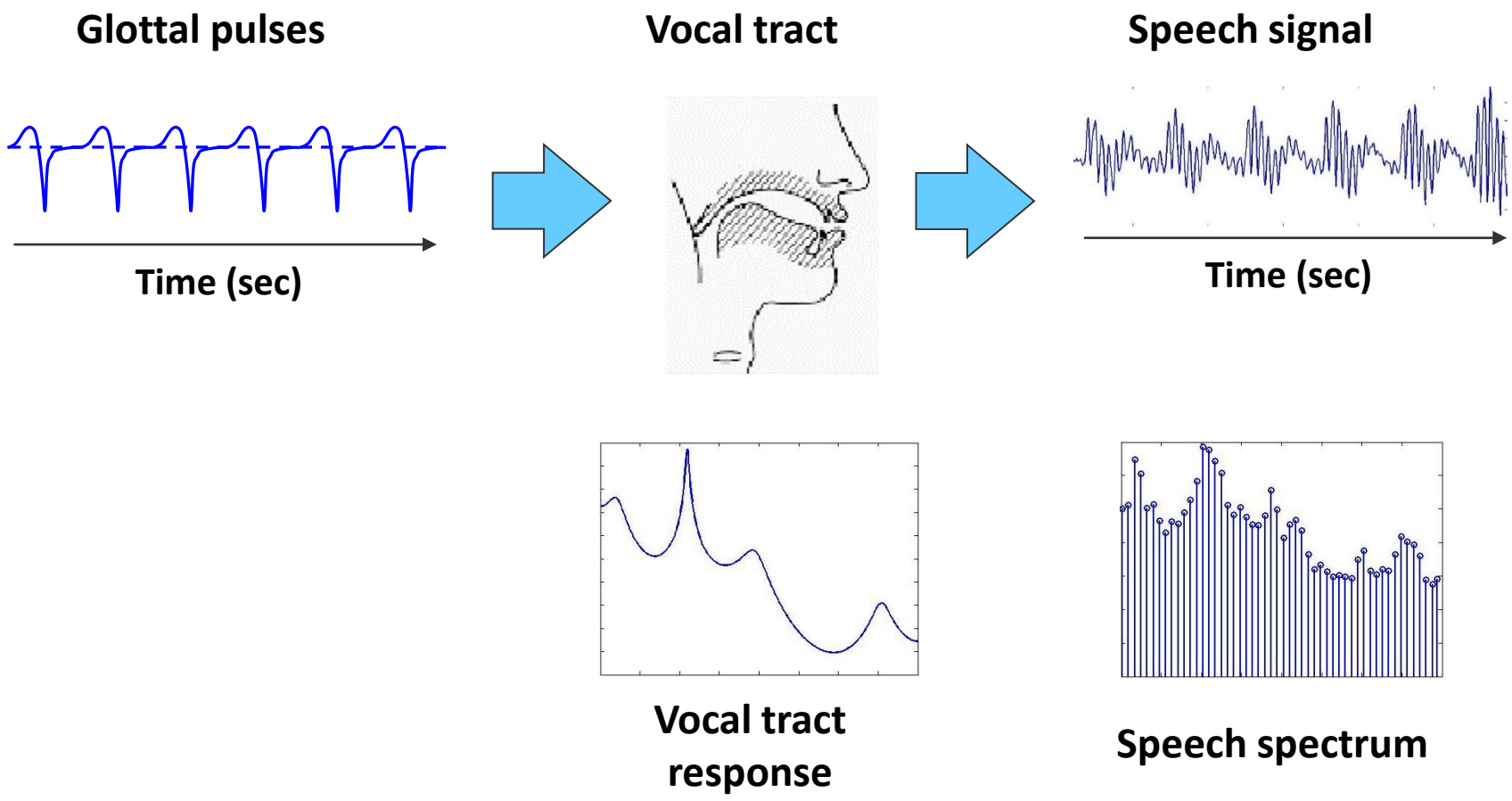


(b)


USC
SPAN

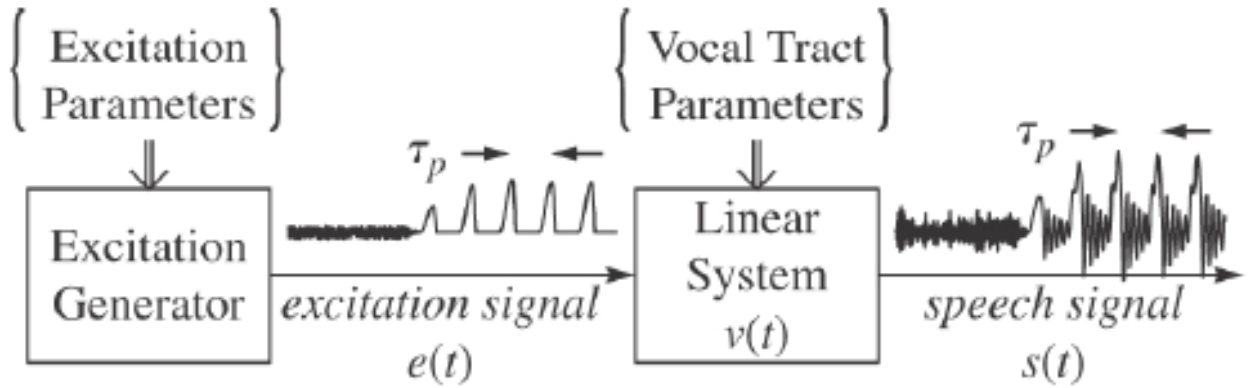
Source-Filter Model

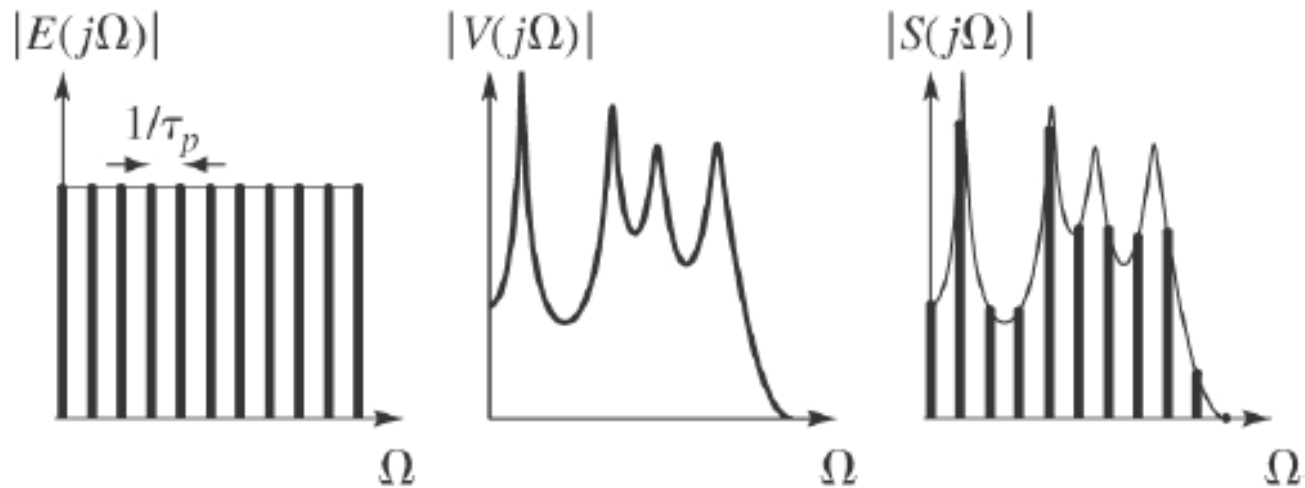
- Features based on speech production model: Source-filter interaction
  - Anatomical structure (vocal tract / glottis) conveyed in speech spectrum



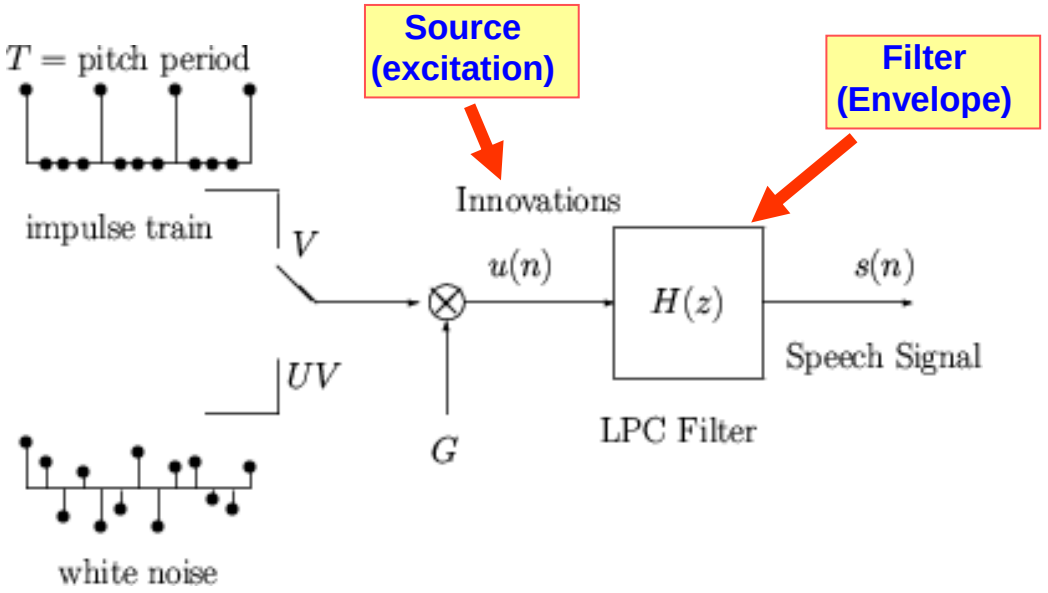
To Quatieri and Rab – Slides ➔

Source-System Model of Speech Production



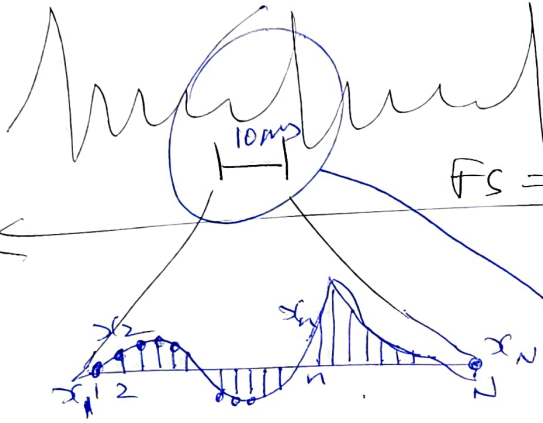


Linear Prediction based Speech Production Model



$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

- Vocal Tract ⟺ H(z) (LPC Filter)
- Air ⟺ u(n) (Innovations)
- Vocal Cord Vibration ⟺ V (voiced)
- Vocal Cord Vibration Period ⟺ T (pitch period)
- Fricatives and Plosives ⟺ UV (unvoiced)
- Air Volume ⟺ G (gain)



$F_S = 40000\,Hz \Rightarrow 40000$ Samples / 1 Sec

$F_S = 40000$ Samples / Sec

1 sec $\simeq$ 40000 samples

1000 ms = 40000 samples

1 ms = 40 samples

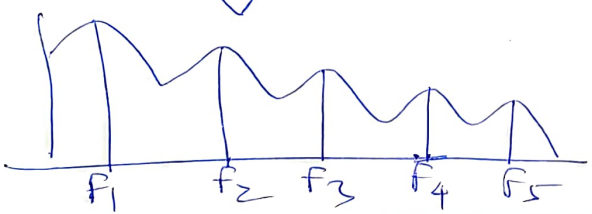
10 ms = 400 samples

400 Samples

$[x_1\ x_2 \cdots x_n \cdots x_N]$

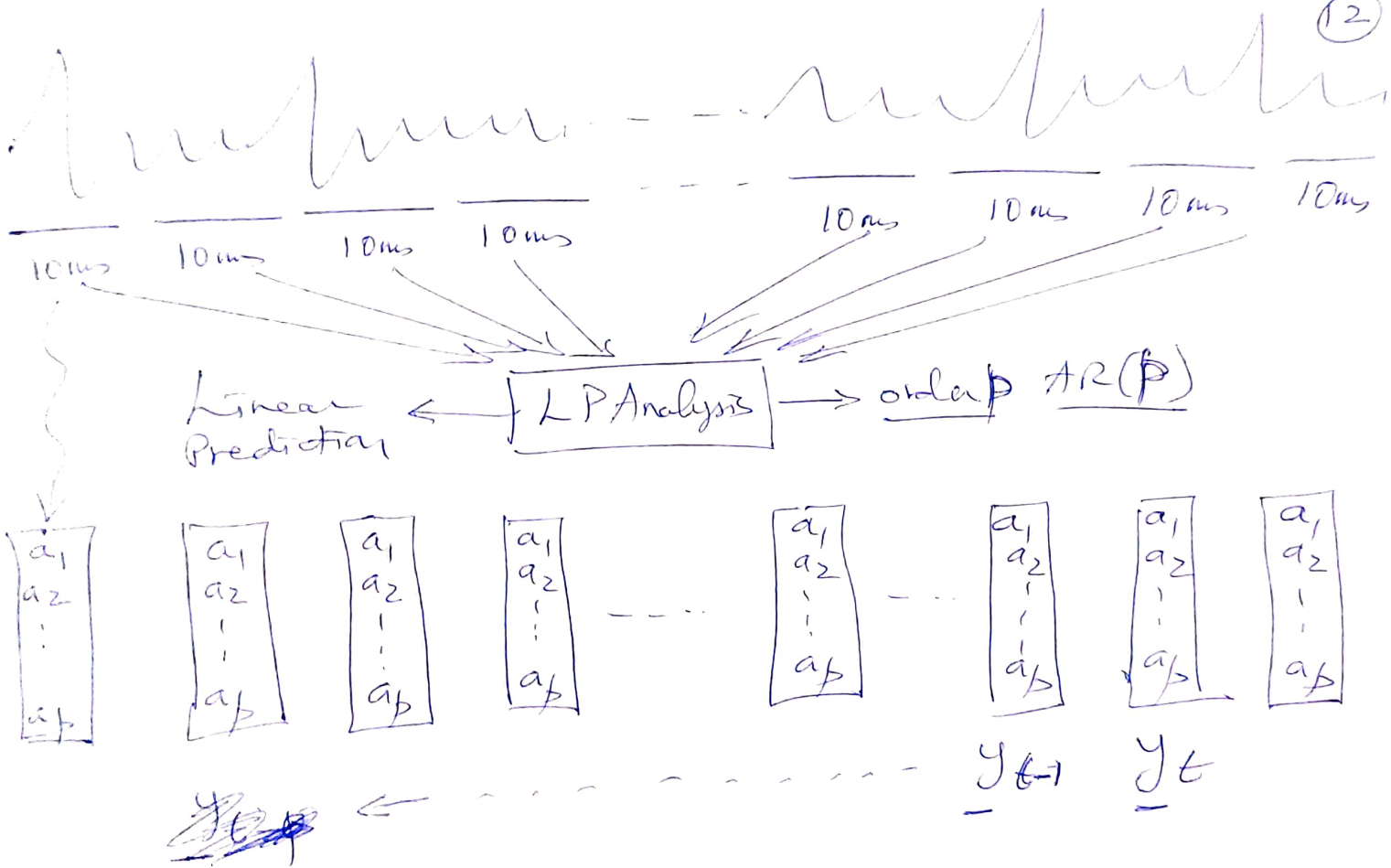
Linear Prediction ⇓ LP Analysis $p = 10$

$[a_0\ a_2 \cdots a_p]$



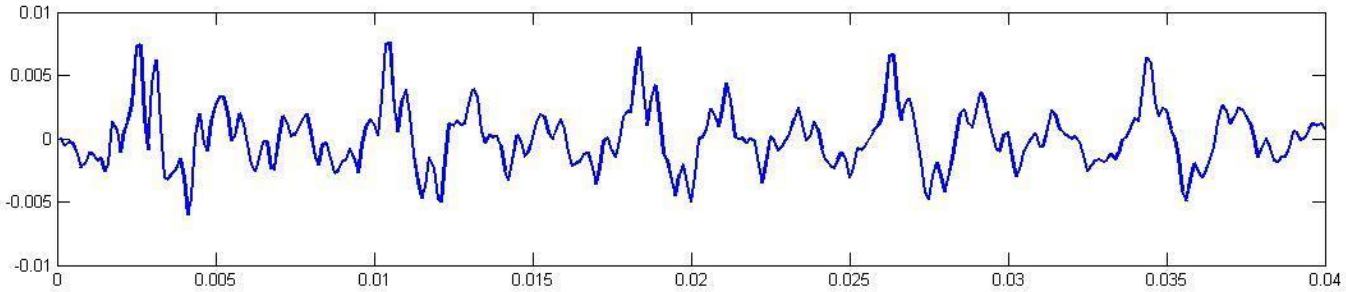
$\Longrightarrow$ p-dim $\in \mathbb{R}^p$ feature representation of $[x_1\ x_2 \cdots x_N]$

10 ms short time frame of speech.

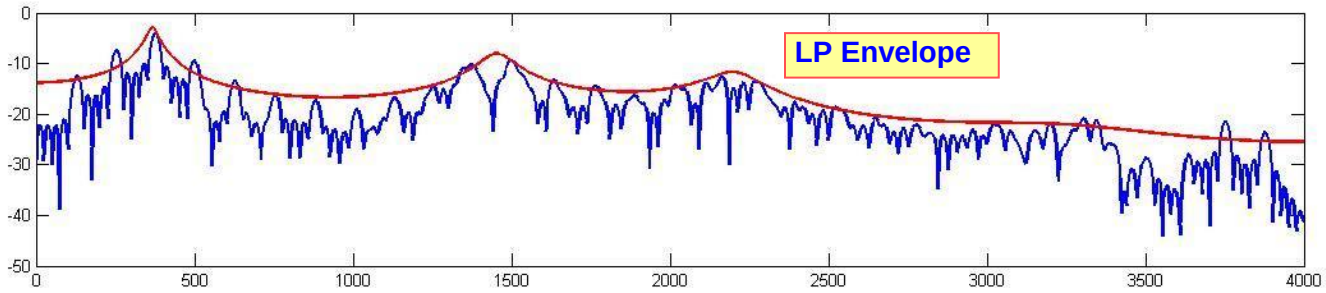

10ms
10ms
10ms
10ms
10ms
10ms
10ms
10ms
LP Analysis
Linear Prediction
order p AR(p)
a_1 a_2 ⋮ a_p
y_{t-1}
y_t

LP Analysis: Envelope (Filter) & Excitation (Source)

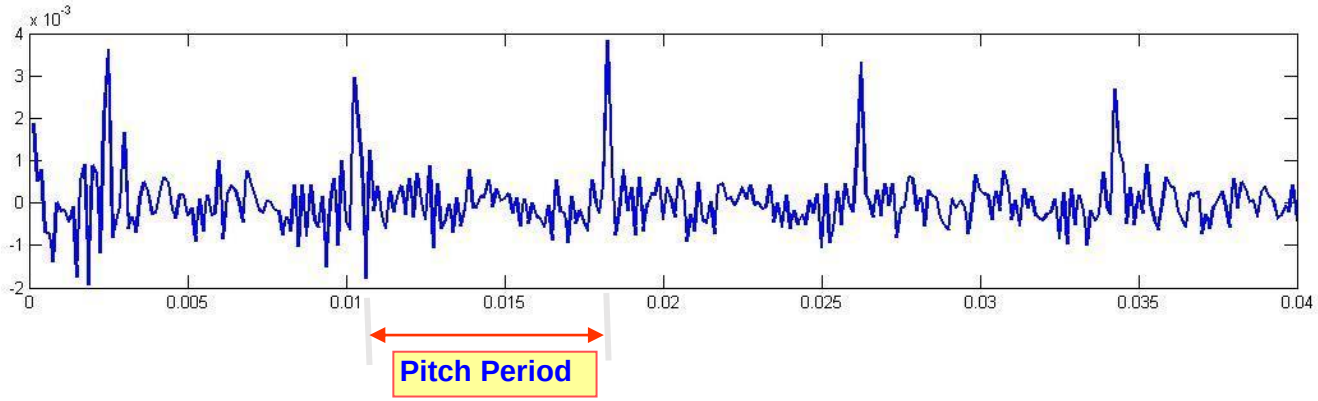
Speech Signal S(n)



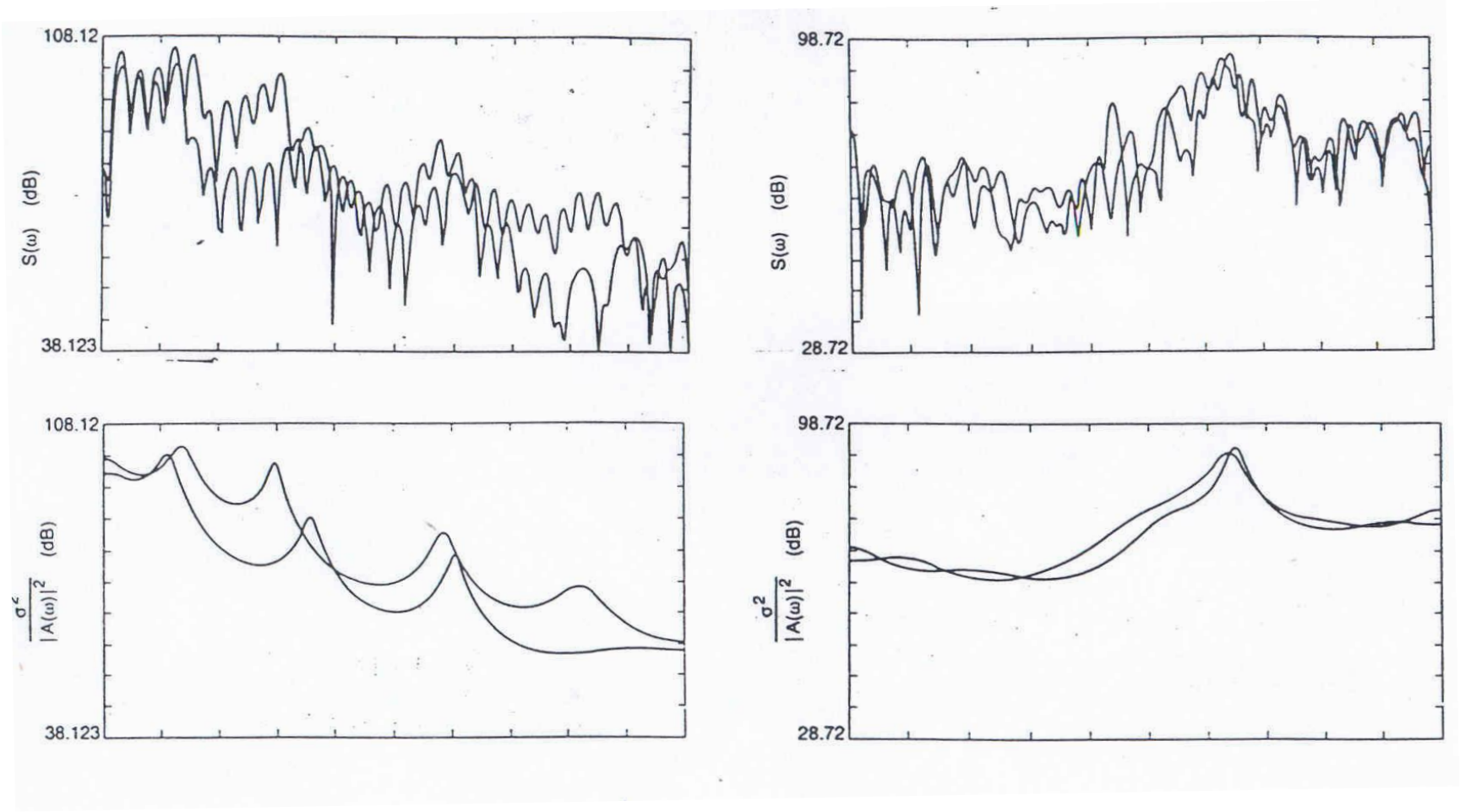
S(w) with LP spectral envelope superimposed



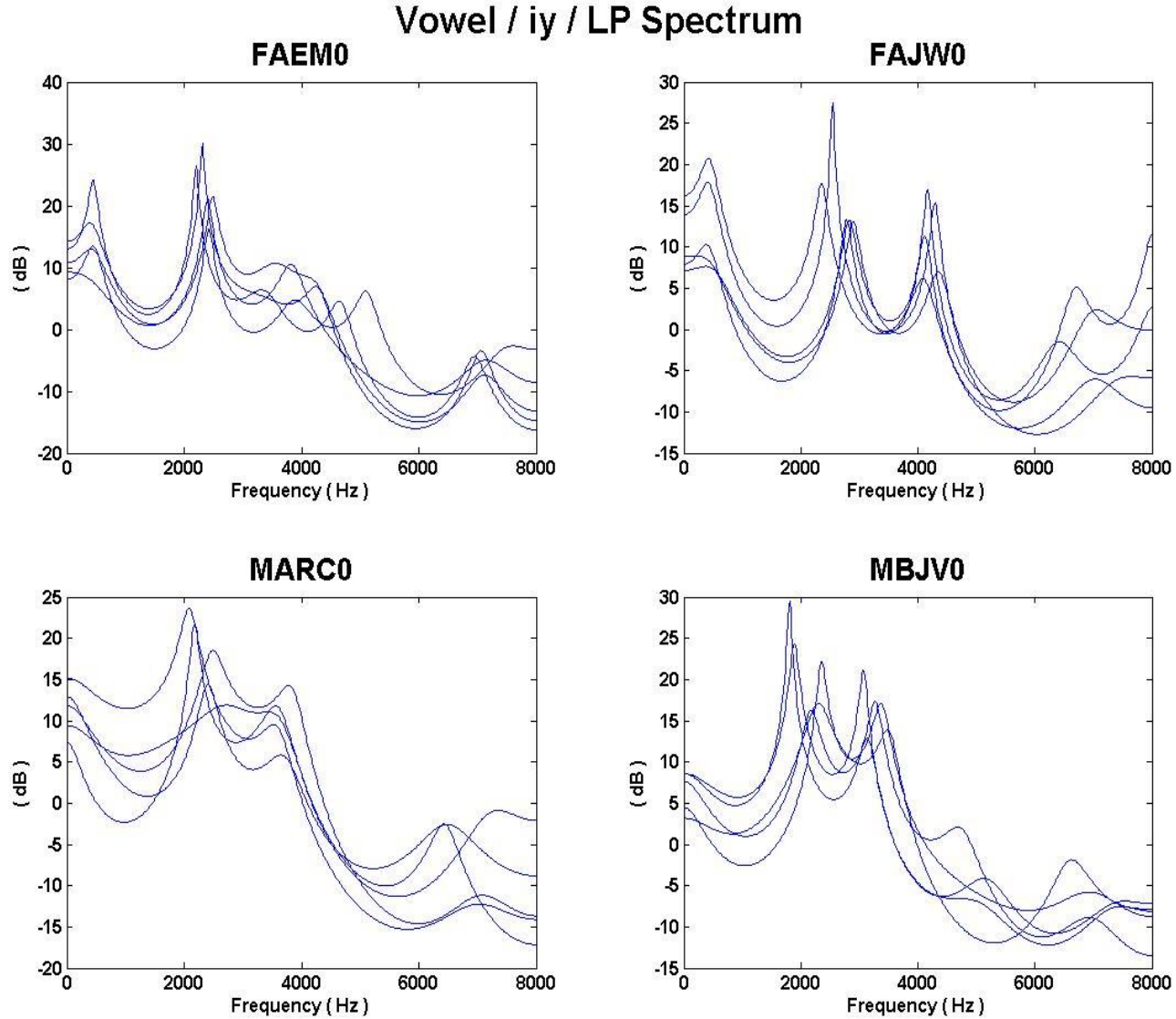
Excitation Signal E(n)



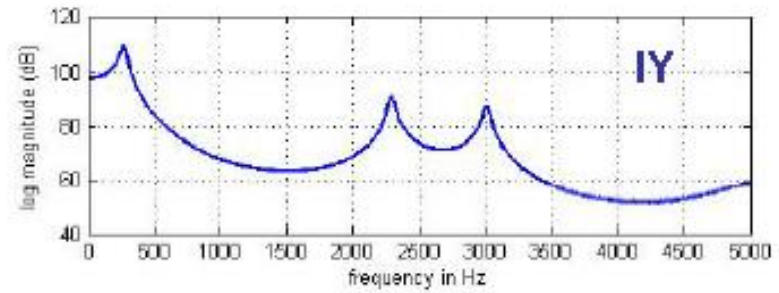
Spectral slices

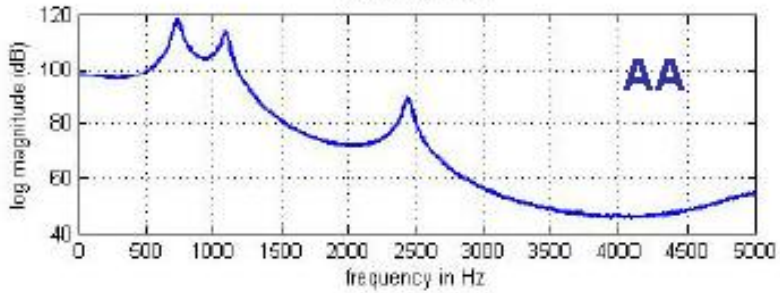


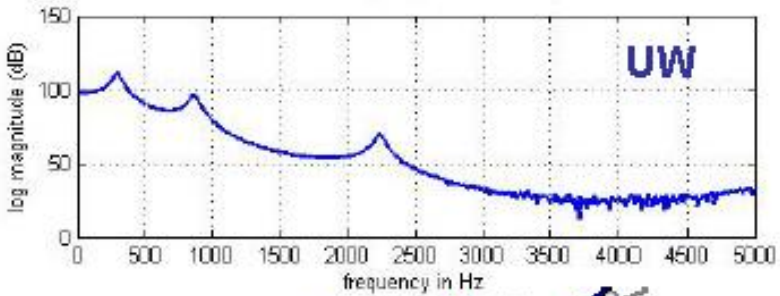
Spectral Envelopes

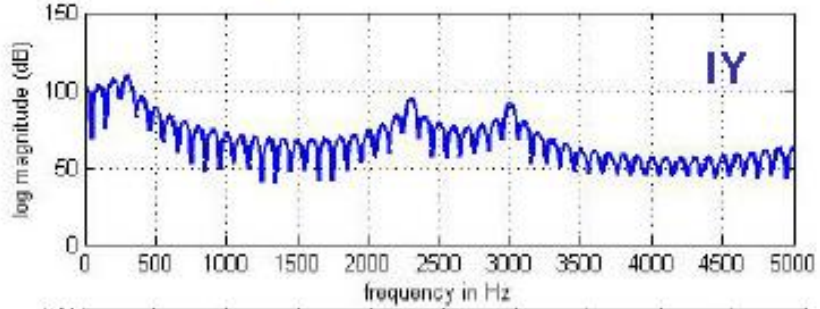


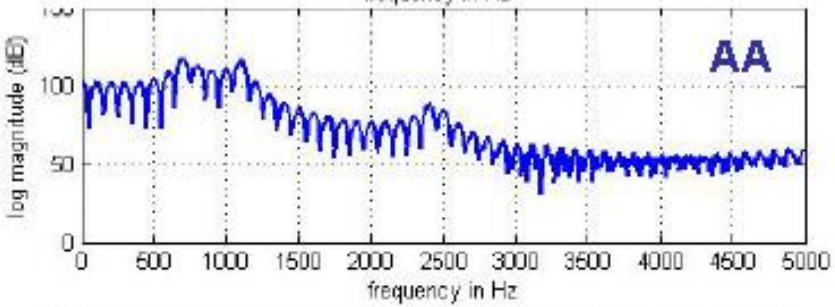
Canonic Vowel Spectra

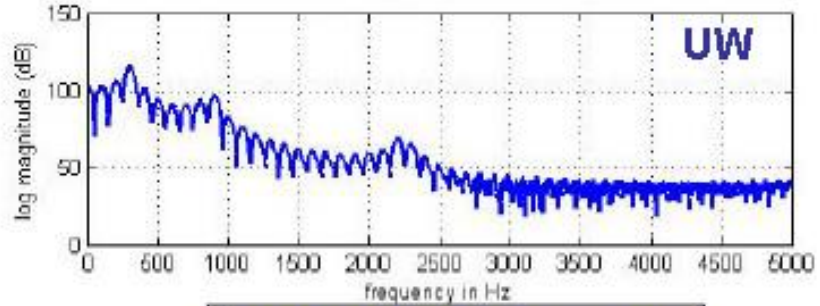








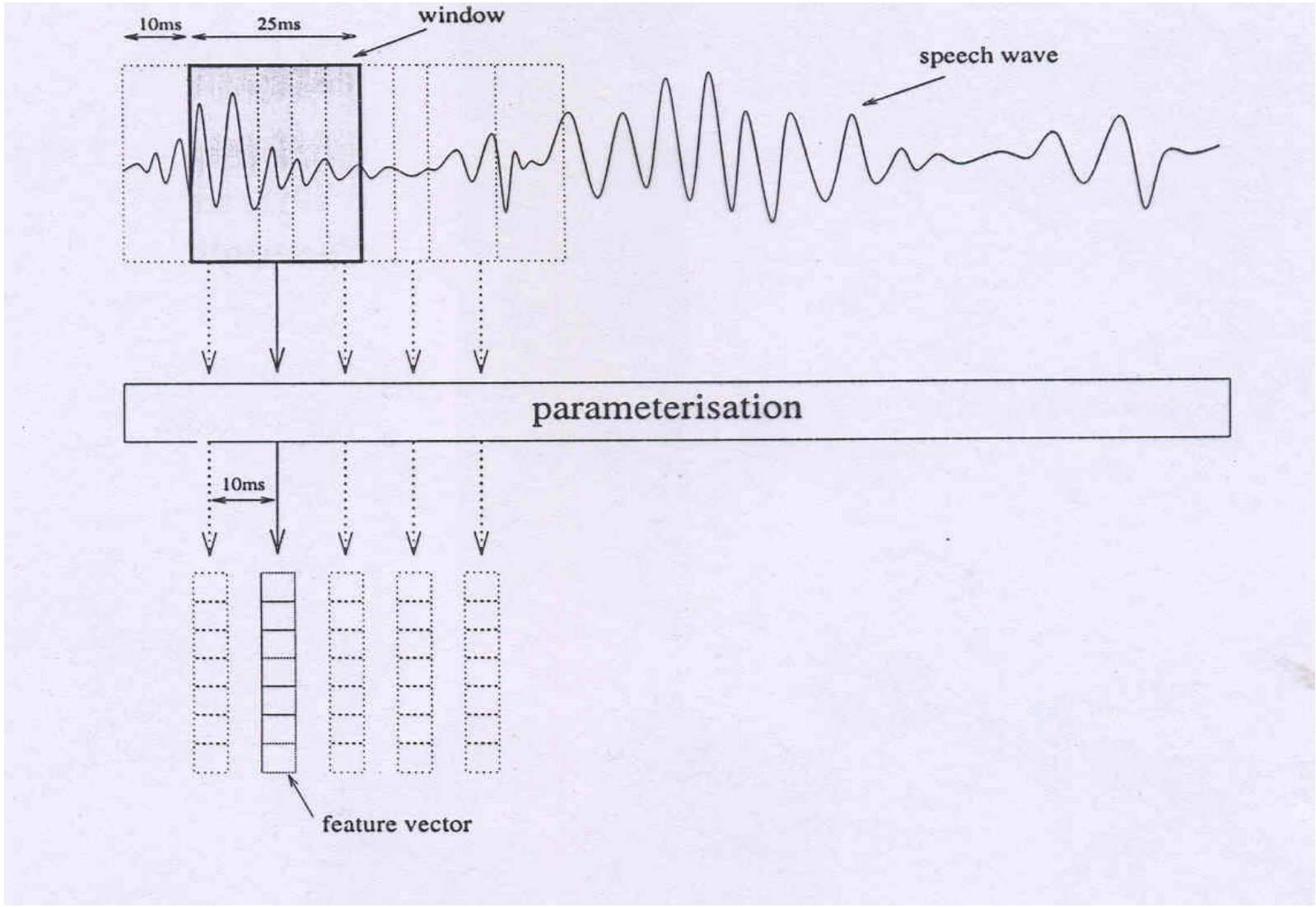




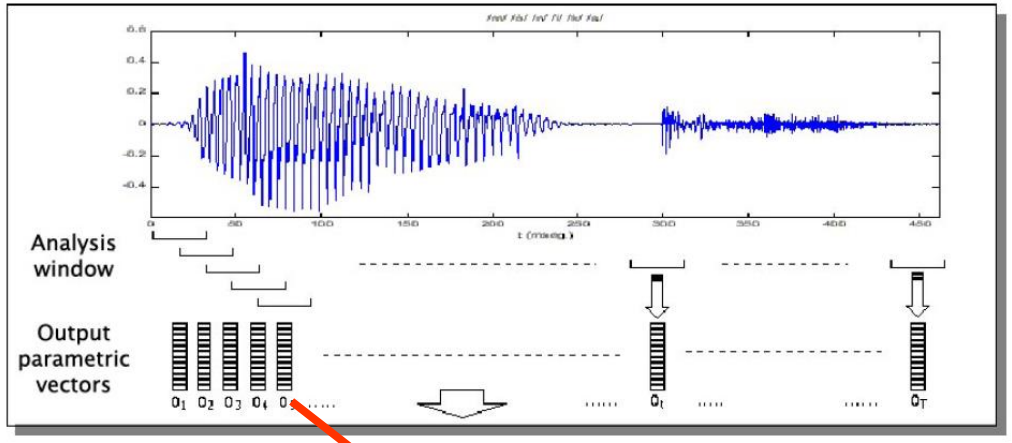
100 Hz Fundamental

54

Short-time Analysis and Parameterization

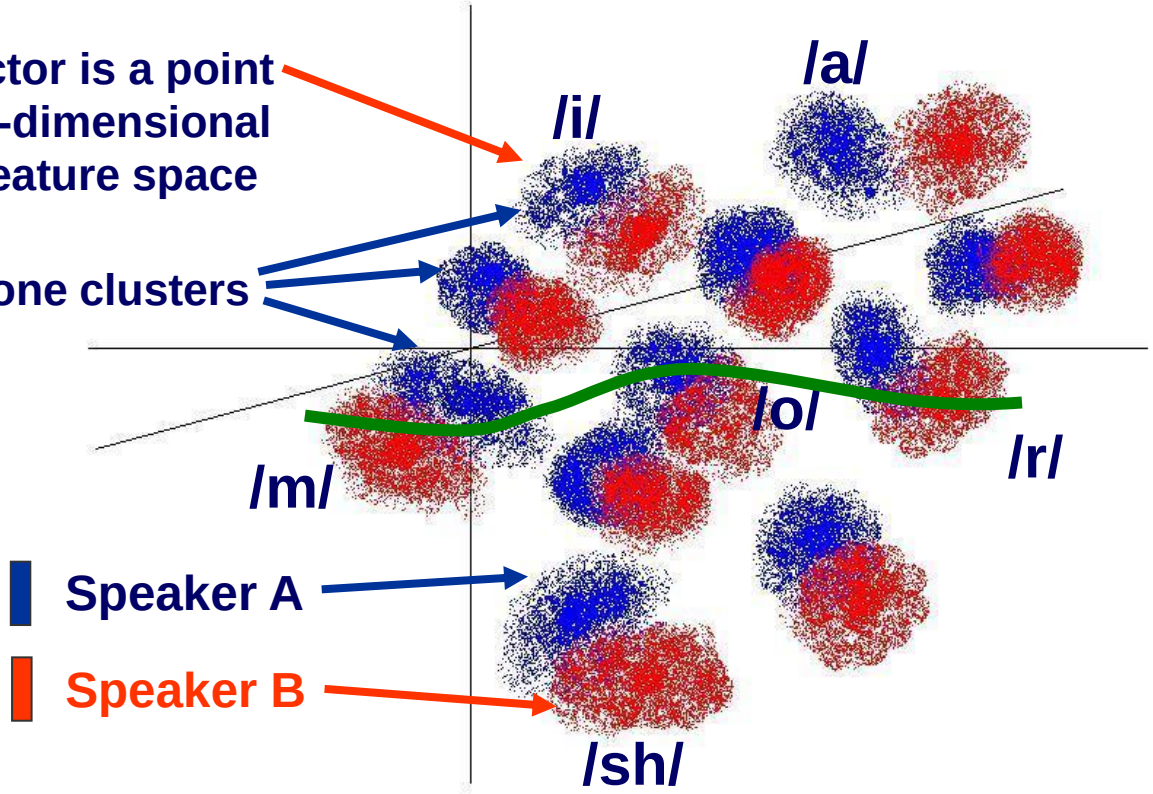


Feature Space

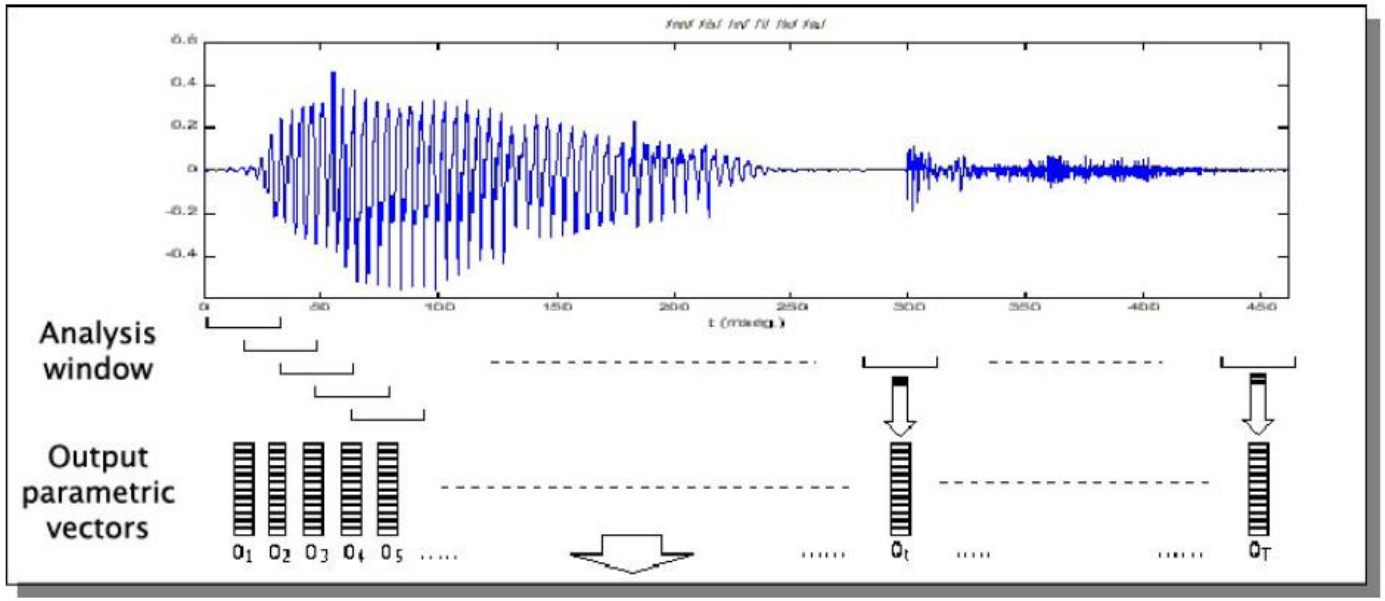


Bag of vectors representation of speaker's acoustic space

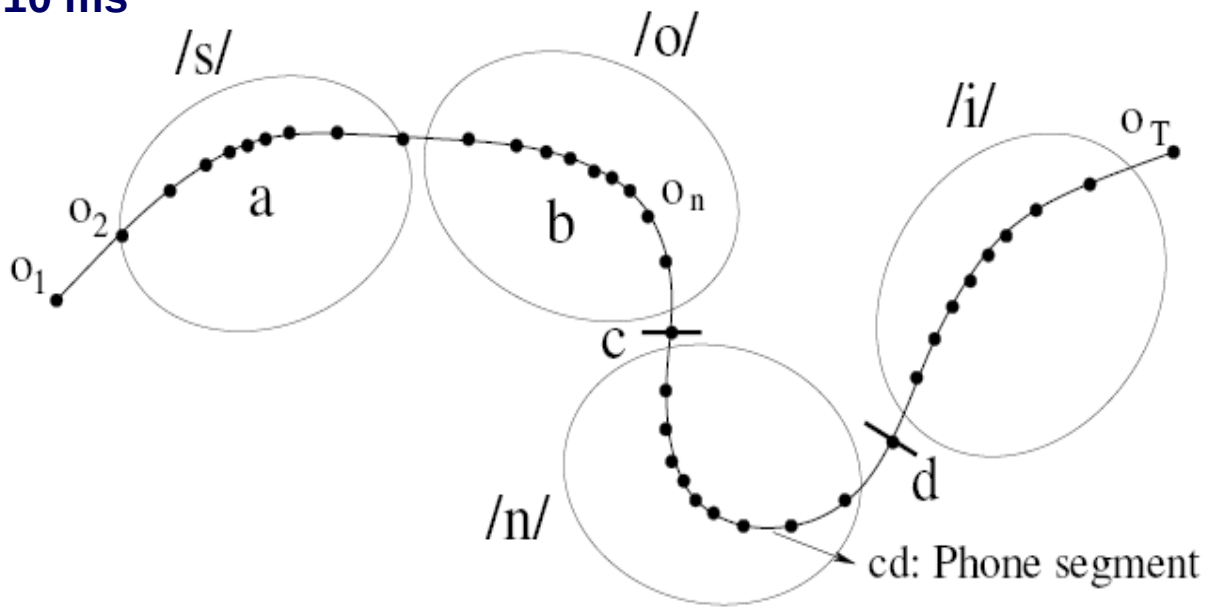
Each vector is a point in the 13-dimensional MFCC feature space



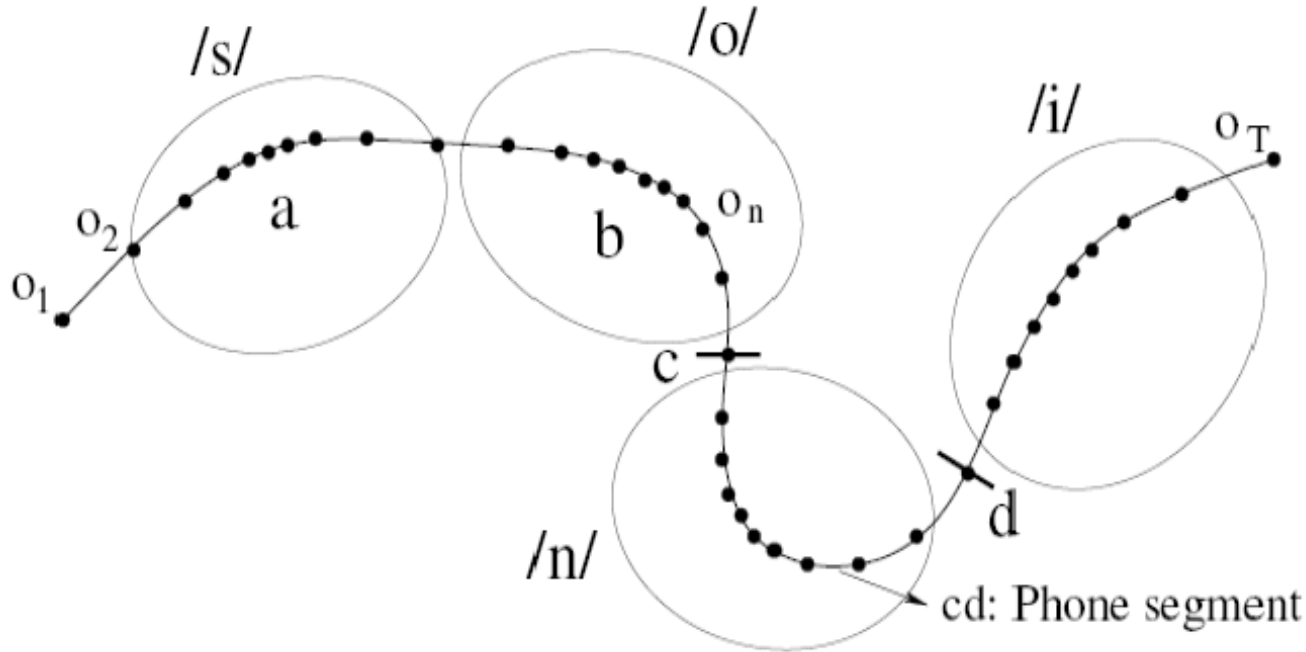
Feature Space



One feature vector every 10 ms



Feature Space



SPEECH RECOGNITION ALGORITHMS

- **TAKE THIS FEATURE VECTOR SEQUENCE**
- **AS INPUT AND DETERMINE "WHAT HAS BEEN SAID"**
- **e.g. SEQUENCE OF PHONES / SEQUENCE OF WORDS etc.**

Thank you !!

Spectrogram

- Speech is a continuous evolution of the vocal tract
- Spectrogram shows time-frequency evolution
- Represented as a time-series of short-time spectra