Introduction to

SPEECH PROCESSING

P. C. Pandey

EE Dept., IIT Bombay

References

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Pearson Education, 2004
2. B. Gold and N. Morgan, *Speech and Audio Signal Processing,* John Wiley, 2002
3. D. O'Shaughnessy, *Speech Communications: Human and Machine,* Universities Press, 2001

Speech communication

- Speech production
  - Propagation of the acoustic wave
    - Speech perception by the listener

Applications of speech processing

- Generation
- Transmission & storage (coding, speech enhancement)
- Reception (speaker & speech recognition, quality assessment)
- Aids for the disabled

Applications of speech processing

- **Generation (speech synthesis)**
  - Machine voice output
  - Reading aid for the blind
  - Testing of speech communication systems
  - Study of speech production & perception system
  - Speech production aids
- **Transmission / storage**
  - Coding
    - reduction in channel capacity requirements
    - encryption
  - Reduction in susceptibility to noise
  - Speech enhancement
  - Voice transformation: change of accent, speaker identity

- **Reception**
  - **Analysis for studying speech signal**
  - **Diagnostic aid for speech disorders**
  - **Speaker identification & quality assessment**
  - **Speech recognition**
  - **Aids for the hearing impaired**
    - **Variable rate playback**
    - **Visual / tactile displays**
    - **Effective use of residual hearing**
    - **Electrical stimulation of auditory nerve**

Speech units

- Words
  - Syllables
    - Phonemes
      - Sub-phonemic events

Speech transcription (scripts and alphabet systems)

- Pictographic
- Phonemic
- Alphabetic
- Syllabic

International Phonetic Alphabet (IPA)

अ /^/, आ /a/, क् /k/, ख् /k^h/, श् /ʃ/, etc.

Efficiency of phonemic transcription

• Measurement of information

Information: mean logarithmic probability (MLP).

For a set of messages $x_1, x_2 \ldots x_n$ **with probabilities** $p_1, p_{2 \ldots} p_n$**, information**

$$I = \mathrm{MLP}\left(x_i\right) = \sum_{i=1}^{n} p_i \log_2 \left(1/p_i\right) \text{ bits}$$

If the messages occur at the rate of M **messages/s, then the rate of information is**

$$C = M\,I \text{ bits/s}$$

• Phonemic transcription of speech

Consider speech generated from phonemic transcription (written text).
No of phonemes in English = 42.
Considering them equiprobable, $p = 1/42$

$$I = \sum_{i=1}^{42} p_i \log_2 \left(1/p_i\right) = 5.39 \text{ bits}$$

For conversational speech, with10 phonemes/s. $C = M\ I$ **= 53.9 bits/s.**

Considering actual phoneme probabilities, I **= 4.9 bits and** C **= 49 bits/s**

Considering relatedness of sequence of phonemes, C **≈ 45 bits/s**

• Channel capacity requirements

Channel capacity of an analog channel, with signal power S, noise power N, and bandwidth B Hz ,

$$C = B \log_2 \left(1 + \frac{S}{N} \right) \text{ bits/s}$$

For conventional analog telephony,

$B = 3$ kHz, SNR = 30 dB → S/N = 1000

$$C = 3 \times 10^3 \log_2 (1 + 1000) \approx 30 \text{ k bits/s}$$

Now consider good quality speech
Digitization (without any data compression techniques)

- 12 bit quantization ($2^{12} = 4096$ levels)
- Sampling rate = 10 k samples/s

Channel capacity required for digital transmission

$$C = (10\times10^3)\log_2 2^{12} = 120 \text{ k bits/s}$$

Thus we have three different estimates for channel capacity requirements for speech

- Analog telephony (3 kHz bandwidth, 30 dB SNR): 30 k bits/s
- Digital transmission (12-bit, 10 k samples/s, no encoding): 120 k bits/s
- Phonemic transcription: $\approx$45 bits/s

Auditory system

- **Peripheral auditory system**

    External ear : pinna and auditory canal (sound collection)

    → Middle ear : ear drum and middle ear bones (impedance matching)

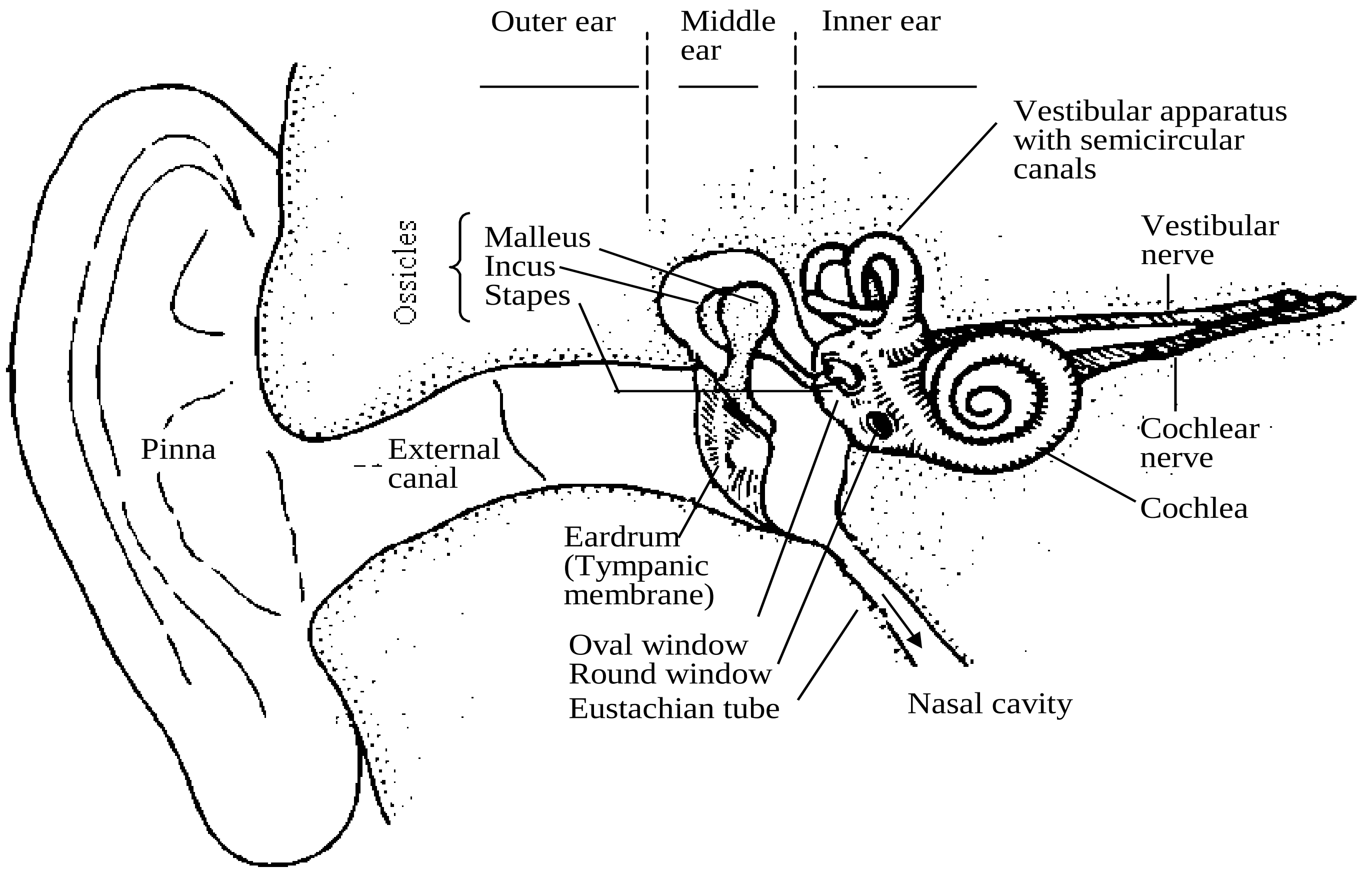
    → Inner ear : cochlea (analysis and transduction)

    → Auditory nerve (transmission of neural impulses)

- **Central auditory system (information interpretation)**

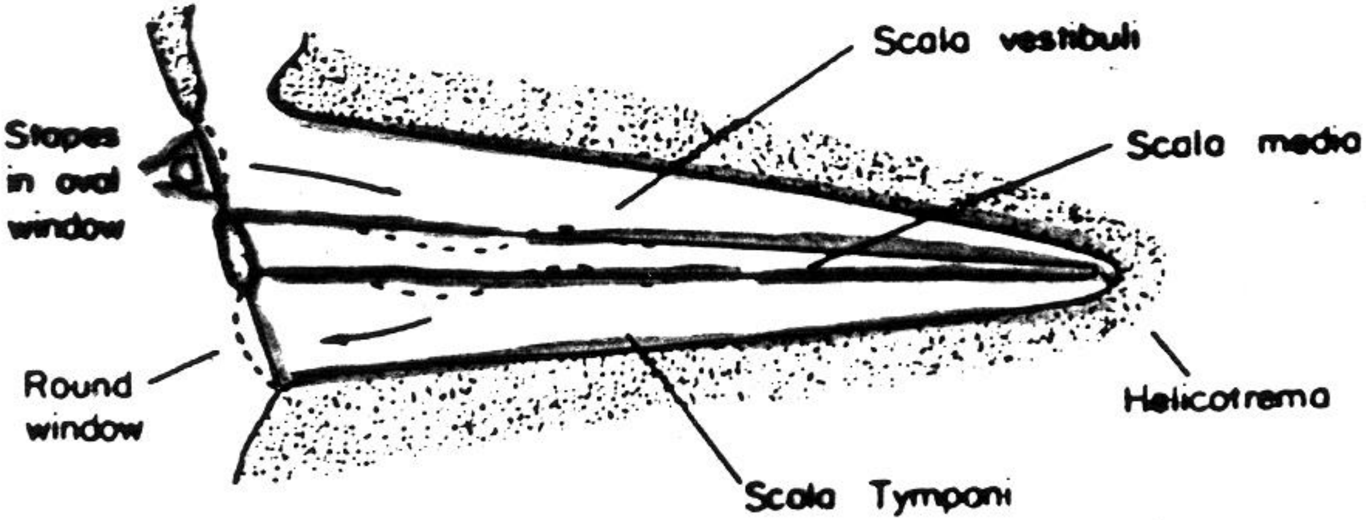
Peripheral auditory system (outer, middle, and inner ear).

(Adapted from Flanagan (1972a), Fig. 4.1.)

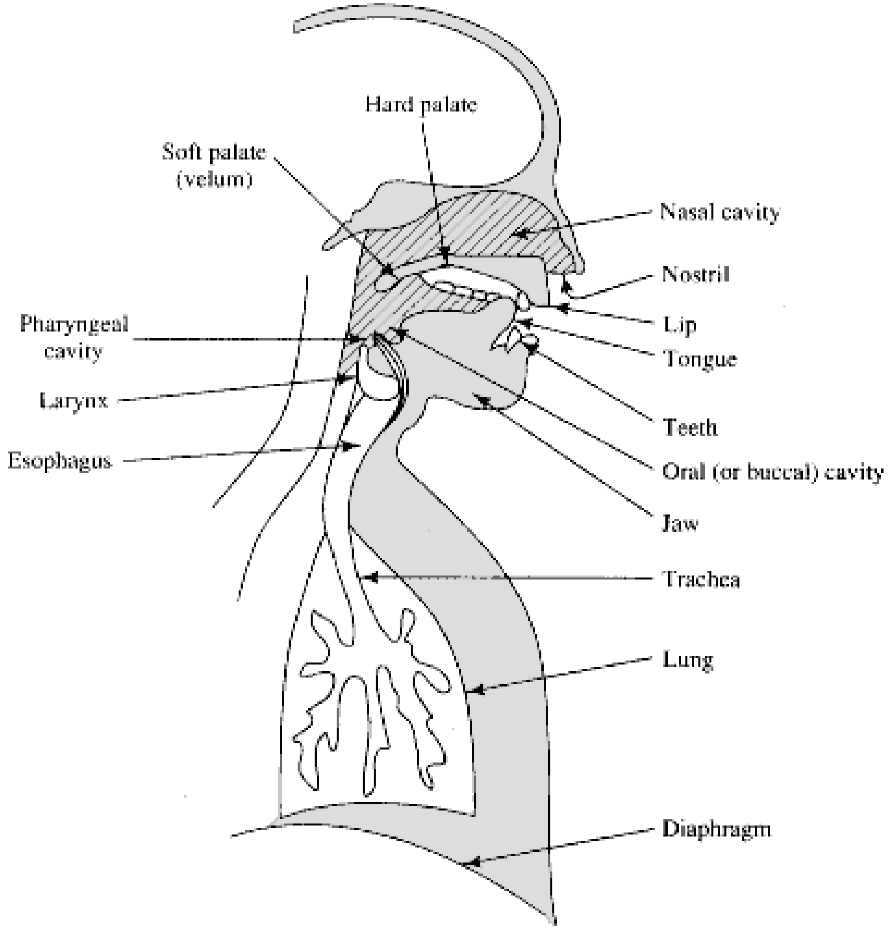


A longitudinal section through the uncoiled cochlea

(Source: Pickles (1982), Fig. 3.1.(C))

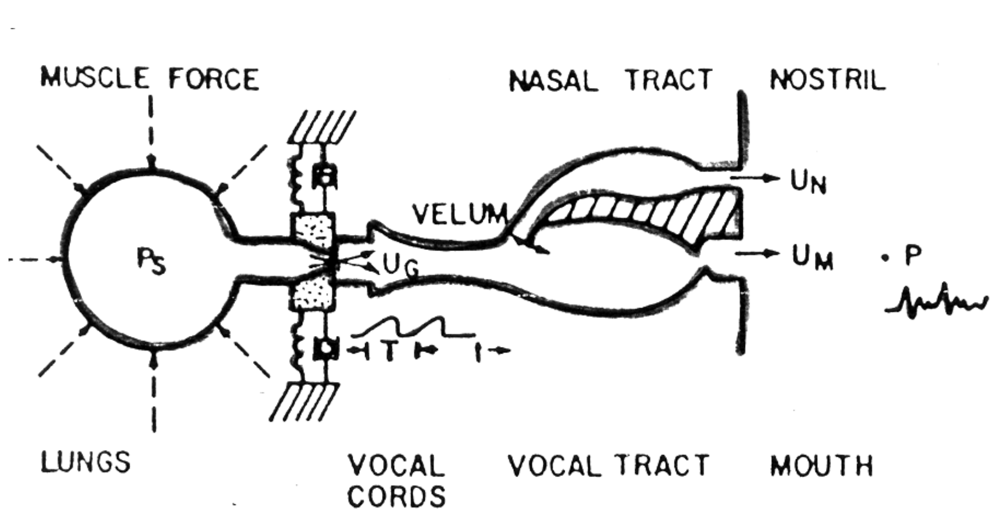


The mechanism of speech production



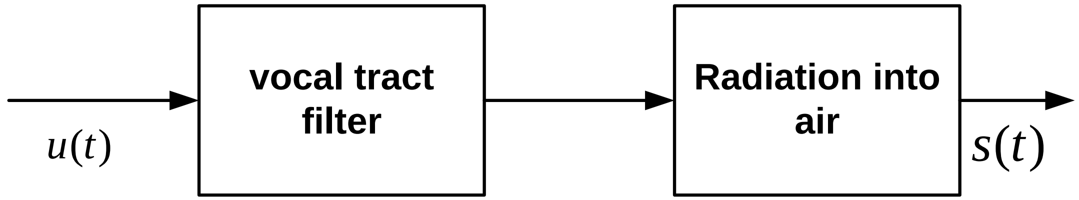
A schematic diagram of the human vocal mechanism

A schematic representation of the vocal apparatus. (Adapted from Rabiner & Schafer (1978), Fig. 3.2)



Pitch : F_0 (rate of vibration of vocal cords)

Formant frequencies: F_1, F_2, .. (resonance frequencies of vocal tract filter)



Phonemic features

- **Modes of excitation**
  - Glottal: ○ voiced ○ unvoiced (aspiration)
  - Frication (constriction in vocal tract) : ○ voiced ○ unvoiced
- **Movement of articulators**
  - Continuant (steady vocal tract): vowels, nasal stops, fricatives
  - Non-continuant (changing vocal tract): diphthongs, semivowels, plosives (oral stops)
- **Place of articulation**
  - bilabial, labio-dental, linguo-dental, alveolar, palatal, velar, gluttoral
- **Changes in F_o**

Classification of English phonemes

Classification of English Phonemes

Vowels	i	I	e	ae		(front)
	a	A	ow			(mid)
	u	U	o			(back)
Diphthongs	aI	oI	aU			
	eI	oU	ju			
Semi-vowels	r	l				(liquid)
	w	y				(glide)
Consonants	m	n	ŋ			nasal stops
	b	d	g			voiced plosives
	p	t	k			? unvoiced plosives
	v	ð	z	ʒ		voiced fricative
	f	θ	s	ʃ	h	unvoiced fricative
			dʒ (ज)			voiced affricate
			tʃ (च)			unvoiced affricate
		h				whisper

Classification of Hindi phonemes

Classification of Hindi Phonemes

short vowels: इ ए अ ओ उ
long vowels: ई आ ऊ

Diphthongs: ऐ औ
(front) (back)

semivowels: व् ल् र् य्

unvoiced plosives	प्	त्	ट्	क्	(unaspirated)
	फ्	थ्	ठ्	ख्	(aspirated)
	ब्	द्	ड्	ग्	(unaspirated)
	भ्	ध्	ढ्	घ्	(aspirated)

nasals म् न् ण् ञ् ङ्

affricates
च् छ् (unvoiced)
ज् झ् (voiced)

fricatives स् ष् श् ह् (unvoiced)

क ख ग घ ङ
च छ ज झ ञ
ट ठ ड ढ ण
त थ द ध न
प फ ब भ म

Suprasegmental features

- **Intonation**
- **Rhythm (syllable stressing)**

  Carriers

  - **Changes in intensity**
  - **Syllable duration**
  - **Changes in voice pitch**
- **Visual features**

  Partial information on place of articulation

Short-time speech analysis

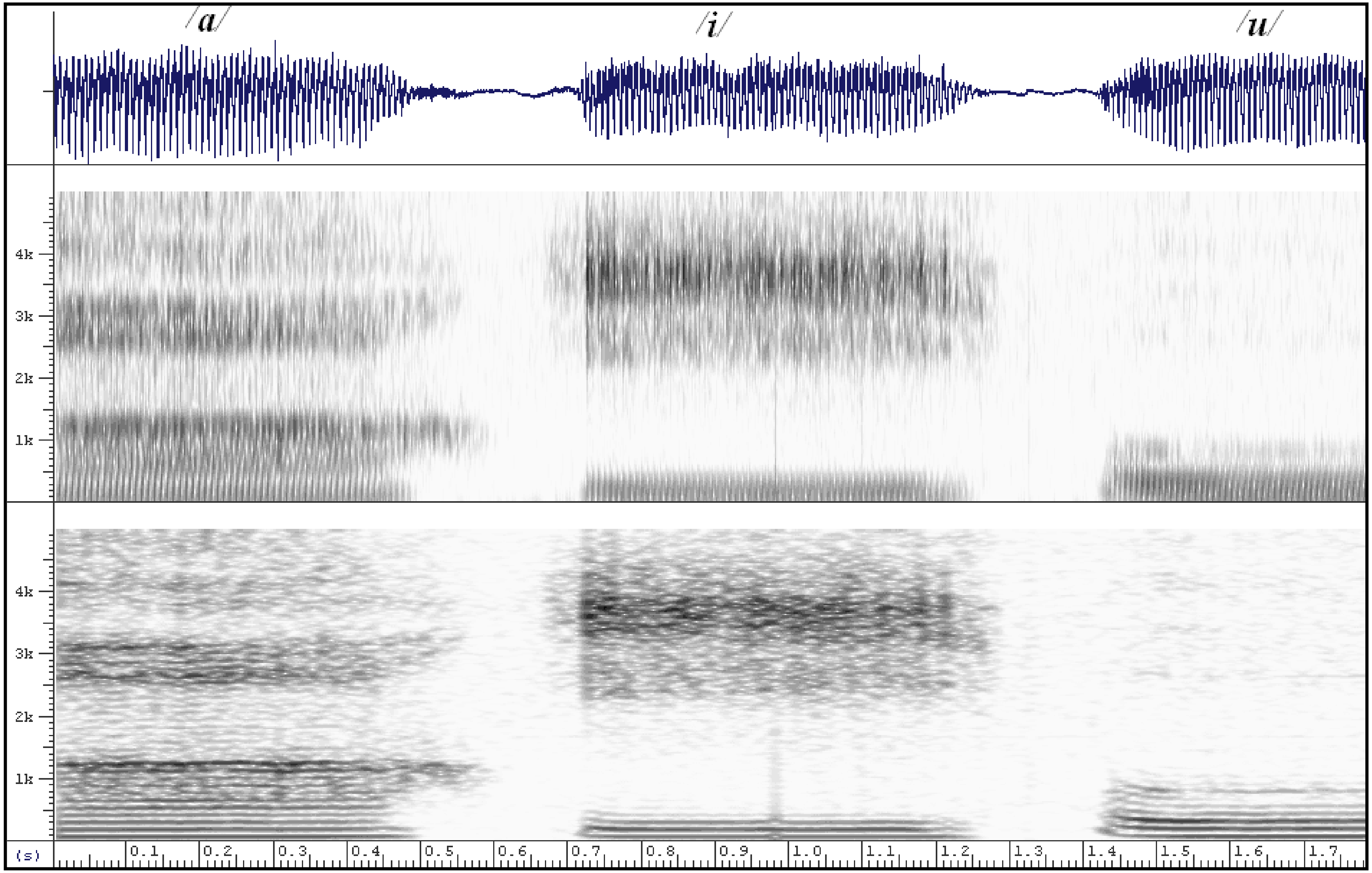
$$\begin{tikzcd} s(n) \arrow[r, "\textit{windowing}"'] & s(n)w(n-m) \arrow[d] \\ \text{Storage/display/transmission} & Q_k(n) \arrow[l, "\textit{decimation}"'] \end{tikzcd}$$

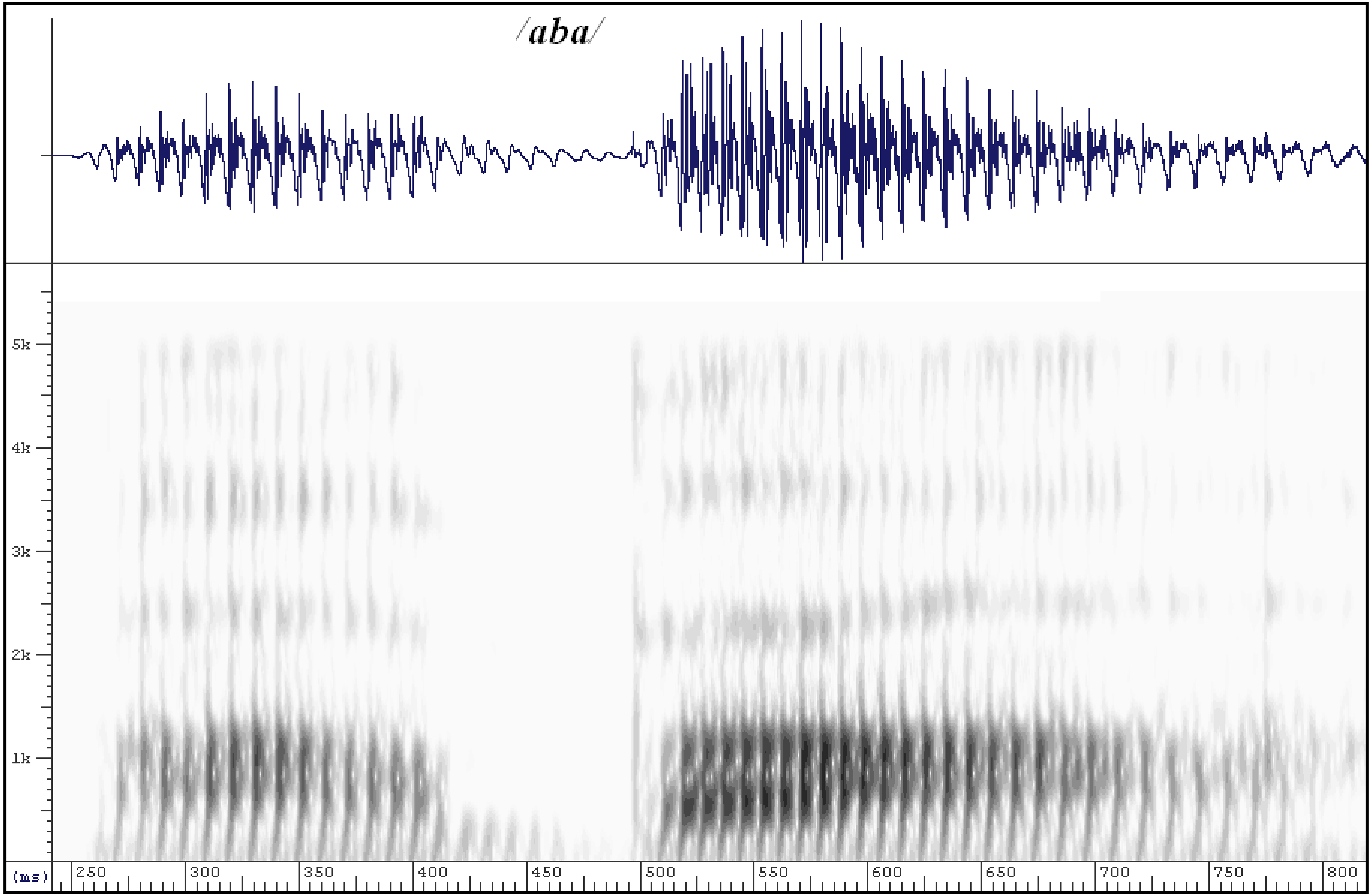
Speech analysis techniques

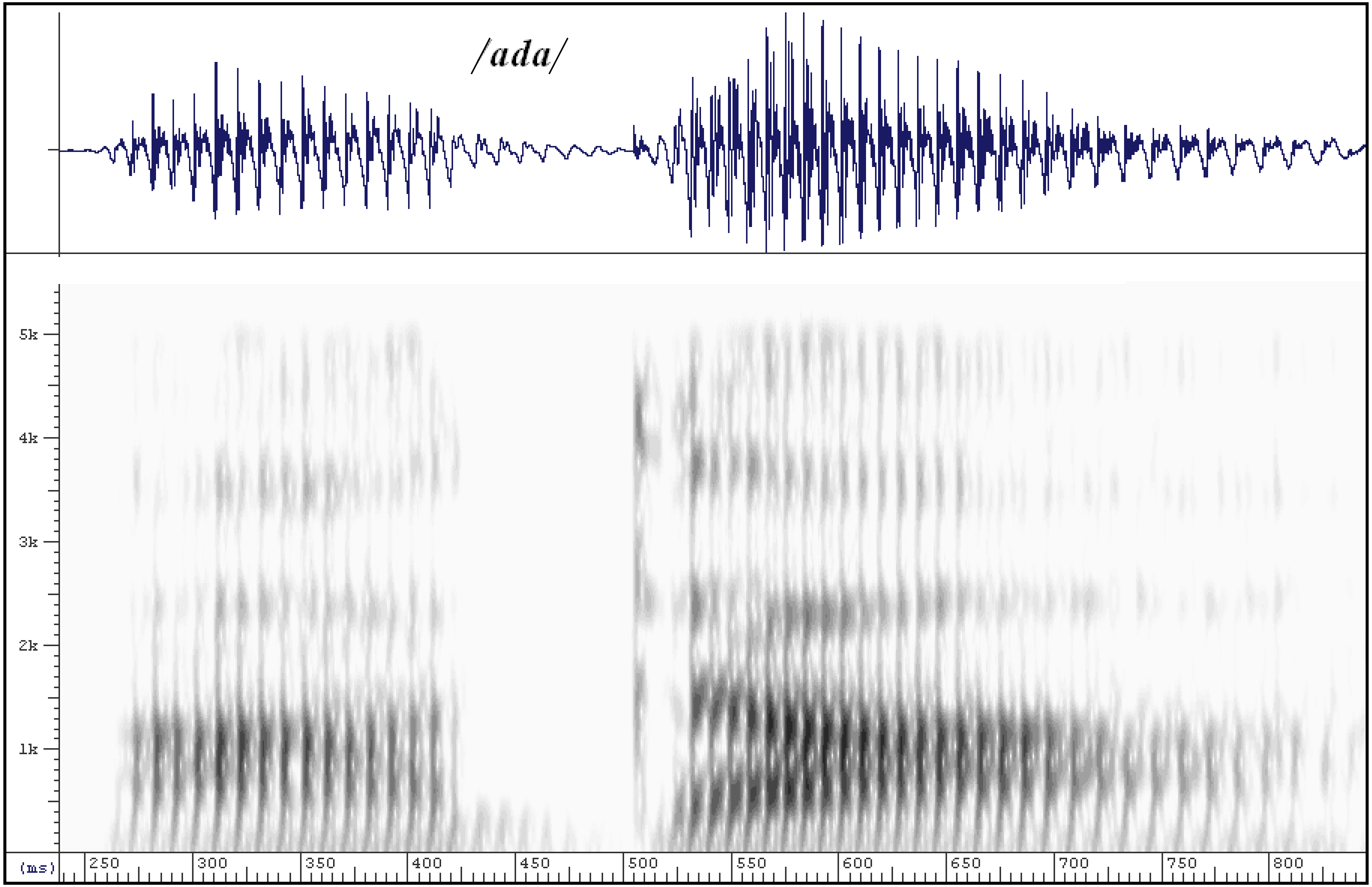
- **Time domain analysis**
  - Intensity
  - Zero crossing rate
  - Autocorrelation
  - AMDF

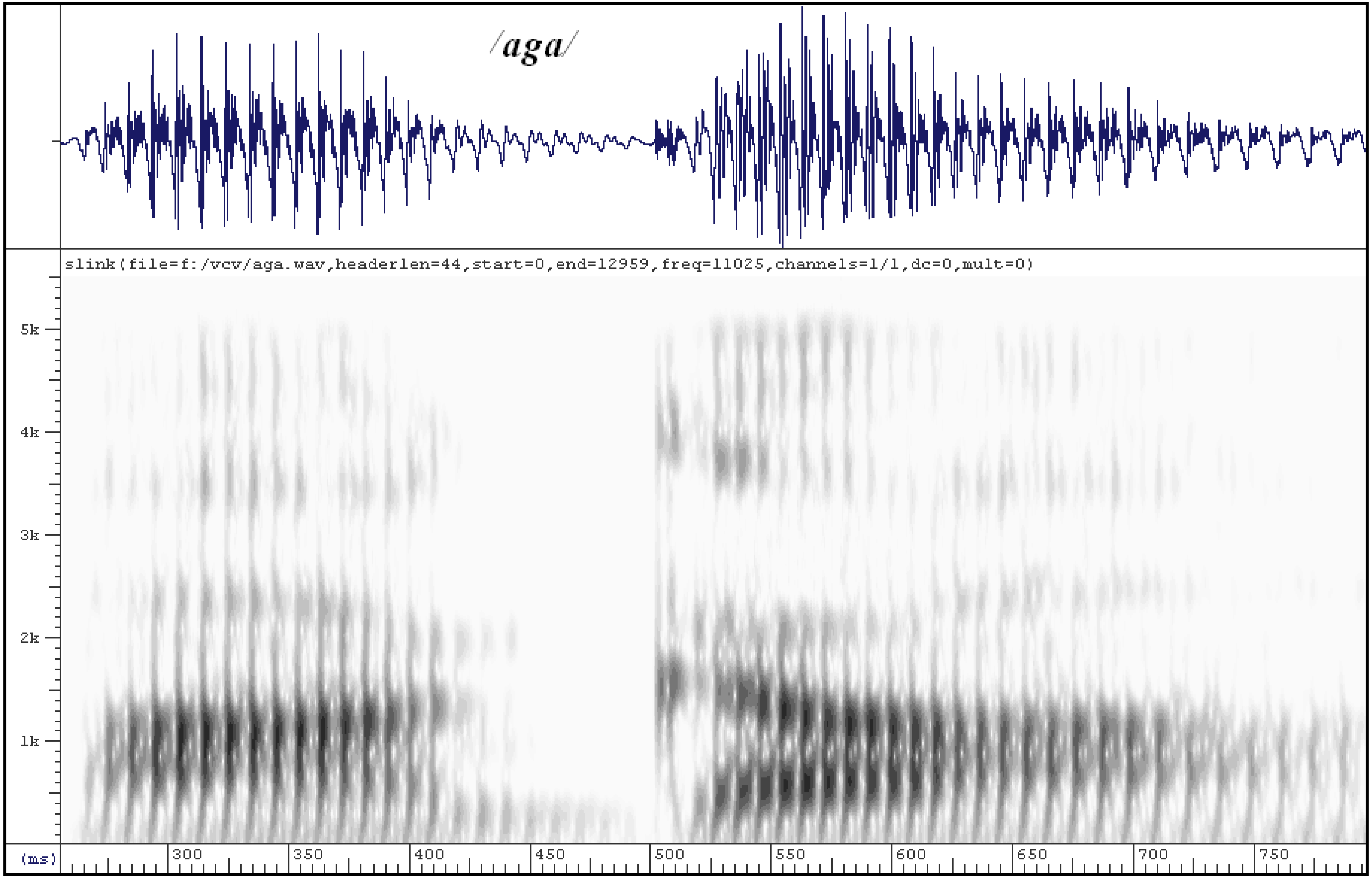
- **Frequency domain analysis**
  - Filter bank analysis
  - Short-time Fourier analysis
  - Spectrographic displays
  - Formant estimation & tracking
  - F_0 estimation and tracking

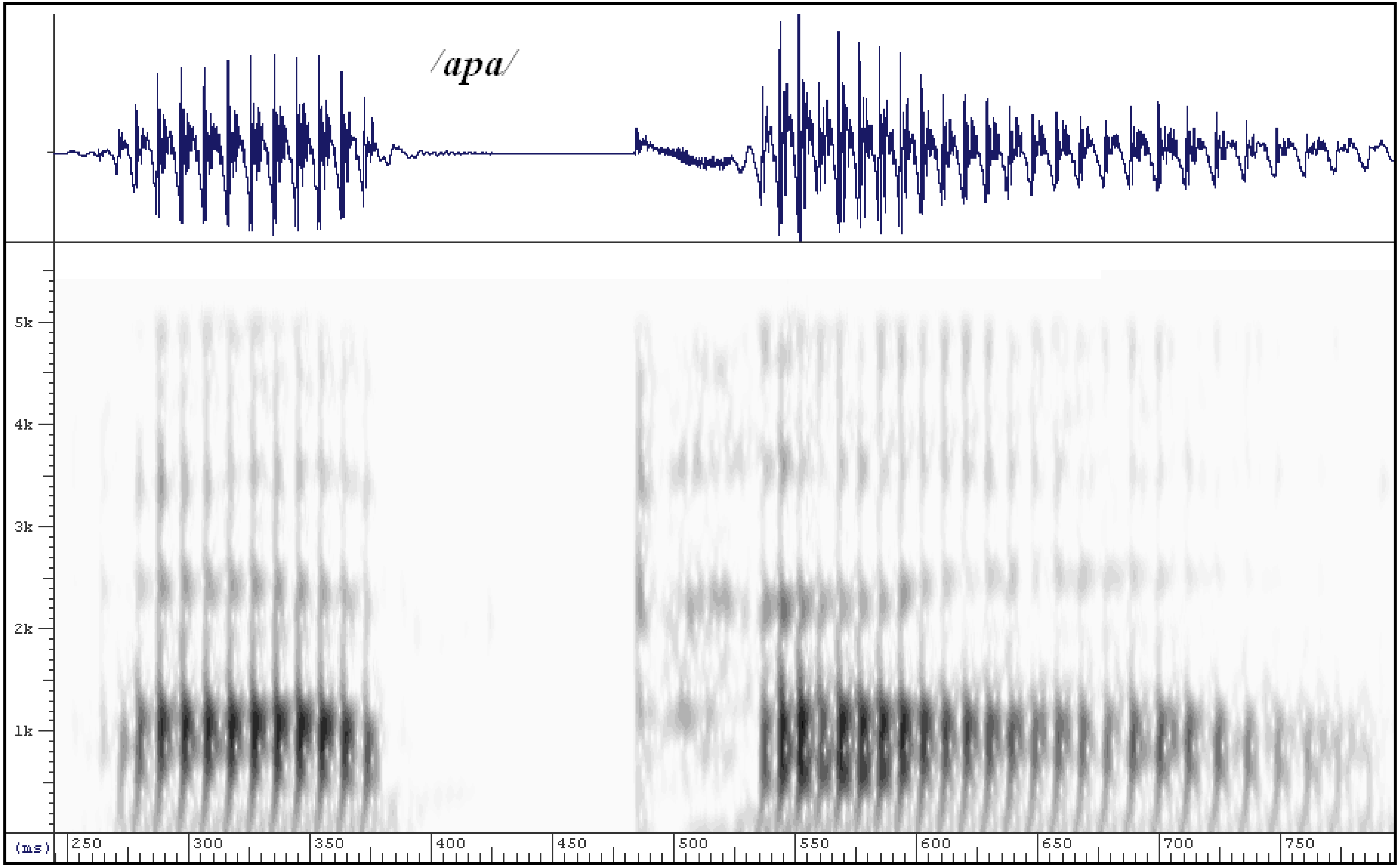
- **Separation of excitation and vocal tract filter effects**
  - Cepstral analysis
  - Linear predictive coding (LPC)

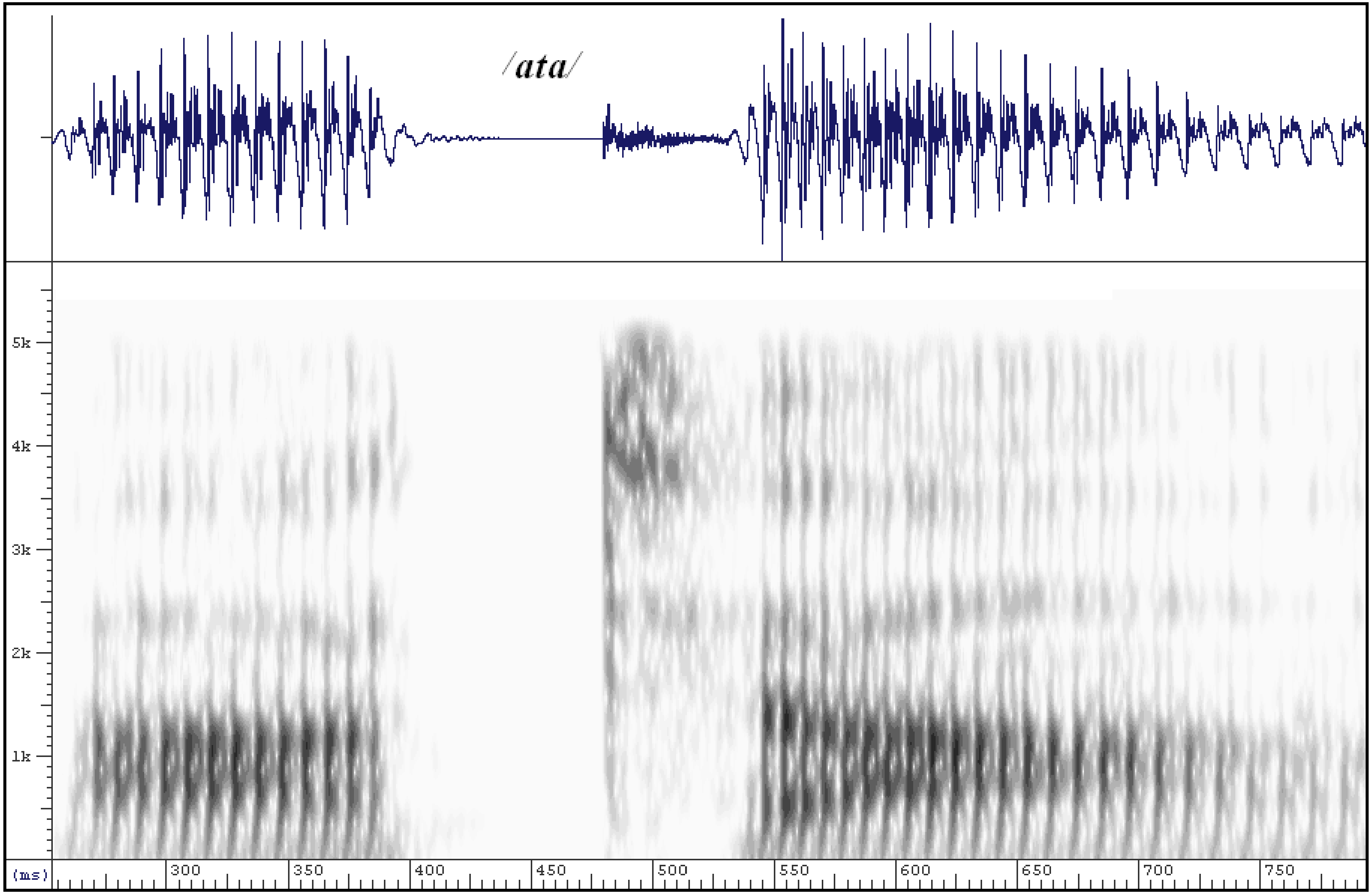

/a/
/i/
/u/
4k
3k
2k
1k
4k
3k
2k
1k
(s)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7

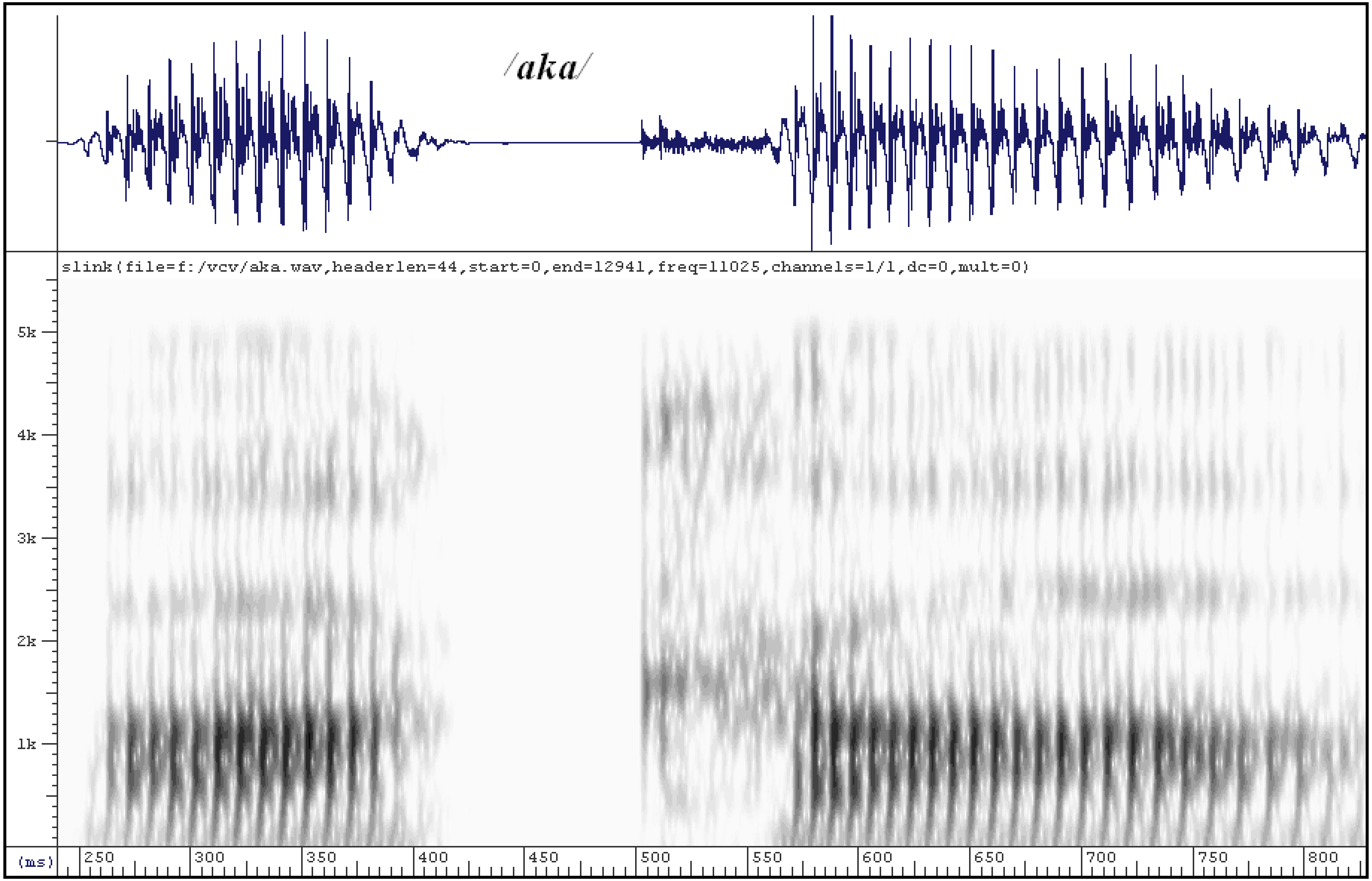

/aba/
5k
4k
3k
2k
1k
(ms)
250
300
350
400
450
500
550
600
650
700
750
800


/ada/
5k
4k
3k
2k
1k
(ms)
250
300
350
400
450
500
550
600
650
700
750
800


/aga/
slink(file=f:/vcv/aga.wav,headerlen=44,start=0,end=12959,freq=11025,channels=1/1,dc=0,mult=0)
5k
4k
3k
2k
1k
(ms)
300
350
400
450
500
550
600
650
700
750


/apa/
5k
4k
3k
2k
1k
(ms)
250
300
350
400
450
500
550
600
650
700
750


/ata/
5k
4k
3k
2k
1k
(ms)
300
350
400
450
500
550
600
650
700
750


/aka/
slink(file=f:/vcv/aka.wav,headerlen=44,start=0,end=12941,freq=11025,channels=1/1,dc=0,mult=0)
5k
4k
3k
2k
1k
(ms)
250
300
350
400
450
500
550
600
650
700
750
800

Multi-resolution spectrographic analysis

Analog analysis **Spectral analysis using bandpass filter & demodulator**

$$s(t) \longrightarrow \left| S_t(f) \right|_{dB}$$

Δf **= 300 Hz (wideband), 45 Hz (narrow band)**

Digital analysis

Short-time Fourier transform: $X(n,k) = \sum_{m=0}^{N-1} w(m)x(n-m)e^{-j2\pi\frac{mk}{N}}$

Hamming window : $w(n) = 0.54 - 0.46\cos\left(2\pi \frac{nt\frac{L}{2}}{L-1} \right), -\frac{L}{2} \le n \le \frac{L}{2}$

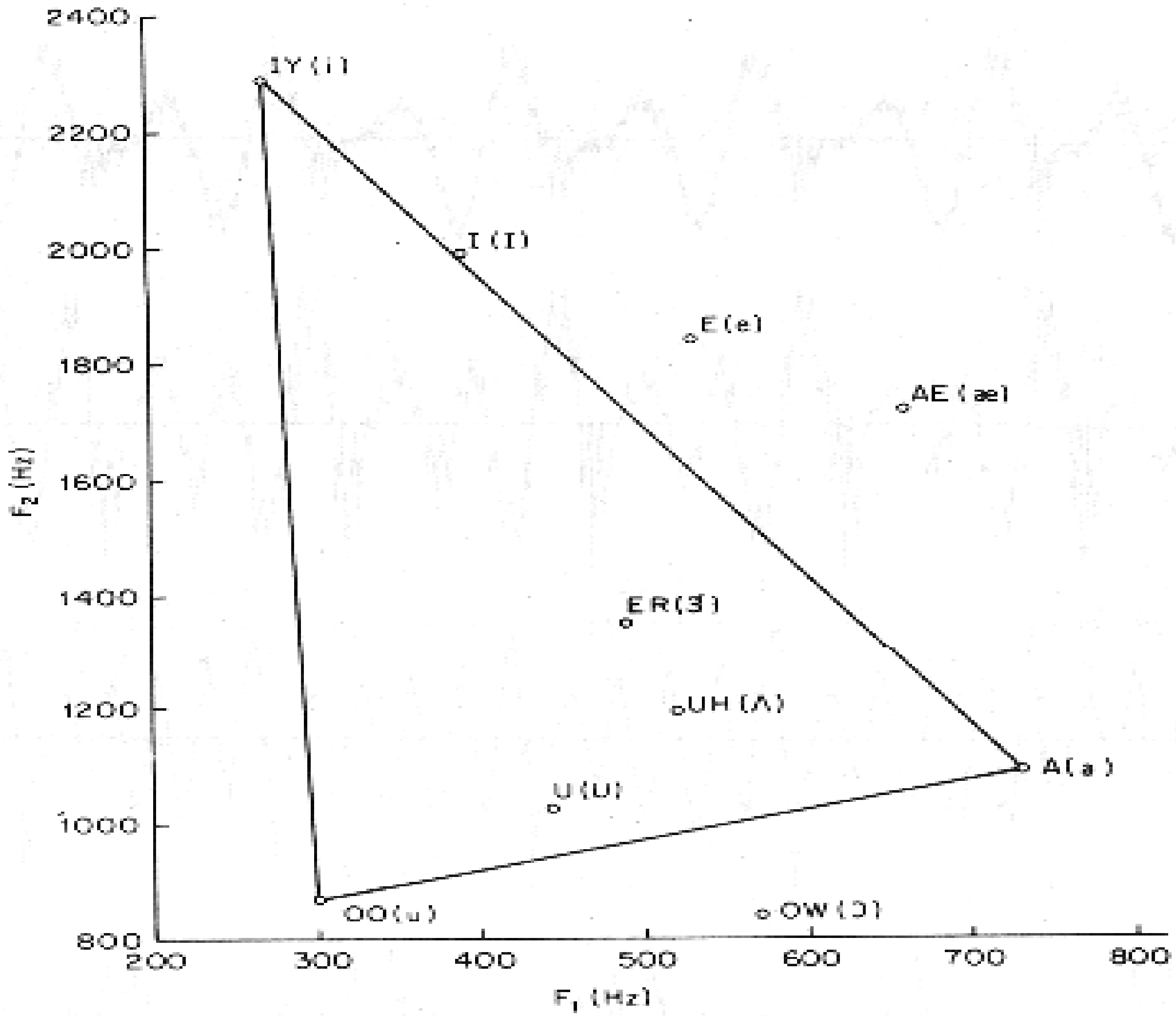
$$S_{dB}(n,k) = 20\log\left(\left| X(n,k) \right| / X_{ref} \right)$$

For $f_s = 10$ kSa/s, $L = 43 \rightarrow \Delta f \approx 300$ Hz

$L = 289 \rightarrow \Delta f = 45$ Hz

• Change in resolution, implemented by padding *L*-sample segment with *N-L* zero valued samples, and using N-point DFT. Pre-emphasis for boosting high-frequency components.

Vowel triangle



The vowel triangle.

Spectrograms for English phonemes
(vowels or /aCa/ syllable)

Manner	Voicing	Place of articulation		
		Back	Mid	Front
Oral stop	unvoiced	k	t	p
	voiced	g	d	b
Nasal stop	unvoiced	×	×	×
	voiced	ng	n	m
Oral fricative	unvoiced	(sh)	s	f
	voiced	(zh)	z	v
Nasal fricative	unvoiced	×	×	×
	voiced	×	×	×

Cepstral analysis

$$e(n) \longrightarrow \boxed{\;v(n)\;} \longrightarrow s(n) = e(n) * v(n)$$

$$S(z) = \underbrace{E(z)}_{\text{Excitation}}\underbrace{V(z)}_{\text{Vocal tract filter response}}$$

We can use "log" transformation for deconvolution and separation of excitation & vocal tract filter parameters.

$$x(n) \xrightarrow[Z]{} X(z) \xrightarrow[\log]{} \hat{X}(z) \xrightarrow[Z^{-1}]{} \hat{x}(n)$$

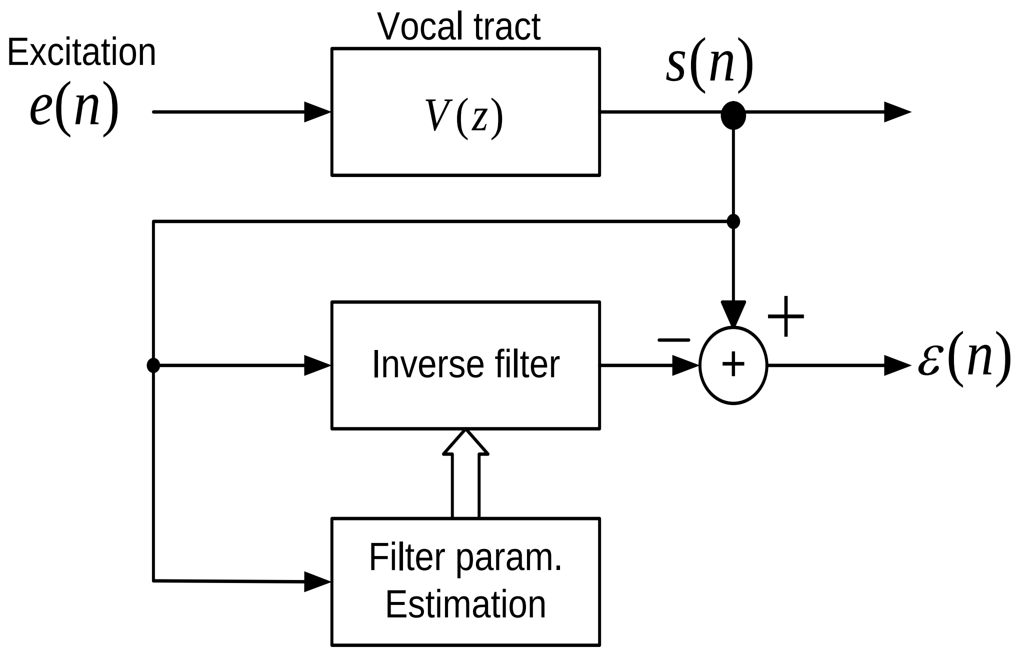
$$\hat{x}(n) = Z^{-1}\left[\log\left\{X(z)\right\}\right]$$

$$\log\left[S(z)\right] = \log E(z) + \log V(z)$$

$$\downarrow Z^{-1}$$

$$\hat{s}(n) = \hat{e}(n) + \hat{v}(n)$$

Linear predictive coding (LPC)

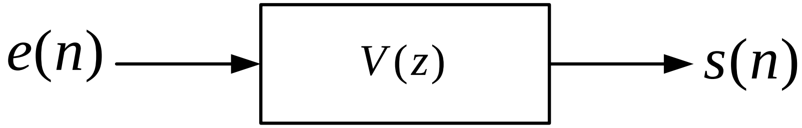


- **Excitation $e(n)$: modeled as an impulse or random noise.**
- **Vocal tract filter transfer function $V(z)$: modeled as an all-pole filter (AR model).**
- **Inverse filter $A(z)$: all-zero filter (FIR or MA filter), with parameters estimated for minimizing $\varepsilon(n)$ in LMS sense.**

Result

- **Error $\varepsilon(n)$ acquires the characteristics of the excitation function**
- **Inverse filter coefficients, obtained by solving a set of linear equations, approximate the vocal tract filter coefficients.**

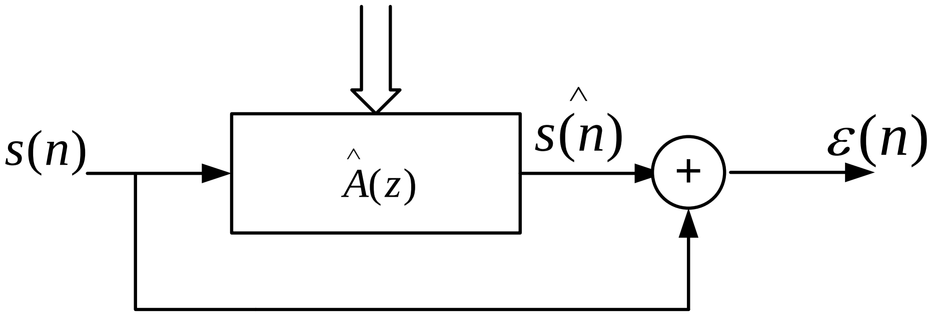
Speech production



$$V(z) = \frac{1}{1 - A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

Linear prediction model



$$\hat{s}(n) = \sum_{k=1}^{p} \hat{a}_k \, s(n-k)$$

$$\varepsilon(n) = s(n) - \hat{s}(n)$$

$$= s(n) - \sum_{k=1}^{p} \hat{a}_k \, s(n-k)$$