Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Επεξεργασία Ομιλίας και Φυσικής Γλώσσας

Ενότητα 4: **Ψηφιακή επεξεργασία ομιλίας στο χρονικό και φασματικό πεδίο**

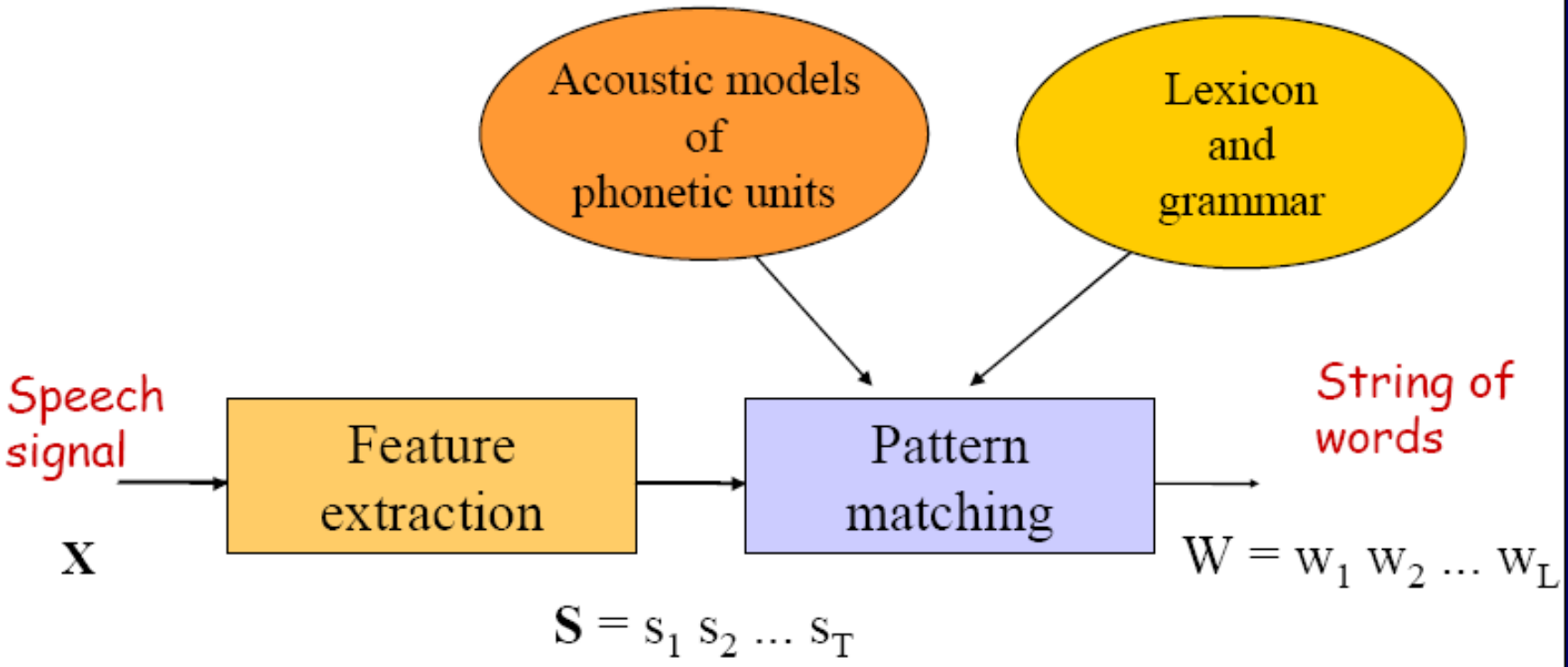
Γεώργιος Κουρουπέτρογλου
koupe@di.uoa.gr

Ανάλυση Ομιλίας

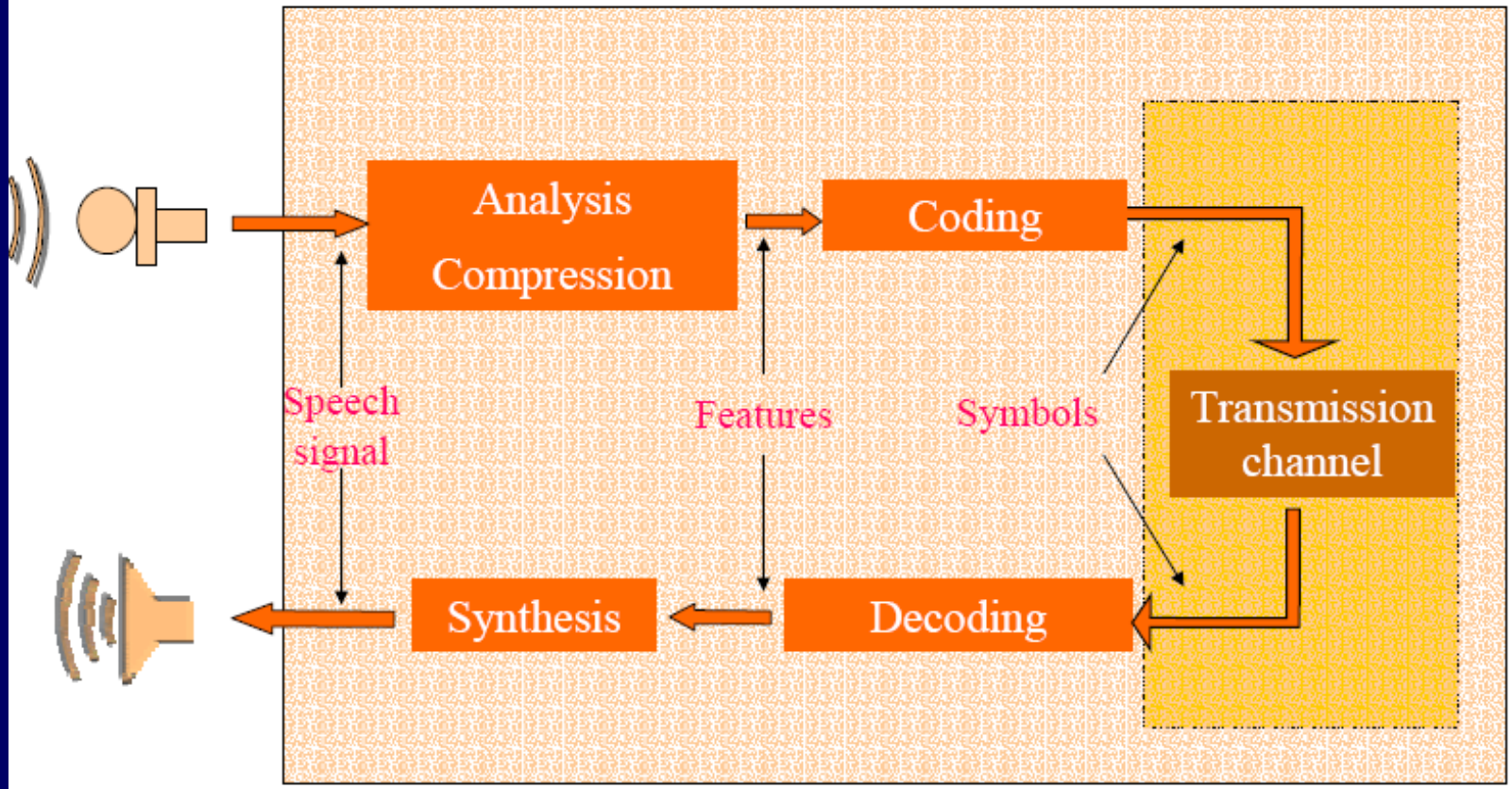
Για αποτελεσματική:

- Σύνθεση ομιλίας
- Αναγνώριση ομιλίας
- Συμπίεση & ψηφιακή μετάδοση ομιλίας

Need of speech analysis in speech recognition

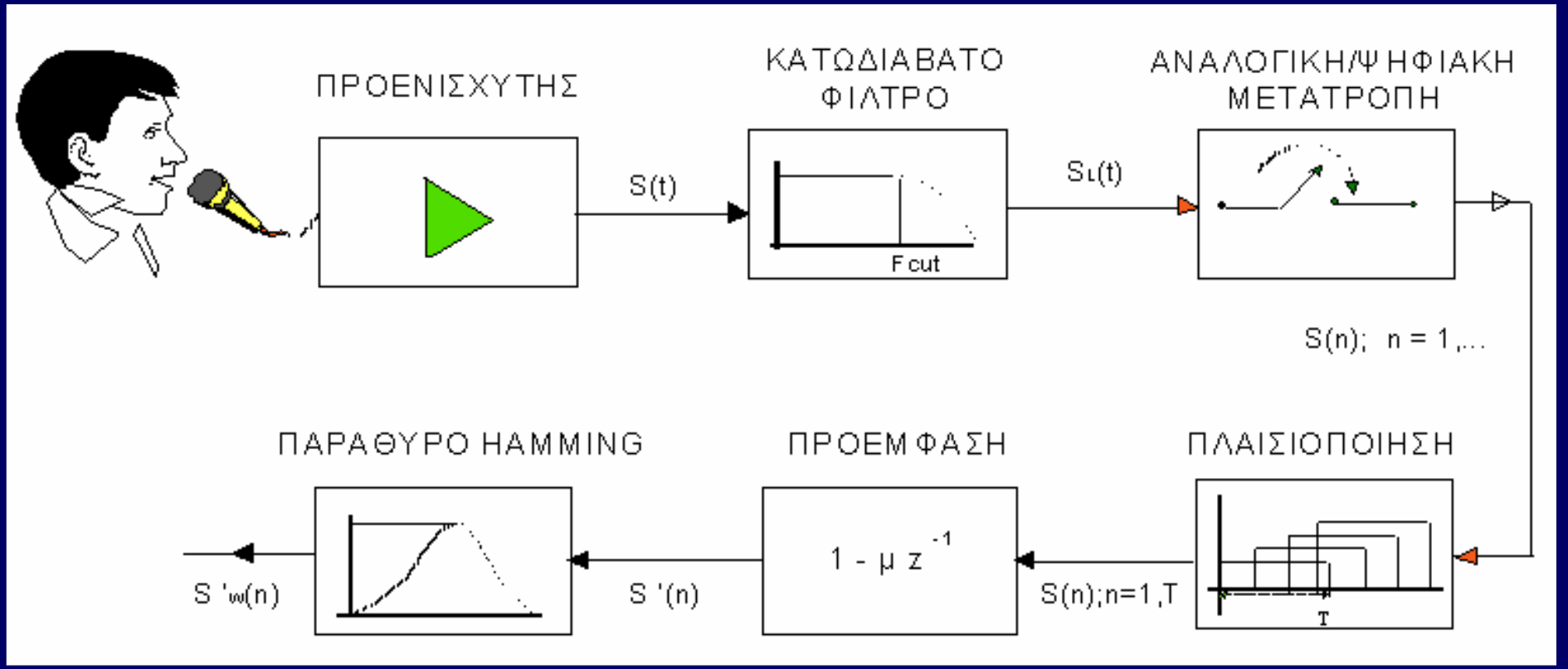


... in digital speech transmission



... and speaker recognition, pronunciation learning, etc

Προ-επεξεργασία Σήματος Ομιλίας



Ψηφιοποίηση Σημάτων Ομιλίας

Δειγματοληψία

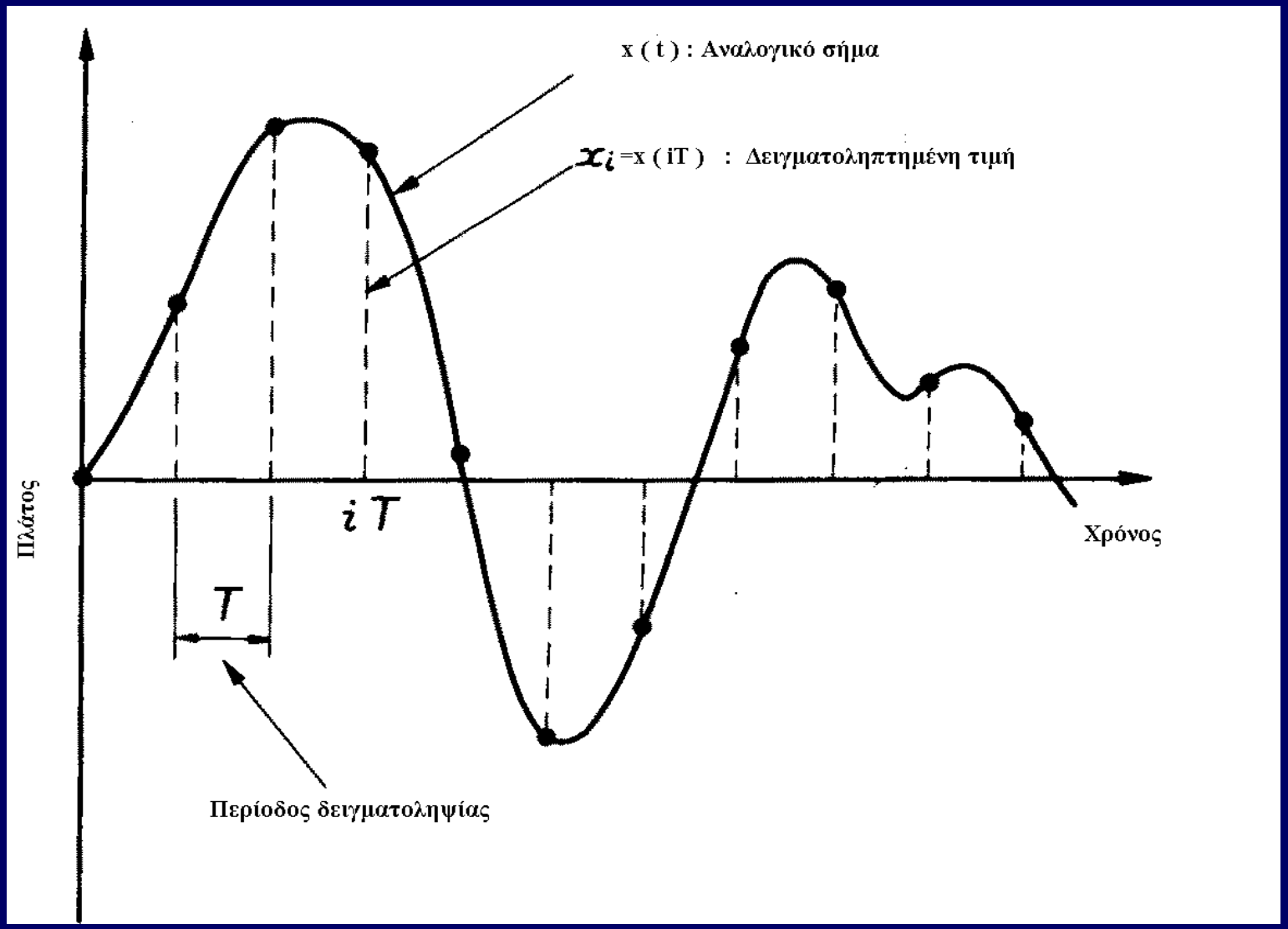
Θεώρημα Shannon: όταν το αναλογικό σήμα x(t) περιορίζεται σε

ζώνη 0 ως W Hz και όταν δειγματοληπτείται κάθε T = 1/2W [sec], το αρχικό σήμα μπορεί να αναπαραχθεί πλήρως ως εξής:

$$x(t) = \sum_{j=-\infty}^{\infty} x(i/2W)(\sin\{2\pi W(t - i/2W)\}/2\pi W(t - i/2W))$$

$x(i/2W)$ x(t) για $t_i = i/2W \ \ (i = ακεραιος) = iT$

$1/T = 2w\ Hz = S$ ονομάζεται ρυθμός Nyquist



Δειγματοληψία στο πεδίο του χρόνου

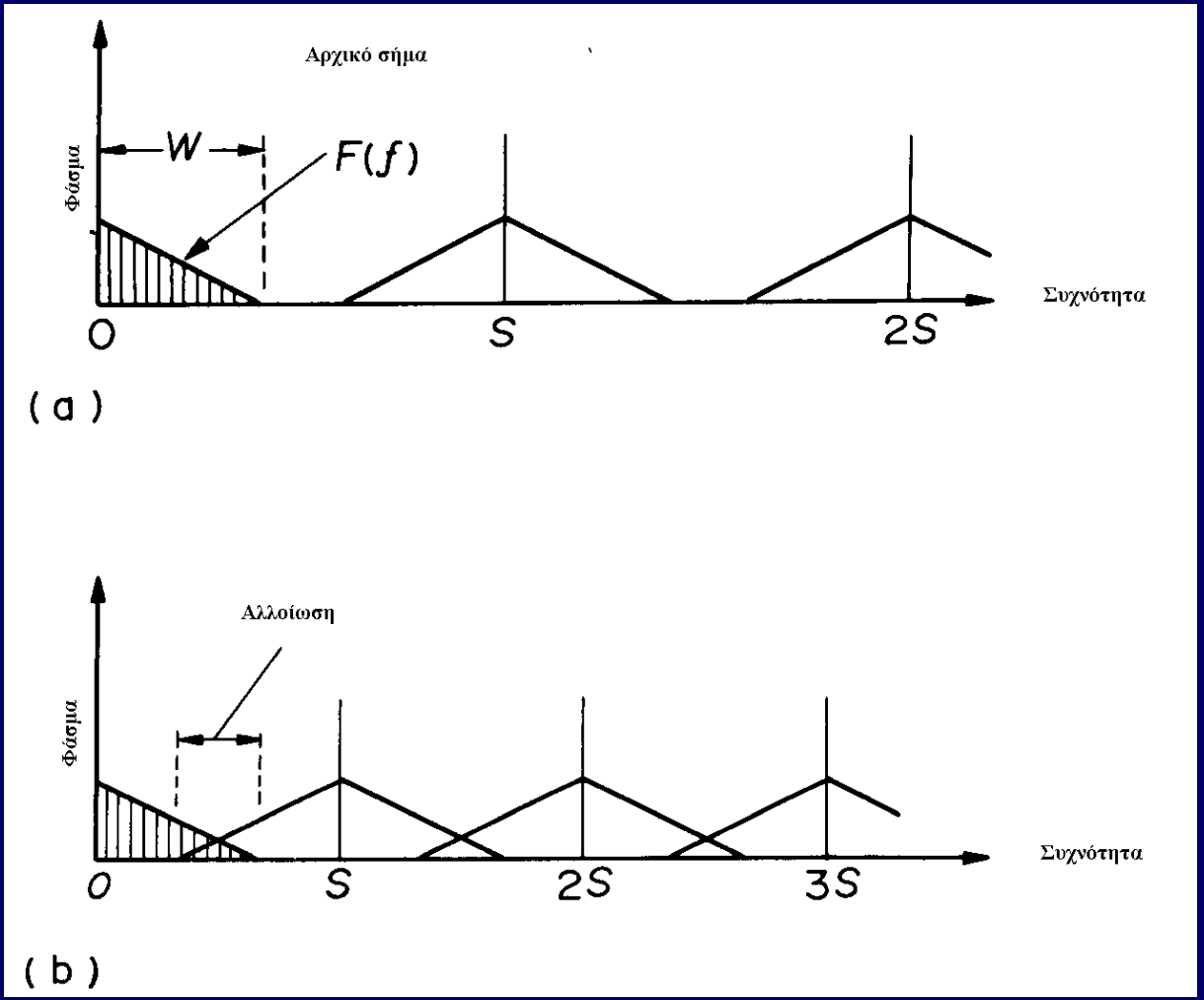
Δειγματοληψία στο πεδίο συχνότητας:

Α) σωστή δειγματοληψία

(S ≥ 2W)

Β) λάθος δειγματοληψία

(S < 2W)



1/T = 2W Hz: Συχνότητα Nyquist

Κβάντιση και Κωδικοποίηση

Κβάντιση:

- Βήμα κβάντισης = Δ
- Αριθμός επιπέδων, ορίζεται συνήθως =2^{B}

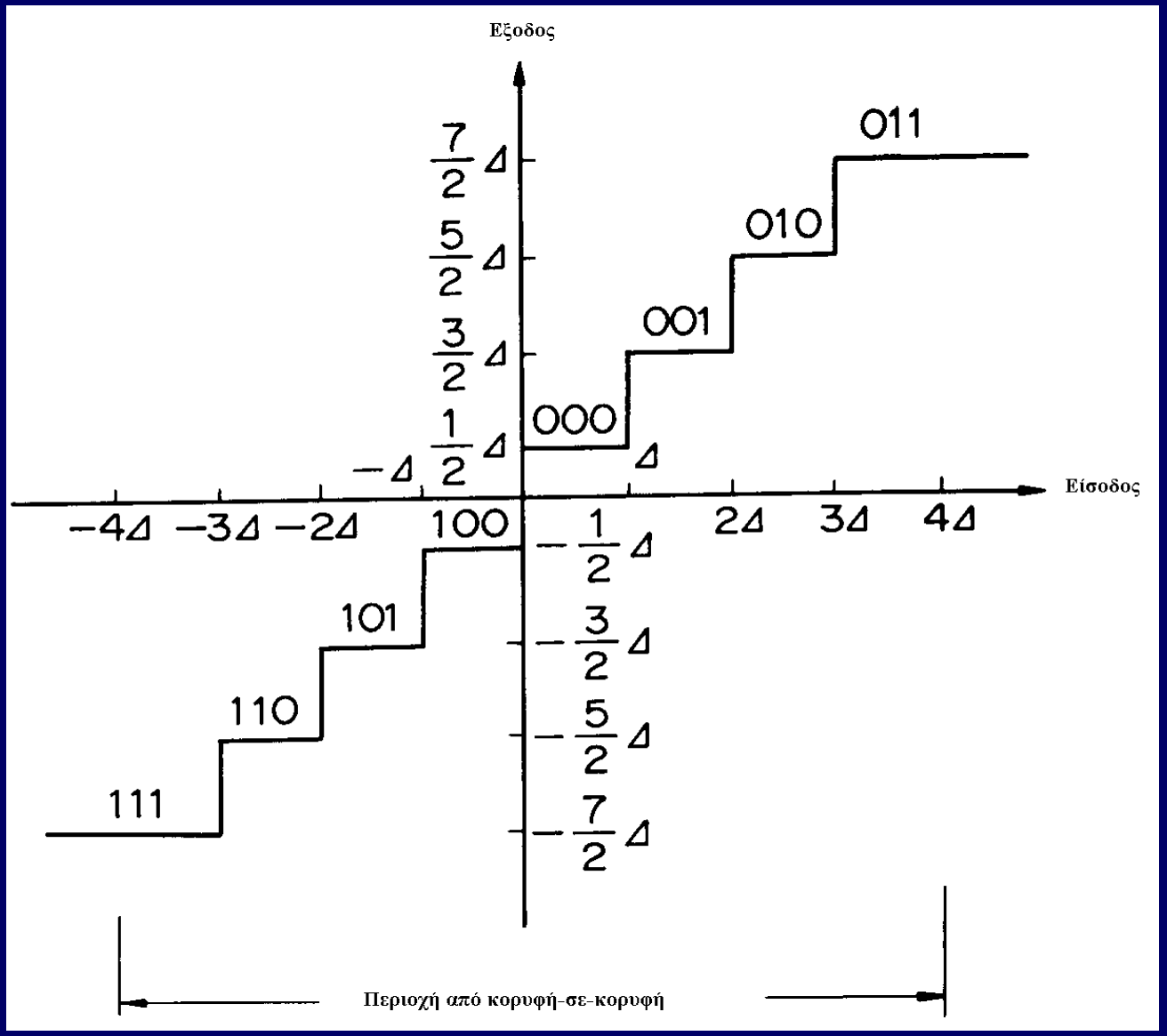
$$2x_{max} = \Delta 2^{\mathrm{B}}$$

- Τιμή μετά την κβάντιση =$\widehat{x_i}$
- Σφάλμα κβάντισης: $e_i = \hat{x} - x_i$

(ή παραμόρφωση κβάντισης ή θόρυβο κβάντισης)

$$-\Delta/2 \leq e_i \leq \Delta/2$$

Παράδειγμα των χαρακτηριστικών εισόδου-εξόδου της κβάντισης οχτώ επιπέδων (3 bit)



Θόρυβος Κβάντισης = ένα στατικό μοντέλο με χαρακτηριστικά:

- Είναι μία στατική διαδικασία λευκού θορύβου
- Δε σχετίζεται με το σήμα εξόδου
- Η κατανομή σφαλμάτων κβάντισης είναι ομοιόμορφη σε όλο το μήκος κάθε διαστήματος κβάντισης.

$$Prob(e_i) = \begin{cases} 1/\Delta, -\Delta/2 \le e_i \le 2 \\ 0, \sigma\varepsilon\ \kappa\alpha\theta\varepsilon\ \alpha\lambda\lambda\eta\ \pi\varepsilon\rho\iota\pi\tau\omega\sigma\eta \end{cases}$$

Λόγος του σήματος προς θόρυβο κβάντισης SNR:

$$SNR = \sigma_x^2/\sigma_e^2 = E[x_i^2]/E[e_i^2] = \sum x_i^2/\sum e_i^2$$

Όταν ικανοποιούνται τα 3 χαρακτηριστικά τότε:

$$\sigma_e^2 = (\Delta/2)/12 = 1/12(2x_{max}/2^B)^2 = x_{max}/3 * 2^B$$

→ Άρα $SNR = 3 * 2^B/(x_{max}/\sigma_x^2)^2$ ή σε dB $SNR = 10\log(\sigma_x^2/\sigma_e^2)$

Συναρτήσεις Παραθύρου (1/2)

Παράθυρο Hamming Wh(n): $W_H(n) = 0{,}54 - 0{,}46\cos(2\pi n/N - 1)$

εξασθένηση πρώτου πλευρικού λοβού: 43dB

Ορθογώνιο Παράθυρο: $W_R(n) = 1\ (0 \le n \le N - 1)$

εξασθένηση πρώτου πλευρικού λοβού: 13dB

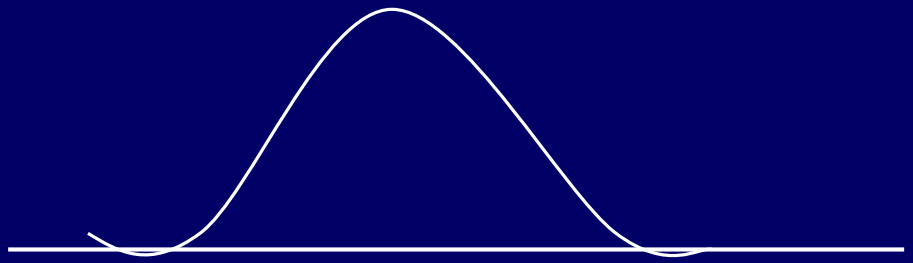
Παράθυρο Hanning: $W_N(n) = 0{,}5 - 0{,}5\cos(2\pi n/N - 1)$

εξασθένηση πρώτου πλευρικού λοβού: 30dB

Τετραγωνικό Παράθυρο

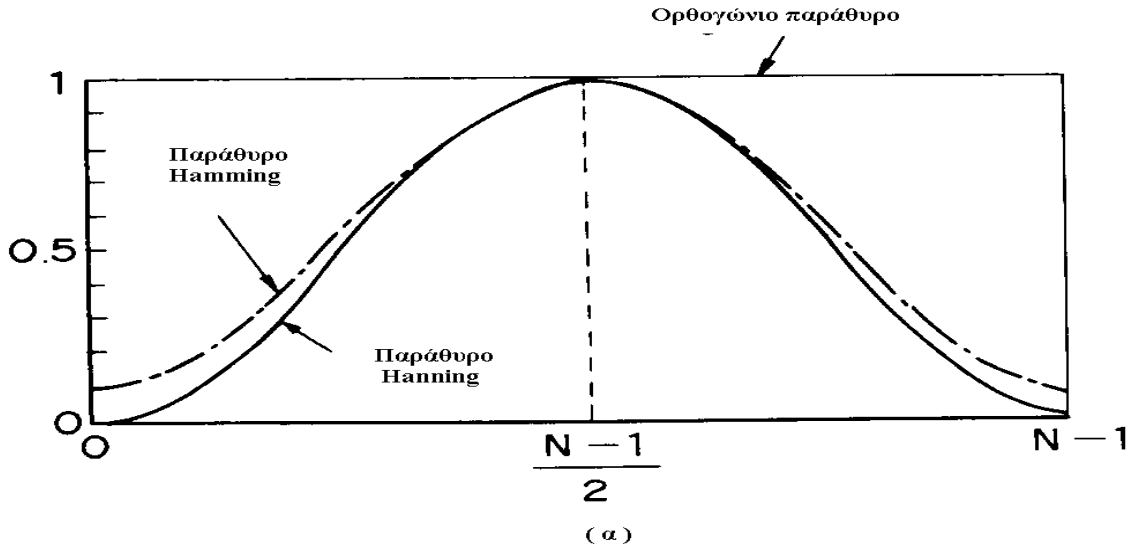
$W(n) = 1\ για\ καθε\ 0 < n < N - 1$

$W(n) = 0\ αλλου$

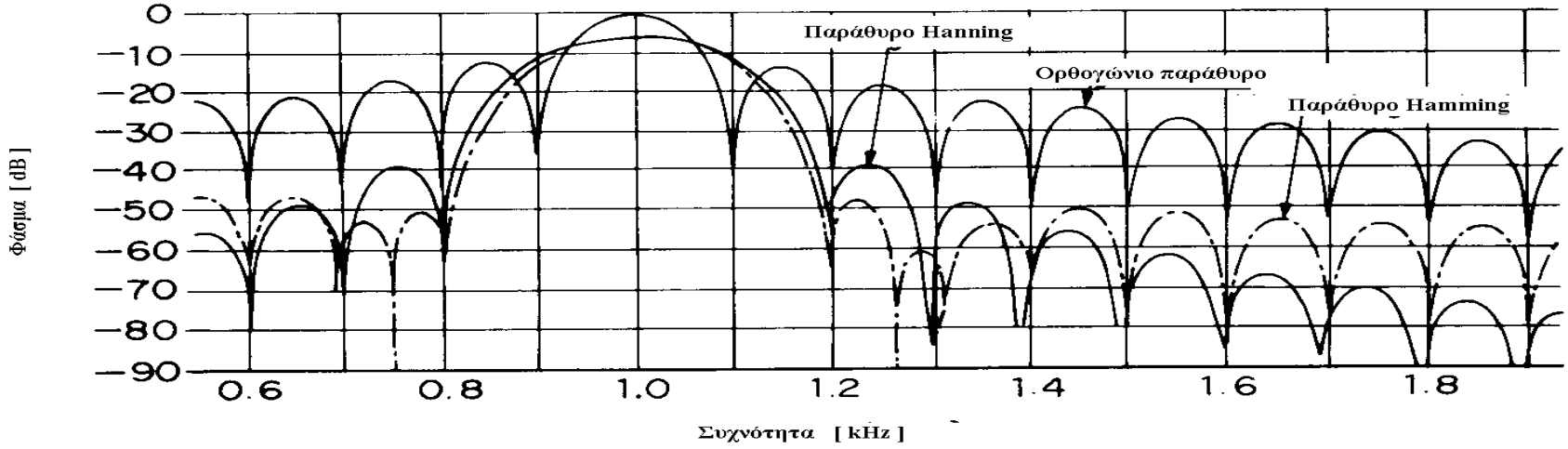


Παράθυρο Τύπου Hamming

Συναρτήσεις Παραθύρου (2/2)



(α)



(β)

Δύο αντικρουόμενες απαιτήσεις για εφαρμογή παραθύρου (windowing):

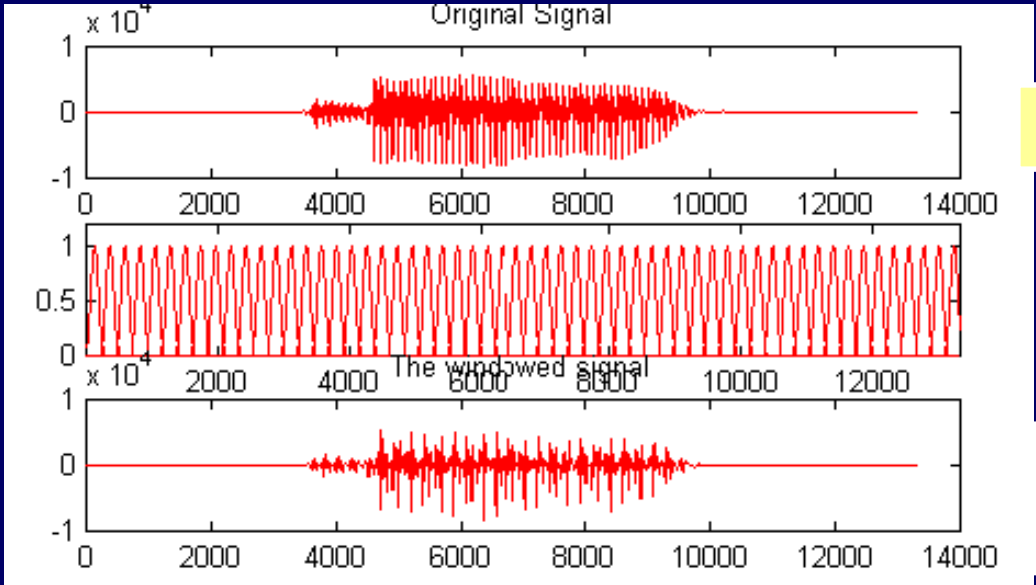
1. Υψηλή διακριτική ικανότητα σε φασματικό επίπεδο, δηλ. κυρίως ένας στενός – ομαλός λοβός

2. Μικρή διαπλάτυνση φασματικών συνιστωσών (leakage), δηλ. μεγάλη εξασθένιση πλευρικών λοβών

Εφαρμογή Χρονικού Παραθύρου

- Ανάλυση ομιλίας-χρονικό πεδίο:

$$s_w(n) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)$$



Αρχικό σήμα ομιλίας

Μετά την εφαρμογή χρονικού παραθύρου

Προέμφαση: συμπίεση της δυναμικής περιοχής του σήματος με εξομάλυνση της φασματικής κλίσης.

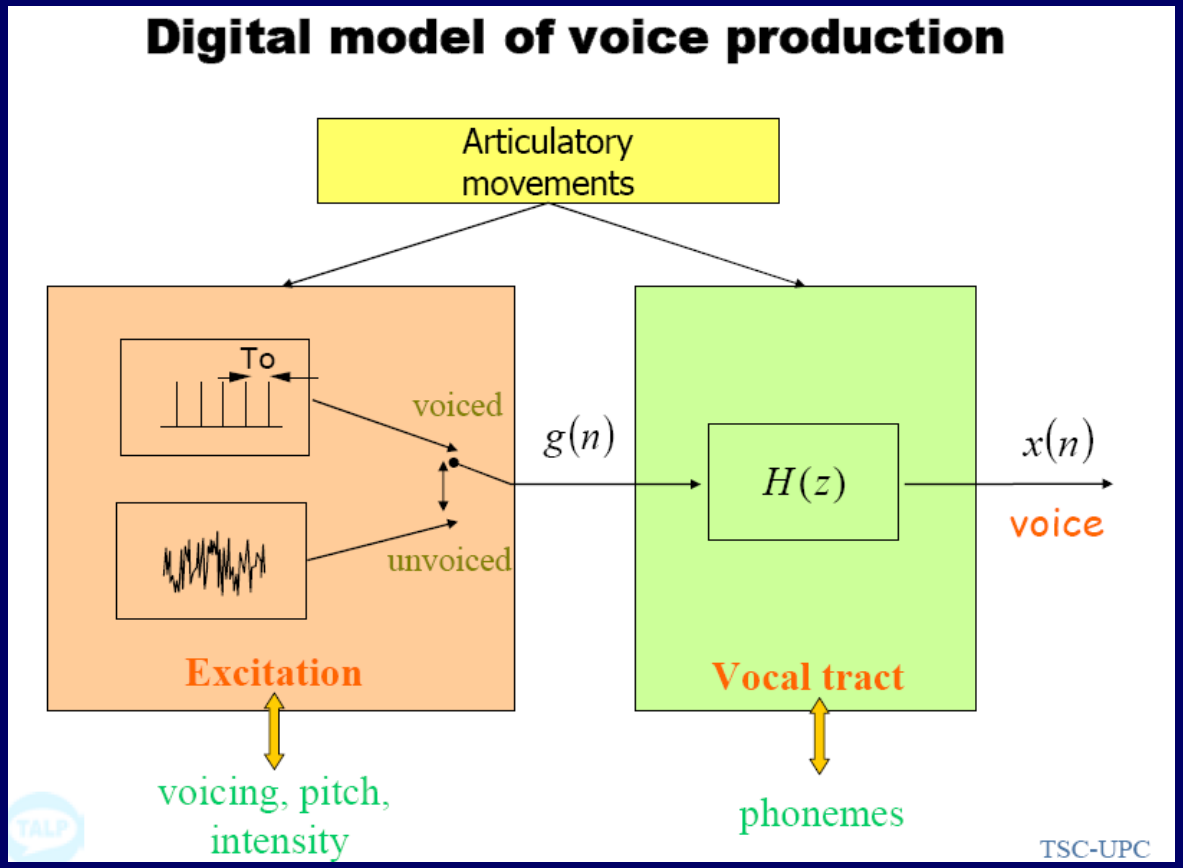
- Είναι αποτελεσματική στην άρση του SNR
- Γίνεται δίνοντας έμφαση στις συνιστώσες υψηλότερων συχνοτήτων 6dB/oct χονδρικά, πριν το φιλτράρισμα χαμηλών συχνοτήτων για τη μετατροπή A/D.
- Μπορεί επίσης να επιτευχθεί μετά τη μετατροπή A/D μέσω διαφορικού υπολογισμού ή μέσω του ψηφιακού φιλτραρίσματος πρώτης τάξης:

$$\mathrm{H}(z) = 1 - \mathrm{a}z^{-1}$$

Όπου a ορίζεται σε μία τιμή γύρω στο 1. Όταν όμως αυξάνουμε το SNR, όσο το δυνατόν περισσότερο, είναι ανάγκη να εφαρμόσουμε την προέμφαση πριν τη μετατροπή A/D.

Αποέμφαση: Η διαδικασία πρόσθεσης μιας κλίσης 6dB/oct ώστε να αναπαράγουμε την αρχική φασματική κλίση.

Δυναμική περιοχή κύματος ομιλίας > 50dB => 10 bits ή περισσότερα για A/D.



$$x(n) = g(n) * h(n)$$

x(n)=speech signal

g(n)=glottal excitation

(h)= vocal tract filter

* convolution

After Fourier Transform FT:

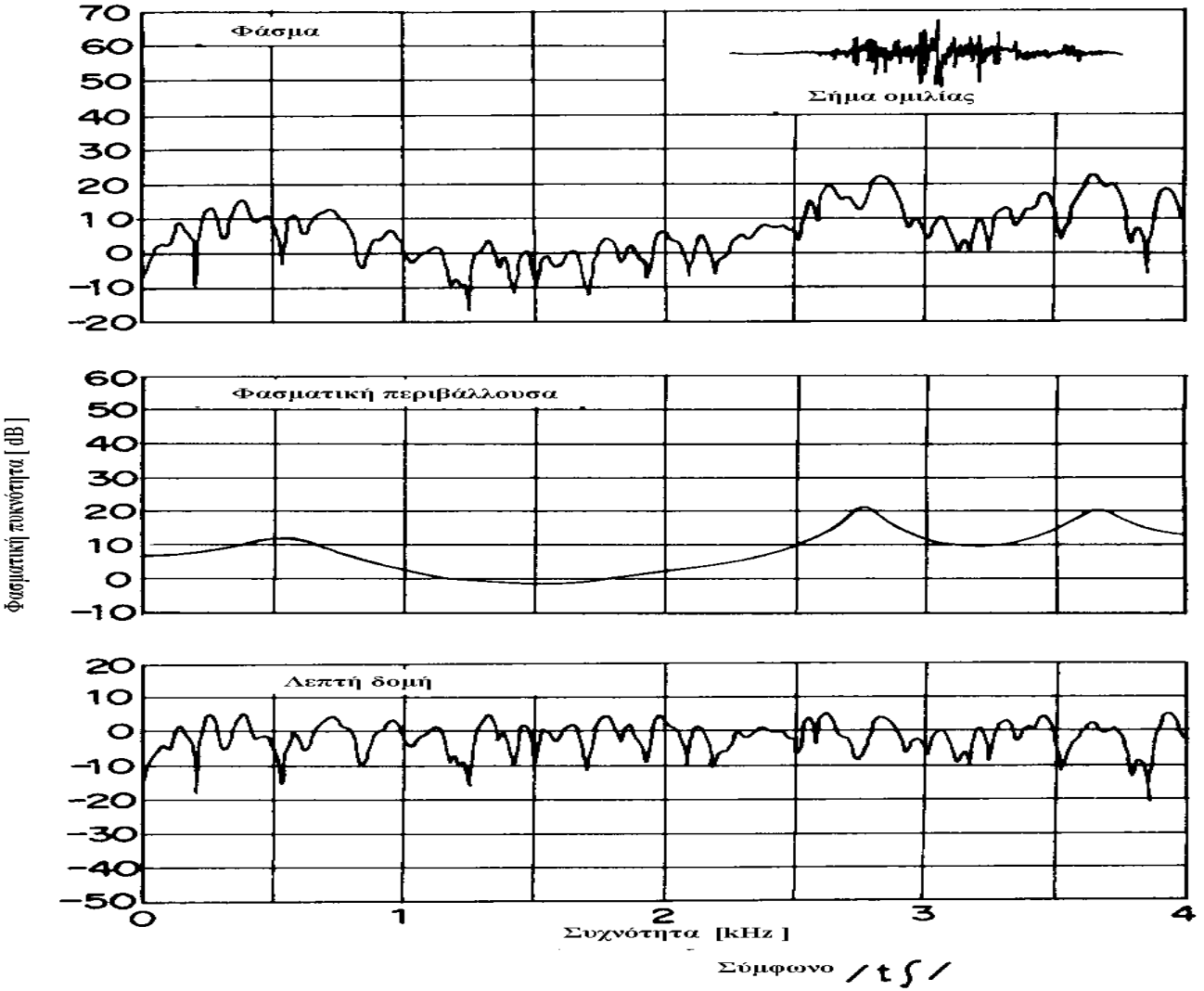
$$FT\{x(n)\} = FT\{g(n) * h(n)\}$$

$$X(\omega) = G(\omega) \cdot \mathrm{H}(\omega)$$

Έξοδος ομιλίας

Φωνητική οδός
(δομή formant)

Πηγή
θόρυβος

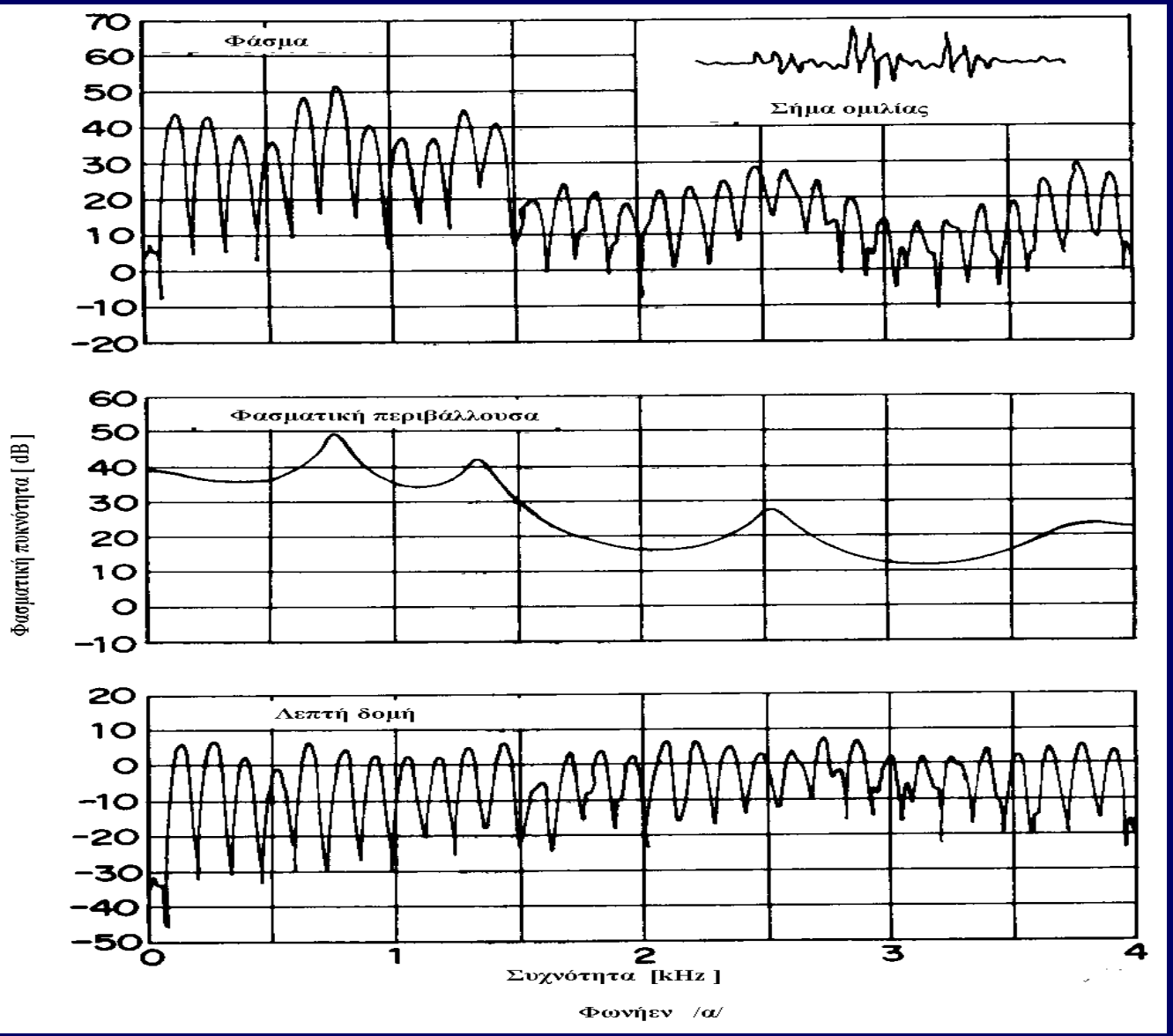


Δομή βραχύχρονων (20 msec) φασμάτων ομιλίας για ανδρικές φωνές κατά την εκφώνηση του συμφώνου |t|

Έξοδος ομιλίας

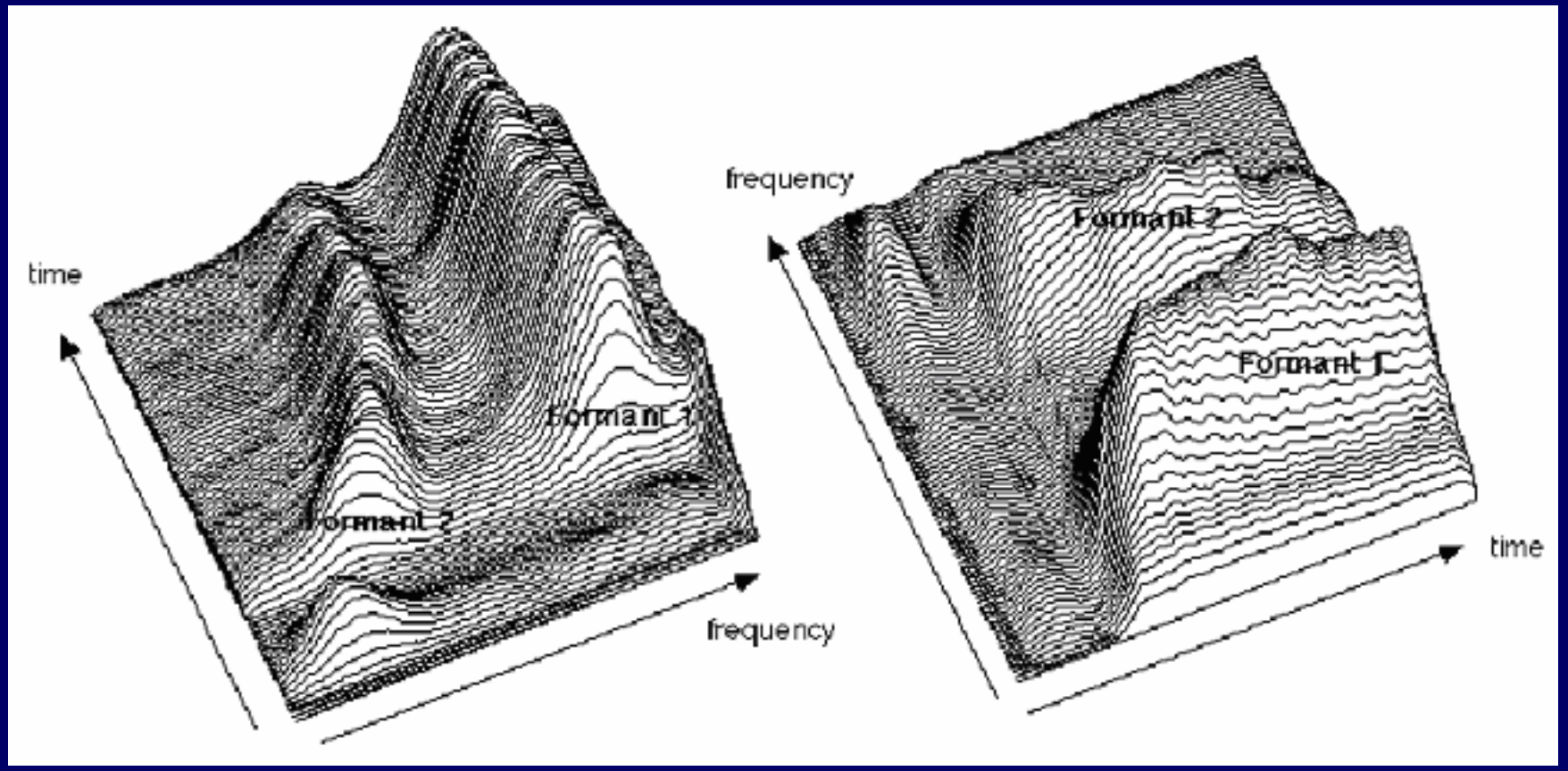
Φωνητική οδός
(δομή formant)

Πηγή
Pitch, σχεδόν περιοδική μορφή

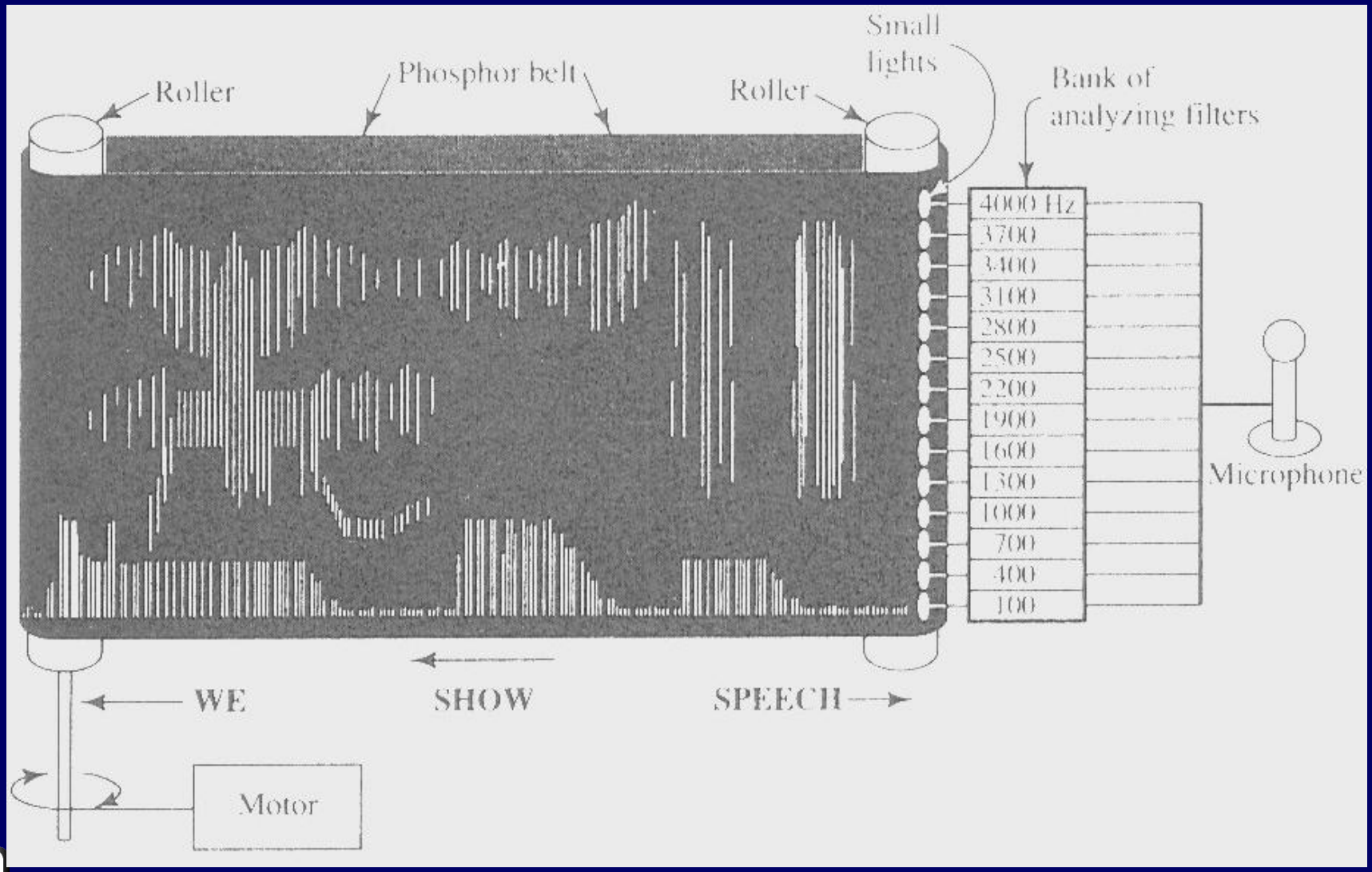


Δομή βραχύχρονων (20 msec) φασμάτων ομιλίας για ανδρικές φωνές κατά την εκφώνηση του φωνήεντος |a|

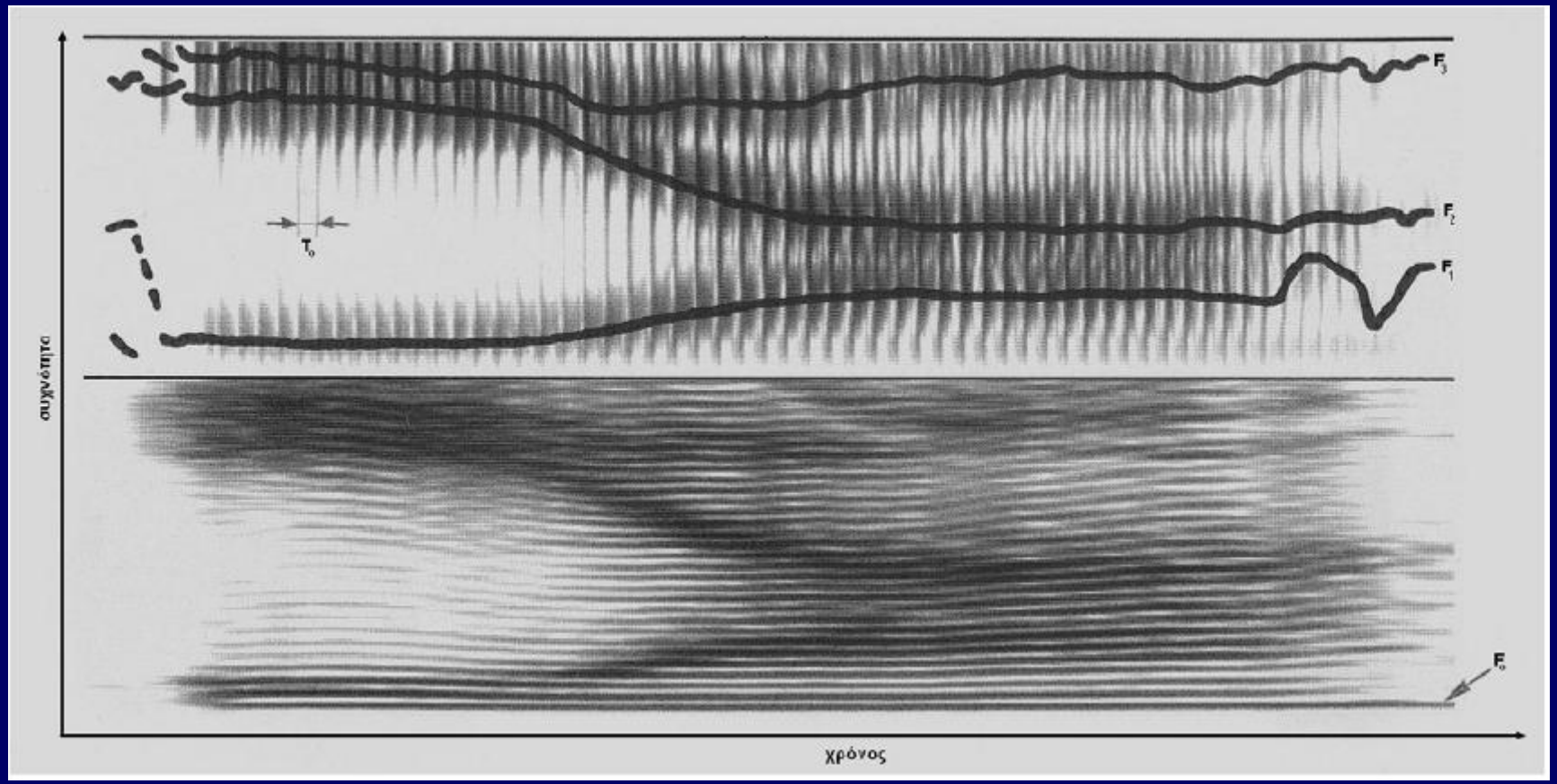
Τρισδιάτατα φασματογραφήματα



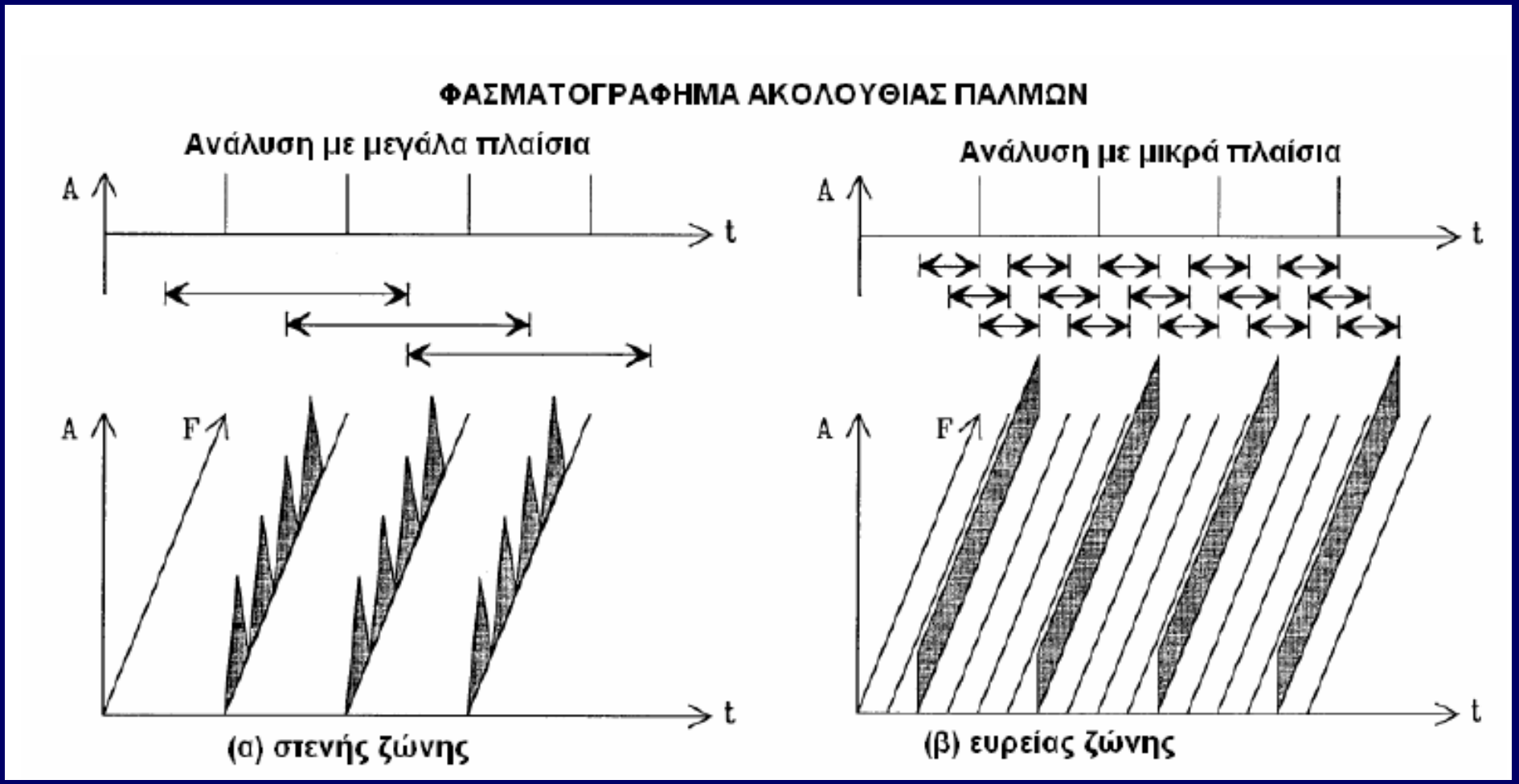
Σύστημα φασματογραφήματος του εργαστηρίου Bell

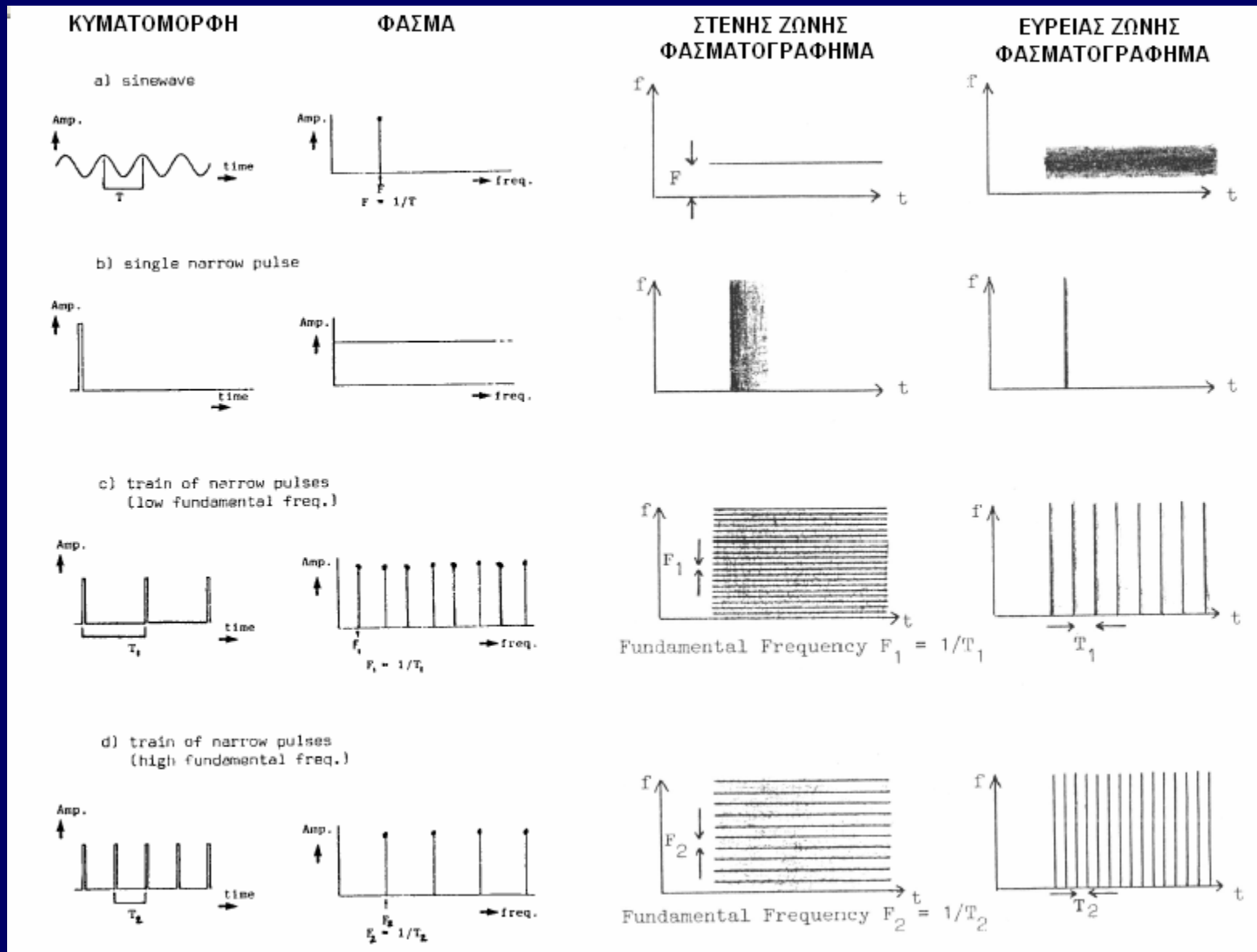


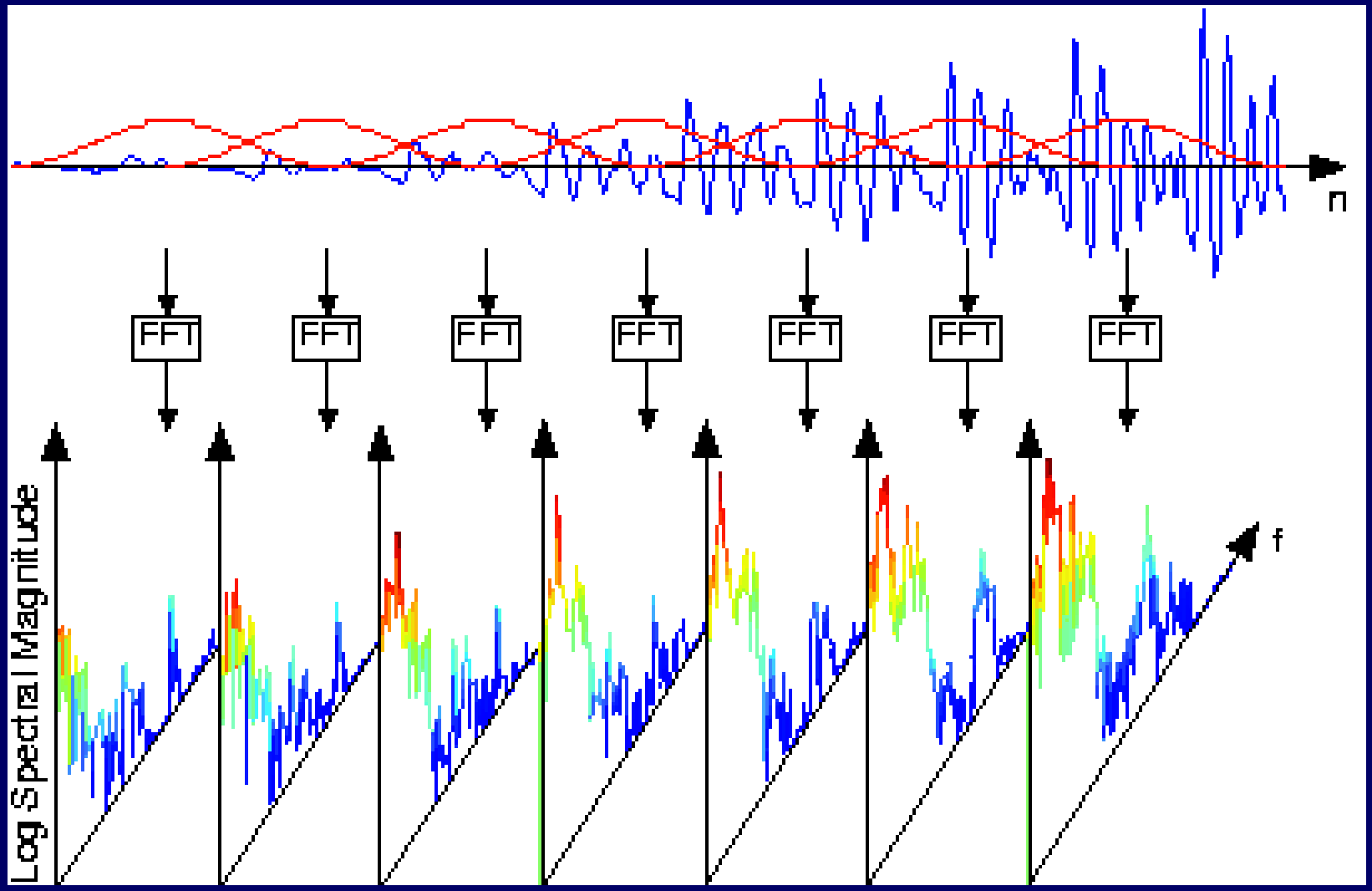
Φασματογράφημα ευρείας ζώνης ανάλυσης πάνω και στενής ζώνης ανάλυσης κάτω



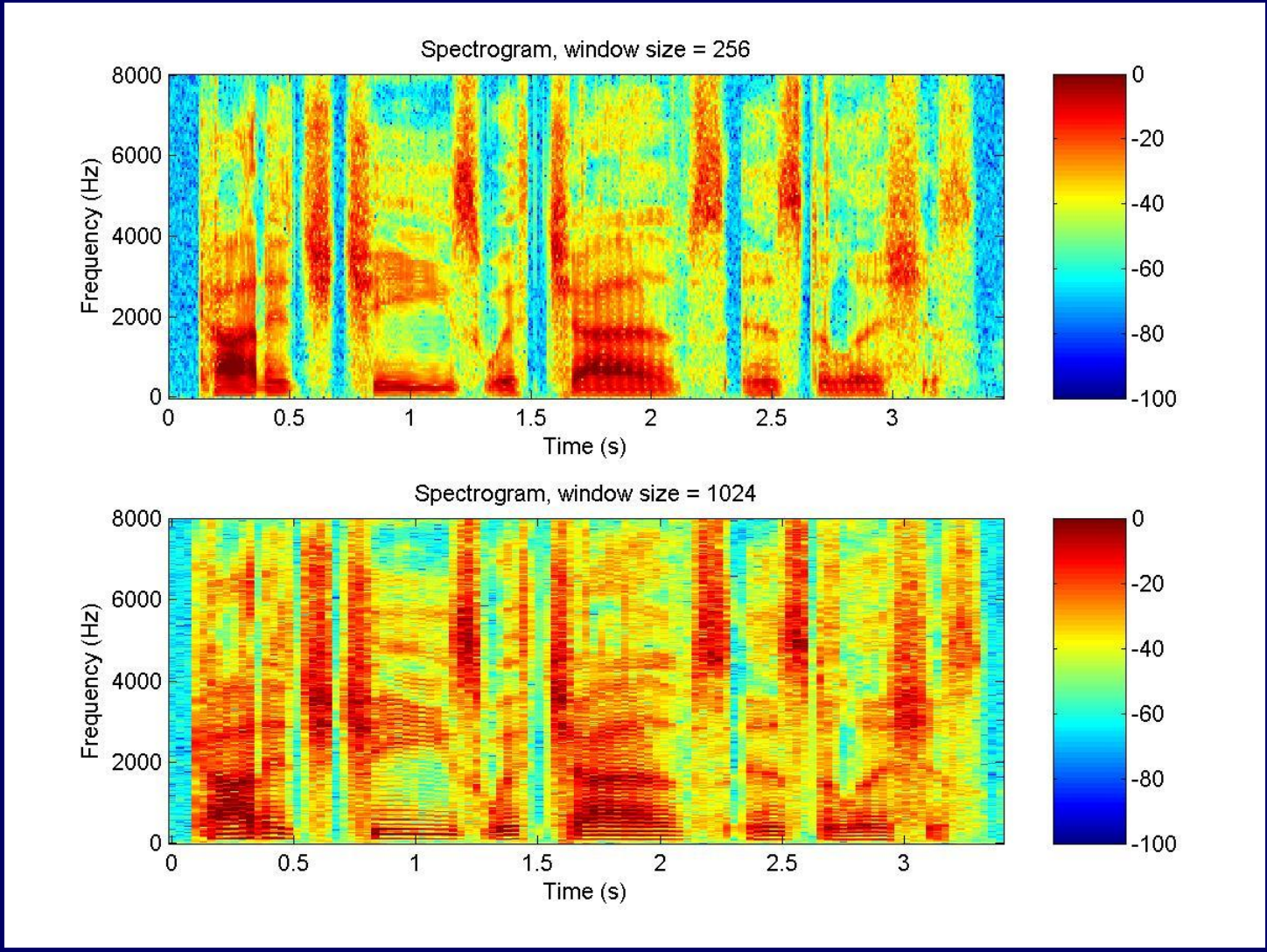
Φασματογράφημα ακολουθίας παλμών στενής και ευρείας ζώνης



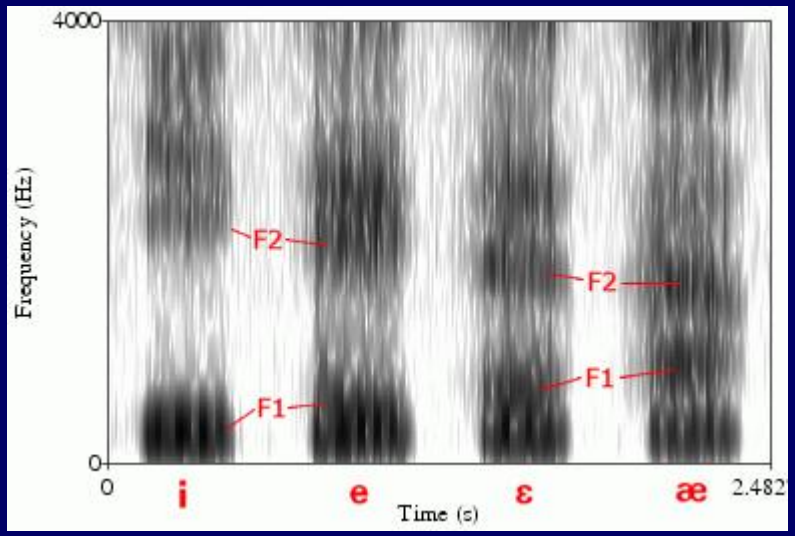

ΚΥΜΑΤΟΜΟΡΦΗ
ΦΑΣΜΑ
ΣΤΕΝΗΣ ΖΩΝΗΣ ΦΑΣΜΑΤΟΓΡΑΦΗΜΑ
ΕΥΡΕΙΑΣ ΖΩΝΗΣ ΦΑΣΜΑΤΟΓΡΑΦΗΜΑ
a) sinewave
Amp.
time
T
Amp.
freq.
F = 1/T
f
F
t
b) single narrow pulse
Amp.
time
Amp.
freq.
c) train of narrow pulses
(low fundamental freq.)
Amp.
time
Amp.
freq.
f
F1
t
Fundamental Frequency F1 = 1/T1
T1
d) train of narrow pulses
(high fundemental freq.)
Amp.
time
Amp.
freq.
F2 = 1/T2
F2
T2
Fundamental Frequency F2 = 1/T2

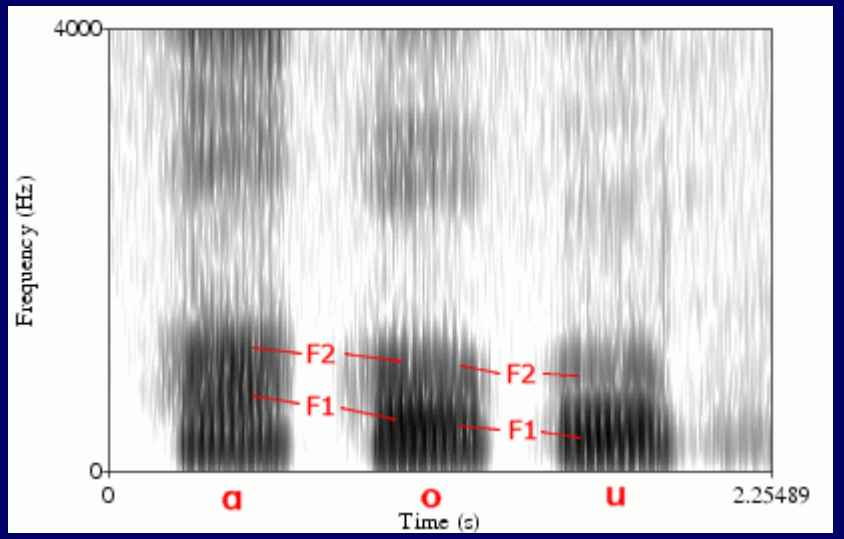

n
FFT
FFT
FFT
FFT
FFT
FFT
FFT
Log Spectral Magnitude
f

Φασματογραφήματα ευρεία και στενής ζώνης

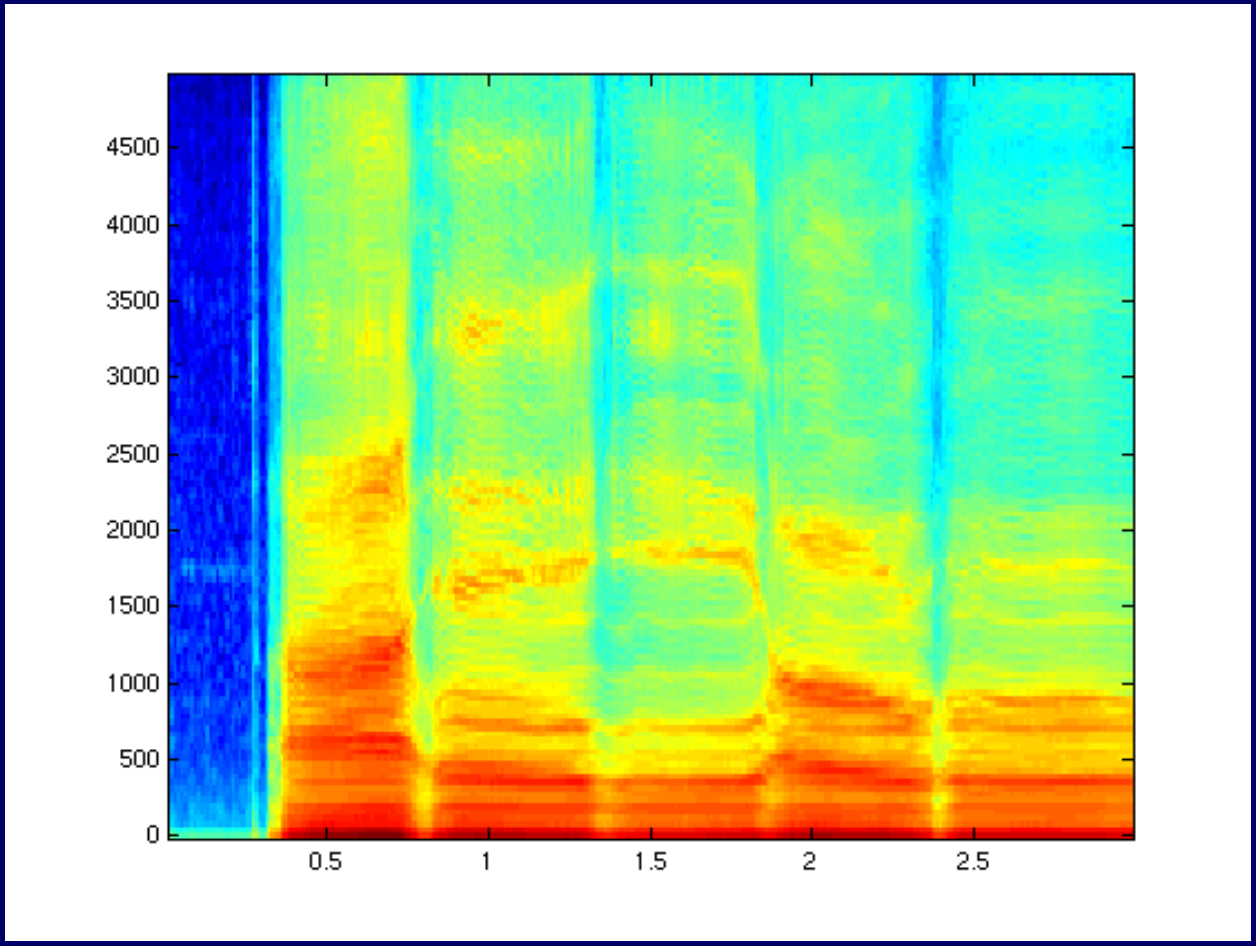


Πρώτος και δεύτερος φωνοσυντονισμός σε φασματογραφήματα ευρείας ζώνης φωνηέντων

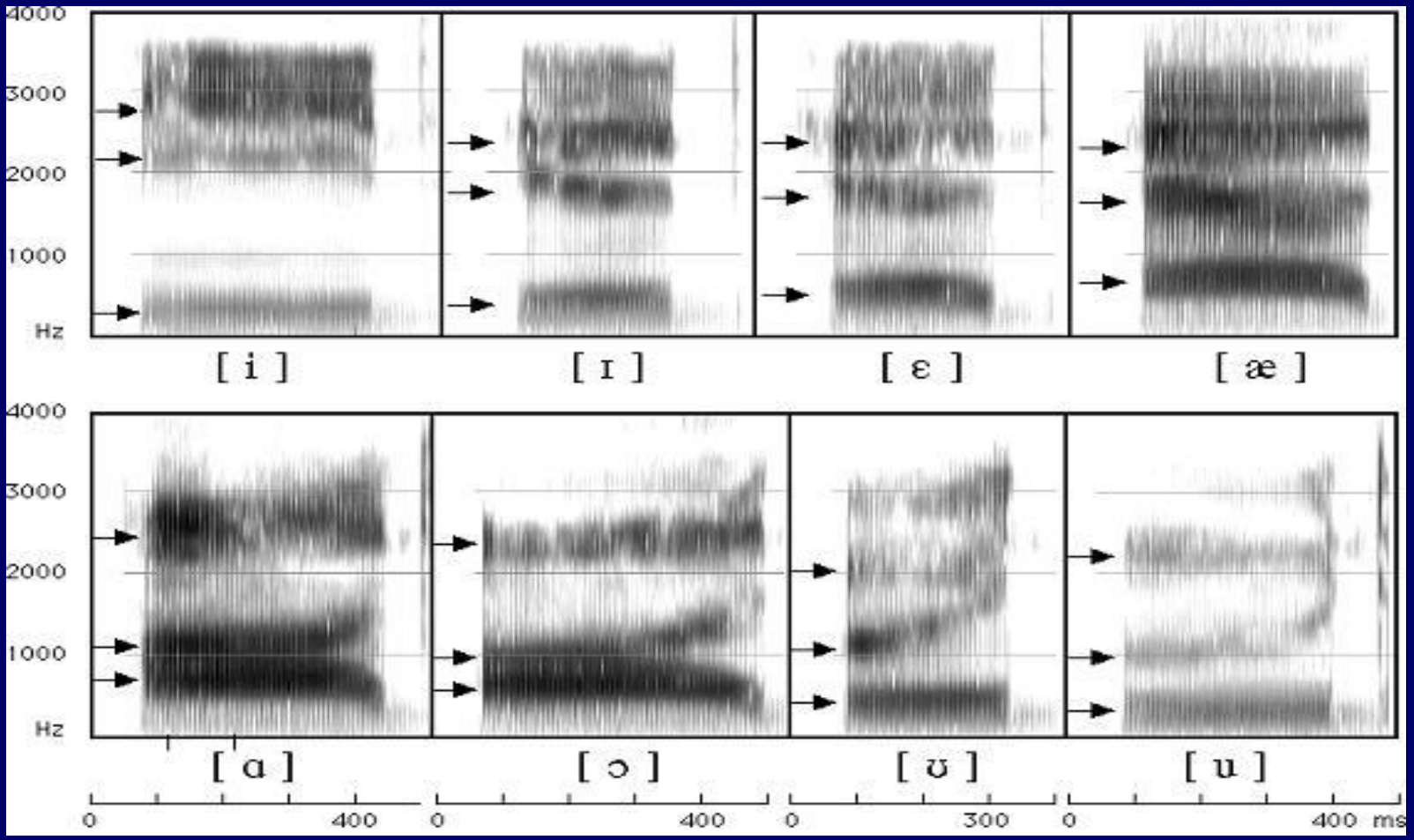




Φασματογραφήματα των φωνηέντων /α, ε, ι, ο, u/



Φασματογραφήματα εκφωνήσεων heed, hid, head, had, hod, hawed, hood, who'd (male speaker, American English)



Παράδειγμα μη-έμφωνων και έμφωνων κλειστών συμφώνων vs voiced stops)

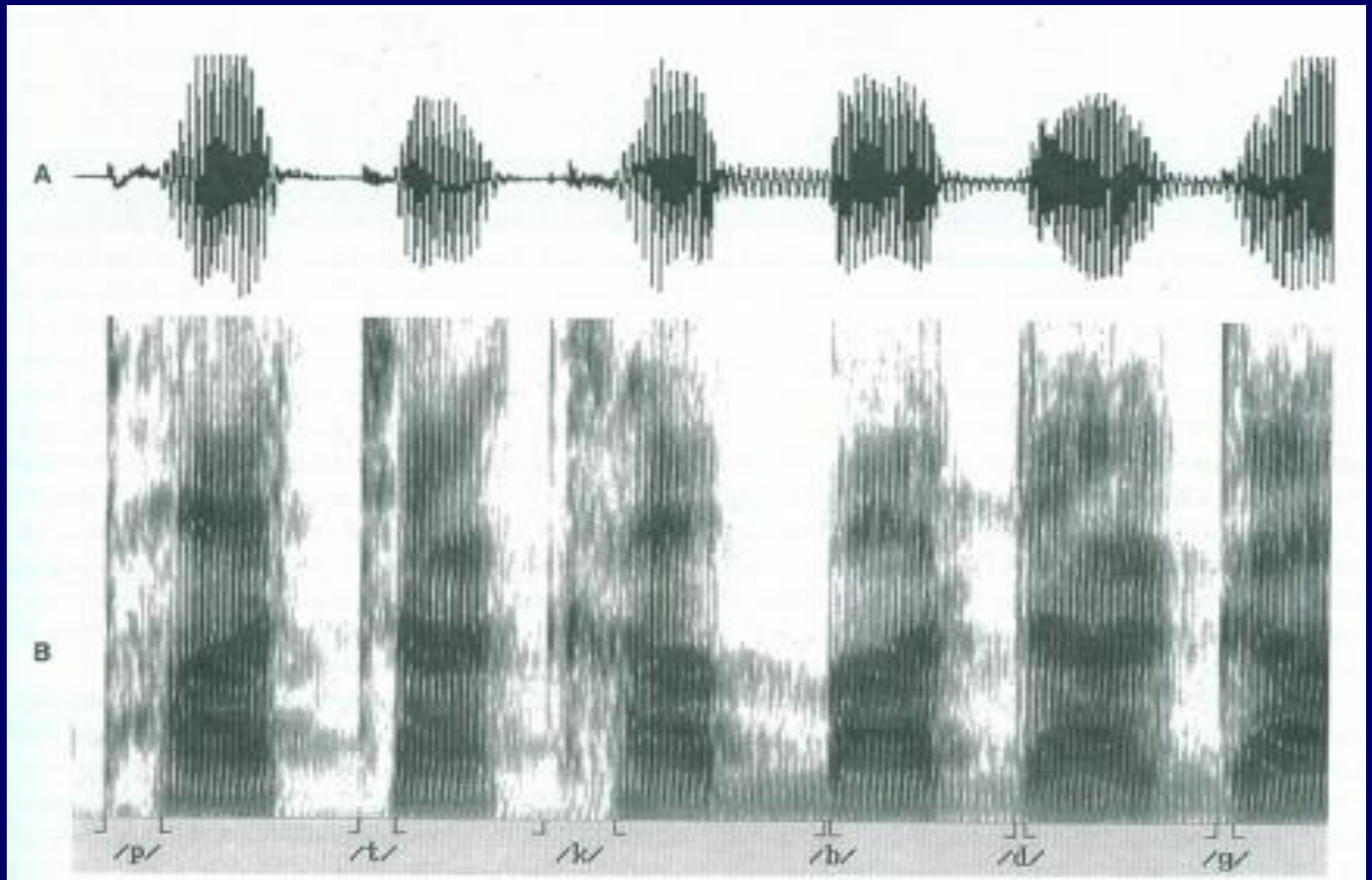
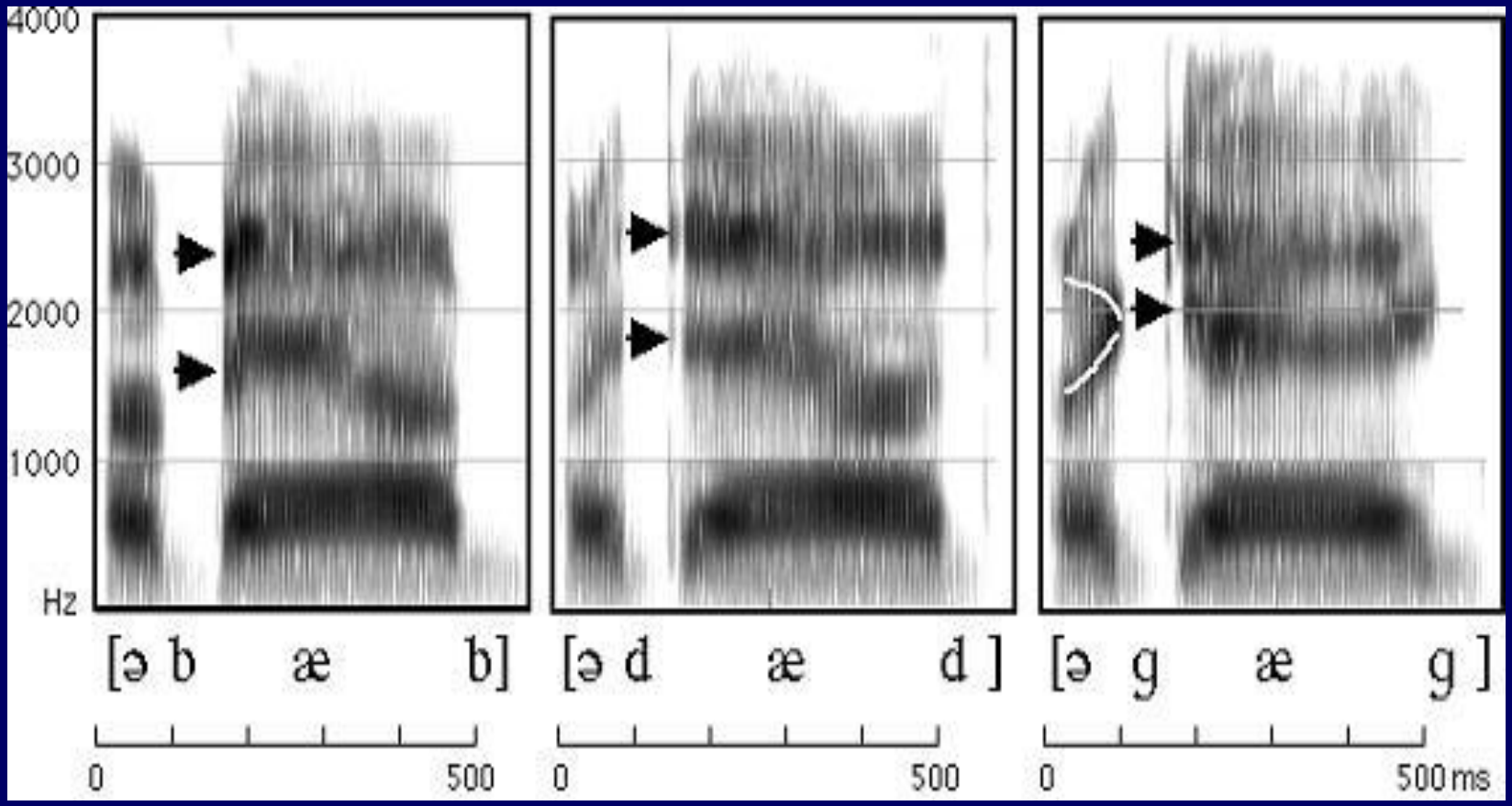
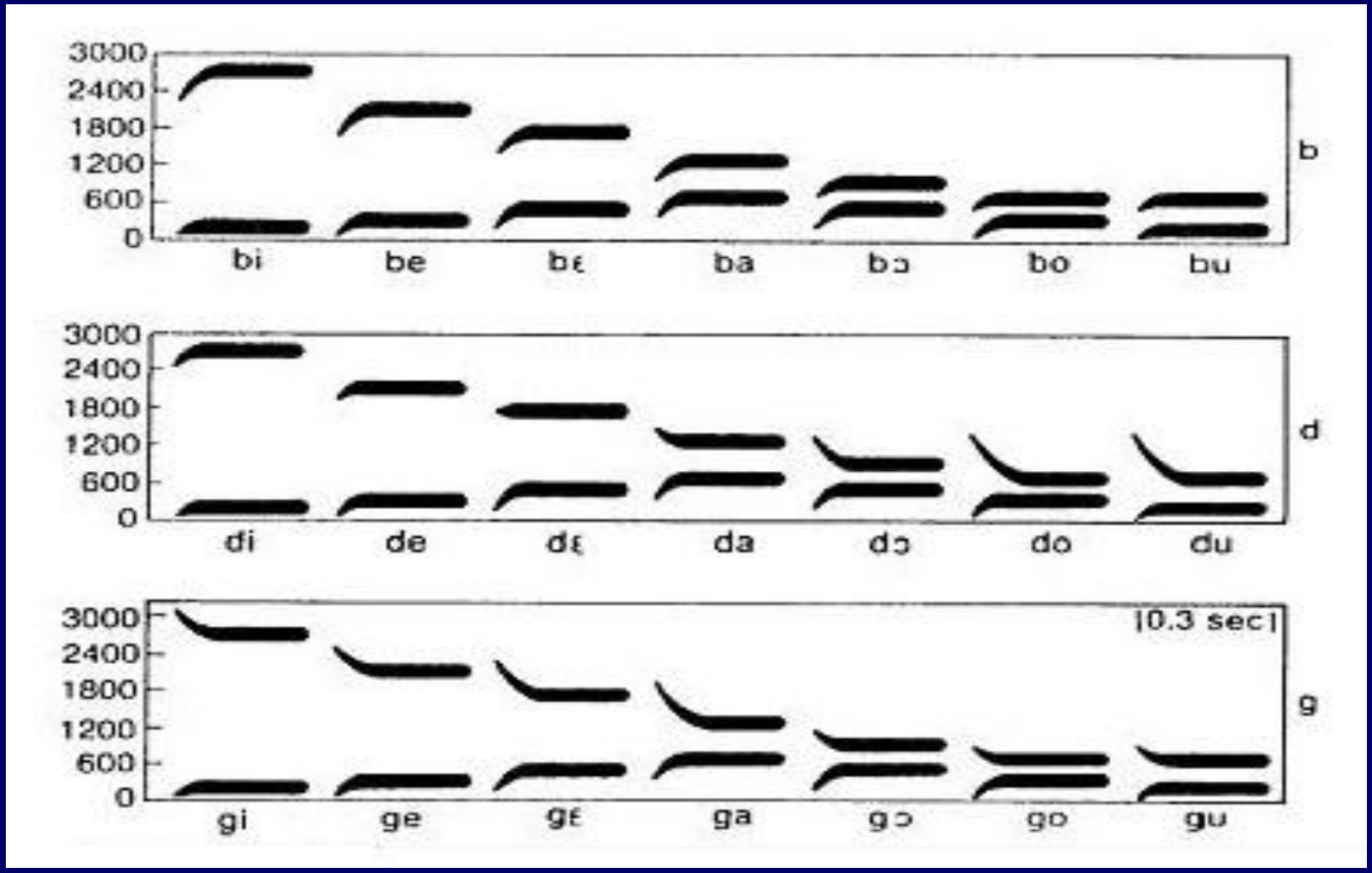


FIGURE 5-16 *A*, Sound pressure waveform and *B*, spectrogram illustrating oral stops, showing noise bursts and formant patterns for the syllables /pʌ tʌ kʌ bʌ dʌ gʌ/.

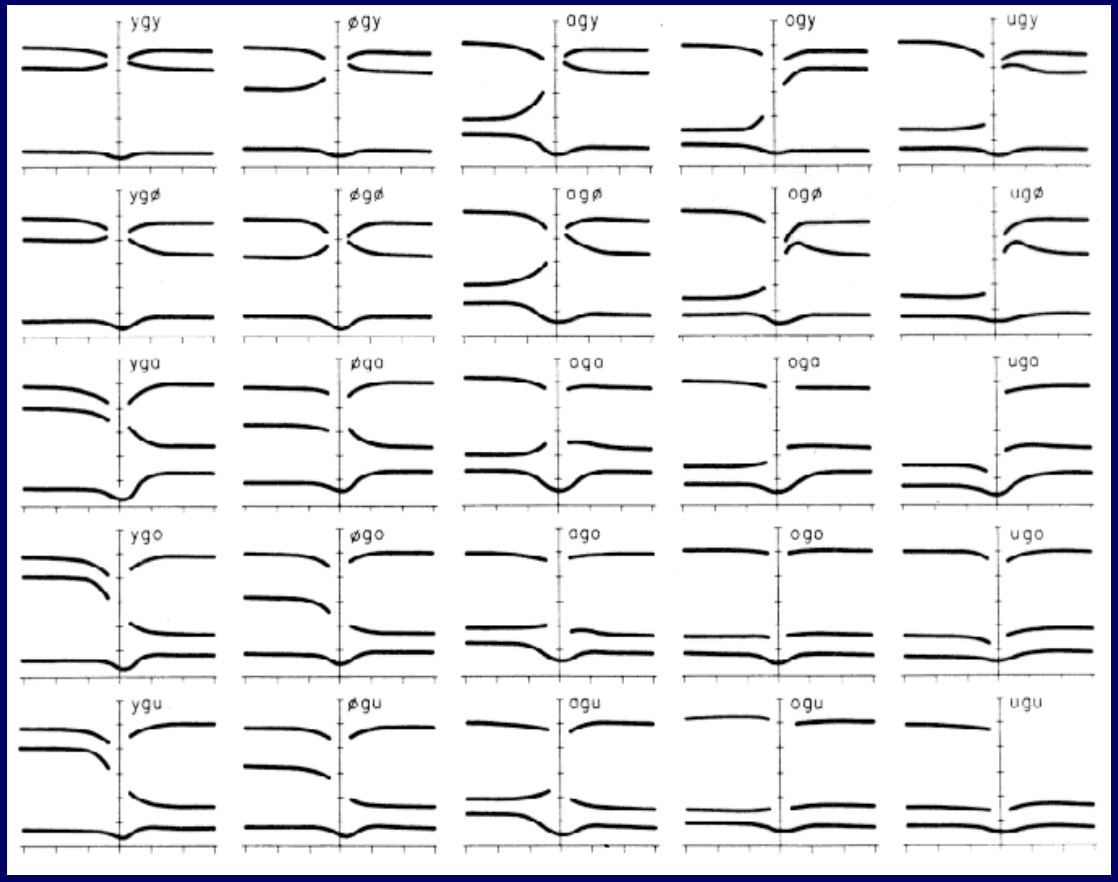
Φασματογραφήματα “a bab, a dad, a gag”



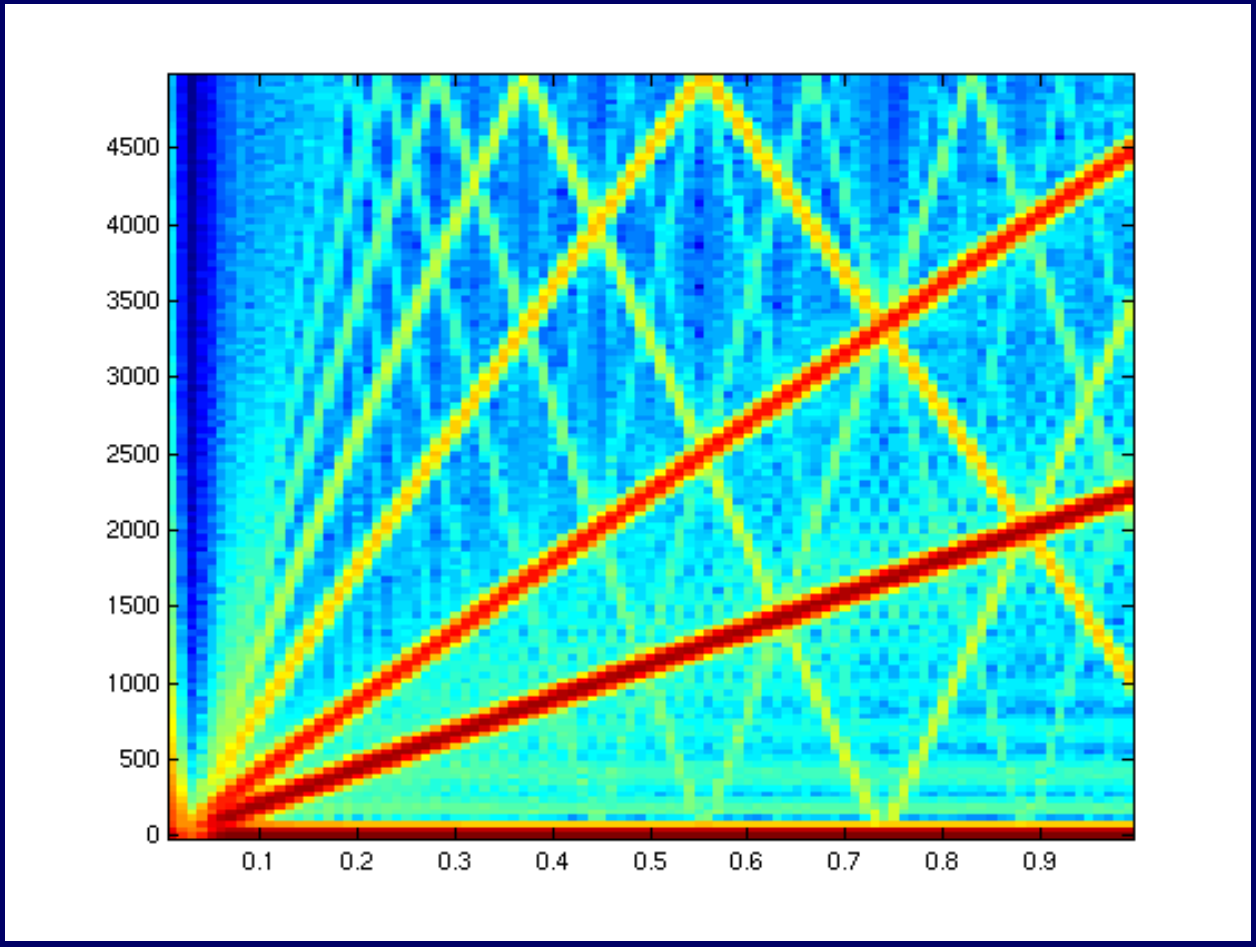
Μεταπτώσεις φωνοσυντονισμών (1/2)



Μεταπτώσεις φωνοσυντονισμών (2/2)



Response of timing model to chirp stimuli



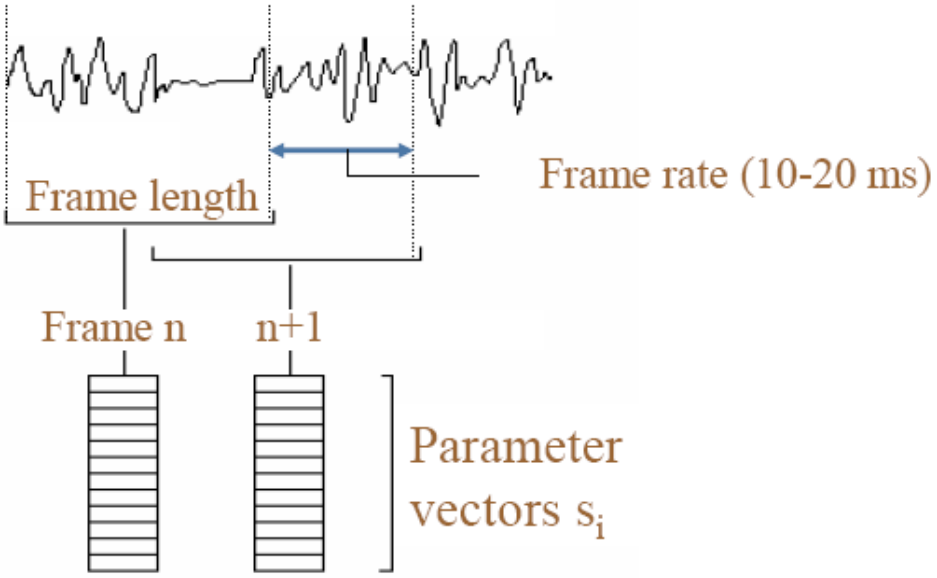
Παραμετροποίηση σήματος ομιλίας

- **Γιατί; Μειονεκτήματα της αναπαράστασης κυματομορφής:**
  - 16.000 δείγματα/sec x 16 bits/sample=256 Kbits/sec!
  - Πλεονασμός.
  - Δεν δείχνει τα χαρακτηριστικά παραγωγής και αντίληψης του σήματος ομιλίας .

- **Στόχος της παραμετροποίησης: η αναπαράσταση ενός τμήματος του σήματος ομιλίας με ένα σύνολο παραμέτρων που αναδεικνύουν τη χρονική μεταβολή της δομής του φάσματος**
  - Μόνο η φασματική περιβάλλουσα χρησιμοποιείται στην αναγνώριση ομιλίας
  - Όταν το σήμα θα πρέπει να επανασυνθεθεί:
    - Χρειαζόμαστε τη φασματική του δομή
    - Λόγω της συνάρθρωσης, το φάσμα που εκτιμάται με την υπόθεση στατικότητας, δεν φέρει όλη τη σχετική πληροφορία.

Parameterization$_2$
Alternative approaches

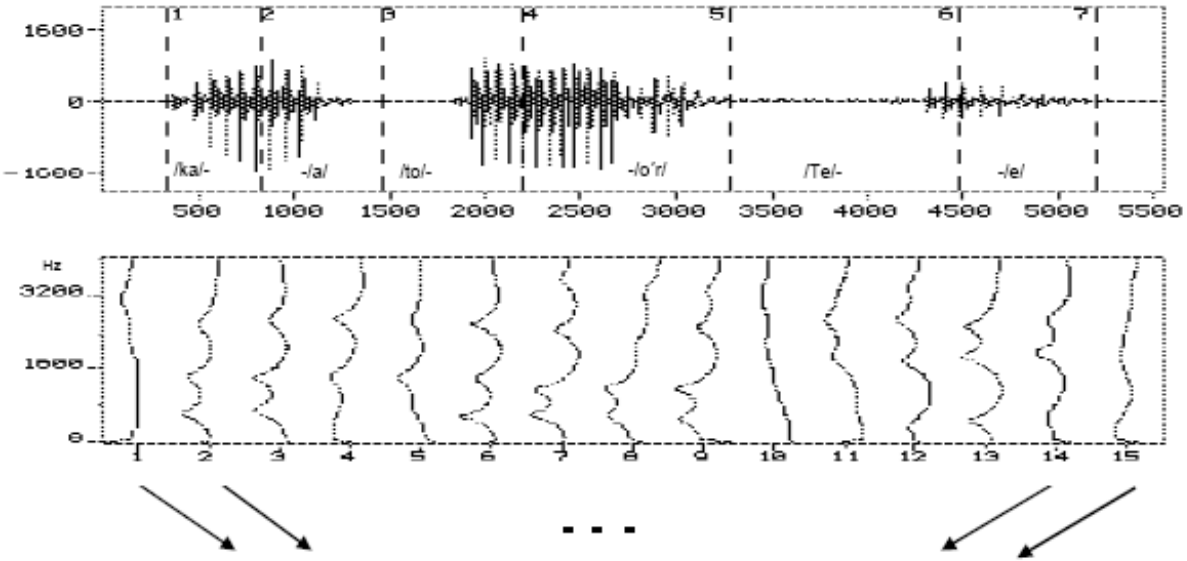
- ✓ Optimal: 1) split the signal in acoustically homogeneous segments; 2) parameterize each segment
- ✓ Usual: ...the signal is parameterized frame-by-frame



Parameterization of the spectral envelope: the problem

To represent the temporal evolution of the speech spectral envelope with a small set of parameters in an efficient way

- ✓ The determintation of the spectral envelope is a deconvolution problem
- ✓ Efficiency: in terms of time-frequency resolution and estimation variance



Κυριότεροι μέθοδοι για την ανάλυση των φασμάτων ομιλίας (1/2)

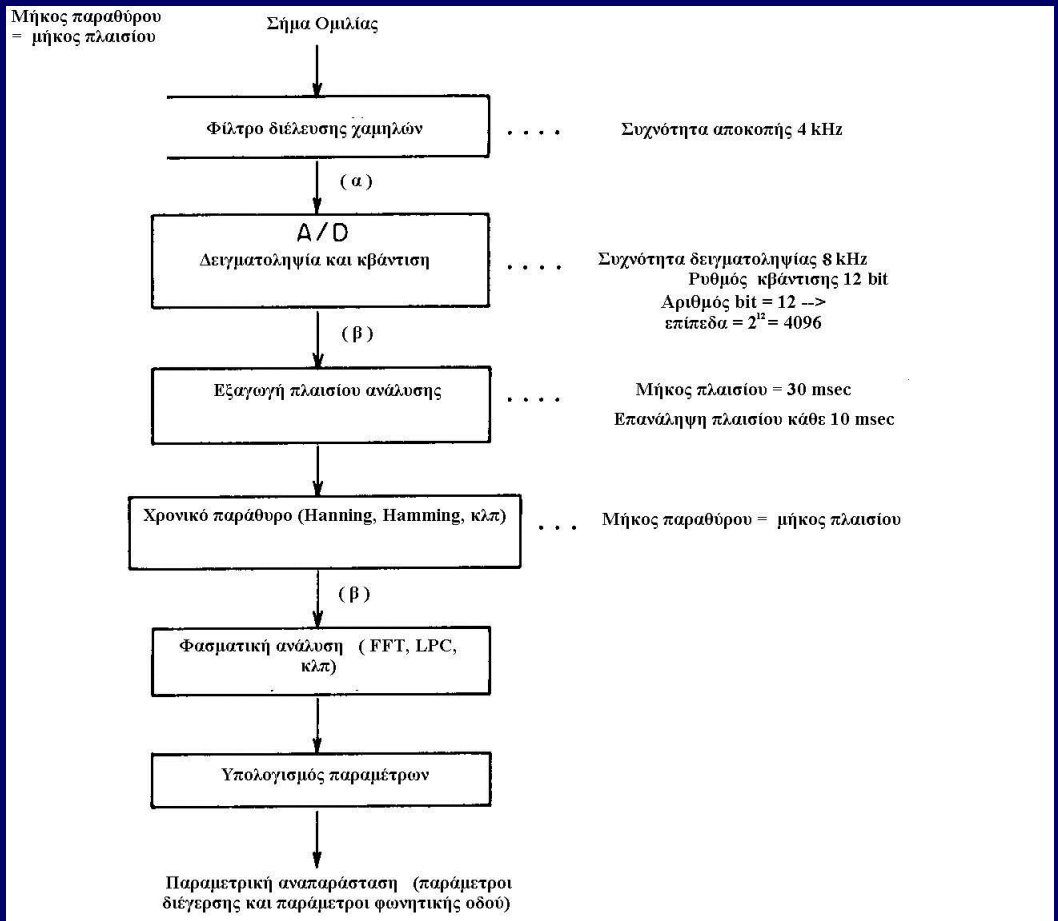
Τύπος	Μέθοδος ανάλυσης	Παράμετροι	Χαρακτηριστικά γνωρίσματα
Μη παραμετρική ανάλυση NPA (non parametric analysis)	Βραχύχρονη αυτοσυ-σχέτιση	φ(m)	Συγκερασμός της φασματικής περιβάλλουσας και της λεπτής δομής. Απλός, εύκολος αλγόριθμος διευκολύνει την πραγμάτωση σε υλική μορφή.
	Βραχύχρονο φάσμα	S(ω)	Πολλαπλασιασμός της φασματικής περιβάλλουσας και της λεπτής δομής. Ταχύς αλγόριθμος μπορεί να πραγματοποιηθεί με FFT.
	Cepstrum	c(τα)	Η φασματική περιβάλλουσα και η λεπτή δομή μπορούν να διαχωριστούν στο πεδίο frequency. Είναι απαραίτητοι 2 FFT και λογαριθμικός μετασχηματισμός.
	Τράπεζα ζωνοπερατών φίλτρων	Rms της εξόδου φίλτρου	Μπορεί να ληφθεί με γενική φασματική περιβάλλουσα. Κατάλληλη για επεξεργασία σε πραγματικό χρόνο.
	Ανάλυση διελεύσεων από το 0	Ρυθμός διελεύσεων από το 0	Μπορούμε να λάβουμε συχνότητα format σε συνδυασμό με την προηγούμενη μέθοδο.

Κυριότεροι μέθοδοι για την ανάλυση των φασμάτων ομιλίας (2/2)

Τύπος	Μέθοδος ανάλυσης	Παρά-μετροι	Χαρακτηριστικά γνωρίσματα
Παραμετρική ανάλυση PA (parametric analysis)	Ανάλυση μέσω σύνθεσης	Forman, εύρος φωνής	Είναι εφικτή η ακριβής μοντελοποίηση. Μπορούμε να λάβουμε συχνότητες formants με ακρίβεια. Είναι απαραίτητη πολύπλοκη επανάληψη
	Κωδικοποίηση γραμμικής πρόβλεψης	α(i)	Απλό ολοπολικό μοντέλο φάσματος. Οι παράμετροι μπορούν να εκτιμηθούν με αυτοσυσχέτιση ή συμμεταβολή χωρίς επανάληψη.
	Μέθοδος μέγιστης πιθανοφάνειας	α(i)	Εγγυημένη σταθερότητα φίλτρου σύνθεσης. Απαραίτητα χρονικά παράθυρα. Αριθμός υπολογισμών ρ^2
	Μέθοδος συμμεταβολής	κ(i)	Η σταθερότητα του φίλτρου σύνθεσης δεν είναι εγγυημένη. Κατάλληλο για βραχύχρονη ανάλυση. Αριθμός υπολογισμών $\alpha\rho^3$
	Μέθοδος PARCOR	k(I)	Κανονική εξίσωση μπορεί να λυθεί με φίλτρο πλέγματος. Αριθμός υπολογισμών $\alpha\rho^2$
	Μέθοδος LSP	w(I)	Καλά χαρακτηριστικά κβάντισης και παρεμβολής. Παρόμοια με formant. Υπολογισμοί περισσότεροι από ότι για PARCOR.

Block διάγραμμα μιας τυπικής διαδικασίας ανάλυσης ομιλίας.

Οι τυπικές τιμές παραμέτρων και παραδείγματα των κυμάτων ομιλίας σε κάθε στάδιο.



Ενέργεια σήματος Ομιλίας

$$\mathrm{E}_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=n}^{n+N-1} [x(m)w(n-m)]^2$$

$$E_n = \sum_{m=-\infty}^{\infty} [x^2(m)w^2(n-m)] = \sum_{m=-\infty}^{\infty} [x^2(m)h(n-m)] = x^2(n) * h(n)$$

$$h(n) = w^2(n)$$

w = Τυπικό χρονικό παράθυρο 10ms ~ 30ms

Αυτοσυσχέτιση και μετασχηματισμός Fourier

Συνάρτηση αυτοσυσχέτισης:

$\varphi(m) = 1/N \sum_{n=0}^{N-1-m} x(n)x(n+|m|)$ (|m| = 0, 1,…, N-1)

όπου N ο αριθμός των δειγμάτων στο βραχύχρονο διάστημα ανάλυσης.

Το μήκος του διαστήματος NT (T =περίοδος δειγματοληψίας) συνήθως τίθεται περίπου στα 30ms. Πιο συγκεκριμένα διαστήματα περίπου των 20 και 40 ms αποφέρουν καλά αποτελέσματα για γυναικείες και ανδρικές φωνές, αντίστοιχα.

Το S(ω) και φ(m) του βραχύχρονου φάσματος συνιστούν το ζεύγος μετασχηματισμού Fourier (θεώρημα Wiener - Khintchine):

$S(\omega) = 1/2\pi \sum_{n=0}^{N-1-m} \varphi(m) \cos \omega m$ και $\varphi(m) = \int_{-\pi}^{\pi} S(\omega) \cos \omega m \, d\omega$

Όπου ω είναι μια κανονικοποιημένη γωνιακή συχνότητα η οποία μπορεί να αναπαρασταθεί από ω = 2πfT (f είναι η πραγματική συχνότητα).

Αυτοσυσχέτιση σήματος ομιλίας

Συνάρτηση αυτοσυσχέτισης:

$$\mathrm{R}_n(k) = \sum_{m=-\infty}^{\infty} [x(m) \cdot w(n-m) \cdot x(m+k) \cdot w(n-(m+k))]$$

$$R_n(k) = R_n(-k) = \sum_{m=-\infty}^{\infty} [x(m) \cdot w(n-m) \cdot x(m-k) \cdot w(n-(m-k))]$$

- Αν $h_k(n) = w(n)w(n+k)$ τότε:

$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(m) \cdot x(m-k)] h_k(n-m) = [x(n)x(n-k)] * h_k(n)$$

Μετασχηματισμός Fourier

Το S(ω) συνήθως υπολογίζεται άμεσα από το σήμα ομιλίας χρησιμοποιώντας το διακριτό μετασχηματισμό Fourier (DFT) ο οποίος πραγματοποιείται από τον αλγόριθμο του γρήγορου μετασχηματισμού Fourier (FFT):

$$S(\omega) = (1/2\pi)^{-\mathrm{T}} \left| \sum_{n=0}^{N-1} x(n) e^{-j\omega n} \right|^2$$

Μετασχηματισμός Fourier και φασματογράφημα:

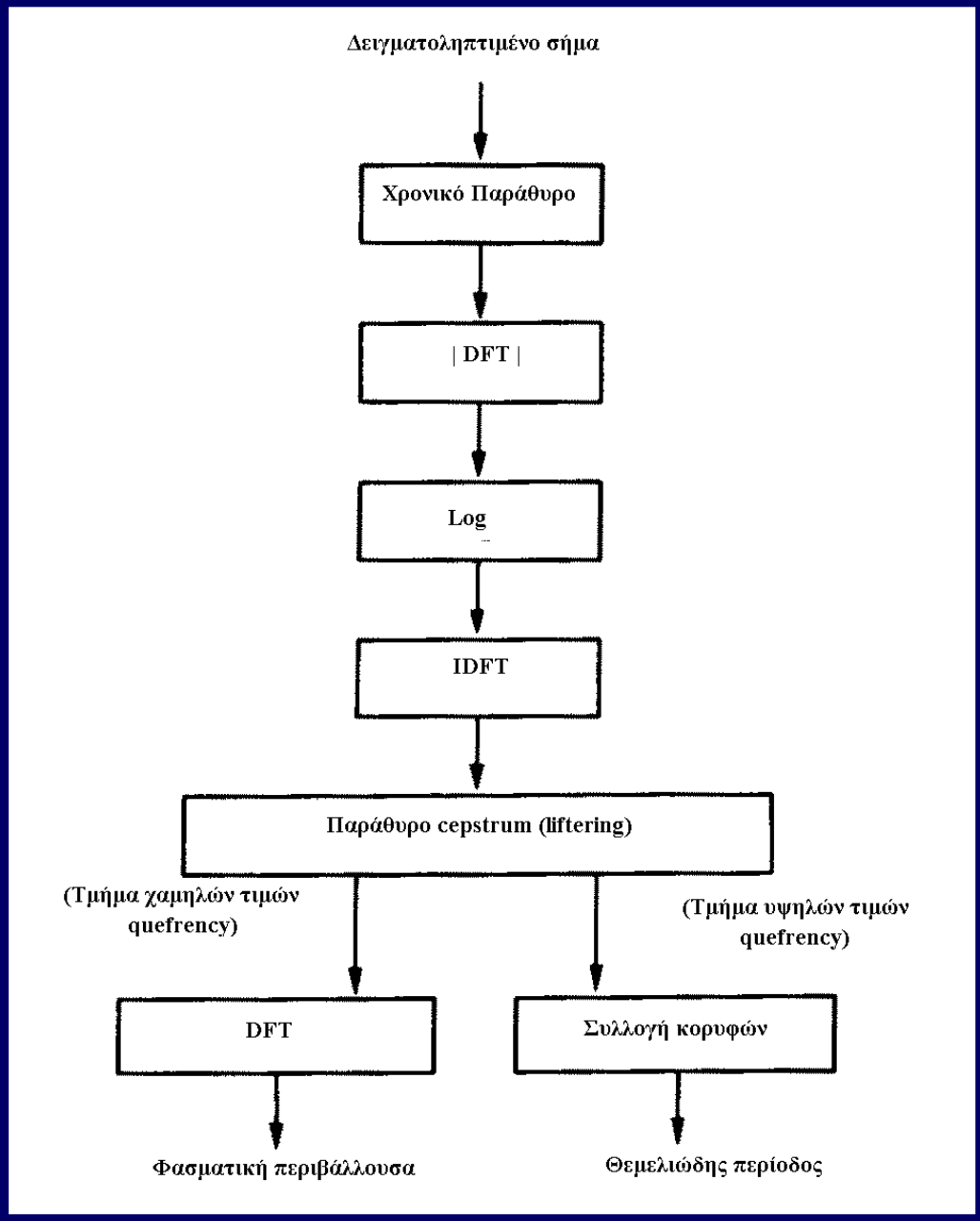
$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m) w(n-m) e^{-j\omega m}$$

Φάσμα Ισχύος:

$$S_n(e^{j\omega}) = \mathrm{X}_n(e^{j\omega}) X_n(e^{j\omega})^* = \left| X_n(e^{j\omega}) \right|^2$$

Cepstrum (1/2)

Block διάγραμμα ανάλυσης cepstrum για την εξαγωγή των φασματικών περιβαλλουσών και της θεμελιώδους περιόδου



Cepstrum (2/2)

$X(\omega) = \mathrm{G}(\omega)\mathrm{H}(\omega) \rightarrow$

$\log|X(\omega)| = \log|\mathrm{G}(\omega)| + \log|\mathrm{H}(\omega)| \rightarrow$

Το cepstrum είναι:

$c(T) = F^{-1}\log|X(\omega)| = \mathrm{F}^{-1}\log|G(\omega)| + \mathrm{F}^{-1}\log|H(\omega)|$

↓

Κορυφή στις υψηλές συχνότητες => liftering

Τεχνική Cepstrum (Homomorphic filtering)

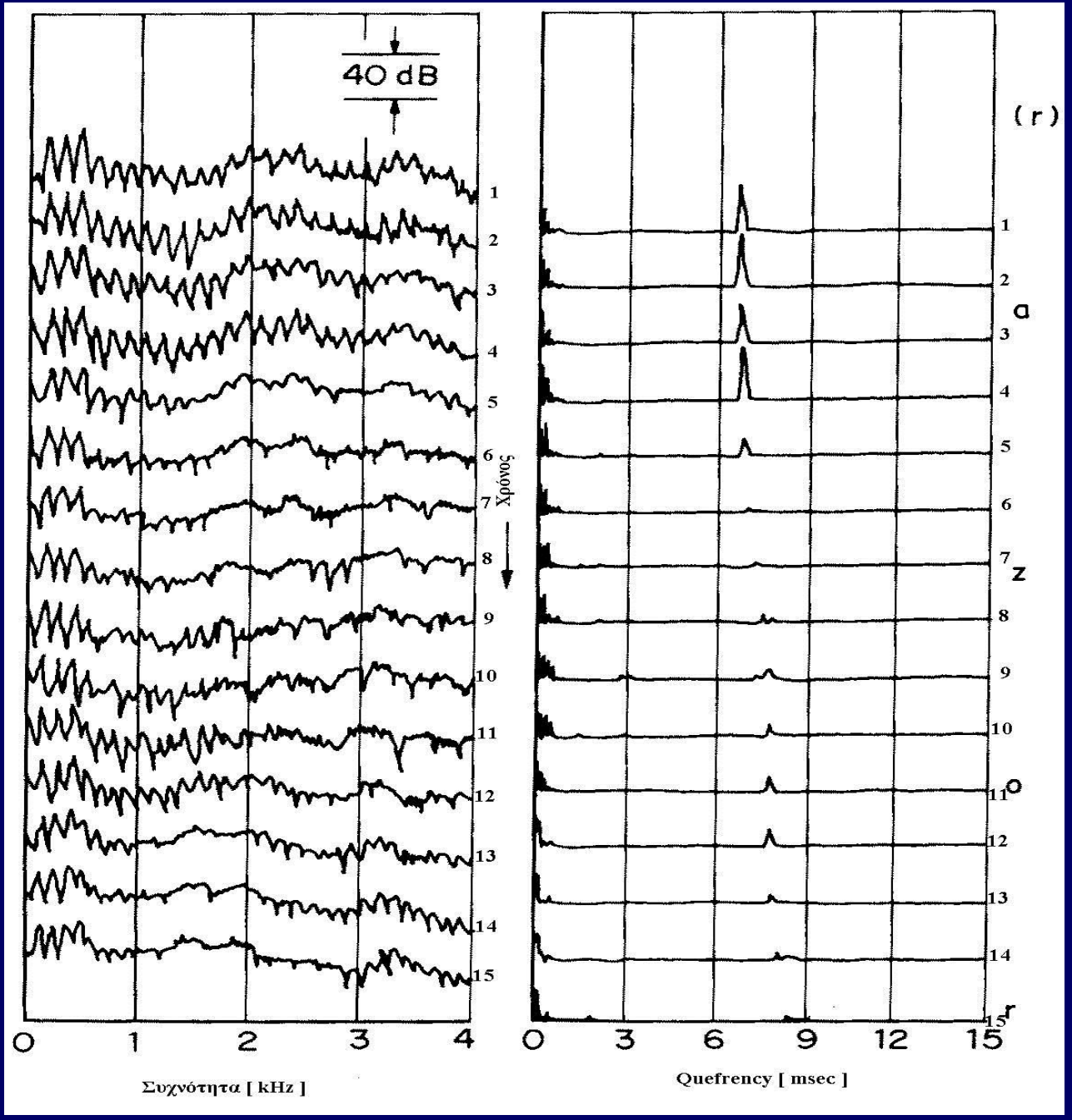
$S(\omega) = \mathrm{H}(\omega)\mathrm{E}(\omega) \rightarrow \log|\mathrm{S}(\omega)|^2 = \log|\mathrm{E}(\omega)|^2 + \log|\mathrm{H}(\omega)|^2$

↓ Μετασχηματισμός Fourier

$F\left[\log|\mathrm{S}(\omega)|^2 = \mathrm{F}[\log|E(\omega)|^2]\right] + \mathrm{F}[\log|\mathrm{H}(\omega)|^2]$

Παραδείγματα βραχύχρονων φασμάτων (αριστερά) και cepstrum (δεξιά) για ανδρική φωνή κατά την εκφώνηση «(r)azor».

Συχνότητα δειγματοληψίας 10 kHz, μήκος παραθύρου Hamming 40ms, διάστημα πλαισίου 10ms.



Liftering

- Low time liftering:

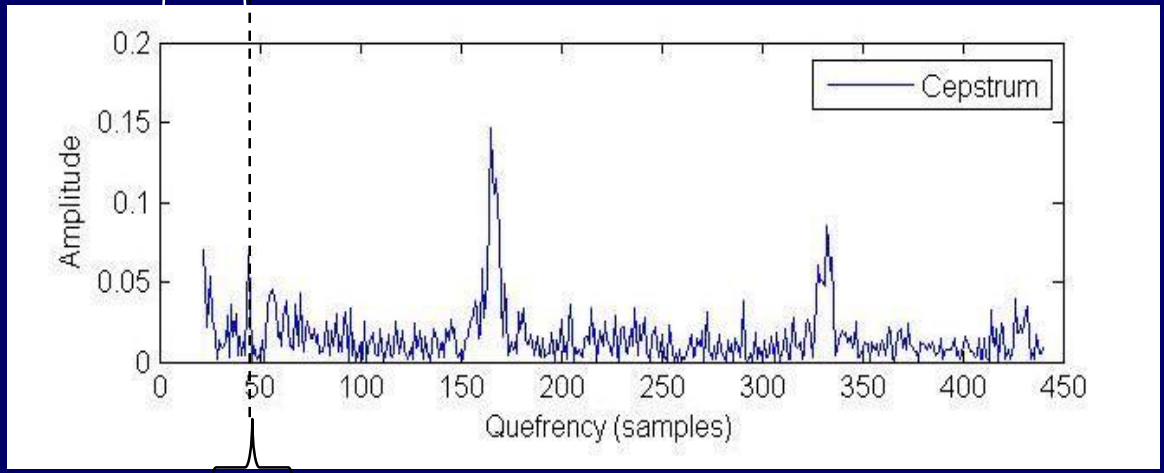
Magnify (or inspect) the low time to find the vocal tract filter cepstrum.

- High time liftering:

Magnify (or inspect) the high time to find the glottal excitation cepstrum (remove this part for speech recognition).

Vocal tract Cepstrum

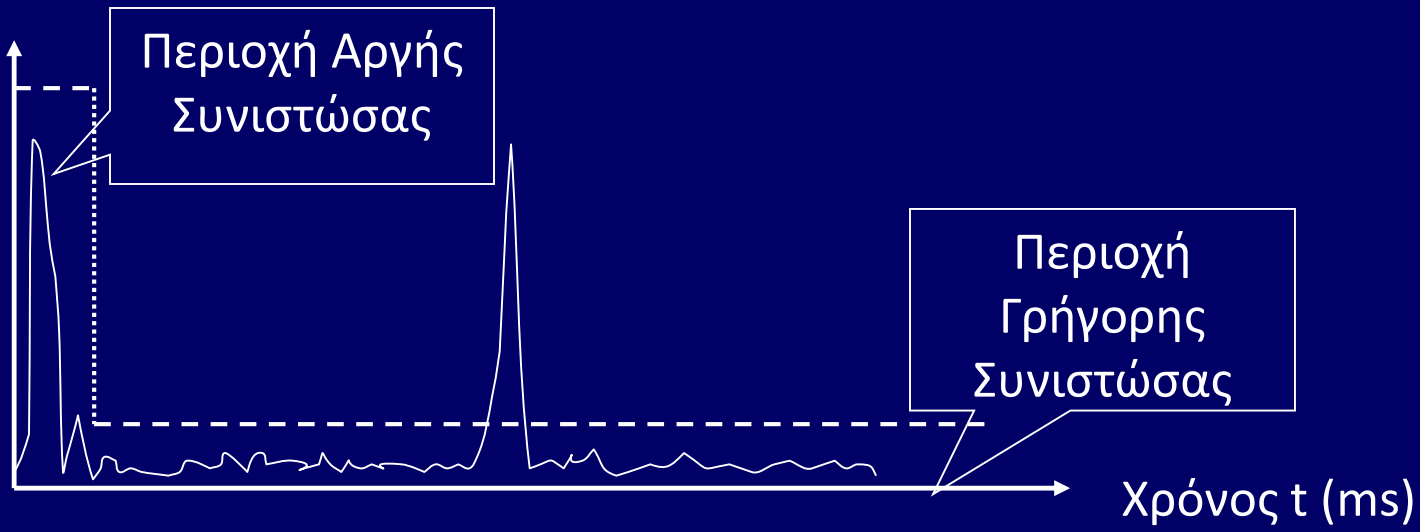
Glottal excitation Cepstrum



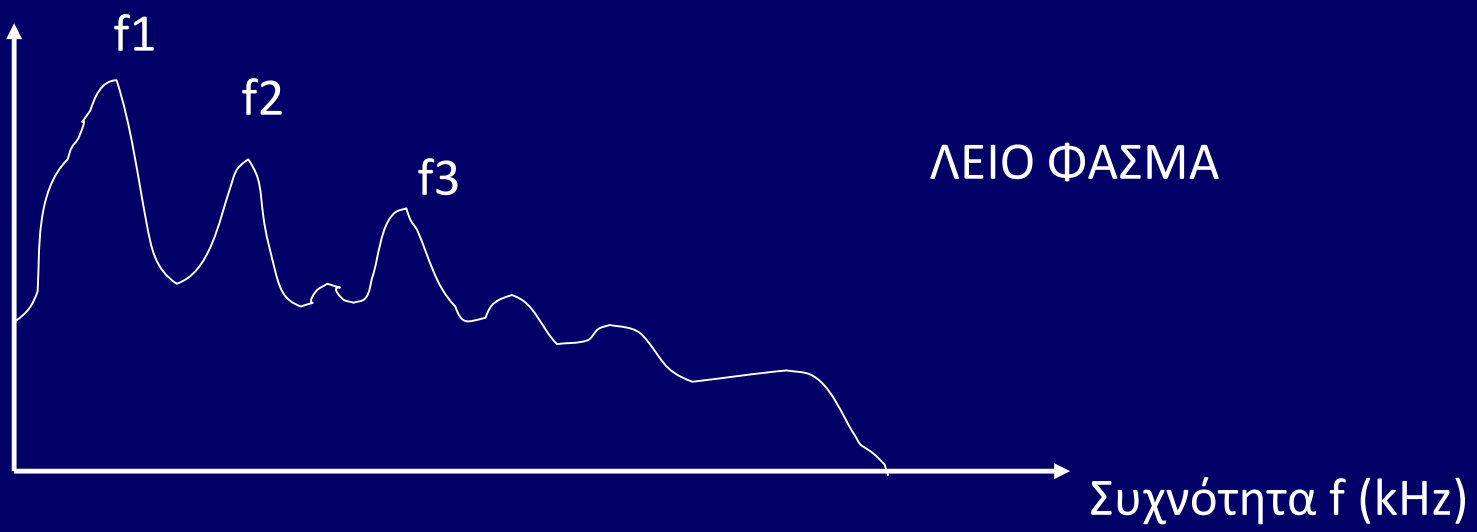
Cut-off Found by experiment

Frequency =FS/ quefrency
FS=sample frequency =22050

a)



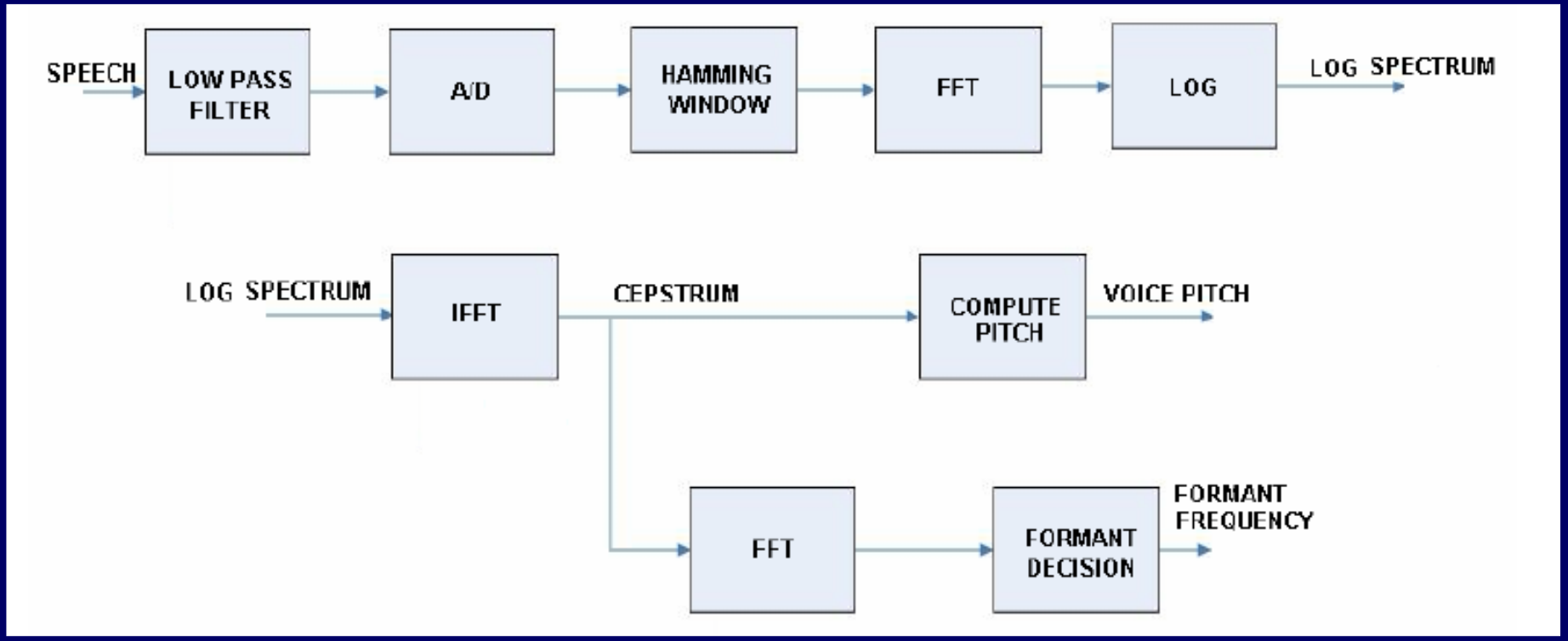
b)



a) το cepstrum ενός φωνήεντος με βάση τον ορισμό

b) λειασμένη εκδοχή του φάσματος μετά από αποκοπή ορισμένων συνιστωσών του cepstrum

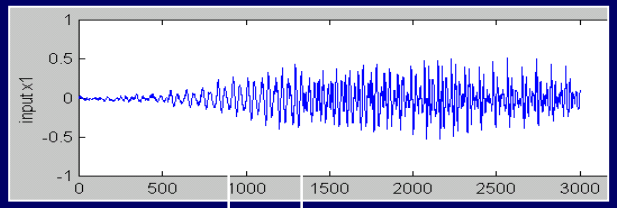
Μπλοκ διάγραμμα της μεθόδου cepstrum



Ανάλυση με Τράπεζα φίλτρων

- Για κάθε χρονικό πλαίσιο (10-30 ms), θα πρέπει να υπολογίζεται ένας αριθμός εξόδων φίλτρων (τυπική επικάλυψη πλαισίων 5ms).
- Υπάρχουν πολλοί διαφορετικοί τρόποι καθορισμού των φίλτρων στην τράπεζα (ομοιόμορφα και μη-ομοιόμορφα φίλτρα).

Σήμα εισόδου



Time frame i — 30ms → Filter outputs (v1,v2,…)

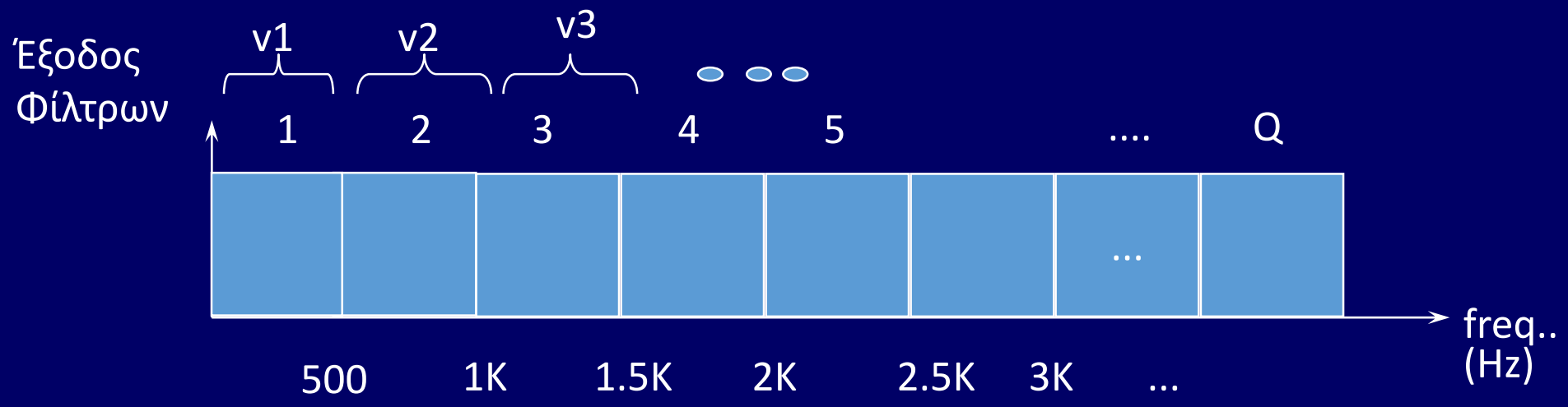
Time frame i+1 — 30ms → Filter outputs (v'1,v'2,…)

Time frame i+2 — 30ms → Filter outputs (v''1,v''2,…)

5ms

Τράπεζα Ομοιόμορφων φίλτρων

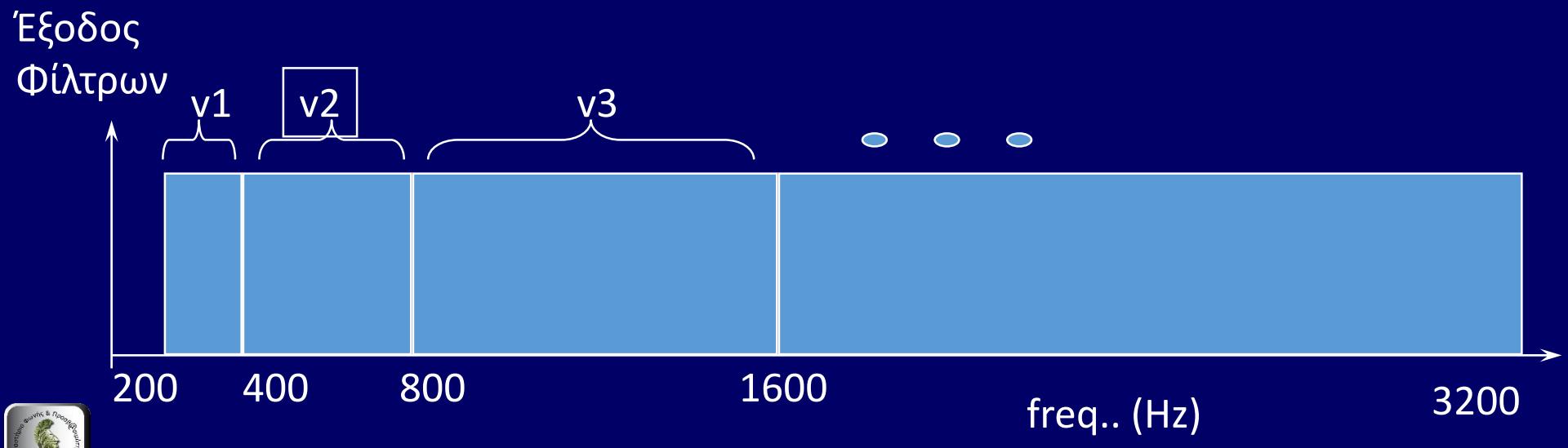
- Εύρος ζώνης B= Συχνότητα Δειγματοληψίας (F_s) / αριθμό φίλτρων στην τράπεζα (N).
- παράδειγμα: F_s =8 Kz, N=20 τότε B=400Hz.
- απλό στην υλοποίηση, αλλά όχι τόσο χρήσιμο.



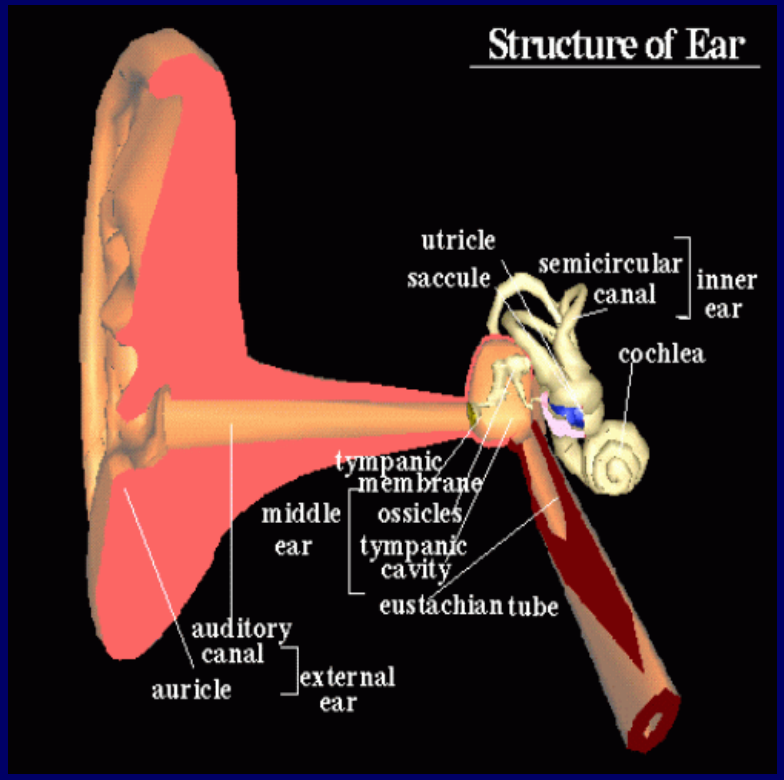
Τράπεζα μη-ομοιόμορφων φίλτρων

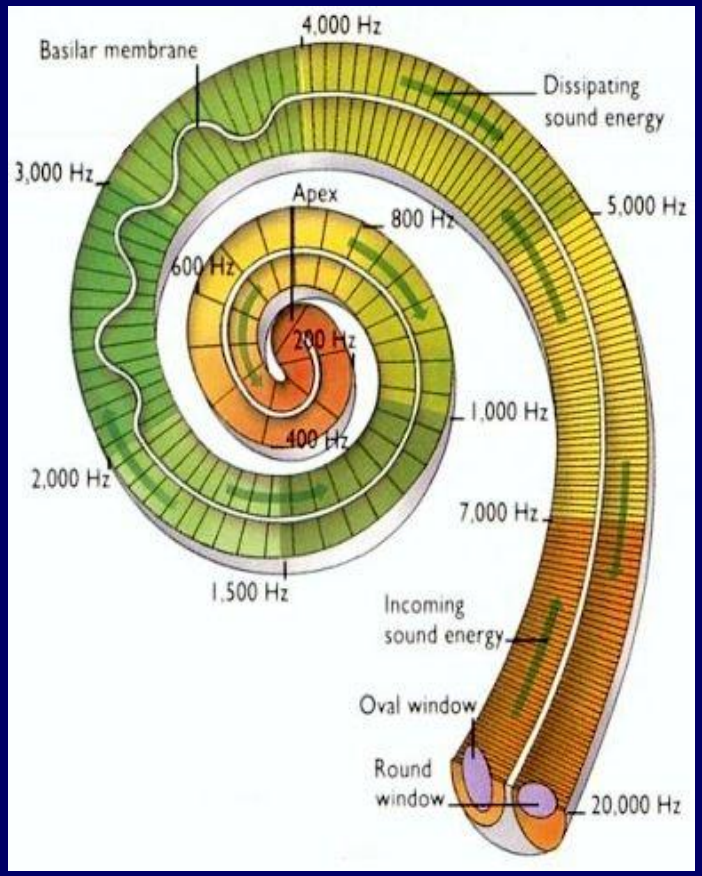
Λογαριθμική κλίμακα συχνοτήτων: κοντά στη λειτουργία του αυτιού παράδειγμα:

Center frequency	300	600	1200	2400
Bandwidth	200	400	800	1600



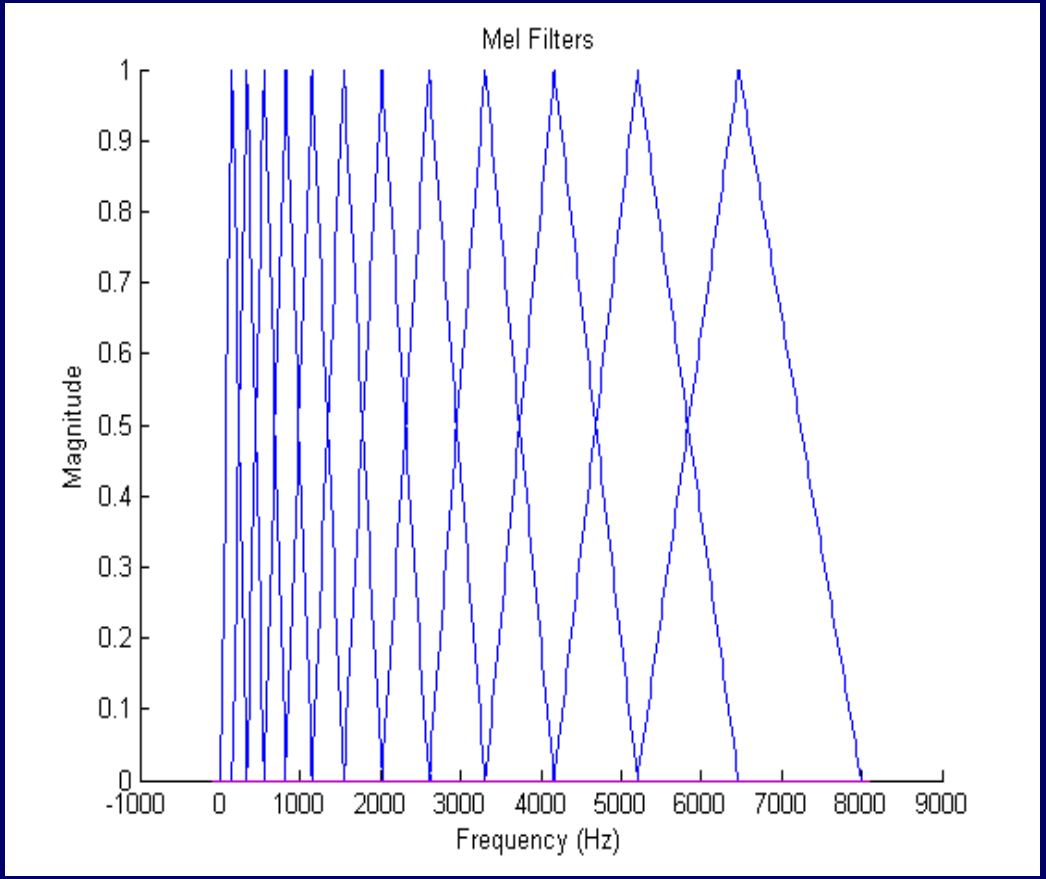
Εσωτερικό αυτιού και κοχλίας





Mel τράπεζα φίλτρων (υπολογίστηκε με ψυχοακουστικά πειράματα)

- Freq. lower than 1 KHz has narrower bands (and in linear scale).
- Higher frequencies have larger bands (and in log scale).
- More filter below 1KHz.
- Less filters above 1KHz.



Mel scale

- Measures relative strength in perception of different frequencies.

- *The mel scale, (named by Stevens, Volkman and Newman in 1937) is a perceptual scale of pitches judged by listeners to be equal in distance from one another.*
- *The reference point between this scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 mels to a 1000Hz tone, 40 dB above the listener's threshold. ….*
- *The name mel comes from the word melody to indicate that the scale is based on pitch comparisons.*

Ακουστική τράπεζα φίλτρων βασισμένη στην κλίμακα Mel

Ψυχοακουστική: το αυτί του ανθρώπου αντιλαμβάνεται την ομιλία με μη γραμμική κλίμακα συχνοτήτων σύμφωνα με τη σχέση:

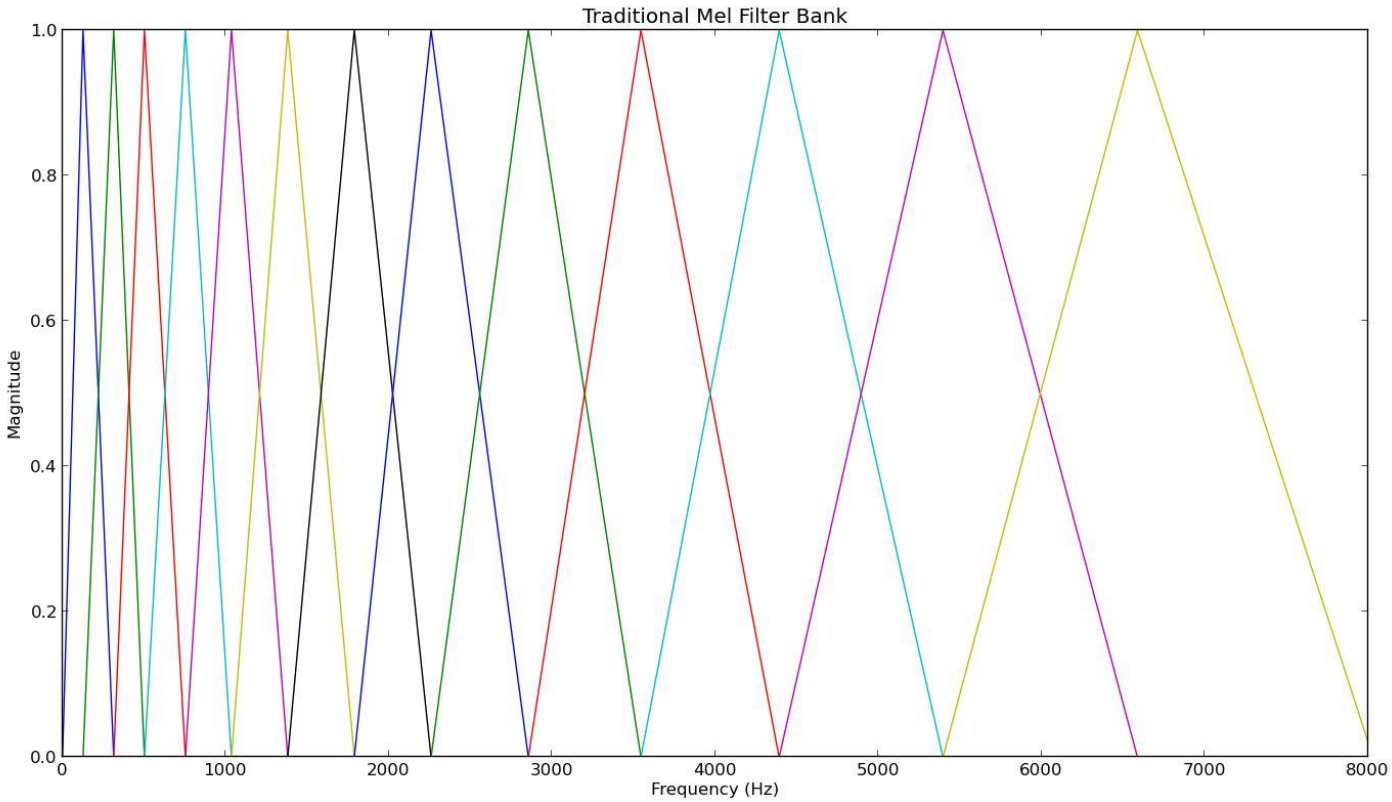
$$mel = 2595 \log_{10}(1 + f/700)$$

Mel: Η υποκειμενική κλίμακα συχνοτήτων σε Hz (αίσθηση).

f: Η αντικειμενική κλίμακα συχνοτήτων σε Hz.

Τράπεζα φίλτρων για ανάλυση ομιλίας στην κλίμακα mel.

20 τριγωνικά ζωνοπερατά φίλτρα στην περιοχή 0 – 4 kHz με μη γραμμική απόσταση και εύρος ζώνης.


Traditional Mel Filter Bank
Magnitude
Frequency (Hz)
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
6000
7000
8000

Κλίμακα Bark

$B = 13 \arctan(0{,}00076\, f) + 3{,}5 \arctan[(f/7500)^2]$

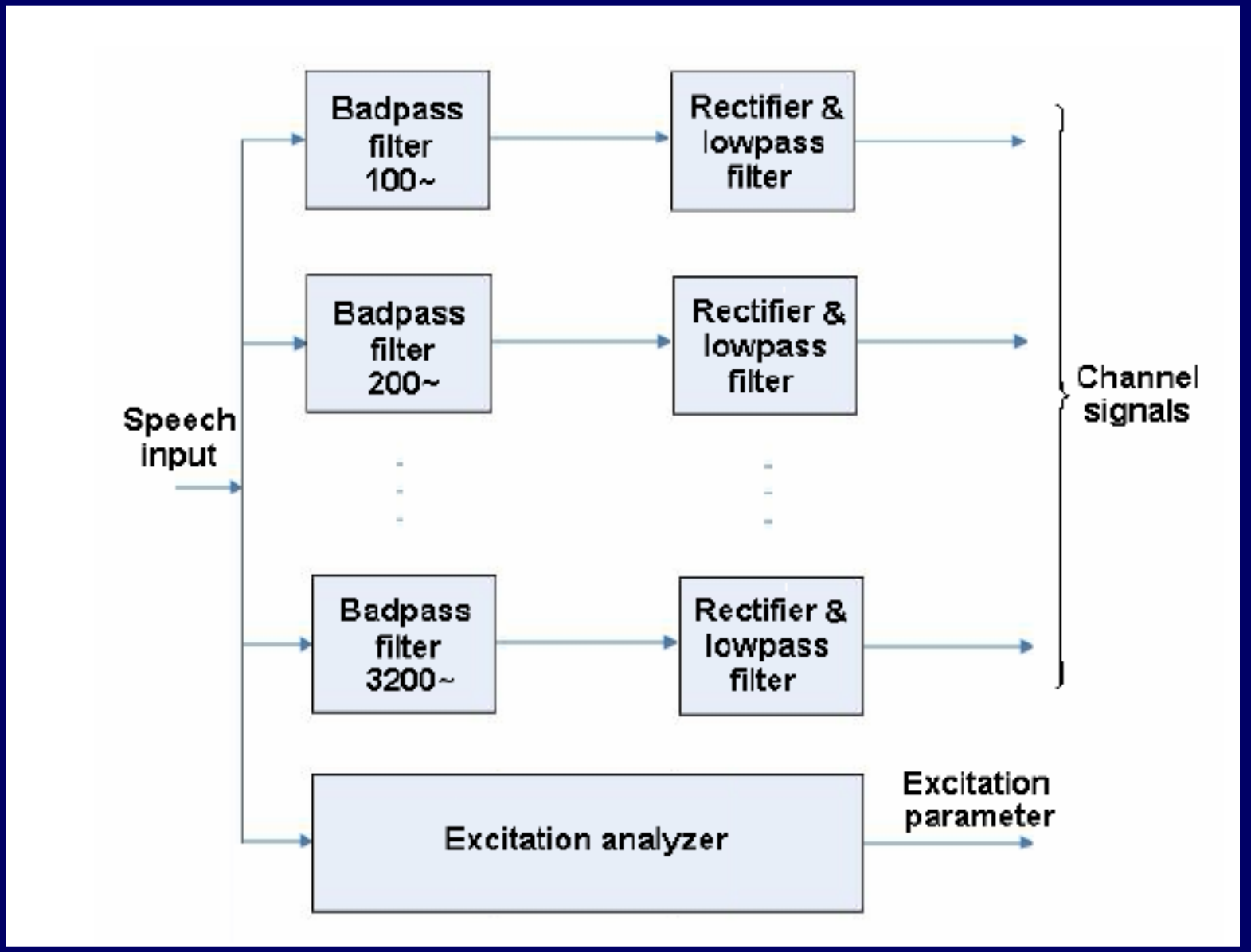
$$Bark(f) = \left[26{,}8/\left(1 + (1960/f)\right) - 0{,}53\right]$$

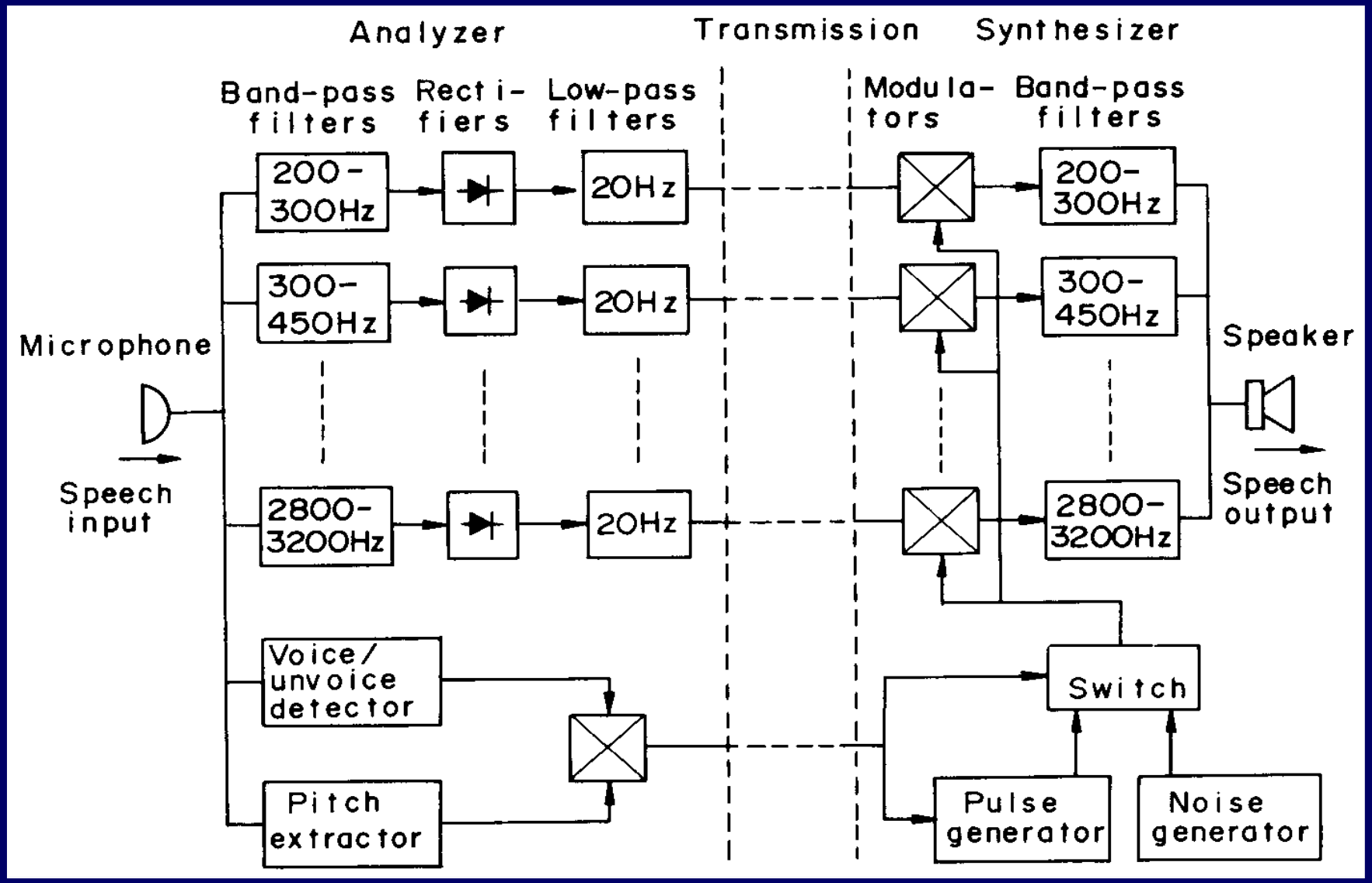
(από το όνομα του Barkhausen, ο οποίος πρότεινε την κλίμακα loudness)

Βασίζεται σε πειράματα ακουστικού masking.

Bark band #	Edge (Hz)	Center (Hz)
1	100	50
2	200	150
3	300	250
4	400	350
5	510	450
6	630	570
7	770	700
8	920	840
9	1080	1000
10	1270	1170
11	1480	1370
12	1720	1600
13	2000	1850
14	2320	2150
15	2700	2500
16	3150	2900
17	3700	3400
18	4400	4000
19	5300	4800
20	6400	5800
21	7700	7000
22	9500	8500
23	12000	10500
24	15500	13500

Ανάλυση Ομιλίας με τράπεζα φίλτρων





Δομή Vocoder Καναλιού

Ανάλυση σήματος ομιλίας βασισμένη στον ρυθμό διελεύσεων από το 0

Zero Crossing Rate (ZCR)

$$Z_n = \sum_{m=-\infty}^{\infty} |sng[x(n)] - sng[x(n-1)]|w(n-m)$$

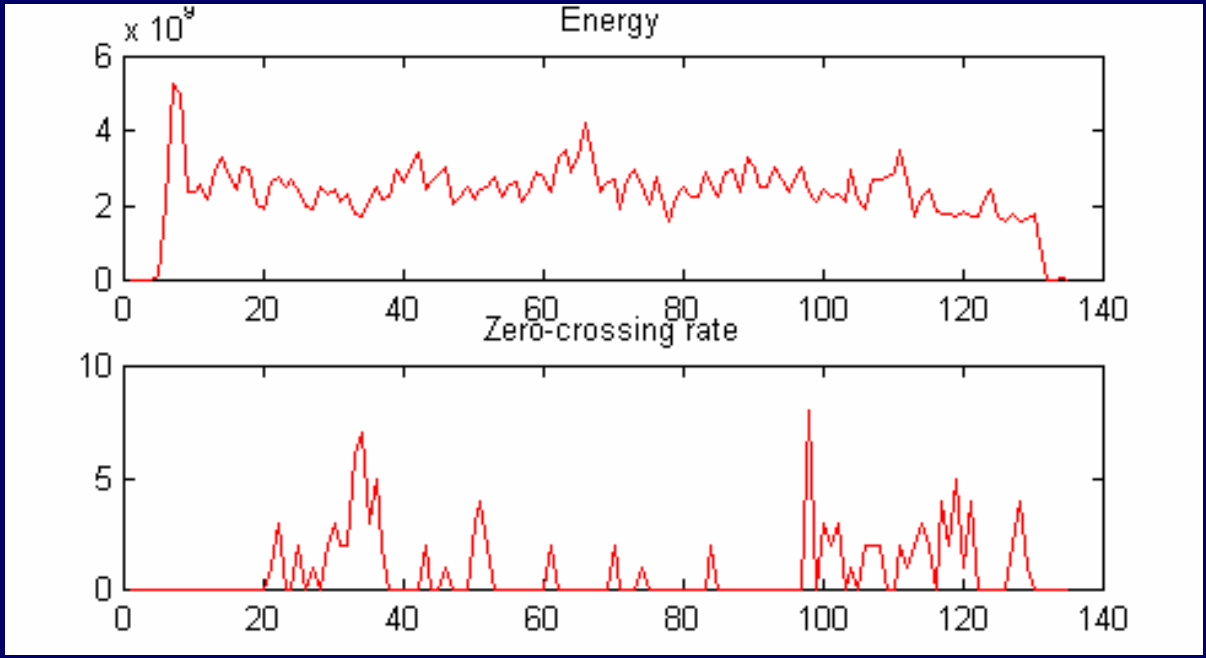
$$= |sng[x(n)] - sng[x(n-1)]|w(n)$$

Όπου:

$$sng[x] = \begin{cases} 1 \ (x \geq 0) \\ 0 \ (x < 0) \end{cases}$$

$$w(n) = \begin{cases} 1/2N \ (0 \leq n \leq N-1) \\ 0 \qquad\qquad other \end{cases}$$

Ενέργεια και ZCR σήματος ομιλίας



Δομή Συστήματος Ανάλυσης μέσω Σύνθεσης

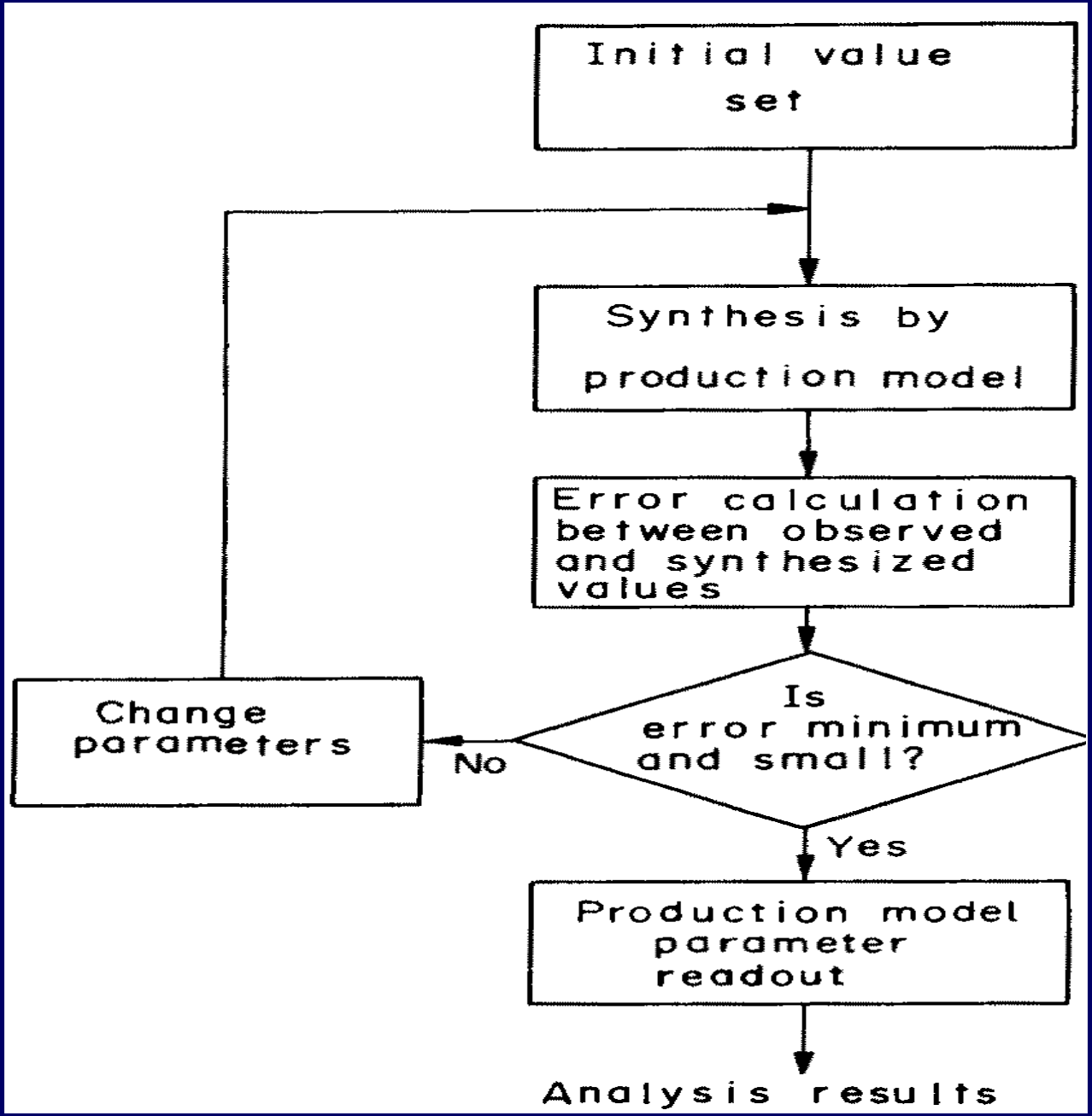
Στα συστήματα Ανάλυσης μέσω Σύνθεσης, το κύμα ομιλίας αναπαράγεται με τη χρήση παραμέτρων πηγής και άρθρωσης οι οποίες εξάγονται με βάση το γραμμικά διαχωρίσιμο ισοδύναμο κύκλωμα για το μηχανισμό παραγωγής ομιλίας.

Αυτές οι παράμετροι ορίζουν 4 είδη πληροφοριών:

1) Διαχωρισμό ανάμεσα σε έμφωνο ήχο (πηγή παλμού) και άφωνο ήχο (πηγή θορύβου).
2) Θεμελιώδης περίοδος ή θεμελιώδης συχνότητα των έμφωνων ήχων.
3) Πλάτος κύματος πηγής.
4) Χαρακτηριστικά συντονισμού του γραμμικού φίλτρου.

Οι πρώτες τρεις παρέχουν πληροφορίες για την πηγή, ενώ το τελευταίο παραμετρικό σύνολο δίνει πληροφορίες φασματικής περιβάλλουσας (δηλ. για την άρθρωση).

Αρχές Μεθόδου Ανάλυση μέσω Σύνθεσης A-b-S



Σύγκριση Κωδικοποίησης Κυματομορφών και της Μεθόδου Ανάλυσης μέσω Σύνθεσης

	Κωδικοποίηση Κυματομορφών	Ανάλυση μέσω Σύνθεσης
Πληροφορίες Κωδικοποίησης	Κυματομορφή	Φασματική Περιβάλλουσα (βραχύχρονο φάσμα) και πληροφορίες πηγής (pitch, πλάτος κύματος και έμφωνη / άφωνη)
Ρυθμός Κβάντισης / bit	9,6–64 kbps (μέσης ή ευρείας ζώνης)	2,4 – 4,8 kbps (στενής ζώνης)
Αντικειμενικός Στόχος Κωδικοποίησης	Κάθε ήχος	Η φωνή ενός ομιλητή
Μέτρο Αξιολόγησης	SNR	Φασματική Παραμόρφωση
Προβλήματα	Δύσκολο να μειωθεί ο ρυθμός ανά bit σε διάστημα στενής ζώνης	Ευάλωτο σε σφάλμα θορύβου και μετάδοσης. Περιορισμένη ποιότητα ομιλίας. Περίπλοκη επεξεργασία.
Παραδείγματα	Κωδικοποίηση πεδίου Χρόνου: PCM, ADPCM, DM, Πεδίο Συχνότητας: SBC, ATC	Vocoder Καναλιού, formant vocoder, vocoder φάσης, LPC (PARCOR, LSP) vocoder, cepstrum vocoder

Κυριότερα Παραδείγματα Ανάλυσης μέσω Σύνθεσης

Είδη Vocoder	Εισηγητής	Μέθοδος Ανάλυσης	Παράμετροι γνωρισμάτων	Χωρητικότητα καναλιού
Vocoder καναλιού	H. Dudley (1939)	Ανάλυση τράπεζας ζωνοπερατού φίλτρου	Πλάτος εξόδου του φίλτρου	300 Hz (αναλογικό) 2400bps(ψηφιακό)
Formant vocoder	W.A. Munson (1950)	Ανάλυση τράπεζας Ζωνοπ. φίλτρου, ανάλ. Zerocrossing	Πλάτος εξόδου του φίλτρου, ρυθμός zero-crossing	300 Hz ή 2400 bps
Vocoder ταύτισης μορφών	C.P. Smith (1957)	Ανάλυση τράπεζας ζωνοπερατού φίλτρου	Φασματική μορφή φωνημάτων	900 bps
Vocoder αυτοσυσχετισμού	M.R.Sshroeder (1959)	Ανάλυση βραχύχρονης Αυτοσυσχέτισης	Συνάρτηση αυτοσυσχέτισης (Μ)	400 Hz
Vocoder φάσης	J.Flanagan (1966)	Ανάλυση τράπεζας ζωνοπερατού φίλτρου	Πλάτος και φάση της εξόδου φίλτρου	1500 Hz 7200-9600bps
Μέγ. πιθανο-φάνειας	F.Itakura S.Saito-1969	Μέθοδος μέγιστης Πιθανοφάνειας	Συντελεστές γραμμικής πρόβλεψης αi	5400 bps
ΟμοιομορΦικός	Oppenheim-1969	Ανάλυση cepstrum	cepstrum c(τ)	7800 bps
PARCOR	Itakura/Saito 1969	Ανάλυση PARCOR	Συντελεστές PARCOR	2400-9600bps
LPCVocoder	B.S.Atal(1971)	Μέθοδος συμμεταβλητών	Συντελεστές γραμμικής πρόβλεψης αi	3600 bps
LSP Vocoder	Itakura/Sugamura1979)	Ανάλυση γραμ.φάσματος	LSP ωi	1600-4800bps

Συνεζευγμένη χρονοσυχνοτική ανάλυση (1/2)

Γενίκευση παραδοσιακού φάσματος ισχύος:

$$P(t,\omega) = \int R(t,\mathrm{T}) e^{-j\omega \mathrm{T}} d\mathrm{T}$$

Αν η χρονικά εξαρτώμενη συνάρτηση αυτοσυσχέτισης είναι:

$$\boldsymbol{R(t,T) = s(t+T/2)s^*(t-T/2)}$$

Τότε λαμβάνουμε το χρονικά εξαρτώμενο φάσμα ισχύος = κατανομή Wigner Ville:

$$WVD(t,\omega) = \int s(t+T/2)s^*(t-T/2)\, e^{-j\omega t} dT$$

Αν λάβουμε το μετασχηματισμό Fourier:

$$\mathrm{AF}(\theta,\mathrm{T}) = \int s(t+T/2)s^*(t-T/2)\, e^{-j\theta T} d\mathrm{t}$$

Που ονομάζεται: Symmetric ambiguity function.

Συνεζευγμένη χρονοσυχνοτική ανάλυση (2/2)

Γενικευμένη χρονικά εξαρτώμενη συνάρτηση αυτοσυσχέτισης:

$$R(t,T) = (1/2\pi) \int \mathrm{AF}(\theta,\mathrm{T})\mathrm{F}(\theta,\mathrm{T})e^{j\theta\mathrm{T}}\, d\theta$$

Όπου Φ(θ,T) η συνάρτηση kernel:

$$R(t,T) = F^{-1}[AF(\theta,T)] \otimes F^{-1}[\Phi(\theta,T)] =$$
$$= [s(t+T/2)s^*(t-T/2)] \otimes \varphi(t,T) =$$
$$= \int s(u+Ta/2)s^*(u-T/2)\varphi(t-u,T)du$$

Όπου ϕ(t,T) ο ανάστροφος μετασχηματισμός Fourier του Φ(θ,Τα).

Cohen's Class:

$$C(t,\omega) = (1/2\pi) \iint AF(\theta,\mathrm{T})\Phi(\theta,\mathrm{T})e^{j(\theta t-\omega\mathrm{T})}d\theta dT$$

Κατανομή Choi – Williams

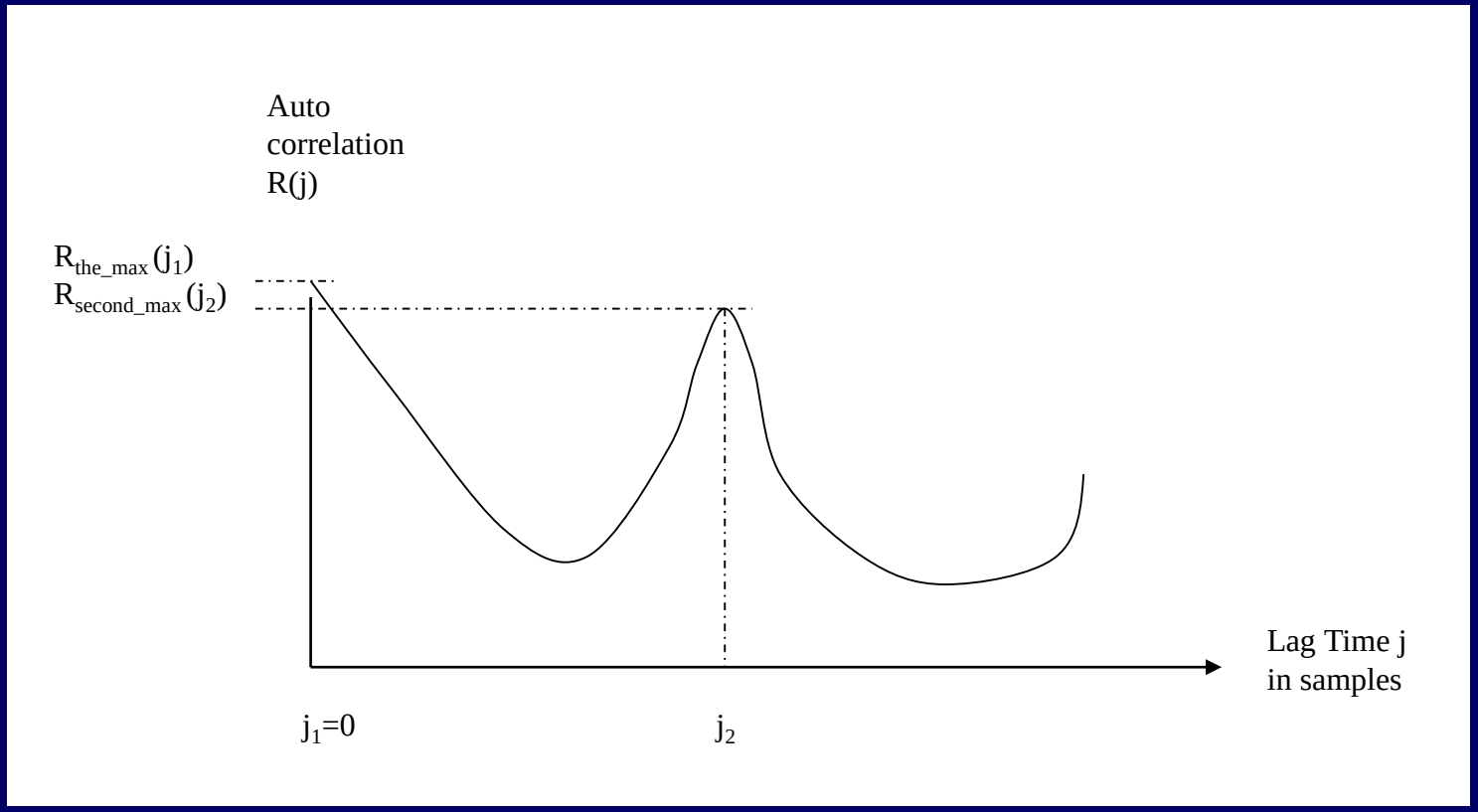
STFT: Short Time Fourier Transform.

Κατηγοριοποίηση των Μεθόδων Εξαγωγής Μουσικού Τόνου (pitch)

Κατηγοριοποίηση	Μέθοδος εξαγωγής pitch	Βασικά Χαρακτηριστικά
Επεξεργασία Κυματομορφών	Μέθοδος παράλληλης επεξεργασίας	Χρησιμοποιεί τον κανόνα πλειοψηφίας για τις περιόδους που εξάγονται με ανιχνευτές κορυφών κυματομορφών.
	Μέθοδος μείωσης δεδομένων	Αφαιρεί επιφανειακά δεδομένα κυματομορφών και αφήνει παλμούς μουσικού τόνου.
	Μέθοδος μετρήσεων διελεύσεων από το 0	Χρησιμοποιεί επαναληπτικές μορφές στο ρυθμό διελεύσεων απ'το 0 των κυματομορφών.
Επεξεργασία Αυτοσυσχέτισης	Μέθοδος αυτοσυσχέτισης	Εφαρμόζει ψαλίδισμα των κέντρων και κορυφών για την εξομάλυνση του φάσματος.
	Μέθοδος τροποποιημένης αυτοσυσχέτισης (MACF)	Χρησιμοποιεί συνάρτηση αυτοσυσχέτισης για το υπολειπόμενο σήμα της ανάλυσης LPC. Ο υπολογισμός υλοποιείται με LPF και πόλωση.
	Αλγόριθμος SIFT	Εφαρμόζει ανάλυση LPC για εξομάλυνση του φάσματος μετά την υποδειγματοληψία στο κύμα.
	Μέθοδος AMDF (Average Magnitude Difference Function)	Χρησιμοποιεί μία μέση διαφορική συνάρτηση.
Επεξεργασία Φάσματος	Μέθοδος cepstrum	Διαχωρίζει τη φασματική περιβάλλουσα και τη λεπτή δομή με αντίστροφο μετασχηματισμό Fourier
	Μέθοδος περιοδικού ιστογράμματος	Χρησιμοποιεί ιστόγραμμα για αρμονικές συνιστώσες στο φασματικό πεδίο.

Μέθοδος αυτοσυσχέτισης για τον υπολογισμό του pitch

When a segment of a signal is correlated with itself, the distance (=Lag_time_in_samples) between the positions of the maximum and the second maximum correlation is defined as the fundamental period (1/pitch_frequency) of the signal.



Η θεμελιώδης συχνότητα μπορεί να υπολογιστεί:

$$f_0 = \frac{1}{Lag\_time\_in\_samples} = \frac{1}{j_2 - j_1}$$

$$\frac{1}{Lag\_time\_in\_samples \times sampling\_priod} = \frac{sampling\_frequency}{j_2 - j_1}$$

Μέθοδος τροποποιημένης αυτοσυσχέτισης για τον υπολογισμό του pitch
Modified Auto-Correlation Function method (MACF): Auto-Correlation Method enhanced by Center clipping

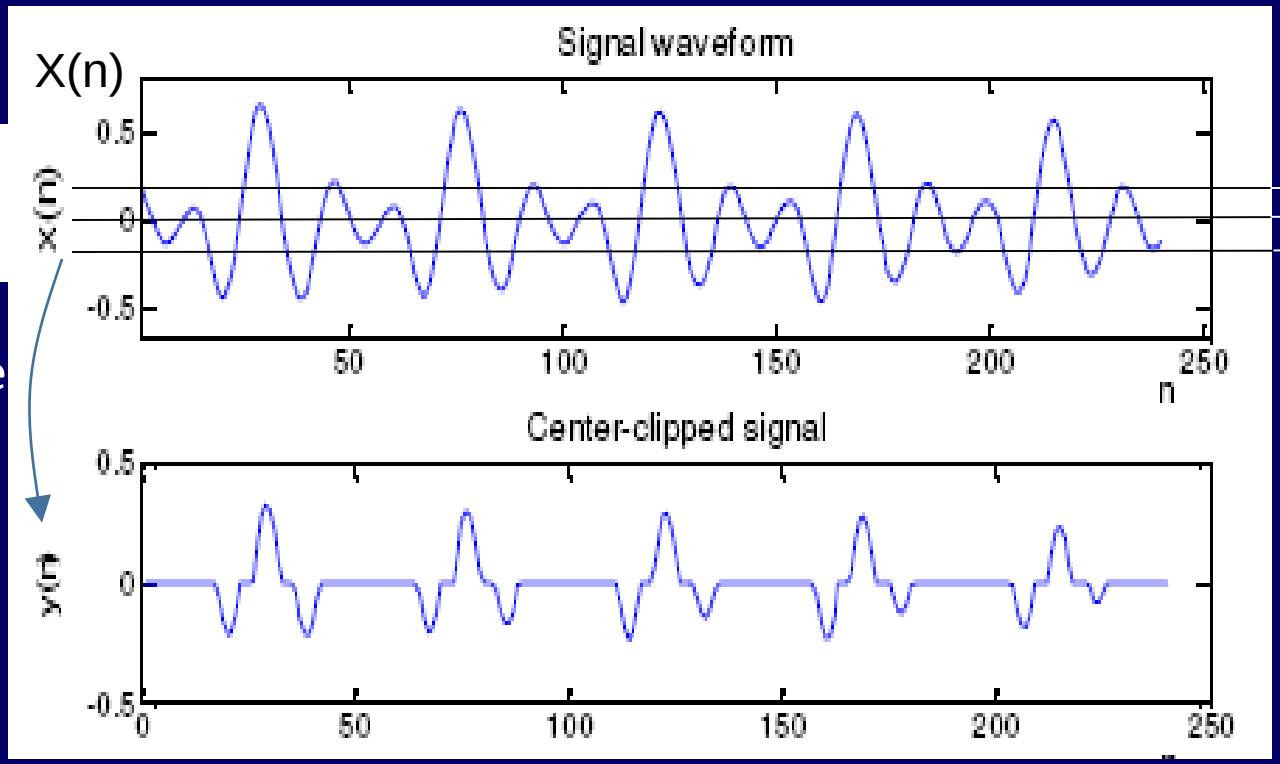
- It will give more accurate result because higher frequency signals will not interfere with the result

$$y(n) = clc[x(n)] \begin{cases} (x(n) - C_L), x(n) \geq C_L \\ 0 \quad , |x(n)| < C_L \\ (x(n) - C_L), x(n) \leq C_L \end{cases}$$

$$R'(m) = \sum_{n=0}^{N-1-m} y(n) \cdot y(n+m), 0 \leq m \leq M_0$$

clc(x)=Cut (remove) the middle part

y(n) =clc(x)



Typical C_L =1/4 peak-to-peak of X

Finding pitch by center clipping

X(n)

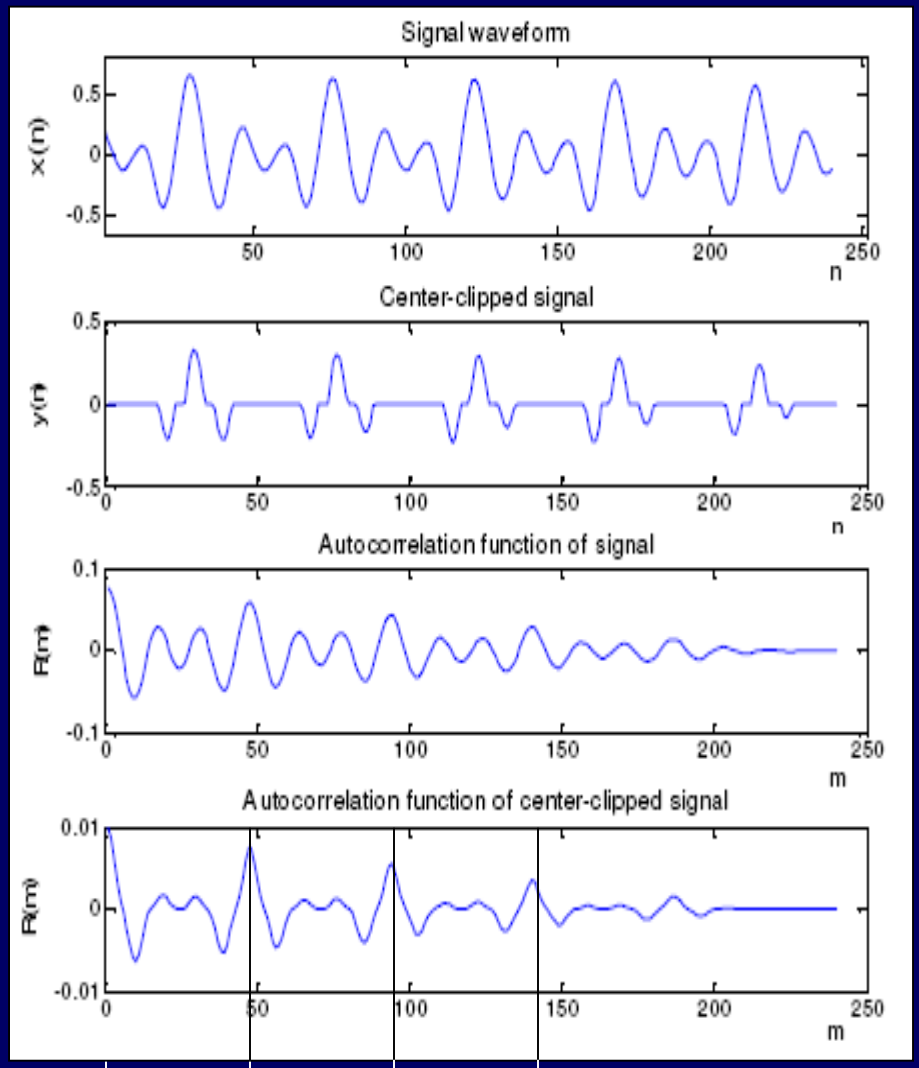
Y(n)= Center Clipped

- In R(m) auto correlation of x(n), it is not easy to pick peaks.

R(m)

- In R'(m), auto correlation of clipped signal y(n)=clc{x(n)}, peaks are easy to pick.

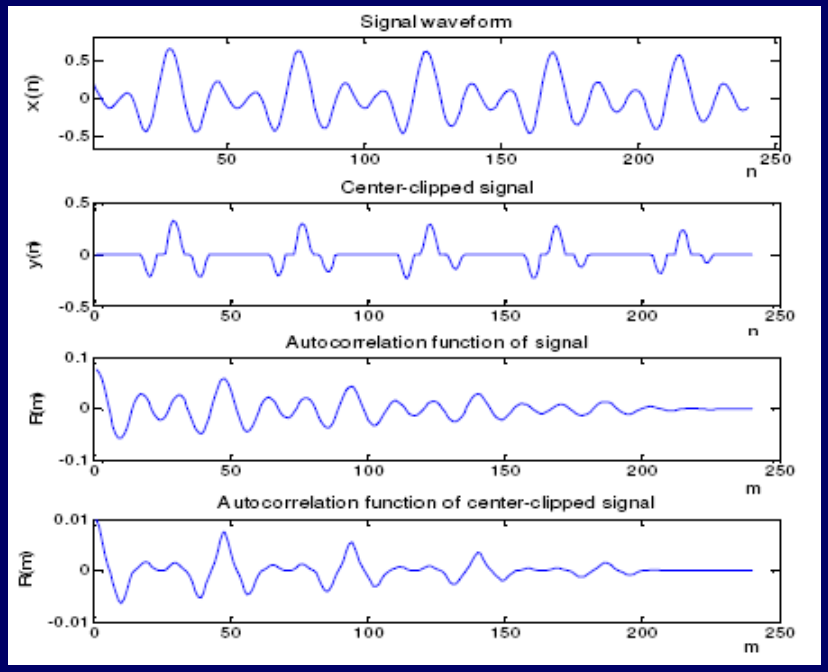
R'(m)

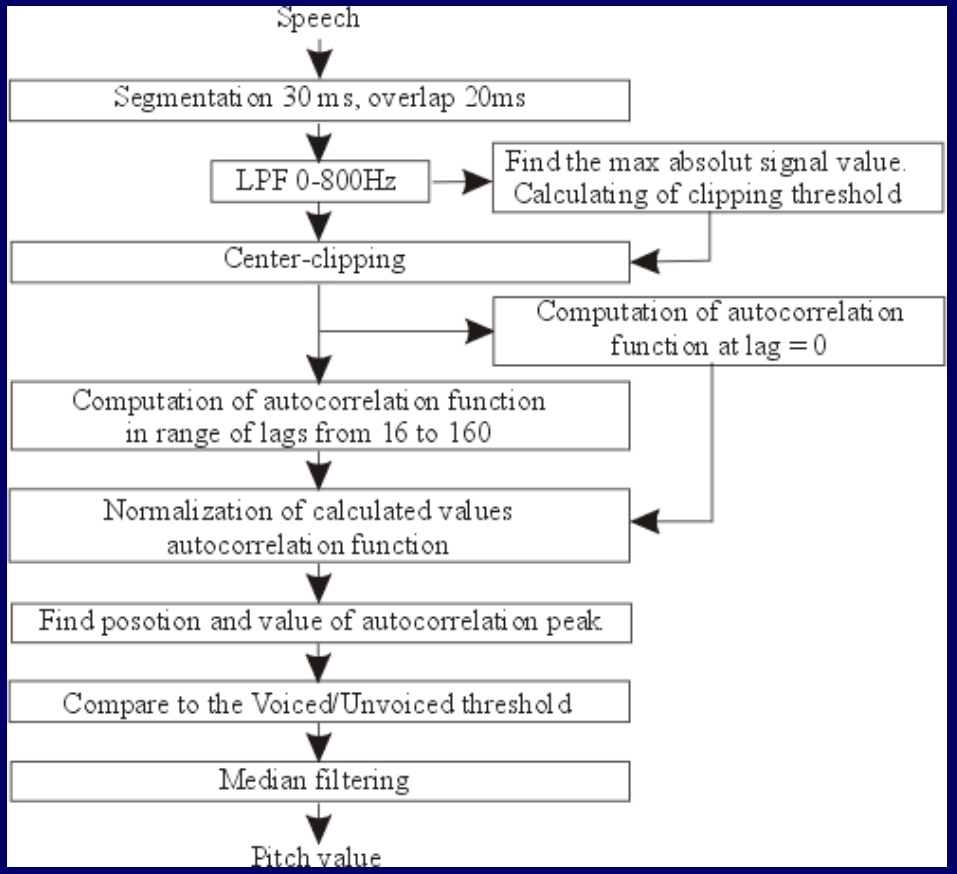


T1 T2 T3

T=mean(T1,T2,T3)=
Period=1/(pitch_frequency)

The MACF (Modified Autocorrelation function) algorithm
για τον υπολογισμό του pitch

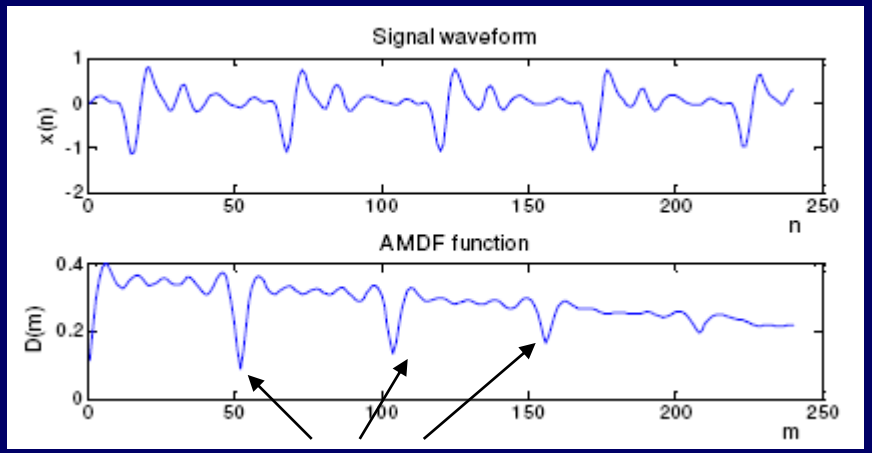




Μέθοδος Average Magnitude Difference Function (AMDF) για τον υπολογισμό του pitch

- An intuitive method, just pick the peaks and find the period.
-

$$D_x(m) = \frac{1}{N} \sum_{N=0}^{N-1-m} |x(m) - x(n+m)|, 0 \leq m \leq M_0$$



Find peaks in D, the estimated period is the average gaps between two neighboring –ve peaks

Cepstrum Pitch Determination (CPD) of pitch

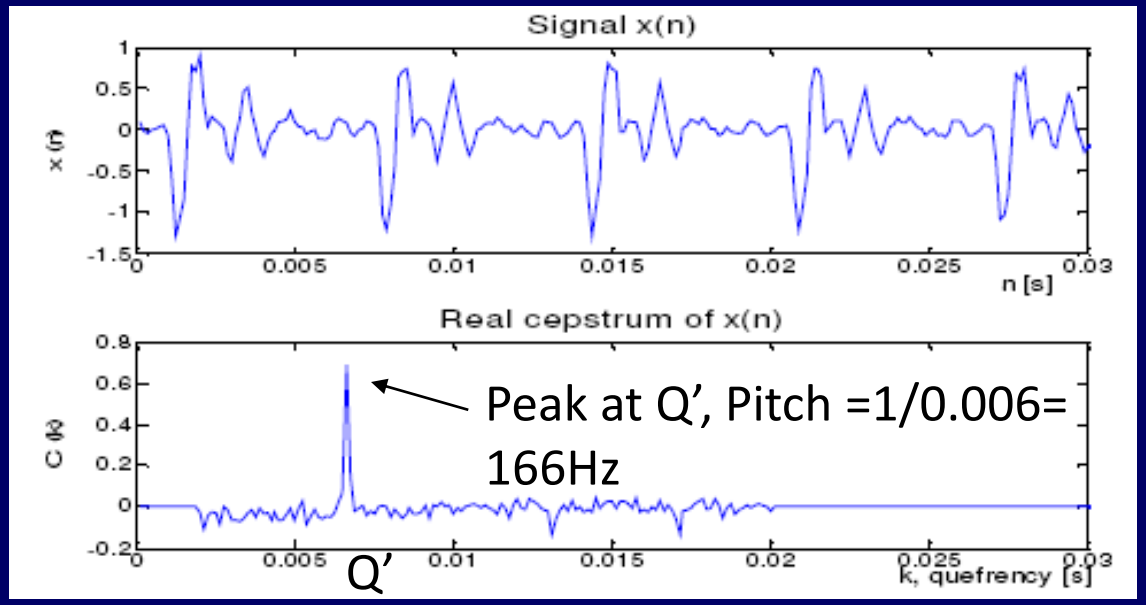
$$s(n) = e(n) * h(n)$$

$$s(w) = E(w) \cdot H(w)$$

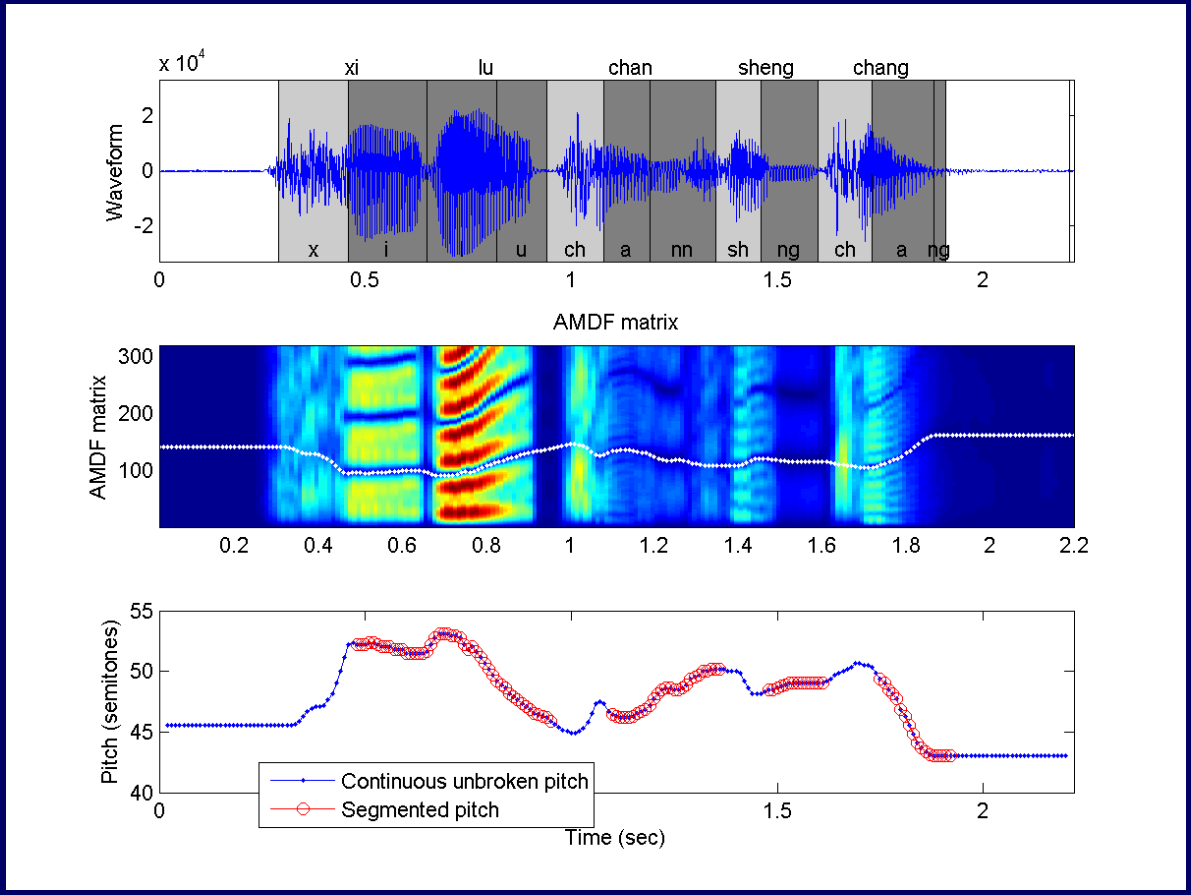
$$F^{-1}\{\log[S(w)]\} = \\ = F^{-1}\{\log[E(w)]\} + F^{-1}\{\log[H(w)]\}$$

$$C(m) = \frac{1}{N}\left\|\sum_{k=0}^{N-1} S(k) \cdot e^{-j\frac{2\pi}{N}mk}\right\|$$

$$C(k) = \log\left\|\sum_{n=0}^{N-1} S(n) \cdot e^{-j\frac{2\pi}{N}nk}\right\|$$



Τυπικό παράδειγμα pitch tracking (χρονικής μεταβολής του pitch)



Frequency Masking – Κάλυψη Συχνοτήτων

Το φαινόμενο κατά το οποίο ένας ήχος δε μπορεί να γίνει αντιληπτός όταν ένας άλλος ήχος με παραπλήσια συχνότητα έχει αρκετά μεγάλη ένταση.

Frequency Masking (1/2)

Α. Θόρυβος Κάλυψης Τόνου (tone masking noise)

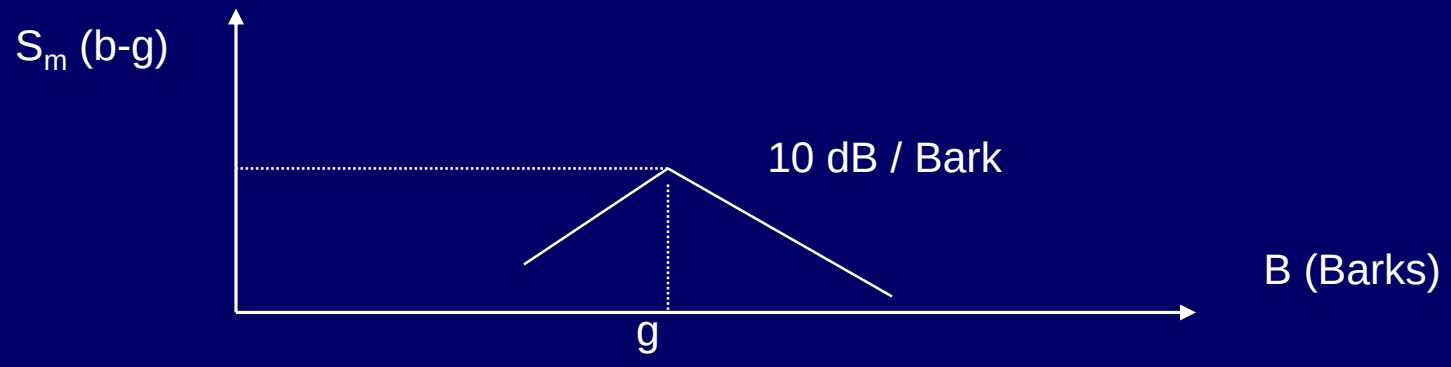
Εμπειρικά αποδεικνύεται ότι:

Θόρυβος ενέργειας EN (dB) στη συχνότητα g (κλίμακας Bark) καλύπτει ένα τόνο συχνότητας b (κλίμακας Bark) αν η ενέργεια του τόνου είναι κάτω από το κατώφλι:

$$T_T(b) = E_N - 6{,}025 - 0{,}275g + S_m(b-g)(dB\ SPL)$$

όπου η συνάρτηση διασποράς κάλυψης Sm είναι:

$$S_m(b) = 15{,}81 + 7{,}5(b + 0{,}474) - 17{,}5(1 + (b + 0{,}474)^2)^{1/2}\ dB$$



Frequency Masking (2/2)

B. Τόνος Κάλυψης Θορύβου (noise masking tone)

Εμπειρικά αποδεικνύεται ότι:

Ένας τόνος συχνότητας g (κλίμακα Bark) ενέργειας ET (dB) καλύπτει θόρυβο σε συχνότητα b (κλίμακα Bark) αν η ενέργεια του θορύβου είναι κάτω από το όριο:

$$T_N(b) = E_T - 2{,}025 - 0{,}175g + S_m(b-g)(dB\ SPL)$$

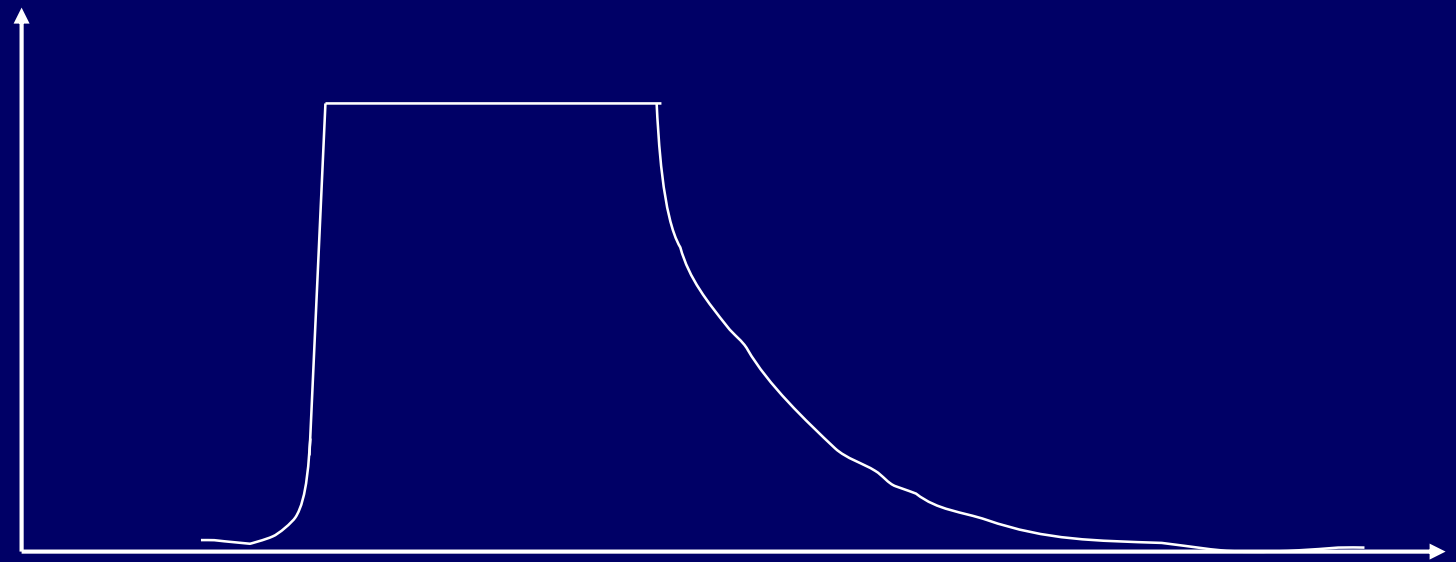
Τα όρια κάλυψης αναφέρονται ως: JND (Just noticeable distortion)

Temporal Masking

Το φαινόμενο κατά το οποίο ένας ήχος, πολύ κοντά χρονικά με έναν άλλο ήχο, δε μπορεί να γίνει αντιληπτός.

Premasking < 5 msec

0 < postmasking <300 msec



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αθηνών» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

Σημείωμα Χρήσης Έργων Τρίτων

- "Η δομή και οργάνωση της παρουσίασης, καθώς και το υπόλοιπο περιεχόμενο, αποτελούν πνευματική ιδιοκτησία της συγγραφέως και του Πανεπιστημίου Αθηνών και διατίθενται με άδεια Creative Commons Αναφορά Μη Εμπορική Χρήση Παρόμοια Διανομή Έκδοση 4.0 ή μεταγενέστερη.

- Οι φωτογραφίες που περιέχονται στην παρουσίαση αποτελούν πνευματική ιδιοκτησία τρίτων. Απαγορεύεται η αναπαραγωγή, αναδημοσίευση και διάθεσή τους στο κοινό με οποιονδήποτε τρόπο χωρίς τη λήψη άδειας από τους δικαιούχους. "