Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Επεξεργασία Ομιλίας και Φυσικής Γλώσσας

Ενότητα 6: Σύνθεση ομιλίας

Γεώργιος Κουρουπέτρογλου

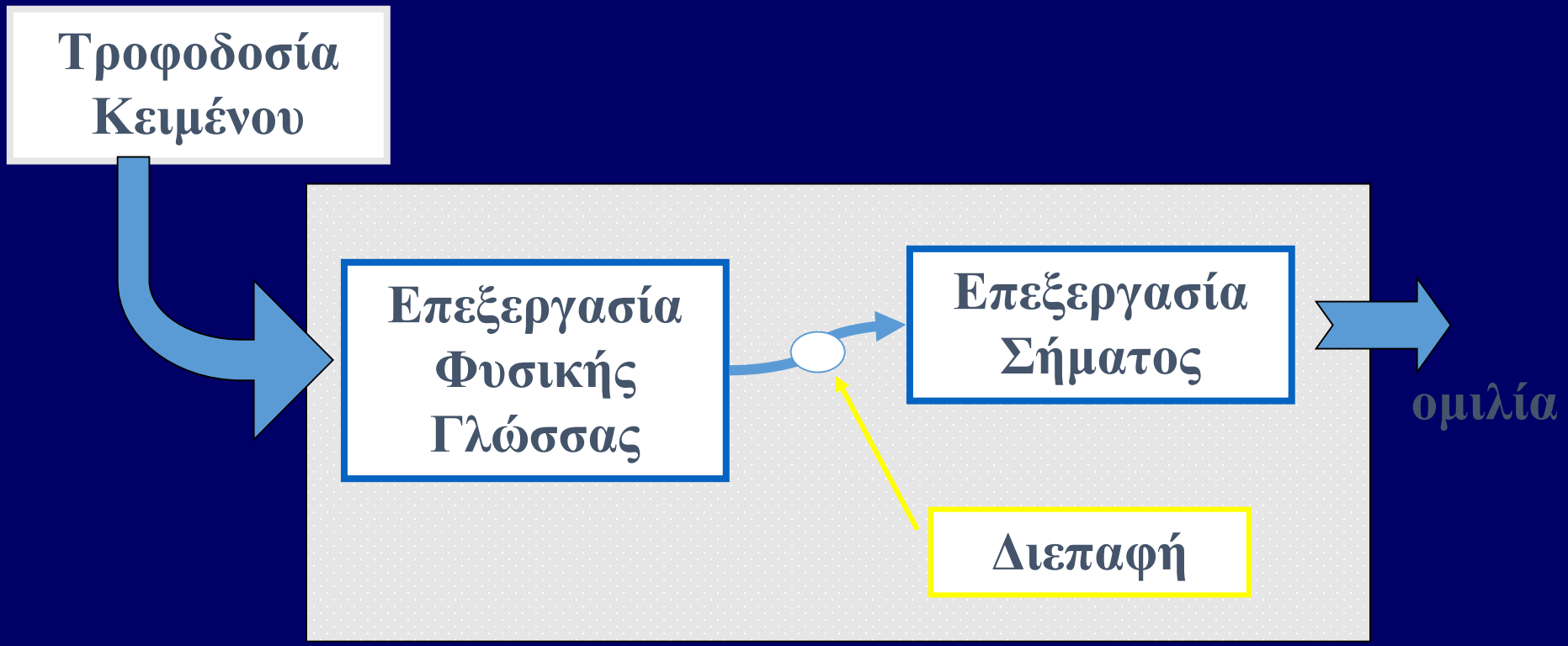
koupe@di.uoa.gr

Μετατροπή κειμένου σε ομιλία (TtS) (1/2)

- Διαδικασία μετατροπής μίας ακολουθίας **κειμενικών συμβόλων** σε ένα αντίστοιχο ακουστικό **σήμα ομιλίας**.

- Απαιτήσεις:
  - Αυτοματοποιημένη
  - Να εκφωνεί ένα οποιοδήποτε κείμενο
  - Η παραγόμενη ομιλία να είναι κατανοητή και όσο το δυνατόν κοντά στην φυσική

Μετατροπή κειμένου σε ομιλία (TtS) (2/2)



Ταξινόμηση συστημάτων σύνθεσης ομιλίας

- Μέθοδος σύνθεσης
  - Βασισμένη σε κανόνες:
    - Σύνθεση με φωνοσυντονισμούς
    - Αρθρωτική σύνθεση
  - Με συρραφή φωνητικών μονάδων
    - δίφωνων
    - Με επιλογή φωνητικών μονάδων (unit selection)
- Τεχνική συρραφής
  - TD-PSOLA, FD-PSOLA
- Κωδικοποίηση φωνητικών μονάδων
  - LPC, hybrid harmonic/stochastic, sinusoidal model, etc.
- Μονο- ή πολύ-γλωσσική

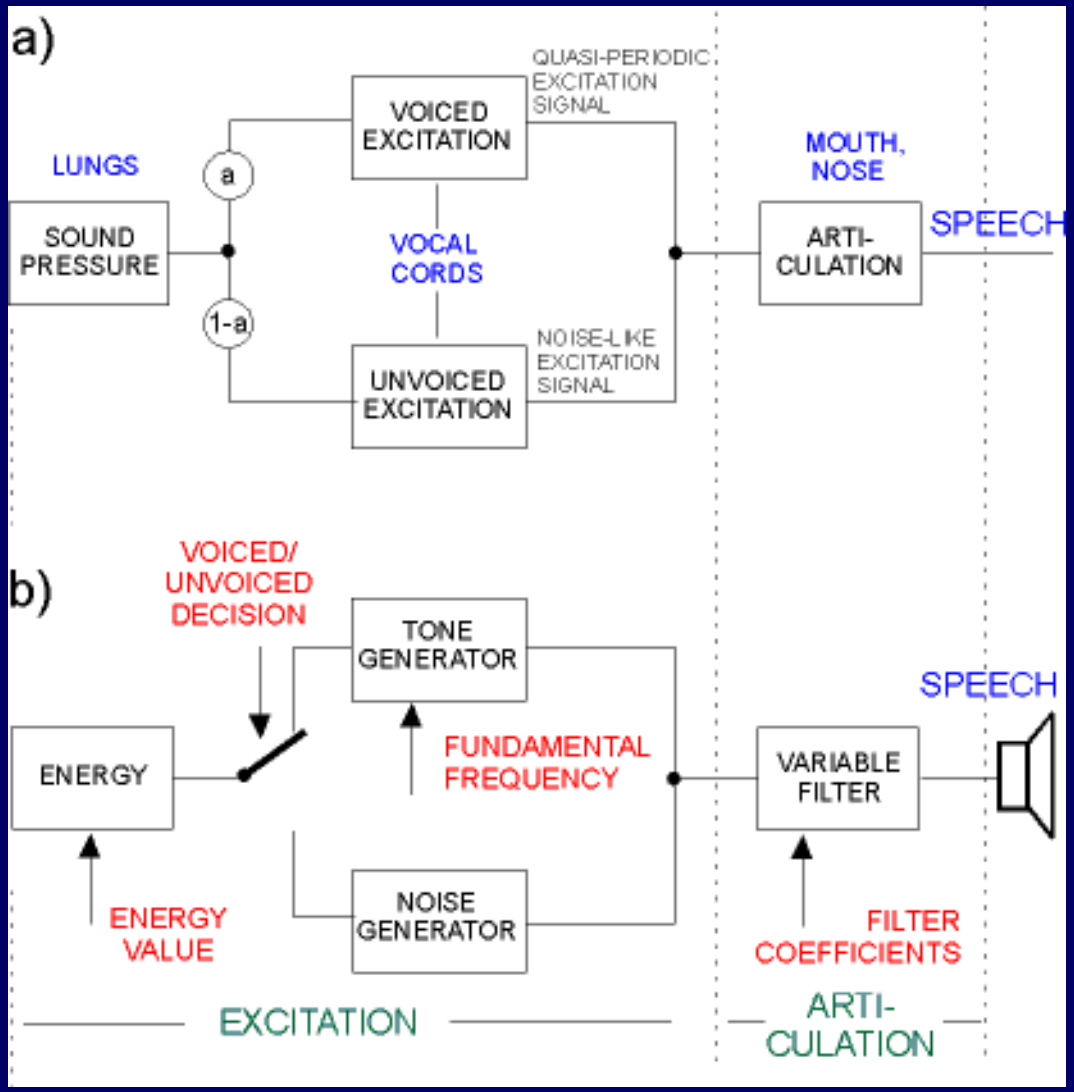
Κατηγορίες τεχνικών σύνθεσης σήματος ομιλίας

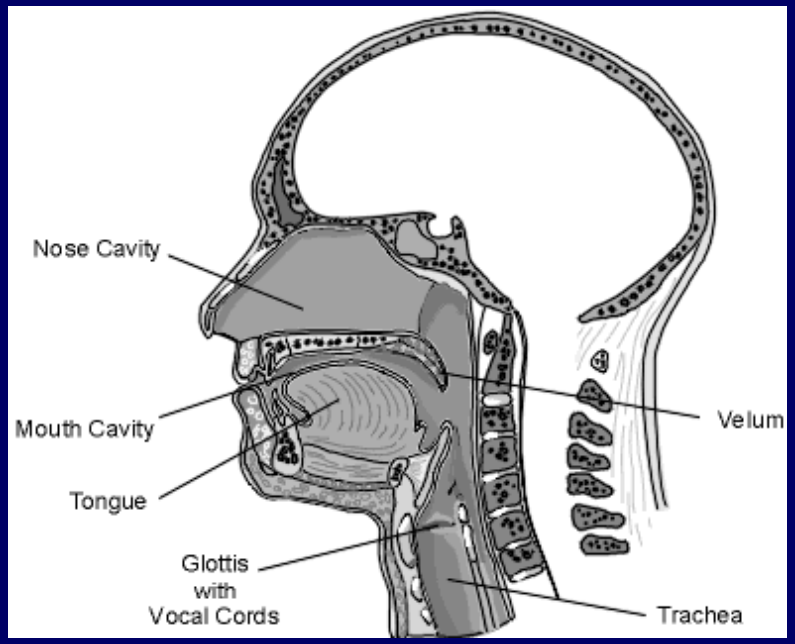
- Σύνθεση με κανόνες (γένεση φωνοσυντονισμών - αποτέλεσμα)
  - Μοντέλο πηγής/φίλτρου
  - Δυσκολία παραμετροποίησης του αποτελέσματος
- Σύνθεση με βάση το μοντέλο των αρθρωτών (αίτιο)
  - Αναπαράσταση κίνησης αρθρωτών
  - Δυσκολία μέτρησης πραγματικής διαδικασία άρθρωσης
  - Υπολογιστική πολυπλοκότητα μοντέλων
- Σύνθεση με συρραφή
  - Κατάτμηση προ-ηχογραφημένης ομιλίας και συγκόλληση των κατάλληλων τμημάτων
  - Επεξεργασία σήματος στα μέρη που συγκολλώνται ώστε:
    - Μεταβολή προσωδιακών χαρακτηριστικών
    - Ομαλές μεταβάσεις
    - Όσο μεγαλύτερη επεξεργασία χρησιμοποιείται τόσο αλλοιώνεται η ποιότητα του σήματος
    - Χρονοβόρα και ενίοτε ακριβή η διαδικασία δημιουργίας της κατάλληλης βάσης

Σύνθεση ομιλίας με κανόνες

- Άμεση αντιστοιχία με τον μηχανισμό παραγωγής ομιλίας.
- Αποτελείται από γεννήτριες παλμών και μία σειρά από χρονικά μεταβαλλόμενα φίλτρα.
- Προσομοιάζεται το φασματογράφημα της φυσικής ομιλίας (δηλ. το αποτέλεσμα της ομιλίας) με μία σειρά από κανόνες που ορίζουν τις τιμές των φίλτρων.

Μοντέλο παραγωγής ομιλίας





Σύνθεση ομιλίας με κανόνες – διαδικασία

- Ηχογράφηση ενός μεγάλου σώματος ακολουθιών Σύμφωνο-Φωνήεν-Σύμφωνο, Φωνήεν-Σύμφωνο-Φωνήεν, ...
- Ανάλυση και εξαγωγή παραμέτρων από ηχογραφήσεις φυσικής ομιλίας σχετικών με την πηγή (φωνητικές χορδές) και το φίλτρο (φωνητική οδός).
- Δημιουργία κανόνων από τις παραμέτρους για CVC, VCV κλπ.
- Κατά την σύνθεση, το παραμετρικό σήμα μετατρέπεται σε ψηφιακό σήμα ομιλίας (κυματομορφή) από έναν συνθέτη που υλοποιεί το μοντέλο που χρησιμοποιήθηκε κατά την ανάλυση.

Σύνθεση φωνοσυντονισμών

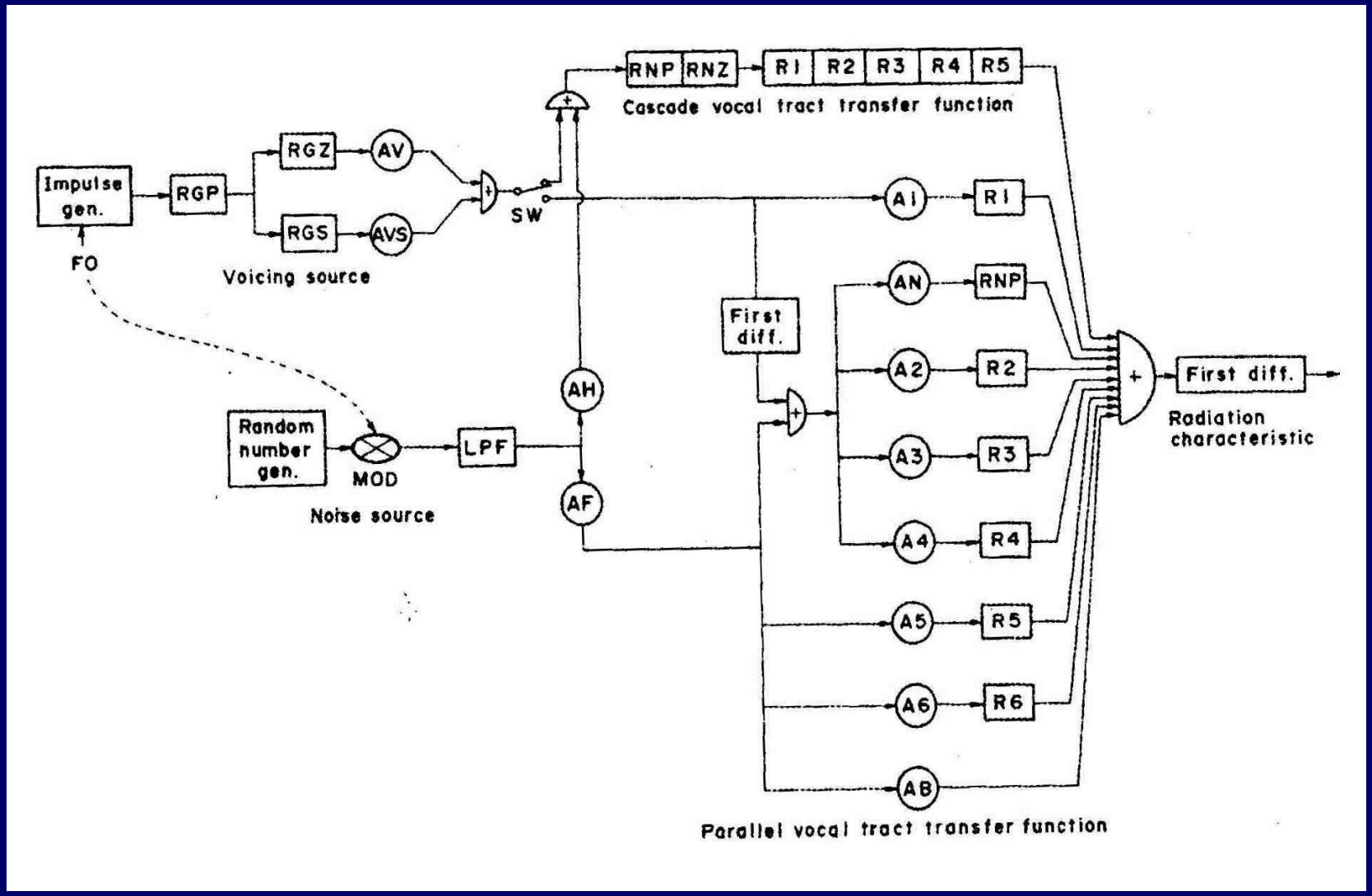
- Ελέγχεται από έναν **πίνακα 40 παραμέτρων** που ανανεώνουν την συμπεριφορά **πηγών** και **φίλτρων** κάθε 5 msec.
- «***Είμαι ο πρώτος συνθέτης ομιλίας του Πανεπιστημίου Αθηνών***» (1998)

  - Μέτρια καταληπτότητα
  - Χαμηλή φυσικότητα

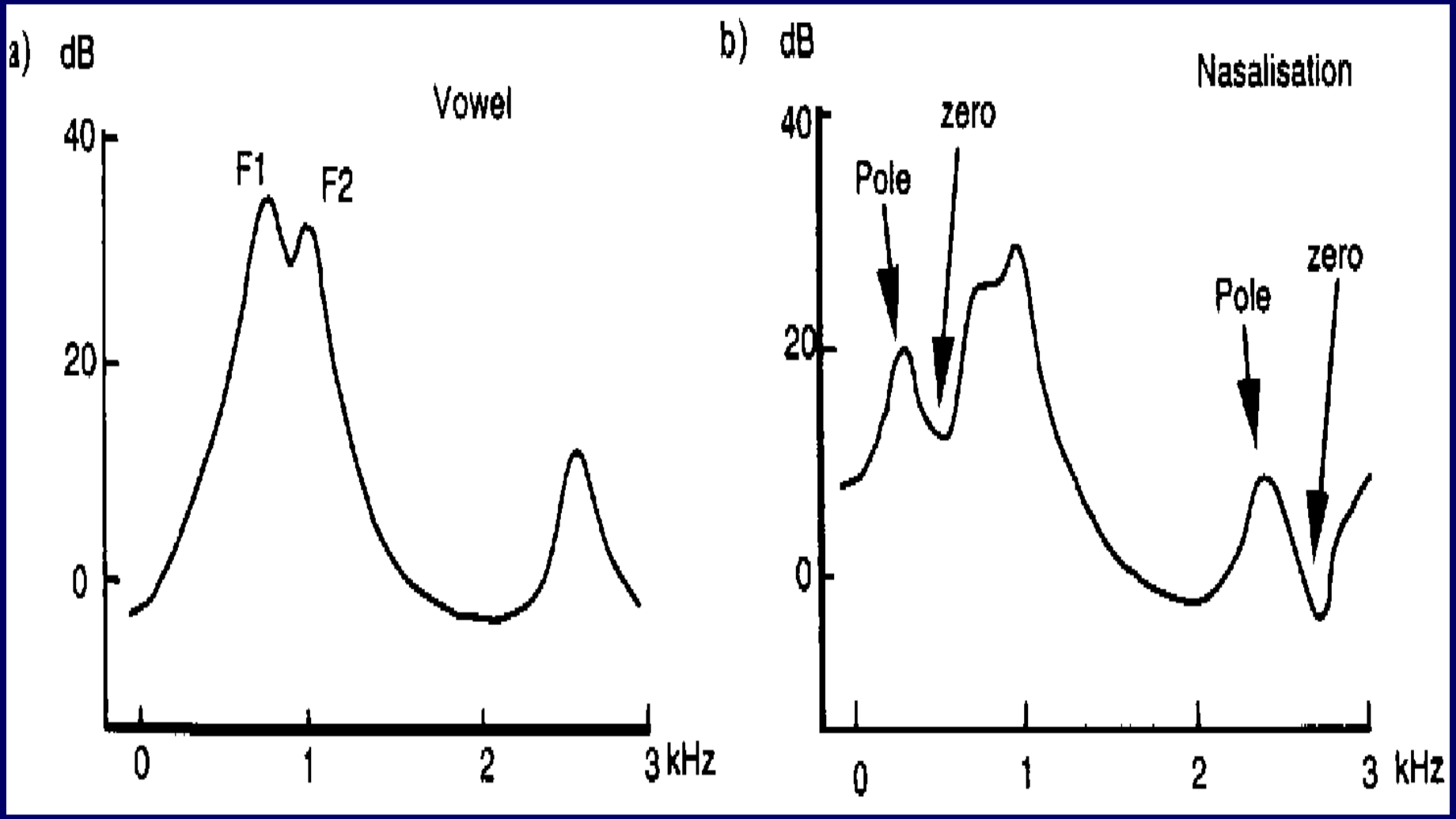
Σύνθεση με κανόνες – ποιότητα

- Η ποιότητα της ομιλίας εξαρτάται από:
  - Την ικανότητα των κανόνων να περιγράψουν το σώμα της ηχογράφησης.
  - Τον τύπο των επιλεγμένων λέξεων και την ποιότητα ηχογράφησης.
  - Την ακρίβεια του μοντέλου που χρησιμοποιείται στην ανάλυση του σώματος. Πολύ απλά μοντέλα αποτυγχάνουν να αποδώσουν τα χαρακτηριστικά ομιλίας.
  - Τον αλγόριθμο που δημιουργεί τους κανόνες.
  - Τις βελτιώσεις μέσω δοκιμής-διόρθωσης

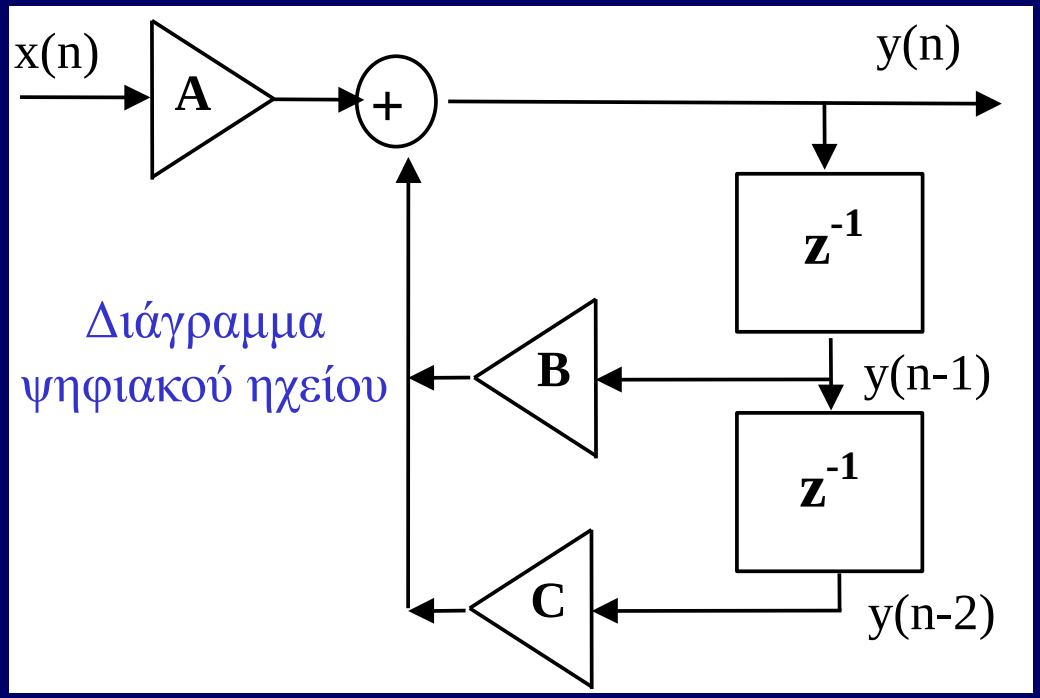
Σύνθεση με κανόνες - διάταξη συνθέτη Klatt



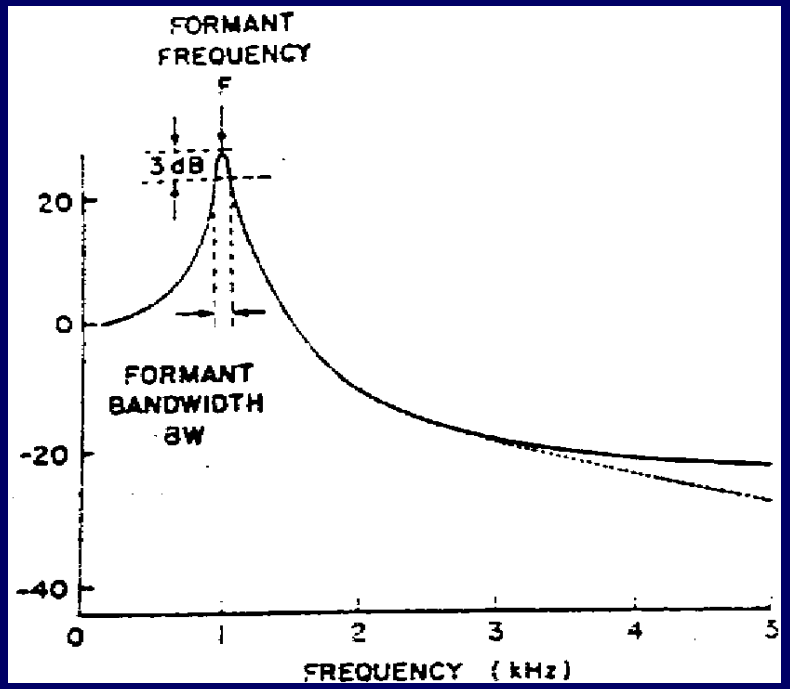
Σύνθεση με κανόνες (1/2)



Σύνθεση με κανόνες - ψηφιακό αντηχείο



Διάγραμμα ψηφιακού ηχείου



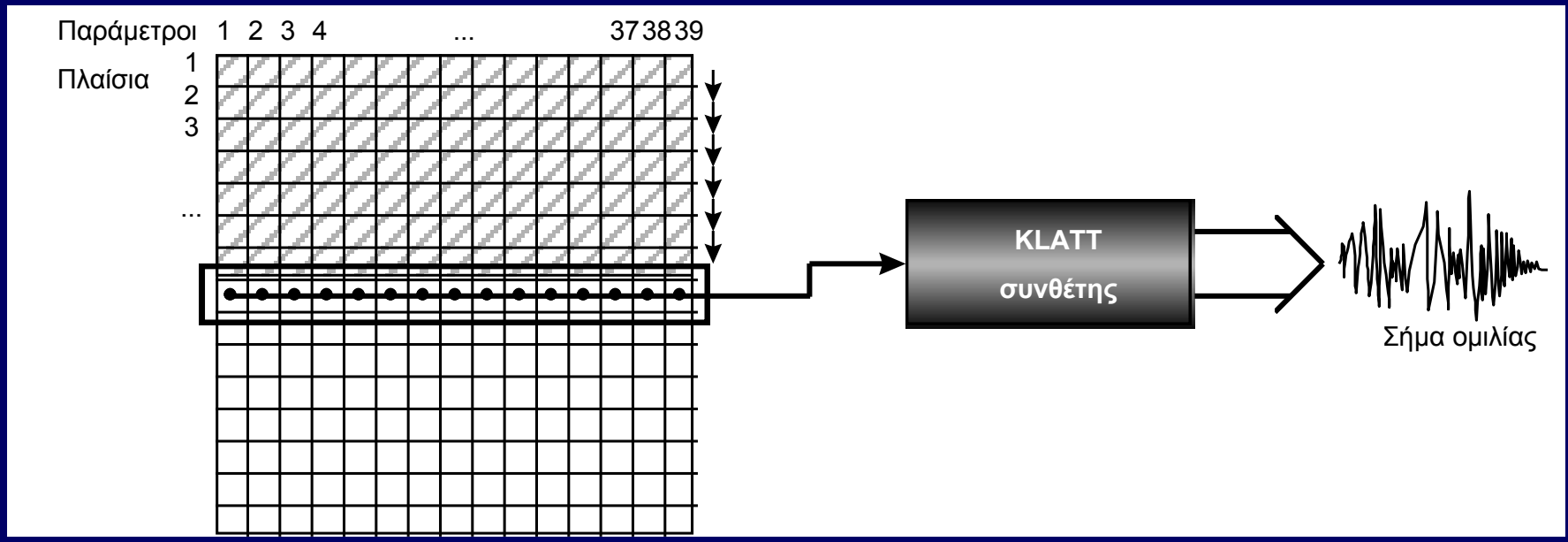
$$y(n) = Ax(n) + By(n-1) + Cy(n-2)$$
$$C = -e^{-2\pi B_w T}$$
$$B = 2e^{-2\pi B_w T}\cos(2\pi FT)$$
$$A = 1 - C - B$$

B_w=εύρος ζώνης

F=συχνότητα συντονισμού

	Σ/Μ	Όνομα	Περιγραφή	Μονάδα	Ελάχιστη τιμή	Μέγιστη τιμή
1	Μ	F0	θεμελιώδης συχνότητα	Hz	0	500
2	Μ	F1	συχνότητα 1ου formant	Hz	150	900
3	Μ	F2	συχνότητα 2ου formant	Hz	500	2500
4	Μ	F3	συχνότητα 3ου formant	Hz	1300	3500
5	Μ	F4	συχνότητα 4ου formant	Hz	2500	4500
6	Σ	F5	συχνότητα 5ου formant	Hz	3500	4900
7	Σ	F6	συχνότητα 6ου formant	Hz	4000	4999
8	Μ	B1	εύρος ζώνης 1ου formant	Hz	40	500
9	Μ	B2	εύρος ζώνης 2ου formant	Hz	40	500
10	Μ	B3	εύρος ζώνης 3ου formant	Hz	40	500
11	Σ	B4	εύρος ζώνης 4ου formant	Hz	100	500
12	Σ	B5	εύρος ζώνης 5ου formant	Hz	150	700
13	Σ	B6	εύρος ζώνης 6ου formant	Hz	200	2000
14	Μ	A1	ένταση 1ου formant	dB	0	80
15	Μ	A2	ένταση 2ου formant	dB	0	80
16	Μ	A3	ένταση 3ου formant	dB	0	80
17	Μ	A4	ένταση 4ου formant	dB	0	80
18	Μ	A5	ένταση 5ου formant	dB	0	80
19	Μ	A6	ένταση 6ου formant	dB	0	80
20	Μ	FNZ	συχνότητα έρρινου μηδενικού	Hz	200	700
21	Σ	FNP	συχνότητα έρρινου πόλου	Hz	200	500
22	Σ	BNZ	εύρος ζώνης έρρινου μηδενικού	Hz	50	500
23	Σ	BNP	εύρος ζώνης έρρινου πόλου	Hz	50	500
24	Σ	FGZ	συχνότητα γλωττιδικού μηδενικού	Hz	0	5000
25	Σ	FGP	συχνότητα 1ου γλωττιδικού ηχείου	Hz	0	600
26	Σ	BGZ	εύρος ζώνης γλωττιδικού μηδενικού	Hz	100	9000
27	Σ	BGP	εύρος ζώνης 1ου γλωττιδικού ηχείου	Hz	100	2000
28	Σ	BGS	εύρος ζώνης 2ου γλωττιδικού ηχείου	Hz	100	1000
29	Μ	AV	ένταση έμφωνης πηγής	dB	0	80
30	Μ	AF	ένταση τυρβώδους πηγής	dB	0	80
31	Μ	AH	ένταση δασώδους πηγής	dB	0	80
32	Μ	AVS	ένταση σχεδόν-ημιτόνοειδούς πηγής	dB	0	80
33	Μ	AN	ένταση έρρινου formant	dB	0	80
34	Μ	AB	ένταση bypass μονοπατιού	dB	0	80
35	Σ	NFC	αριθμός σειριακών formants	-	4	6
36	Σ	SR	ρυθμός δειγματοληψίας	δείγματα /sec	5000	20000
37	Σ	NWS	αριθμός δειγμάτων ανά πλαίσιο	-	1	200
38	Σ	G0	ρυθμιστικό συνολικής έπολαβής	dB	0	80
39	Σ	SW	διακόπτης επιλογής σειριακού/παράλληλου	-	0(σειρ.)	1(παραλ.)

Σύνθεση με κανόνες (2/2)

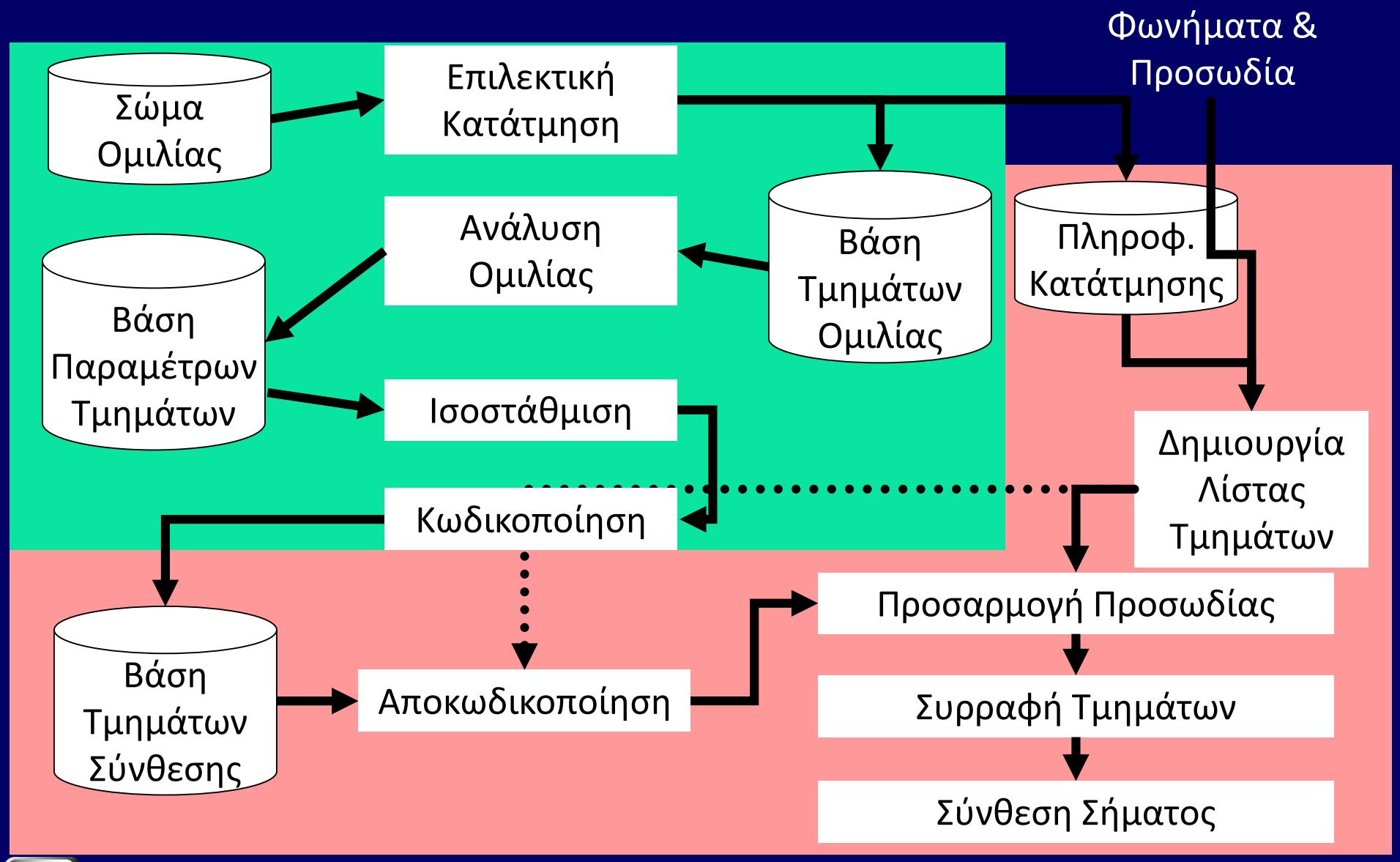


Σύνθεση με συρραφή (1/4)

- Σε αντίθεση με το συνθέτη formant, δεν απαιτείται λεπτομερής γνώση και ανάλυση:
  - Του μηχανισμού παραγωγής
  - Των συνιστωσών του φασματογραφήματος

- Τα στοιχεία αυτά εμπεριέχονται στα τμήματα που θα συρραφούν.

Σύνθεση με συρραφή (2/4)



Σύνθεση με συρραφή – δίφωνα

- Σύνθεση με δίφωνα:
  - Δίφωνο είναι ένα τμήμα ομιλίας αποτελούμενο από 2 φωνήματα
  - Η πιο απλή περίπτωση σύνθεσης με συρραφή.
  - Μία γλώσσα με N φωνήματα έχει το πολύ N2 δίφωνα.
  - Πολλά όμως δεν εμφανίζονται στην (συγκεκριμένη) γλώσσα
  - Για τα ελληνικά, τα δίφωνα είναι 500-1100 (αντί 1369), ανάλογα με την υλοποίηση και τις απαιτήσεις.

- Τα δίφωνα προτιμώνται από τα απλά φωνήματα λόγω:
  - Της σταθερής κατάστασης που εμφανίζεται στα κέντρα των επιμέρους φωνημάτων
  - Κατά την συρραφή συγκολλώνται όμοια φωνήματα λιγότερες ασυνέχειες

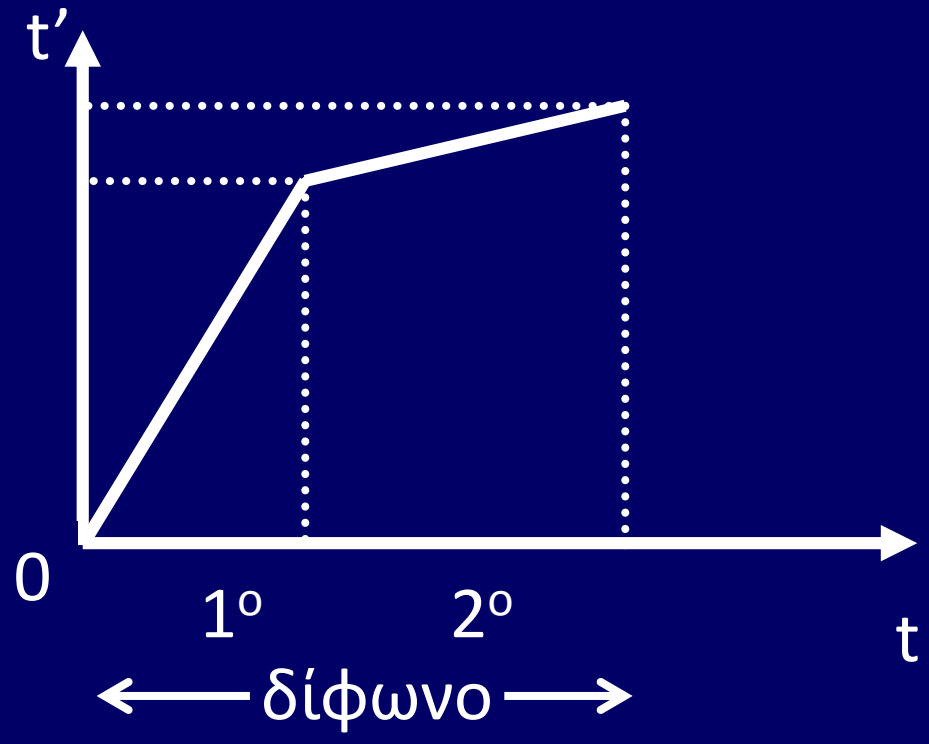
Τεχνικές συρραφής

- Ένα στιγμιότυπο από κάθε δίφωνο:
  - Μονότονη ηχογράφηση σε περιβάλλοντα ανόητων λέξεων.
  - Δημιουργείται η ανάγκη για προσαρμογή των προσωδιακών στοιχείων του σήματος κατά τη σύνθεση → **επεξεργασία σήματος.**
- Πολλά στιγμιότυπα από κάθε δίφωνο (unit-selection):
  - Πολλά στιγμιότυπα από **ρέοντα λόγο**, με ποικίλα ακουστικά και μορφολογικά χαρακτηριστικά.
  - Δημιουργείται η ανάγκη αναζήτησης του κατάλληλου στιγμιότυπου που ταιριάζει μορφολογικά και ακουστικά με τα γειτονικά του κατά τη σύνθεση → **αλγόριθμος αναζήτησης**.
  - Συχνά συνοδεύεται και από επεξεργασία σήματος, όταν η αναζήτηση αστοχεί αρκετά.

Τονικές και Χρονικές Μεταβολές

- Προκειμένου να είναι δυνατές οι μεταβολές στο **pitch** (καθώς και σε άλλα προσωδιακά διανύσματα), τα ακουστικά σήματα **επεξεργάζονται**.

Σύνθεση με συρραφή (3/4)



Προσαρμογή διάρκειας: $t'(t)$ προσαρμόζει ένα σημείο t της ανάλυσης σε ένα t'της σύνθεσης.

Προσαρμογή τόνου: $\omega_0(t')$, παρέχει την συχνότητα σε ένα σημείο t' της σύνθεσης.

Σύνθεση με συρραφή (4/4)

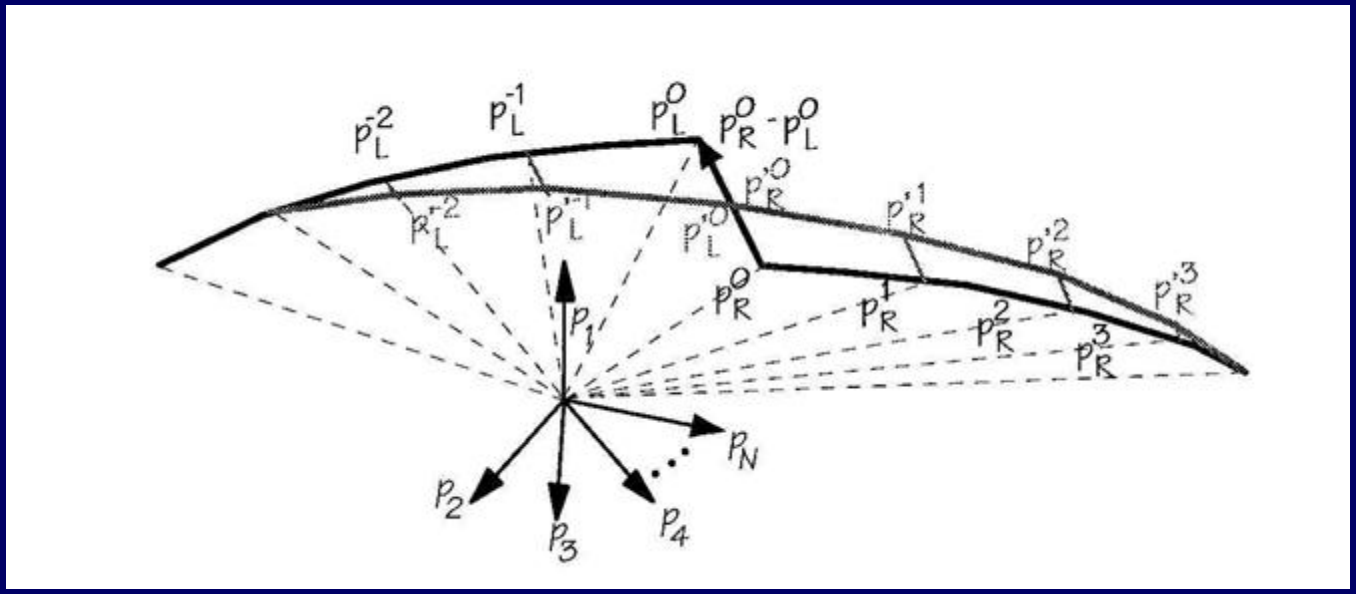
Ομαλοποίηση (smoothing) εφαρμόζεται στα σημεία συρραφής.

Σε 2 σύνολα p_L και p_R από παραμέτρους, εφαρμόζουμε γραμμική παρεμβολή σε M_L σημεία από τα αριστερά και M_R από τα δεξιά.

$$p_L'^{-i} = p_L^{-i} + (p_R^0 - p_L^0)\frac{(M_L - i)}{2M_L}$$

$$p_R'^{j} = p_R^{j} + (p_L^0 - p_R^0)\frac{(M_R - j)}{2M_R}$$

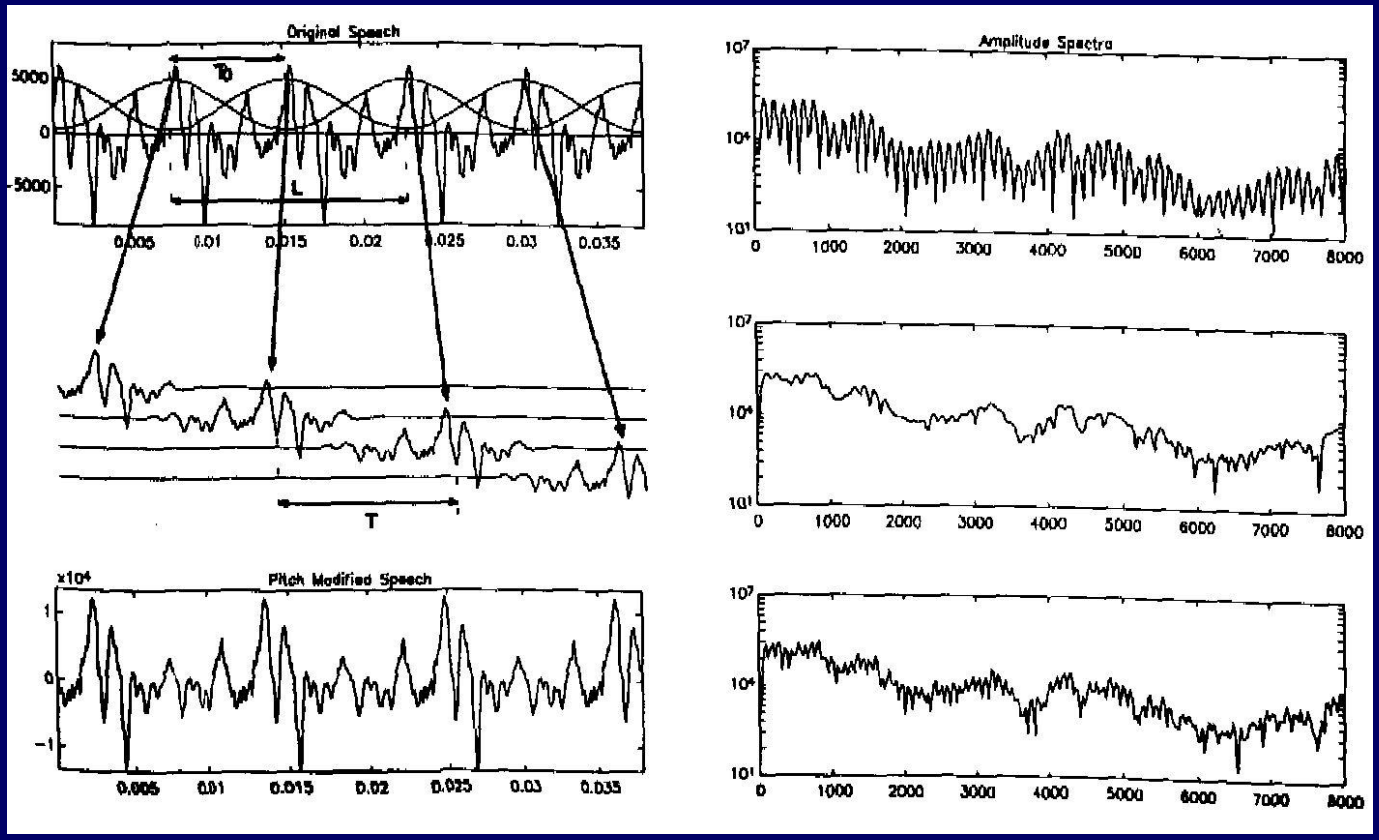
για $i = 0 \ldots M_L - 1$ και $j = 0 \ldots M_R - 1$



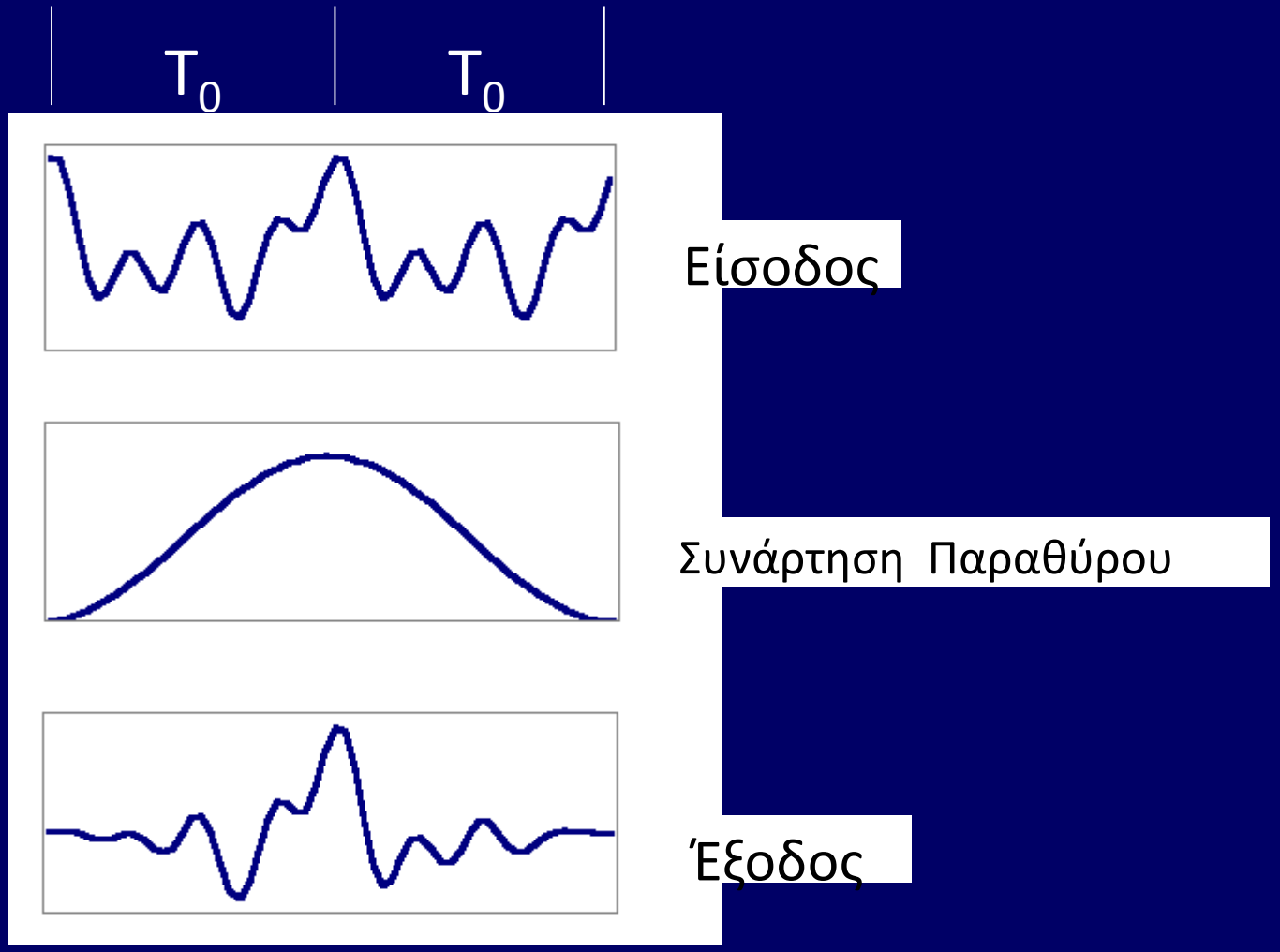
PSOLA (1/10)

Pitch **S**ynchronous **O**ver**L**ap & **A**dd

- Υπολογισμός του φάσματος σε ρυθμό σύγχρονο του pitch
- Για τον υπολογισμό του φάσματος: γραμμική πρόβλεψη
- Pitch marks

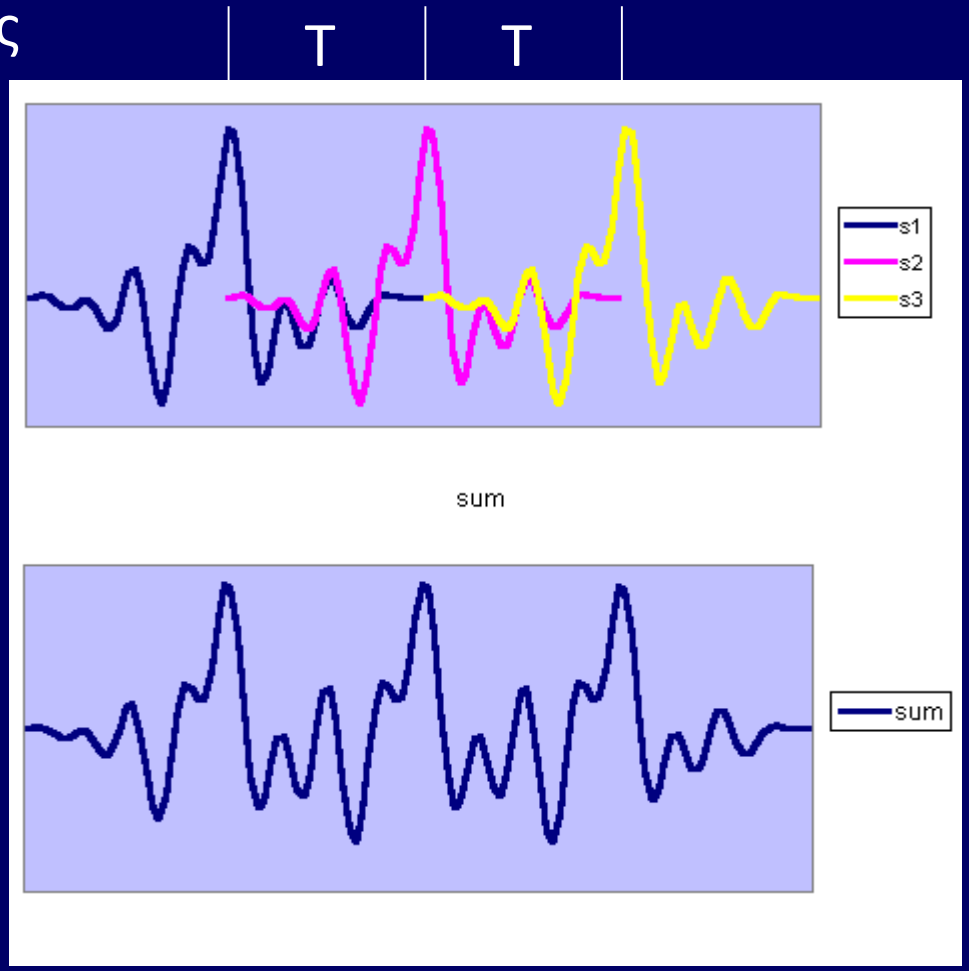


PSOLA (2/10)

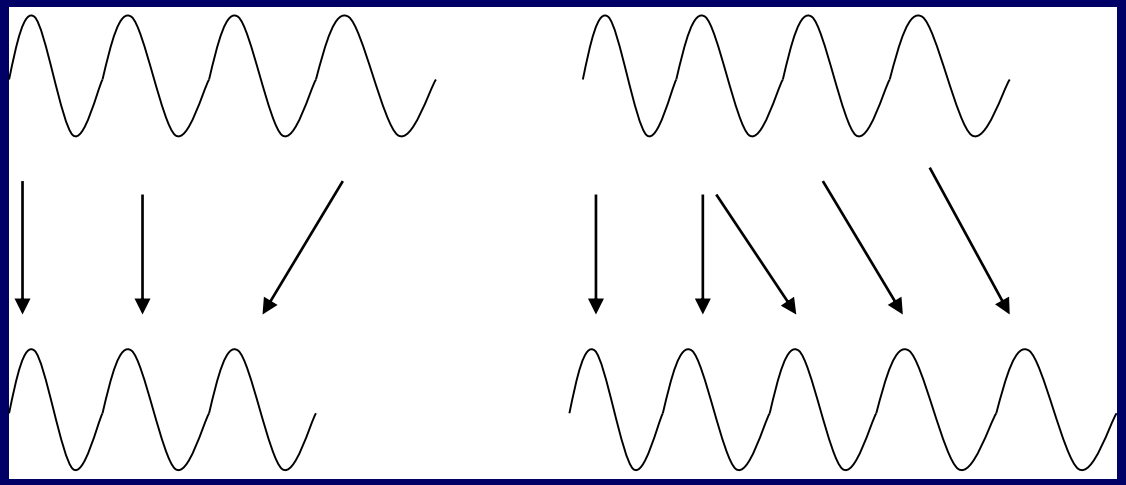


PSOLA (3/10)

Τ = νέα περίοδος



PSOLA (4/10)



Χρονική συμπίεση

Χρονική επέκταση

PSOLA (5/10)

- Ανάλυση
  - υπολογισμός pitch marks ανάλυσης
  - αποσύνθεση του αρχικού σήματος $s(n)$ σε μια ακολουθία μικρών σημάτων ανάλυσης $s_i(n)$ με εφαρμογή παραθύρου με ρυθμό σύγχρονο του pitch
  - προσδιορισμός pitch marks σύνθεσης

$s_i(n) = s(n)w(n - iT_0)$, όπου $s(n)$ περιοδικό

$$\tilde{s}(n) = \sum_{i=-\infty}^{\infty} s_i\left(n - (T - T_0)\right)$$

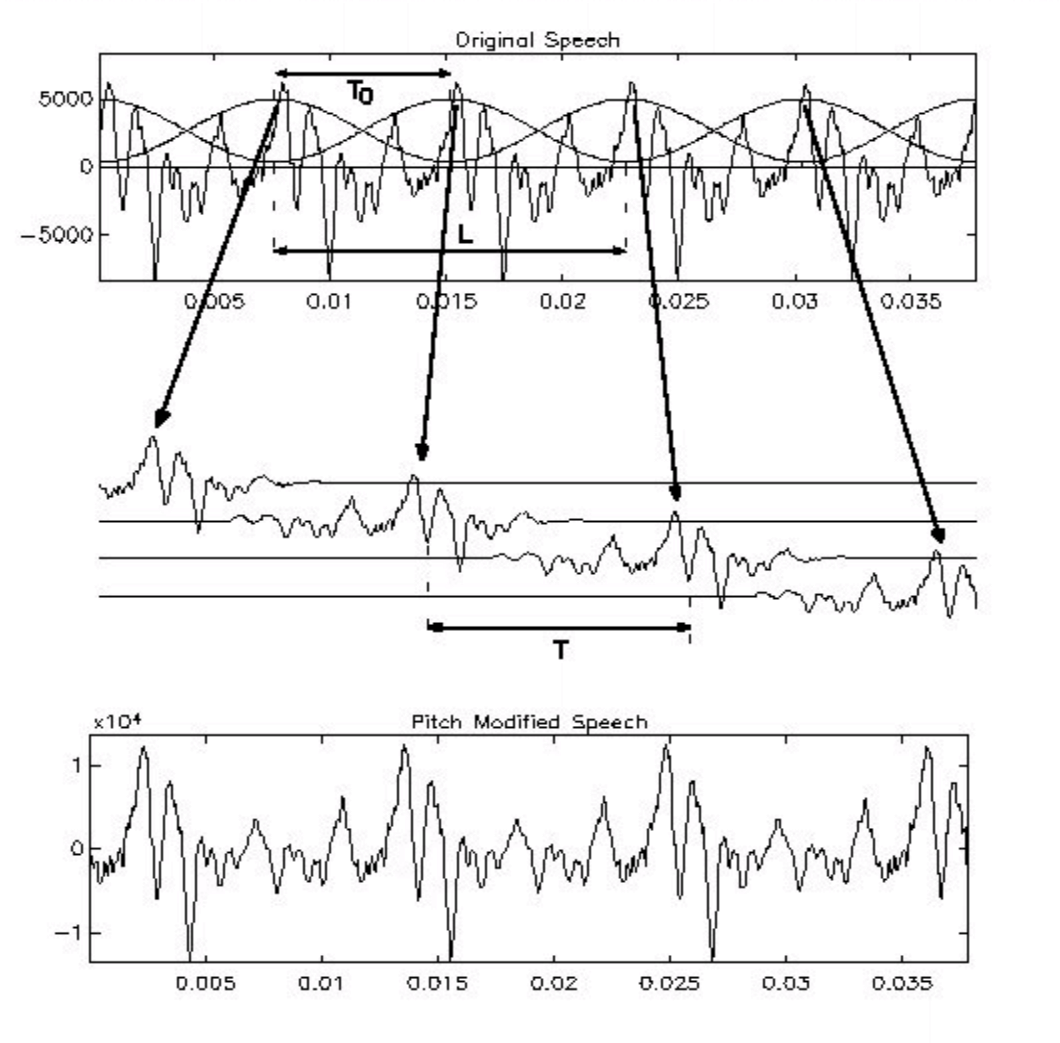
PSOLA (6/10)

- $T = T_0$, το συνθετικό σήμα ~ανάλογο του αρχικού (λόγω επίδρασης παραθύρου)

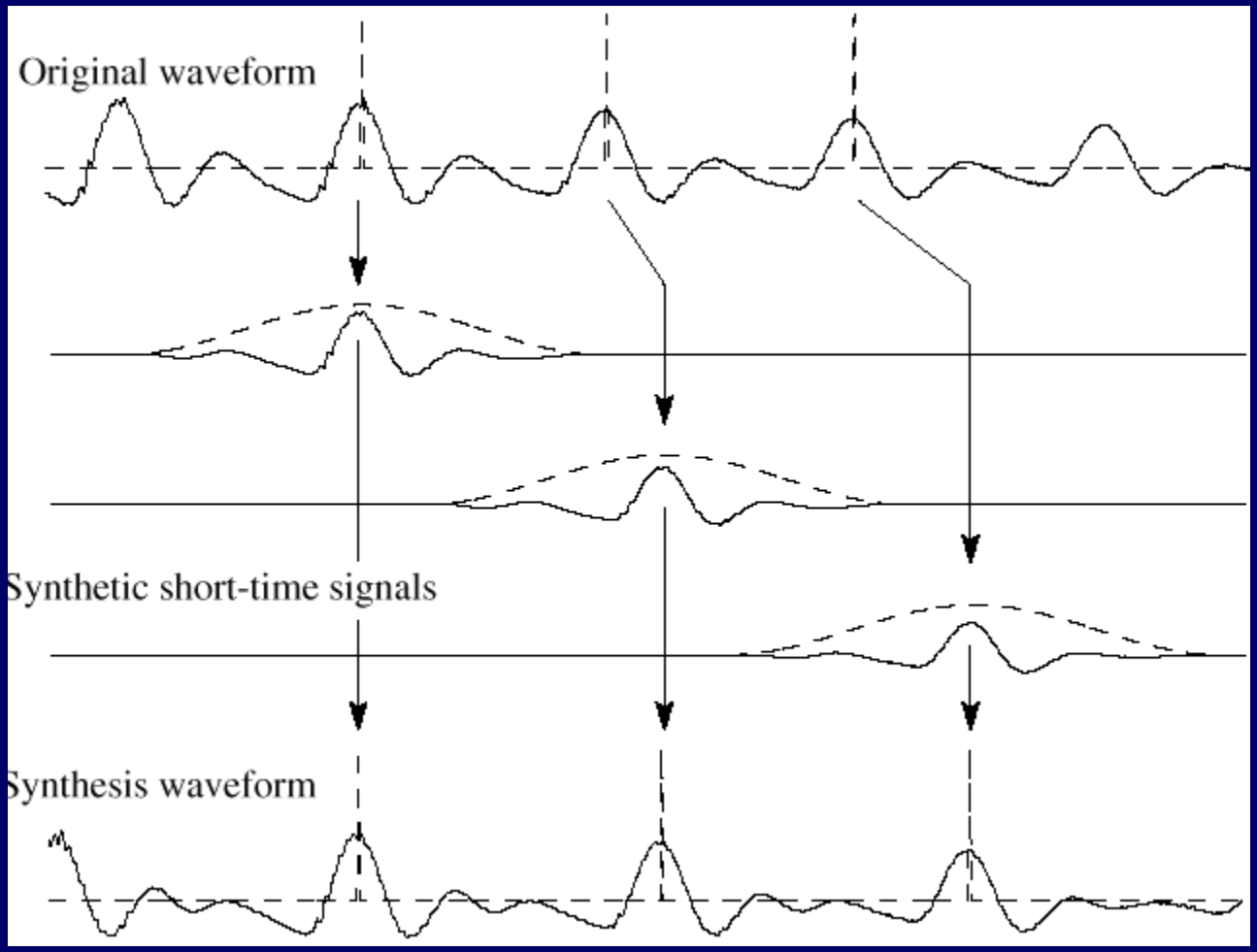
$$\tilde{s}(n) = \sum_{i=-\infty}^{\infty} s(n)w(n - iT) = s(n)w^*(n) \cong ax(n)$$

- w(n): μεγάλη διάρκεια → φάσμα $w^*(n)$ με μικρό εύρος → σε συνεξέλιξη με το φάσμα s(n) δίνει ~ φάσμα του s(n)

PSOLA (7/10)

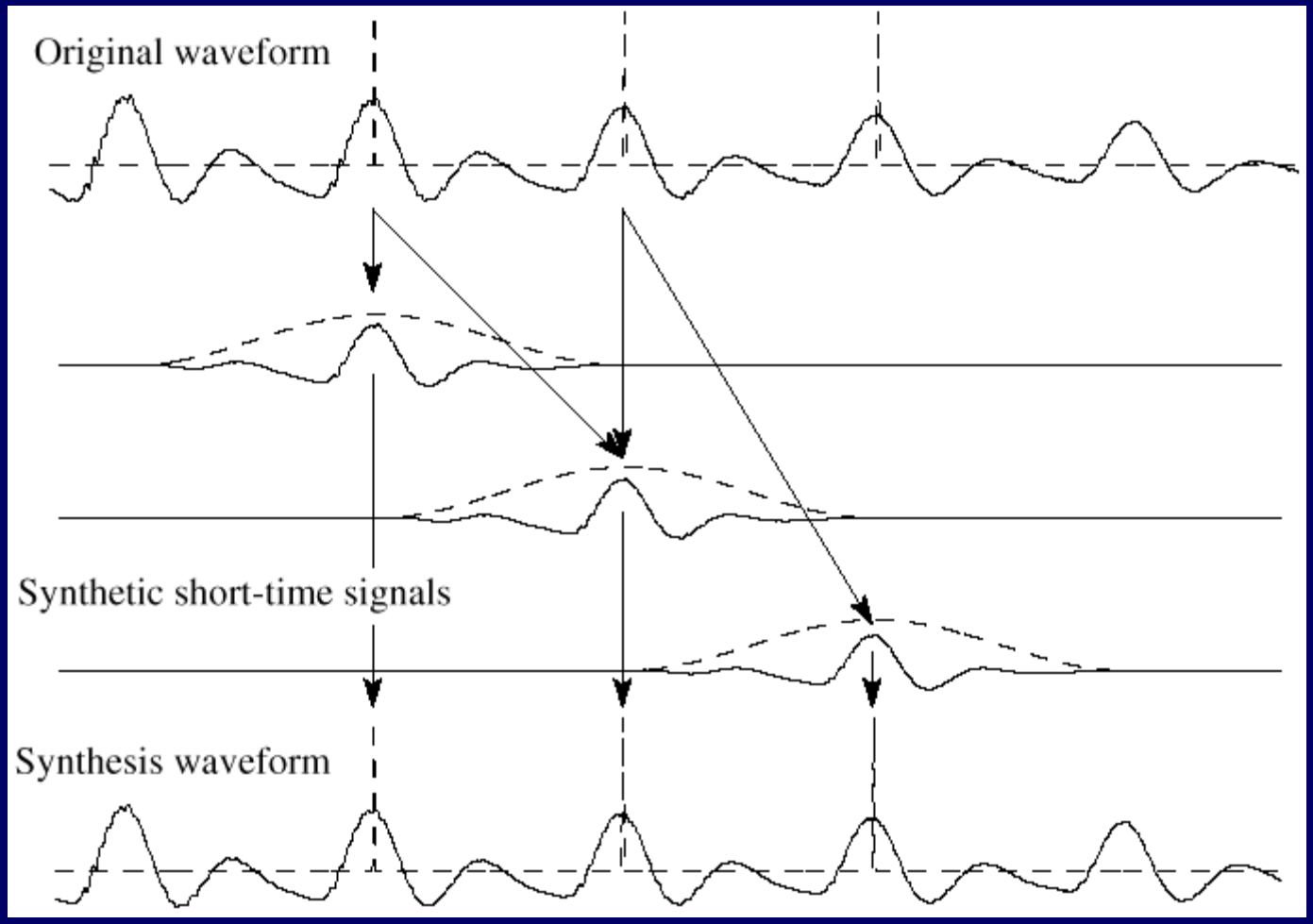

Original Speech
5000
0
−5000
T0
L
0.005
0.01
0.015
0.02
0.025
0.03
0.035
T
x10^4
Pitch Modified Speech
1
0
−1
0.005
0.01
0.015
0.02
0.025
0.03
0.035

PSOLA (8/10)



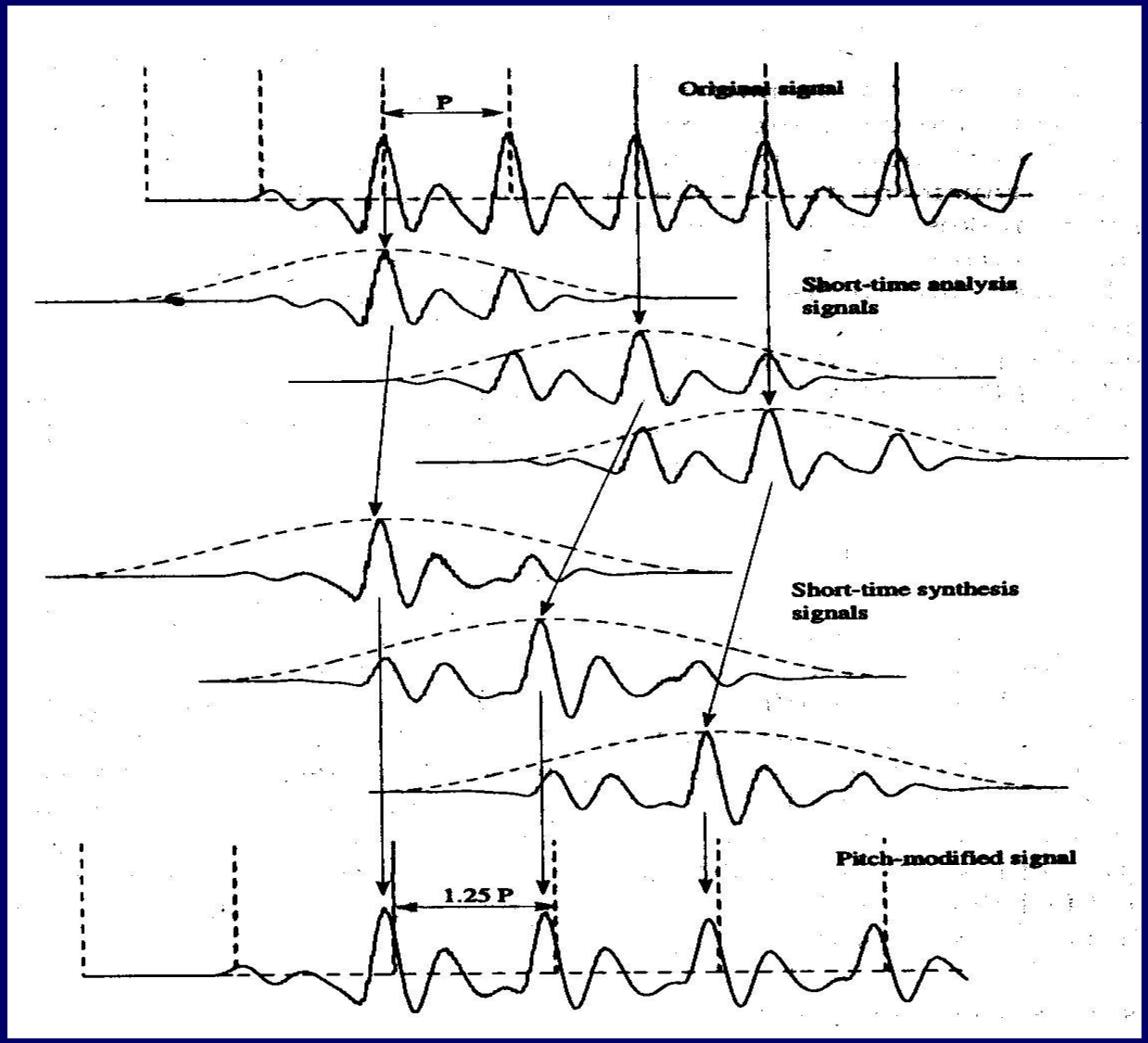
Pitch-scale modification with TD-PSOLA method.

PSOLA (9/10)



Time-scale modification with TD-PSOLA method.

PSOLA (10/10)



Μετατροπή στην χρονική και τονική κλίμακα ταυτόχρονα.

Σύνθεση συρραφής διφώνων

- Οι παράμετροι της ομιλίας εμπεριέχονται σε μικρά ηχογραφημένα τμήματα (φωνήματα, δίφωνα, λέξεις, φράσεις).
- ***«Είμαι η πρώτη φωνητική βάση διφώνων του Πανεπιστημίου Αθηνών»*** (2001)

  - Υψηλή καταληπτότητα
  - Μέτρια φυσικότητα

Προσωδία (1/5)

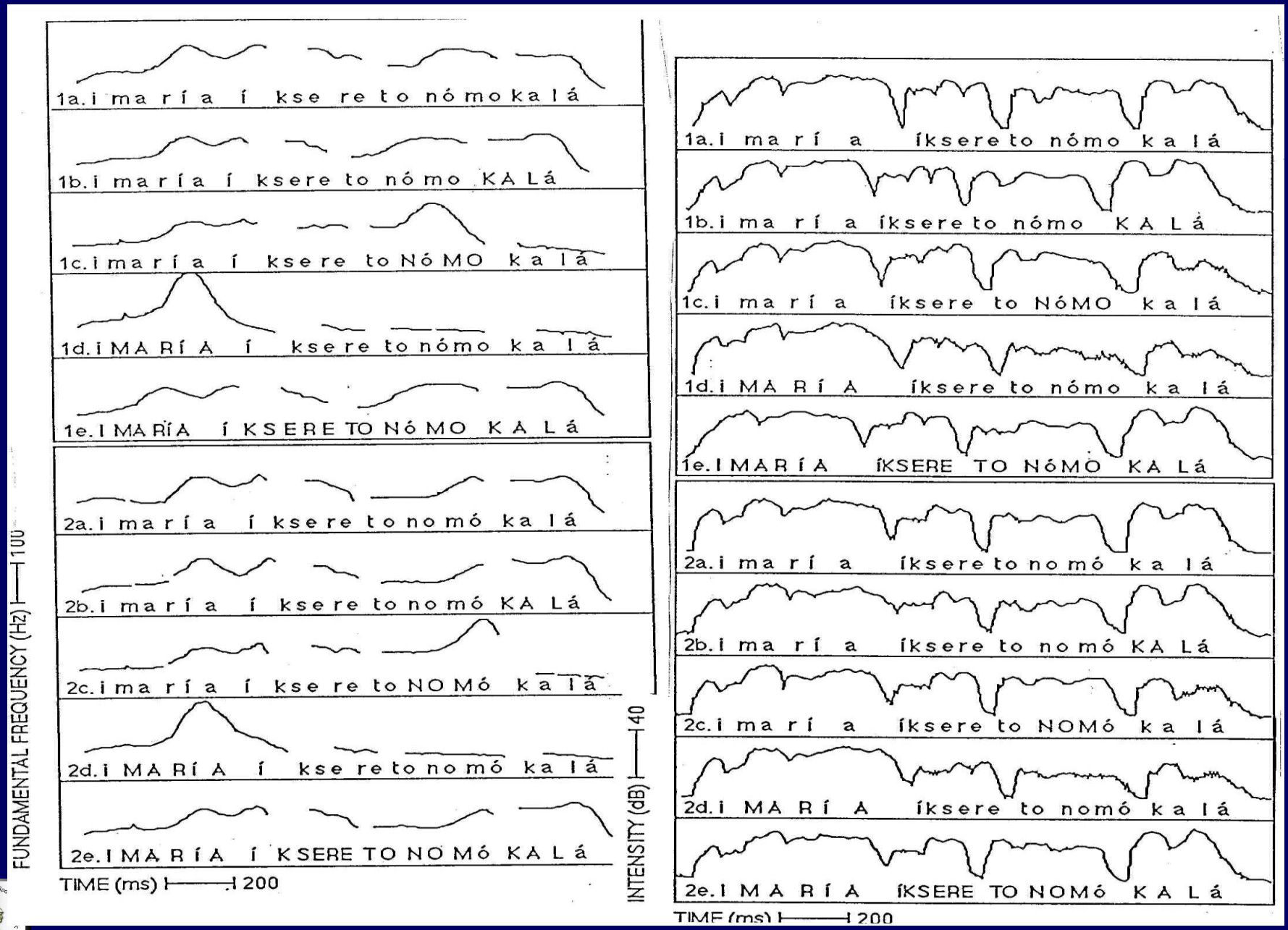
- Τονικές (κινέζικα, αφρικάνικα) & επιτονικές γλώσσες (ευρωπαϊκές).
- Τονικές: διαφορετικά επίπεδα τόνου αλλάζουν το νόημα μίας λέξης.
- Επιτονικές γλώσσες: free-stress (π.χ. Αγγλικά) και fixed-stress (π.χ. Γαλλικά).
- Ο μελωδικός τονισμός και η καμπύλη επιτονισμού εξαρτώνται από την οργάνωση των λέξεων σε μονάδες υψηλότερου επιπέδου (προσωδιακές λέξεις, ονοματικά σύνολα, ενδιάμεσες φράσεις κλπ).

Προσωδία (2/5)

- Εκδηλώνεται σε επίπεδο **φωνημάτων, συλλαβών, φράσεων, προτάσεων, ...**

- Κάθε ένα επίπεδο είναι μια επεξεργασμένη έκδοση του προηγουμένου.

- Σε επίπεδο πρότασεις μπορεί να υποδείξει την εστίαση (focus).
  - “Μου αρέσει να ακούω τους αγαπημενους **μου** δίσκους”.
  - “**Μου αρέσει** να ακούω τους αγαπημενους μου δίσκους”.
  - “Μου αρέσει να ακούω τους **αγαπημενους** μου δίσκους”.
  - “Μου αρέσει να ακούω τους αγαπημενους μου **δίσκους**”.


1a.i ma rí a í kse re to nómo ka lá
1b.i ma ría í ksere to nómo KA Lá
1c.i ma rí a í ksere to NóMO ka lá
1d.i MA RÍ A í ksere to nómo ka lá
1e.I MA RÍA Í KSERE TO NÓMO KA LÁ
2a.i ma rí a í ksere to no mó ka lá
2b.i ma rí a í ksere to no mó KA Lá
2c.i ma rí a í kse re to NO Mó ka lá
2d.i MA RÍ A í kse re to no mó ka lá
2e.I MA RÍ A Í KSERE TO NO Mó KA LÁ
FUNDAMENTAL FREQUENCY (Hz) 100
TIME (ms) 200
INTENSITY (dB) 40
1a.i ma rí a íksere to nómo ka lá
1b.i ma rí a íksere to nómo KA Lá
1c.i ma rí a íksere to NóMO ka lá
1d.i MA RÍ A íksere to nómo ka lá
1e.I MA RÍ A ÍKSERE TO NÓMO KA Lá
2a.i ma rí a íksere to no mó ka lá
2b.i ma rí a íksere to no mó KA Lá
2c.i ma rí a íksere to NOMó ka lá
2d.i MA RÍ A íksere to nomó ka lá
2e.I MA RÍ A ÍKSERE TO NOMó KA Lá
TIME (ms) 200

Προσωδία (3/5)

- Εστίαση (focus): «Ο Νίκος ήρθε στη Νάξο με πλοίο».
  - Σωστό μήκος συλλαβών και επιτονισμός.
  - Λάθος εστίαση οδηγεί σε παρεξήγηση.
  - Το TtS πρέπει να καταλαβαίνει όχι απομονωμένες προτάσεις αλλά όλο το κείμενο...

Προβλήματα κατά τον επιτονισμό

- Που μπαίνουν οι μελωδικοί τόνοι;
- Που μπαίνουν τα τονικά όρια;
- Τι σχήμα έχουν;
- Τι μέγεθος, μήκος και θέση έχουν στη φράση;

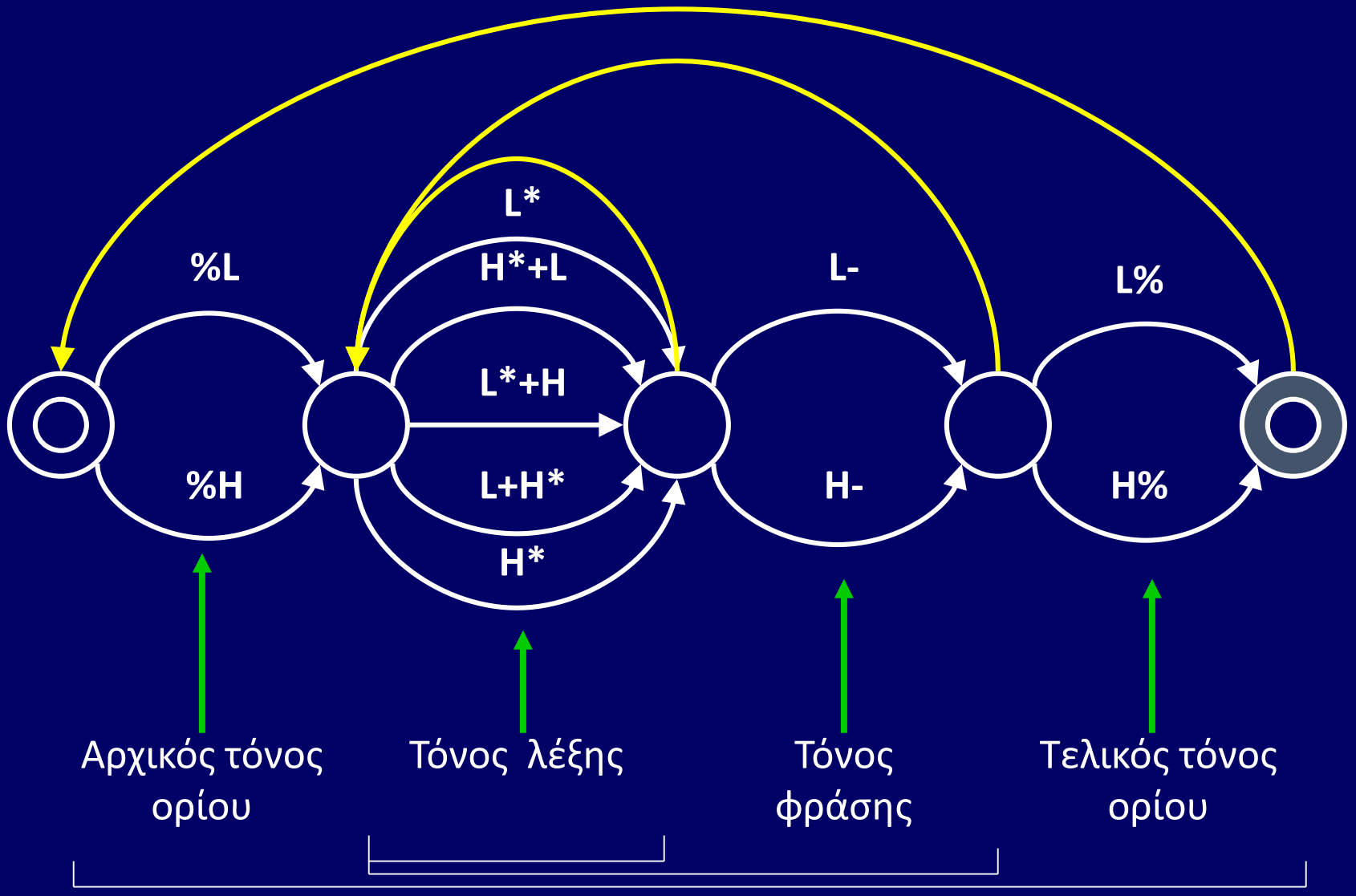
ToBI labelling system

- Το ToBI (**To**nes and **B**reak **I**ndices) http://www.ling.ohio-state.edu/~tobi είναι ένα πλαίσιο για την ανάπτυξη συμβάσεων ευρείας αποδοχής για την μετεγγραφή της επιτονικής και προσωδιακής δομής των προφορικών εκφράσεων σε μία ποικιλία γλωσσών.

- Υποθέτει μία στενή σχέση ανάμεσα στον επιτονισμό και σε ένα ιεραρχικό μοντέλο από προσωδιακές συνιστώσες.

- Δεν αποτελεί ένα διεθνές φωνητικό αλφάβητο για την προσωδία καθώς η προσωδιακή οργάνωση διαφέρει από γλώσσα σε γλώσσα και συχνά από διάλεκτο σε διάλεκτο μέσα στην ίδια γλώσσα.

Προσωδία με βάση το ToBi



Προσωδία (4/5)

Τα βήματα που ακολουθεί ένα TtS προκειμένου να γεννήσει τα προσωδιακά διανύσματα μίας φράσης:

1. Αναγνώριση και κατηγοριοποίηση φράσεων.
2. Πρόβλεψη προσωδιακής δομής φράσεων.
    1. Θέση, διάρκεια και τύπος **παύσεων** (break indices).
    2. Θέση και τύπος **μελωδικών τόνων** (pitch accents).
    3. Θέση και τύπος **τόνων στα όρια** (endtones – prosodic phrase tones + boundary tones).
3. Υπολογισμός διανύσματος διάρκειας.
4. Απόδοση καμπύλης F_0.
5. Υπολογισμός διανύσματος έντασης.

Προσωδία (5/5)

- ***«Η θέλησή του Δημοσθένη ήταν τόσο μεγάλη, ώστε, όπως μας αναφέρει ο Πλούταρχος, έβαζε στο στόμα του μικρά χαλίκια την ώρα που απήγγειλε λόγους, προκειμένου να βελτιώσει την άρθρωσή του. Οι προσπάθειες του αυτές, απέδωσαν καρπούς και εξελίχθηκε σε σπουδαίο ρήτορα και πολιτικό.»***
- Από εμπειρικά μοντέλα... σε εκπαιδευόμενα...

 (2002)

 (2003)

Σύνθεση με μεγάλα σώματα φωνητικών βάσεων (Corpus based – Unit Selection)

Unit Selection (1/4)

- Units: ήμι-φωνήματα, φωνήματα, συλλαβές,...
- Μεγαλύτερα σώματα (ηχογραφήσεις) ομιλίας.
  - 3-5 ώρες από έναν ομιλητή.
- Κατά την σύνθεση, επιλογή των units που:
  - ταιριάζουν καλύτερα στα προσωδιακά χαρακτηριστικά της φράσης.
  - συρράπτονται πιο ομαλά με τα ήδη επιλεγμένα units.
- Κάθε **unit στη βάση** έχει **χαρακτηριστικά** όπως pitch, διάρκεια, θέση σε σχέση με άλλα units.
- Κάθε **επιθυμητό (target) unit σύνθεσης** αποχτάει επίσης ανάλογα **χαρακτηριστικά**.

Unit Selection (2/4)

- Μία συνάρτηση κόστους (target cost) υπολογίζει την απόσταση μεταξύ των επιθυμητών τιμών και των αποθηκευμένων για κάθε υποψήφιο προς επιλογή unit.
- Στόχος η αναζήτηση μίας συνολικά ελάχιστης απόστασης:
  - Ελαχιστοποίηση συνολικού κόστους C για μία ακολουθία n units.

$$C = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i)$$

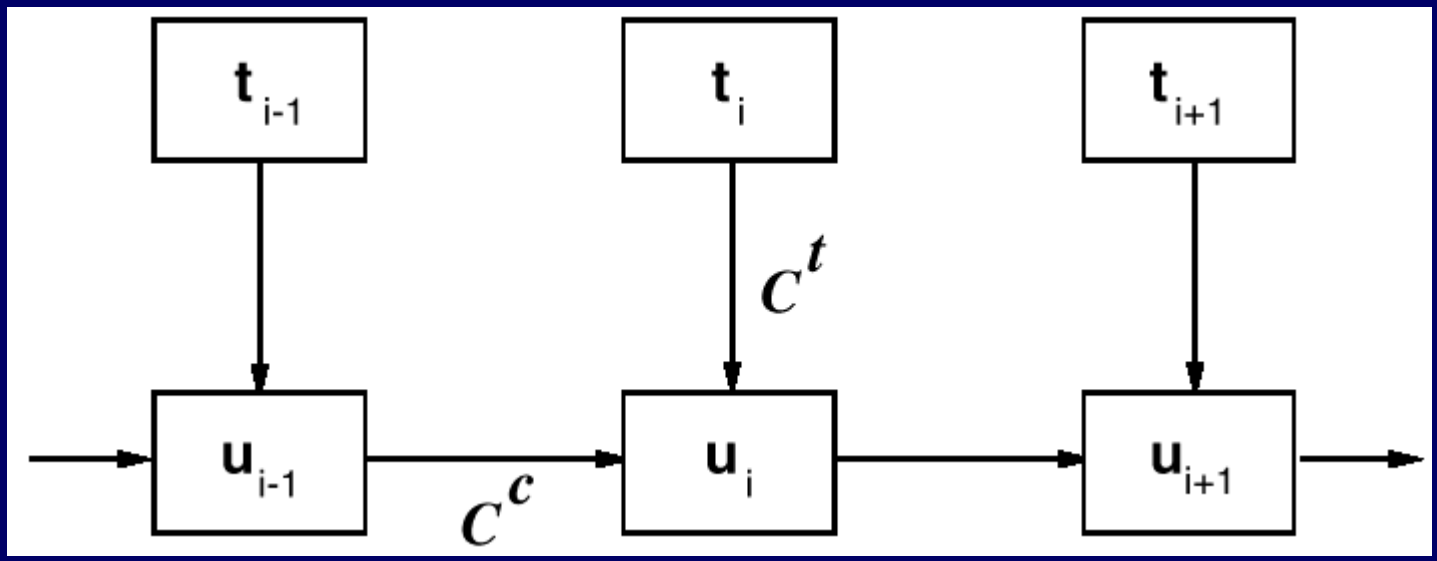
C^t, κόστος για κάθε unit σε σχέση με το επιθυμητό t_i.

C^c, κόστος συρραφής (ακουστικές ιδιότητες) για κάθε ζεύγος unit

Unit Selection (3/4)

Δεδομένης μίας ακολουθίας $t_i^n = (t_1, \dots, t_n)$ ζητάμε να υπολογίσουμε ένα σύνολο από units, $u_i^n = (u_1, \dots, u_n)$.

t_i = επιθυμητά (target) προσωδιακά και ακουστικά χαρακτηριστικά.



Οι παράμετροι της ομιλίας εμπεριέχονται σε μικρά ηχογραφημένα τμήματα (φωνήματα, δίφωνα, λέξεις, φράσεις).

Unit Selection (4/4)

- **Πλεονεκτήματα**
  - Ελαχιστοποιεί την ανάγκη επεξεργασίας των δειγμάτων (units) κατά την σύνθεση.
  - Μεγάλη φυσικότητα.
- **Μειονεκτήματα**
  - Οι αλγόριθμοι αναζήτησης μερικές φορές αστοχούν.
  - Συχνά «αφύσικη» προσωδία που όμως δεν αξιολογείται αρνητικά λόγω της υψηλής ποιότητας του ηχοχρώματος.
  - Μεγάλος όγκος βάσης, δυσκολία κατασκευής της και ανάλυσης των δεδομένων της.

Θεματικά πεδία (1/2)

- Πεδία εφαρμογής TtS στα οποία η θεματολογία είναι συγκεκριμένη.
  - Περιορισμένο λεξιλόγιο.
  - Περιορισμένα γλωσσολογικά φαινόμενα.
  - Περιορισμένα προσωδιακά φαινόμενα (από επαγγελματίες εκφωνητές του πεδίου).
  - Άρα, μικρότερο πεδίο προσομοίωσης και λιγότερα λάθη κατά την επεξεργασία
- Ευκολύνεται η δημιουργία στατιστικών μοντέλων από πραγματικές μετρήσεις (π.χ. Μορφοσυντακτική ανάλυση, προσωδία → ~95%)
- Ακουστικό σήμα: γίνεται προσιτή η χρήση μεγαλύτερων ακουστικών μονάδων συρραφής → μεγαλύτερη φυσικότητα

Θεματικά πεδία (2/2)

- Παραδείγματα:
  - Υπηρεσίες καταλόγου (131, σινεμά, βενζινάδικα κλπ)
  - Ανακοινώσεις (δελτίο καιρού, οικονομικό δελτίο κλπ)
  - Περιήγηση σε μουσεία
  - Τεχνικά ή νομικά κείμενα
  - Ειδήσεις
- Με βάση τους περιορισμούς που πρέπει να θέτει ένα θεματικό πεδίο, δεν θεωρούνται θεματικά πεδία:
  - Λογοτεχνικά κείμενα
  - Ποίηση

...γενικών και περιορισμένων θεματικών πεδίων

- ***«Αυτό το έκθεμα είναι ένας αμφορέας και σήμερα βρίσκεται στο μουσείο της Αθήνας»***
- Rhetorical – rVoice 1 (γενικό)
- Rhetorical – rVoice 2 (γενικό)
- Loquendo – Actor 1 (γενικό)
- ΔΗΜΟΣΘέΝΗΣ – (περιορισμένο)

Τι είναι «κείμενο»;

- *«Ο Νίκος παίζει πιάνο.»*
- *«Στις 21/12 ο Νίκος θα παίξει 3 από τις 4 μπαλάντες του στο ΚΨΜ.»*
- *«Ο Νίκος (ο αδελφός του κ. Πέτρου) δουλεύει στο ΚΨΜ.»*
- *«Ο Νίκος είναι:*
  - *Ψηλός*
  - *Έξυπνος*
  - *Αφηρημένος»*
- Ο Νίκος έρχεται με τα πόδια, ενώ η Μαρία με αυτοκίνητο.
- <table><tr><td>Μοντέλο</td><td>Κυβικά</td></tr><tr><td>Ibiza</td><td>1600</td></tr><tr><td>Amazon</td><td>2000</td></tr>...</table>
- ...

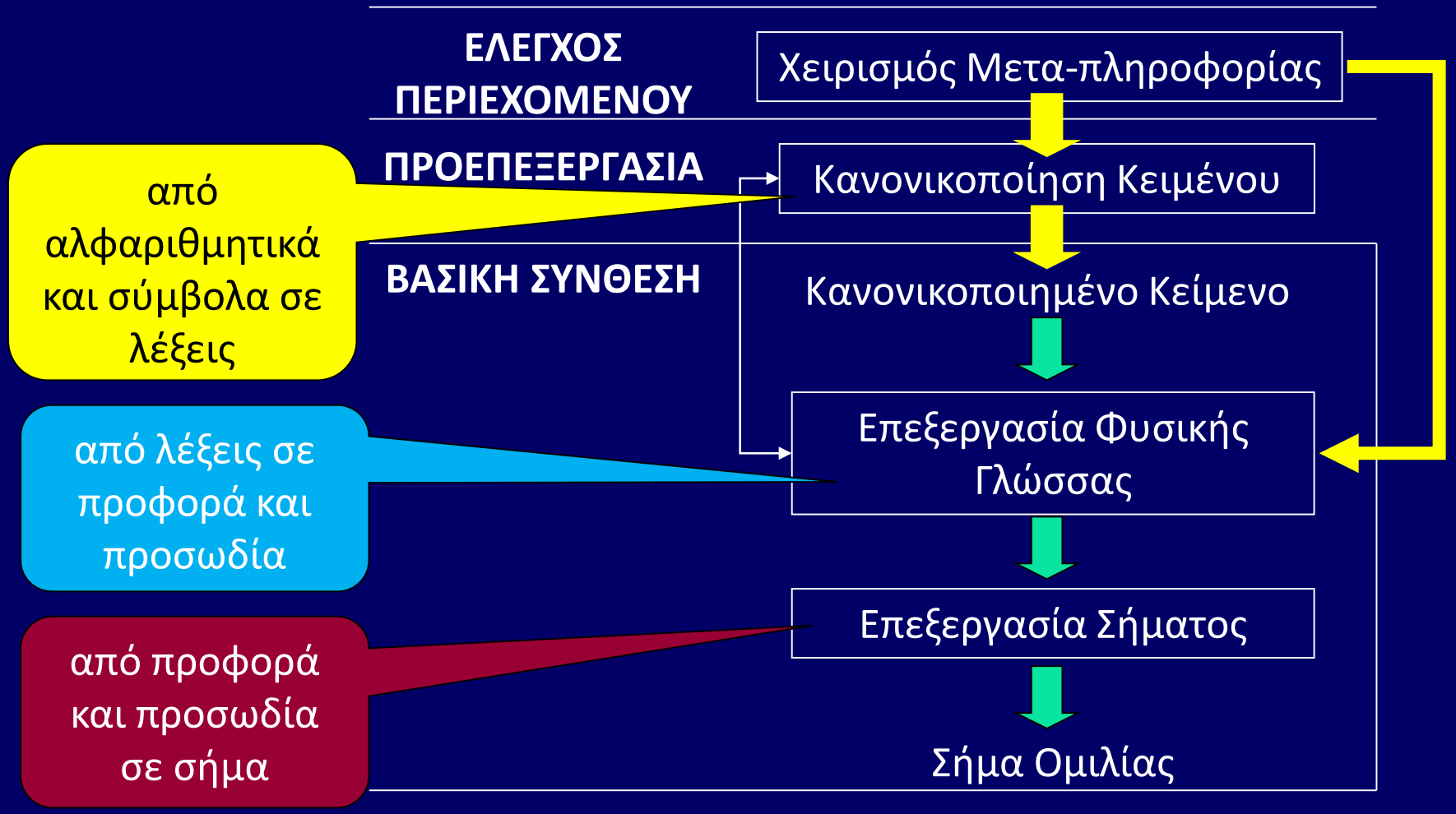
Τι είναι «συνθετική ομιλία»;

- Ομιλία που παράγεται με αυτόματες διαδικασίες από έναν ηλεκτρονικό υπολογιστή και προσομοιάζει την συμπεριφορά της ανθρώπινης ομιλίας.
- Σαν τεχνολογία υφίσταται πειραματικά από τα τέλη ’50 και εμπορικά από το ’70.
- Διεπαφή με τον χρήστη τελευταίας γενιάς.
- Από hardware (κουτάκια, χαμηλή ρομποτική ποιότητα) σε software (ευέλικτα, υψηλή ποιότητα)

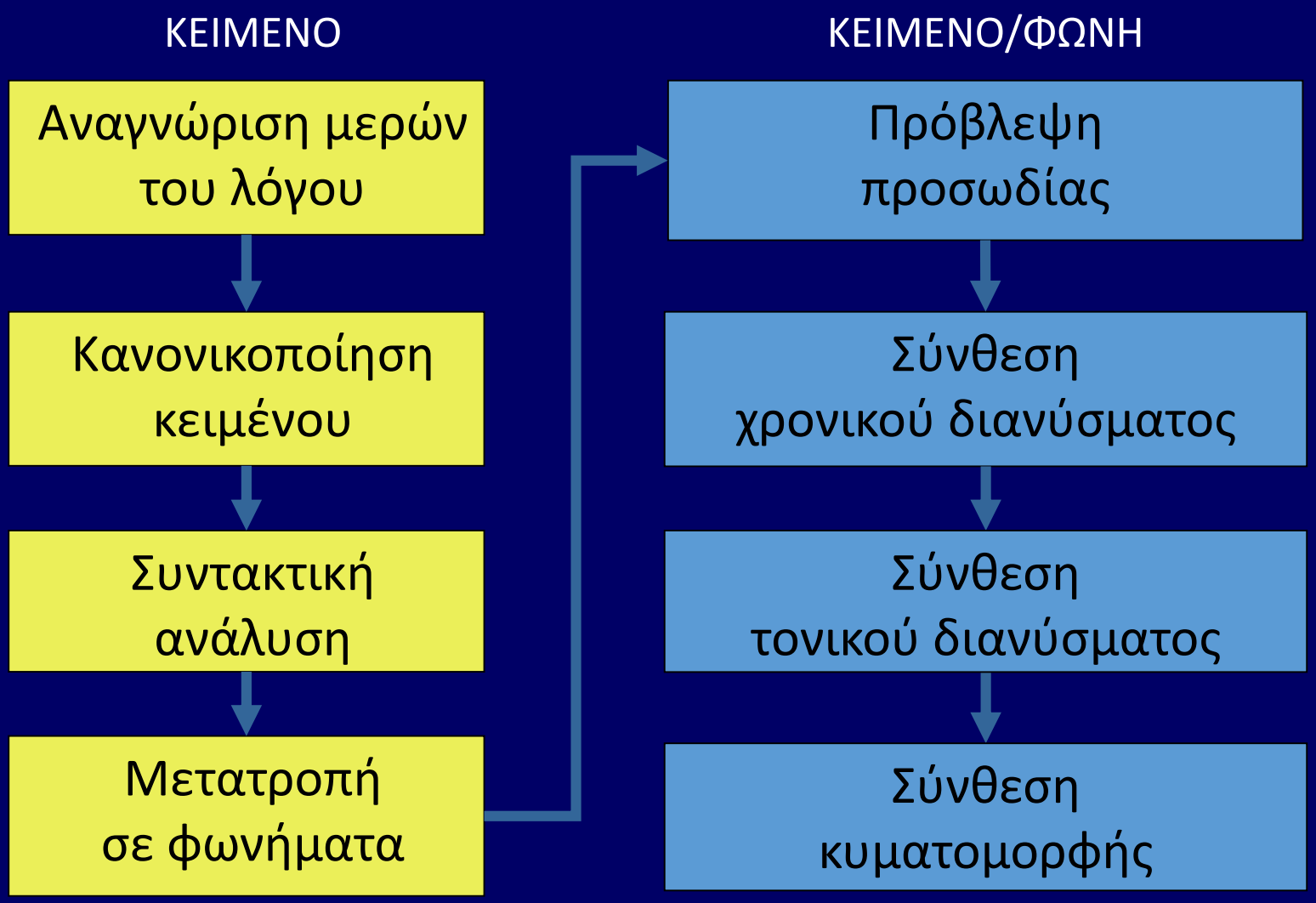
Τι είναι ένα «Σύστημα Μετατροπής Κειμένου σε Ομιλία»;

- Ένα σύστημα το οποίο δέχεται μία ακολουθία συμβόλων και με τεχνικές νοημοσύνης συνθέτει μία ή περισσότερες εκδοχές από αντίστοιχα ακουστικά σήματα ομιλίας.
- Ποιό είναι το πεδίο ορισμού των συμβόλων;
- Πόσο εξελιγμένες είναι οι τεχνικές νοημοσύνης;
- Πως επιλέγεται η κατάλληλη εκδοχή;
- Πόσο φυσικά είναι τα ακουστικά σήματα που παράγονται;

Διαδικασία Σύνθεσης



Κλασσική Μετατροπή ΚσΟ



Κανονικοποίηση κειμένου (1/5)

- «Μένω Πατησίων και 3ης Σεπτεμβρίου».
- «Πάρε με στο 210 7275320».
- «Ήταν 13 Ιανουαρίου».
- «Ήταν 13 του μηνός».
- «Ήταν 1 γυναίκα, 1 άντρας και 1 παιδί».
- «Εξετάστηκαν οι αιτήσεις 1654 φοιτητών».
- «Ελάτε μεταξύ 3 και 4 η ώρα το απόγευμα».
- «1500 λουλούδια ήταν οι υποψήφιες».
- «1500 ήταν οι υποψήφιες».

Κανονικοποίηση κειμένου (2/5)

	EXPN	abbreviation	*adv, N.Y, mph, gov't*
alpha	LSEQ	letter sequence	*CTA, D.C, CDs*
	ASWD	read as word	*CAT*, proper names
	MSPL	misspelling	*geogaphy*
	NUM	number (cardinal)	*12, 45, 1/2, 0.6*
	NORD	number (ordinal)	*May 7, 3rd, Bill Gates III*
	NTEL	telephone (or part of)	*212 555-4523*
	NDIG	number as digits	*Room 101*
N	NIDE	identifier	*747, 386, I5, pc110, 3A*
U	NADDR	number as street address	*5000 Pennsylvania, 4523 Forbes*
M	NZIP	zip code or PO Box	*91020*
B	NTIME	a (compound) time	*3.20, 11:45*
E	NDATE	a (compound) date	*2/2/99, 14/03/87* (or US) *03/14/87*
R	NYER	year(s)	*1998, 80s, 1900s, 2003*
S	MONEY	money (US or other)	*$3.45, HK$300, Y20,000, $200K*
	BMONEY	money tr/m/billions	*$3.45 billion*
	PRCT	percentage	*75%, 3.4%*
	SPLT	mixed or "split"	*WS99, x220, 2-car* (see also SLNT and PUNC examples)
	SLNT	not spoken,	word boundary or emphasis character:
M		word boundary	*M.bath, KENT*RLTY, _really_*
I	PUNC	not spoken,	non-standard punctuation: "***" in
S		phrase boundary	*$99,9K***Whites*, "..." in *DECIDE...Year*
C	FNSP	funny spelling	*sllooooowww, sh*t*
	URL	url, pathname or email	*http://apj.co.uk, /usr/local, phj@tpt.com*
	NONE	should be ignored	ascii art, formatting junk

Συντακτικές συμφωνίες (3/5)

- Οι κλιτές (inflected) γλώσσες έχουν επιπλέον το πρόβλημα της κλίσης των Non Standard Words (NSW).
- Για την αντιμετώπιση του προβλήματος χρησιμοποιούμε (ΔΗΜΟΣΘέΝΗΣ) γλωσσολογική επεξεργασία που στηρίζεται στη Γραμματική της Νέας Ελληνικής (Μπαμπινιώτη – Χρήστου).
- Φροντίζουν για τη γραμματική συνέπεια ανάμεσα σε αριθμητικά και ουσιαστικά.
- 36,34% των περιπτώσεων βασίζονται σε αυτές τις συμφωνίες για την ορθή κανονικοποίηση τους.
- Συμφωνία ονοματικών συνόλων: *1636 φοιτητών*
- Αντικείμενο στην αιτιατική: *Το μουσείο δέχεται καθημερινά 1500 επισκέπτες*
- Κατηγορούμενο στην ονομαστική: *Οι επιτυχόντες είναι 1501*
- ...

Κανονικοποίηση κειμένου (5/5)

- «Μένω Πατησίων και 3ης Σεπτεμβρίου».
- «Ήταν 13 Ιανουαρίου».
- «Ήταν 13 του μηνός».
- «Πάρε με στο 210 7275320».
- «Ήταν 1 γυναίκα, 1 άντρας και 1 παιδί».
- «Εξετάστηκαν οι αιτήσεις 1654 φοιτητών».
- «Ελάτε μεταξύ 3 και 4 η ώρα το απόγευμα».
- «1500 λουλούδια ήταν οι υποψήφιες».
- «1500 ήταν οι υποψήφιες».

Μετατροπή σε φωνήματα

- Ταύτιση της προφορικής αναπαράστασης με τη γραπτή.
- Εξαρτάται από ιδιώματα κάθε περιοχής.

Τα παιδιά είδαν για μιά στιγμή τον Κώστα και την Χαρά σε μία ταβέρνα.»

«ta pε□iά i□αn jια □iά stiγmi ton cόστα κε ti xará σε mια tavérna»

- «χαρά» «xarÜ»
- «χέρι» «÷Ýri»

Αναγνώριση μερών του λόγου

- Απαραίτητη διαδικασία για:
  - Κανονικοποίηση κειμένου σε κλιτές γλώσσες (π.χ. Ελληνικά)
  - Μετατροπή σε φωνήματα (π.χ. «χρόνια», «record»)
  - Πρόβλεψη προσωδιακής δομής μίας πρότασης:
    - Προσδιορισμός προσωδιακών φράσεων
    - Προσδιορισμός παύσεων στις προσωδιακές φράσεις
    - Προσδιορισμός μελωδικού τονισμού (pitch accent) σε φράσεις
    - Προσδιορισμός τόνων ορίων (boundary tones) σε φράσεις
- Ποσοστό επιτυχίας (ελληνικά): ~95%

«Φωνή»

- Ο καθένας θέλει μία φωνή με προσωποποιημένα χαρακτηριστικά.
- Τα ομιλούντα προϊόντα (θα) θέλουν να έχουν την δική τους ξεχωριστή φωνή.
- Τι είναι φωνή;
  - Καταρχήν, ένα συγκεκριμένο ηχόχρωμα
  - Μία ακολουθία από διαδικασίες επεξεργασίας φυσικής γλώσσας: διαφορετικές τοπολαλιές, ιδιώματα, τρόποι προφοράς συνεπτηγμένων μορφών, προσωδιακή συμπεριφορά.

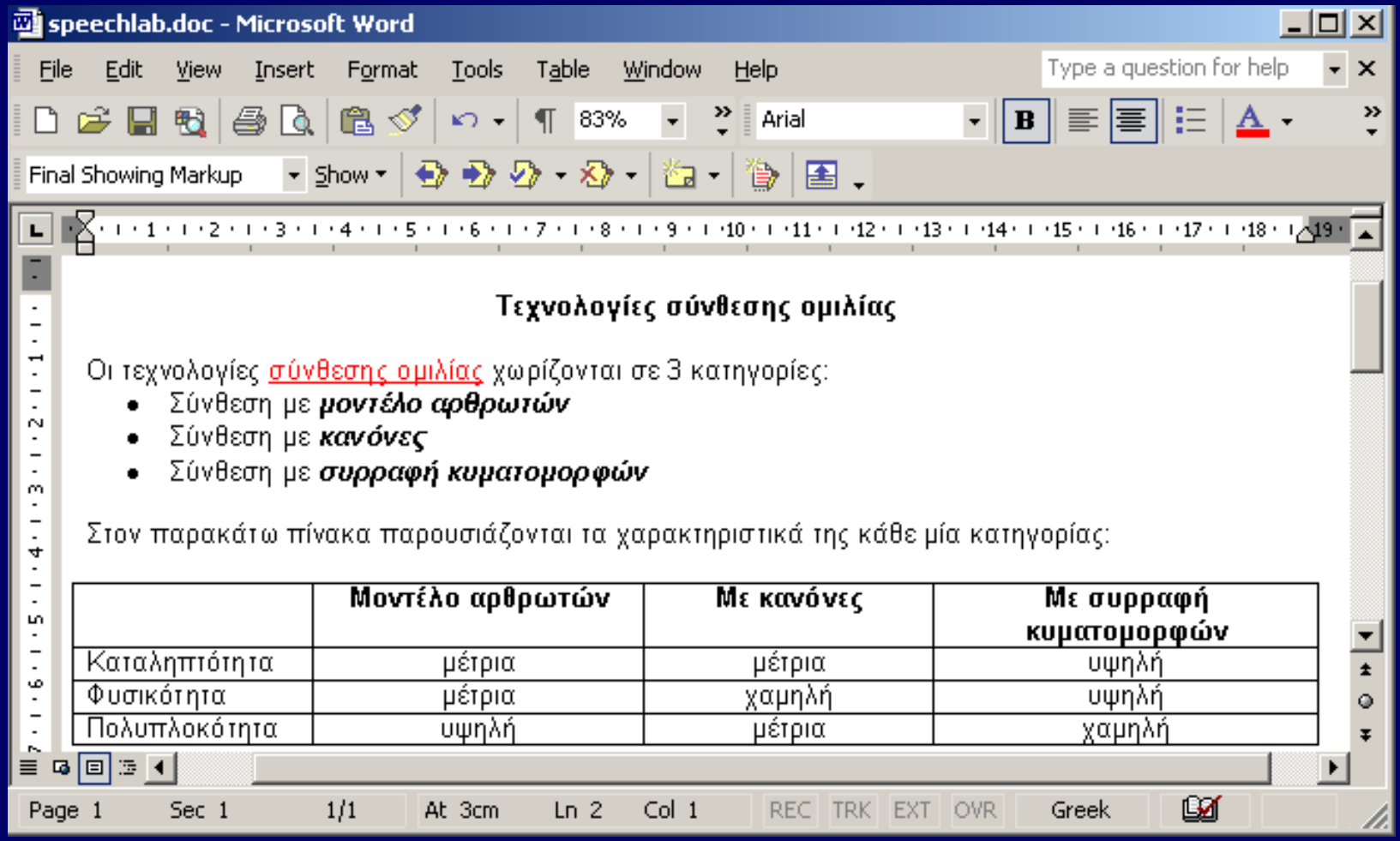
Ποιότητα συνθετικής ομιλίας

- **Καταληπτότητα**: ήταν κατανοητό το περιεχόμενο της ομιλίας;
- **Φυσικότητα**: πόσο κοντά στην φυσική χροιά ήταν η ομιλία;
- Επιπλέον, μεταδόθηκε σωστά η **πληροφορία**;

Μετα-πληροφορία (1/2)

- **Οδηγίες οπτικοποίησης**: **bold**, *italics*, tables, bullets, big-small letters κλπ (π.χ. HTML, MS-Word)
- **Οδηγίες δομής**: header, title, section, record content κλπ (π.χ. HTML, XML, SQL)
- **Δομή κειμένου**: παρενθέσεις, σημάδια, κλπ
- **Γλωσσολογική πληροφορία**: ρητορικές σχέσεις, σύνταξη, γραμματική, μορφολογία (π.χ. SOLE, plain κείμενο)
- **Οδηγίες ομιλίας**: prosody, emp, rate, pitch κλπ (π.χ. SABLE, VoiceXML, SSML, ACSS)

Μετα-πληροφορία (2/2)



	Μοντέλο αρθρωτών	Με κανόνες	Με συρραφή κυματομορφών
Καταληπτότητα	μέτρια	μέτρια	υψηλή
Φυσικότητα	μέτρια	χαμηλή	υψηλή
Πολυπλοκότητα	υψηλή	μέτρια	χαμηλή

Text-to-Speech

Doc-to-Speech

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. "