ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Αλληλεπίδραση Ανθρώπου–Υπολογιστή

Β10. Αναγνώριση ομιλίας

(2022–23)

Ίων Ανδρουτσόπουλος
http://www.aueb.gr/users/ion/

Οι διαφάνειες αυτές βασίζονται στην ύλη του βιβλίου *Speech and Language Processing* των D. Jurafsky και J.H. Martin, 2η έκδοση, Pearson Education, 2009 και 3η έκδοση (υπό προετοιμασία).

Τι θα ακούσετε

- **Εισαγωγή** στην αναγνώριση ομιλίας.
- **Δημιουργία παραστάσεων τμημάτων** ομιλίας με **προ-εκπαιδευμένους Transformers**.
  - wav2vec, HuBERT.
- **Μοντέλα αναγνώρισης ομιλίας**.
  - Μοντέλα κωδικοποιητή/αποκωδικοποιητή.
  - Χρήση γλωσσικών μοντέλων.
- **Μέτρα αξιολόγησης** αναγνώρισης ομιλίας.

Γιατί αναγνώριση ομιλίας;

- Χρήστες με **δυσκολίες ακοής, κίνησης, όρασης**.
  - Π.χ. αυτόματη **παραγωγή υποτίτλων**.
  - **Χειρισμός συσκευών** μέσω **προφορικών εντολών**.
- Όταν τα **χέρια** ή τα **μάτια** είναι **απασχολημένα**.
  - Π.χ. **περπάτημα, οδήγηση**.
- Ιδιαίτερα σε συστήματα **προφορικών διαλόγων**.
  - Π.χ. **κλείσιμο εισιτηρίων** μέσω **τηλεφώνου**.
- **Εξαγωγή πληροφοριών** ή **γνώμης**.
  - Π.χ. από τηλεφωνικές **συνδιαλέξεις** ή **εκπομπές**.
- **Φυσικότερη** ή **εντυπωσιακότερη** επικοινωνία.
  - Π.χ. **υπαγόρευση** μηνυμάτων ή κειμένου.
  - Π.χ. αλληλεπίδραση με **ρομπότ** ή **παιχνίδια**.
  - Προσοχή στις **λανθασμένες προσδοκίες!**

Τι επηρεάζει την αναγνώριση;

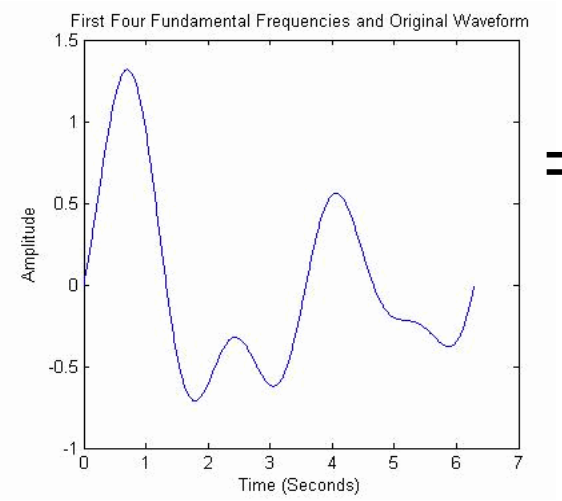
- **Μέγεθος λεξιλογίου**.
  - **Εύκολο**: αναγνώριση **αριθμών** ή **δεκάδων λέξεων**.
  - **Πιο δύσκολο**: αναγνώριση **δεκάδων χιλιάδων λέξεων** (π.χ. στην υπαγόρευση κειμένου).
- **Μεμονωμένες λέξεις** ή **συνεχής ομιλία**.
  - Σε συνομιλίες **μεταξύ ανθρώπων** συνήθως **δεν υπάρχουν κενά** μεταξύ των λέξεων.
  - Η αναγνώριση **μεμονωμένων** λέξεων είναι πιο **εύκολη**.
- Για **συγκεκριμένο χρήστη** ή όχι;
  - Π.χ. τα συστήματα υπαγόρευσης συχνά βελτιώνονται με **δείγματα ομιλίας** του **συγκεκριμένου χρήστη**.
  - Τα περισσότερα συστήματα πλέον **δεν απαιτούν** ειδική **εκπαίδευση ανά χρήστη**.

Τι επηρεάζει την αναγνώριση;

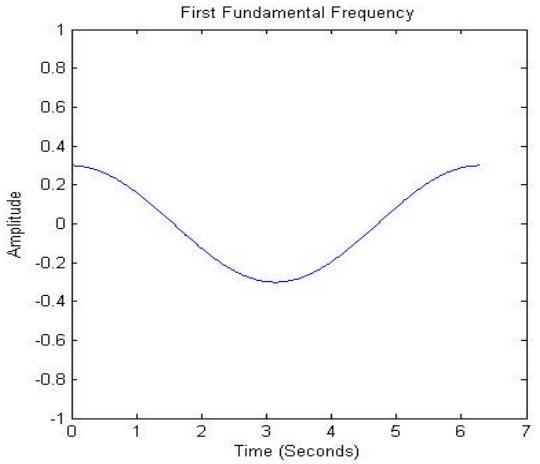
- **Μητρική γλώσσα** ή όχι; **Διάλεκτοι**…. **Ηλικία**…
  - Συνήθως υποστηρίζονται καλύτερα **συγκεκριμένες γλώσσες** και **διάλεκτοι,** κυρίως για **ενήλικες**.
- **Μικρόφωνα**, **πλήθος χρηστών**, **θόρυβος**.
  - **Ευκολότερο**: **ένας χρήστης** με **ακουστικό κεφαλής** σε **ήσυχο γραφείο**.
  - **Πολύ δυσκολότερο**: **πολλοί χρήστες** σε **θορυβώδες περιβάλλον** (π.χ. συνεδρίαση) με **μακρινά μικρόφωνα**.
- **Είδος συνομιλίας**.
  - Η αυτόματη αναγνώριση ομιλίας **μεταξύ ανθρώπων** (π.χ. **πρακτικά συνεδριάσεων**) είναι πολύ πιο **δύσκολη**.
  - Οι **άνθρωποι απλοποιούν** την ομιλία τους όταν μιλούν σε **μηχανές** (ή σε **παιδιά** ή σε μαθητές **ξένων γλωσσών**).

Μετασχηματισμός Fourier

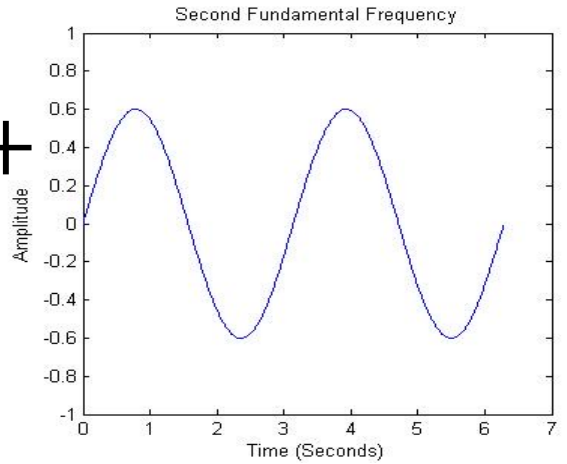
- Μπορούμε να σκεφτούμε **κάθε ήχο** (ή σήμα) ως **άθροισμα πολλών** (γενικά άπειρων) **ημιτονοειδών**.



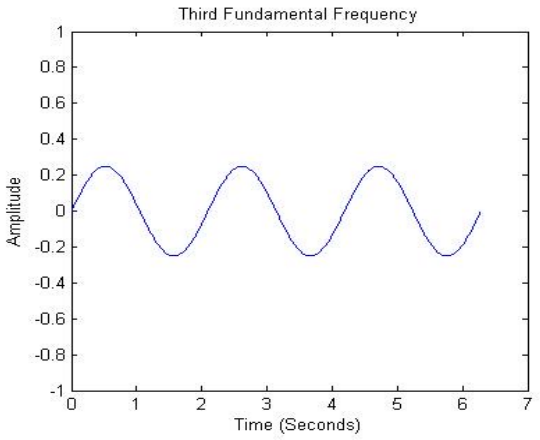
=



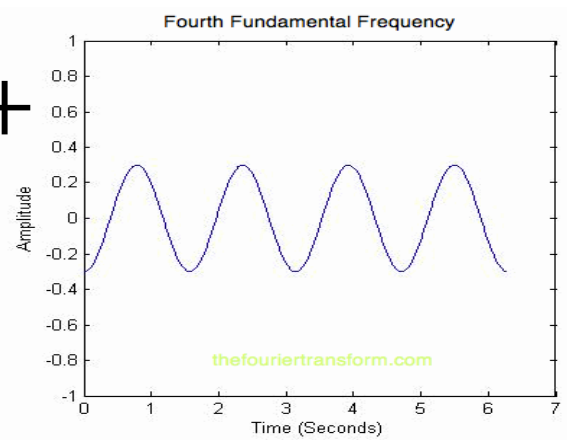
+



+

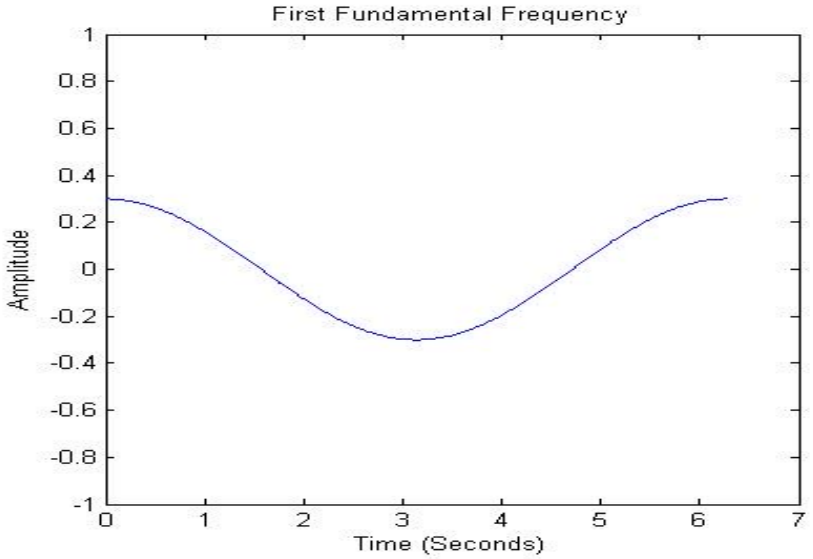


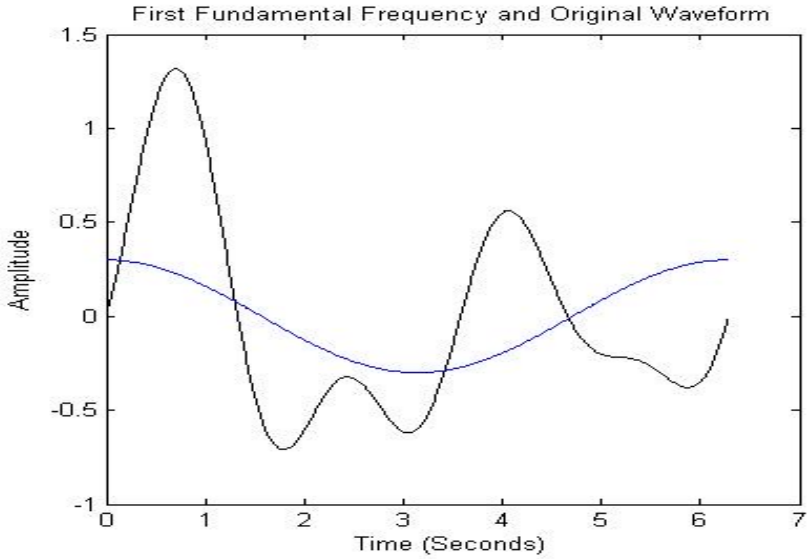
+

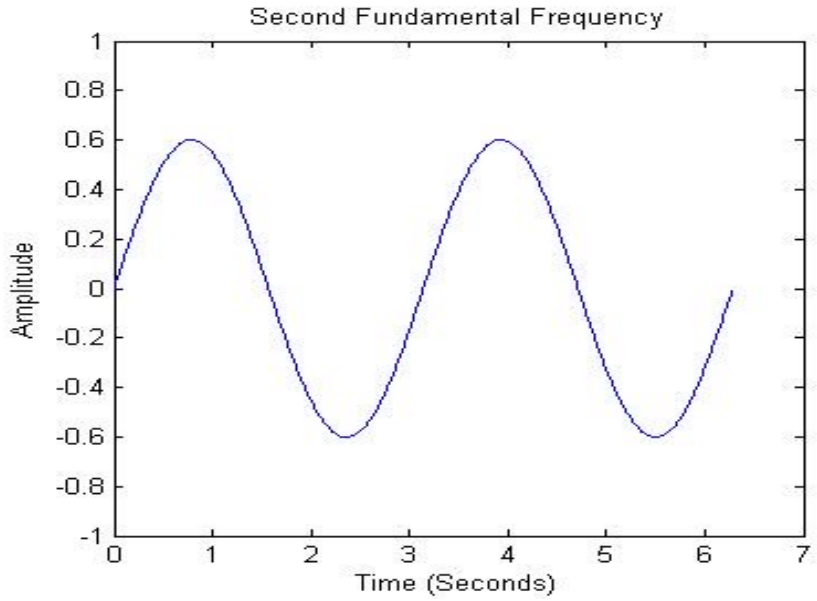


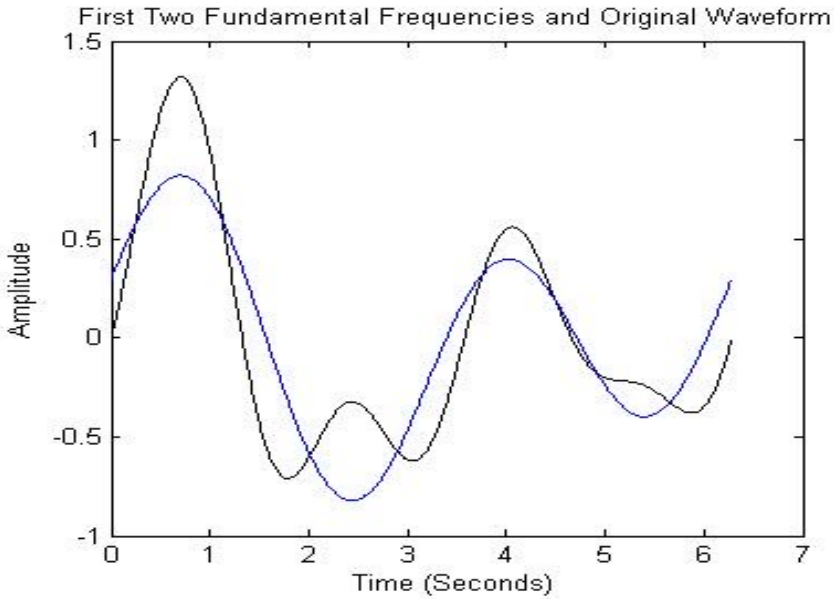
Σχήματα από τον ιστότοπο
http://www.thefouriertransform.com/.

Μετασχηματισμός Fourier



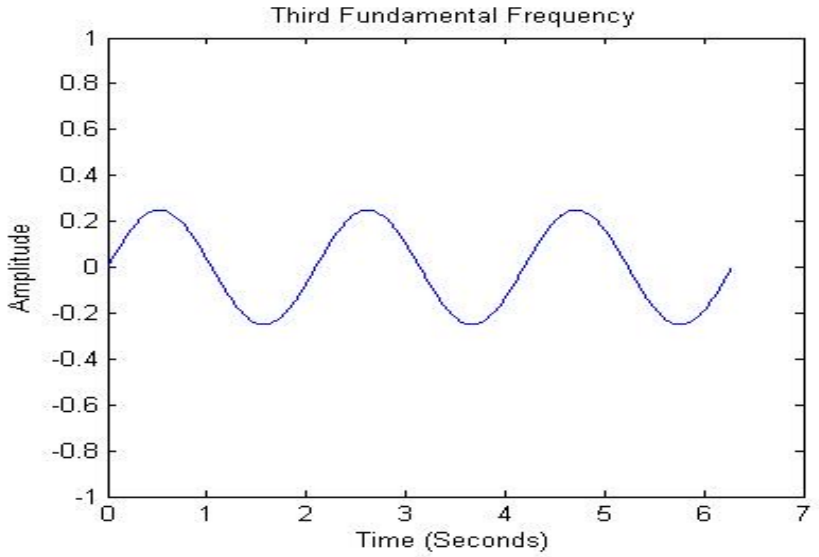


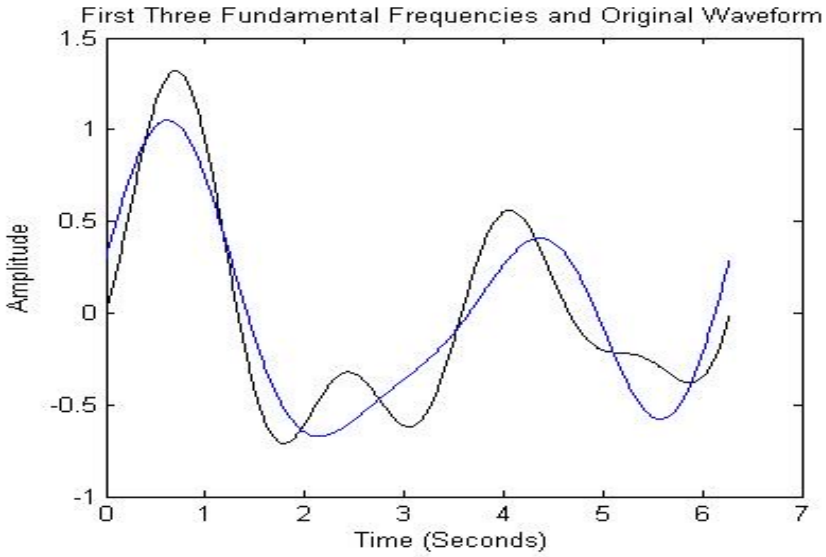


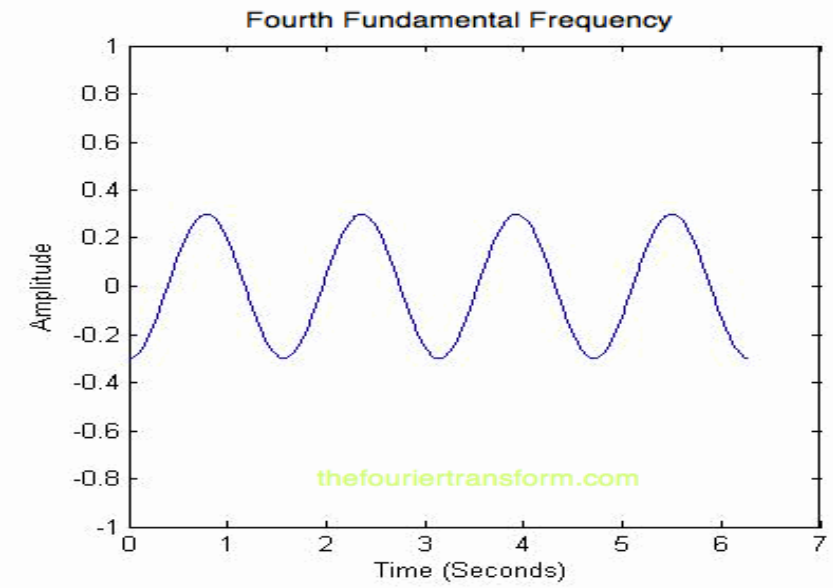


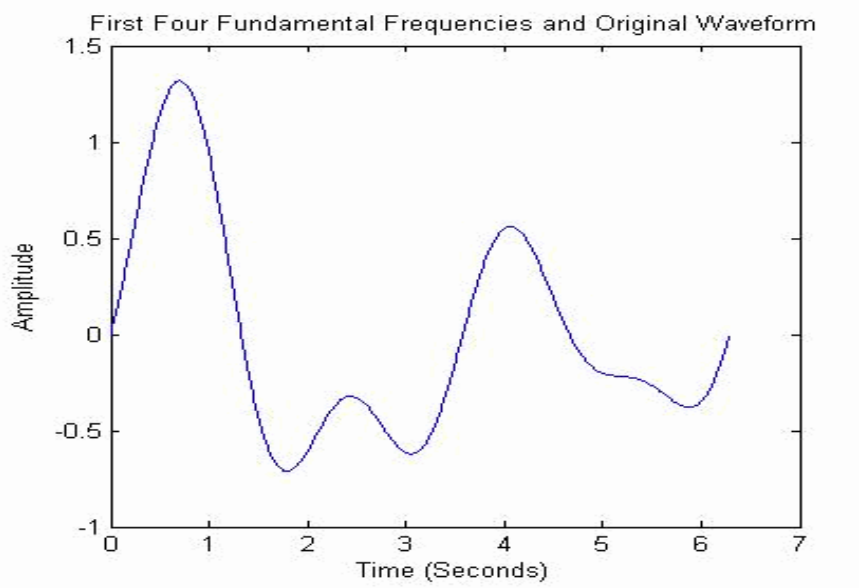
Σχήματα από τον ιστότοπο http://www.thefouriertransform.com/.

Μετασχηματισμός Fourier





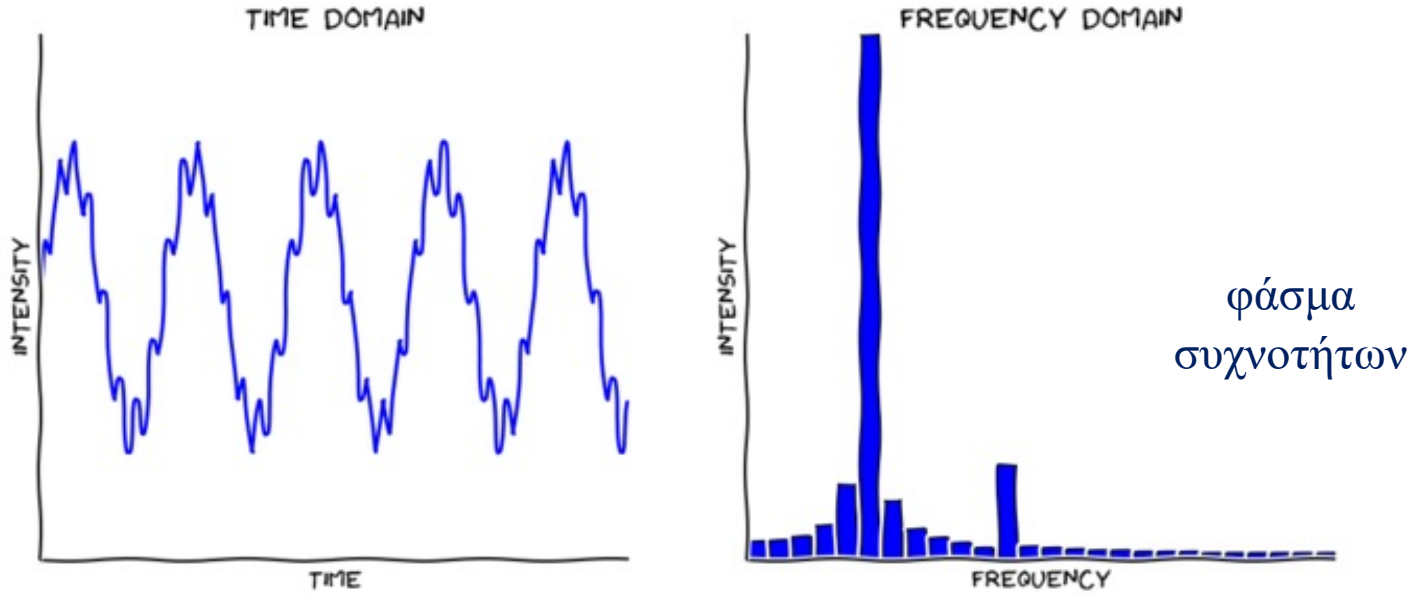




Σχήματα από τον ιστότοπο http://www.thefouriertransform.com/.

Μετασχηματισμός Fourier

- **Μετατρέπει** το **αρχικό σήμα** $f(t)$ (συνάρτηση του **χρόνου** t) σε **μιγαδική συνάρτηση** $\hat{f}(\xi)$ της **συχνότητας** (ξ).
  - $\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(t) \cdot e^{-2\pi \cdot i \cdot t \cdot \xi} dt \quad (e^{i \cdot \theta} = \cos\theta + i \cdot \sin\theta)$
  - Το **μέτρο** του **μιγαδικού** $|\hat{f}(\xi)|$ δείχνει **πόσο συμμετέχει η συχνότητα** ξ στο αρχικό σήμα.

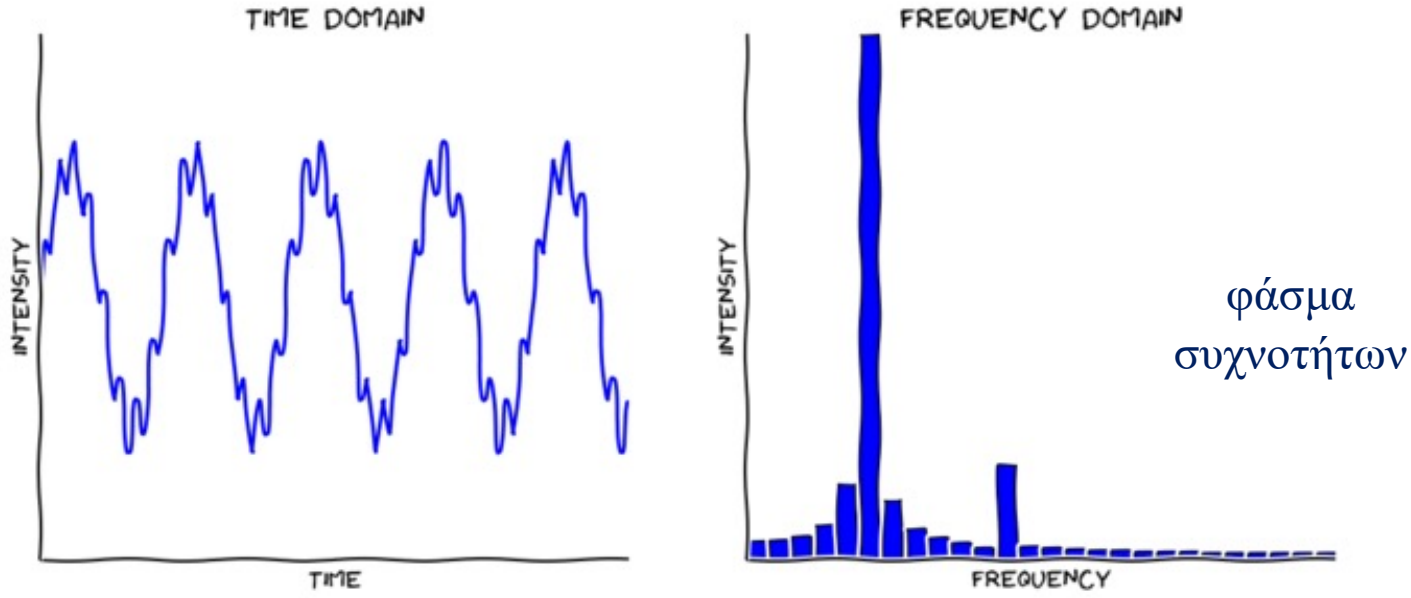


φάσμα συχνοτήτων

Σχήματα από την ιστοσελίδα https://learn.adafruit.com/fft-fun-with-fourier-transforms/background

Διακριτός μετασχηματισμός Fourier (DFT)

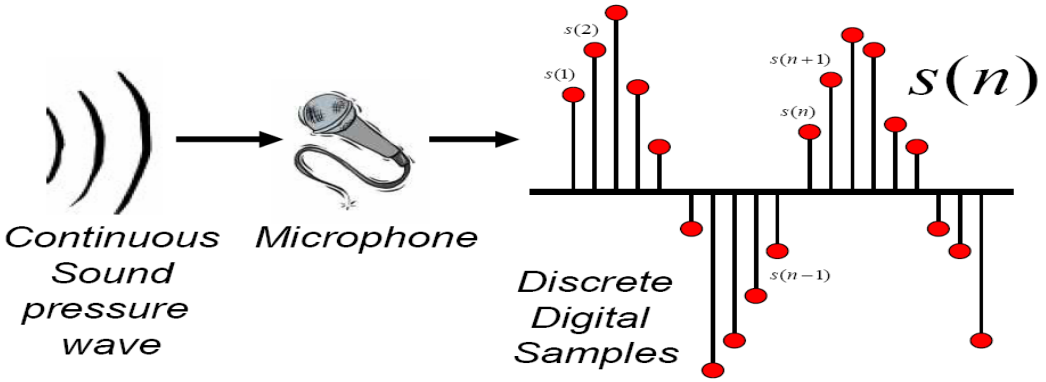
- Για **διακριτό σήμα** $x[0], \dots, x[N-1]$ και N **διακριτές συχνότητες** ξ:
  - $\hat{x}(\xi) = \sum_{n=0}^{N-1} x[n] \cdot e^{\frac{-2\pi \cdot i \cdot n \cdot \xi}{N}}$ $\quad (e^{i\cdot\theta} = \cos\theta + i \cdot \sin\theta)$
  - Αν $N = 2^{\mathrm{m}}$ (δύναμη του 2), μπορούμε να χρησιμοποιήσουμε τον **αλγόριθμο FFT** (Fast Fourier Transform).



φάσμα συχνοτήτων

Σχήματα από την ιστοσελίδα https://learn.adafruit.com/fft-fun-with-fourier-transforms/background

Ψηφιακή παράσταση σήματος

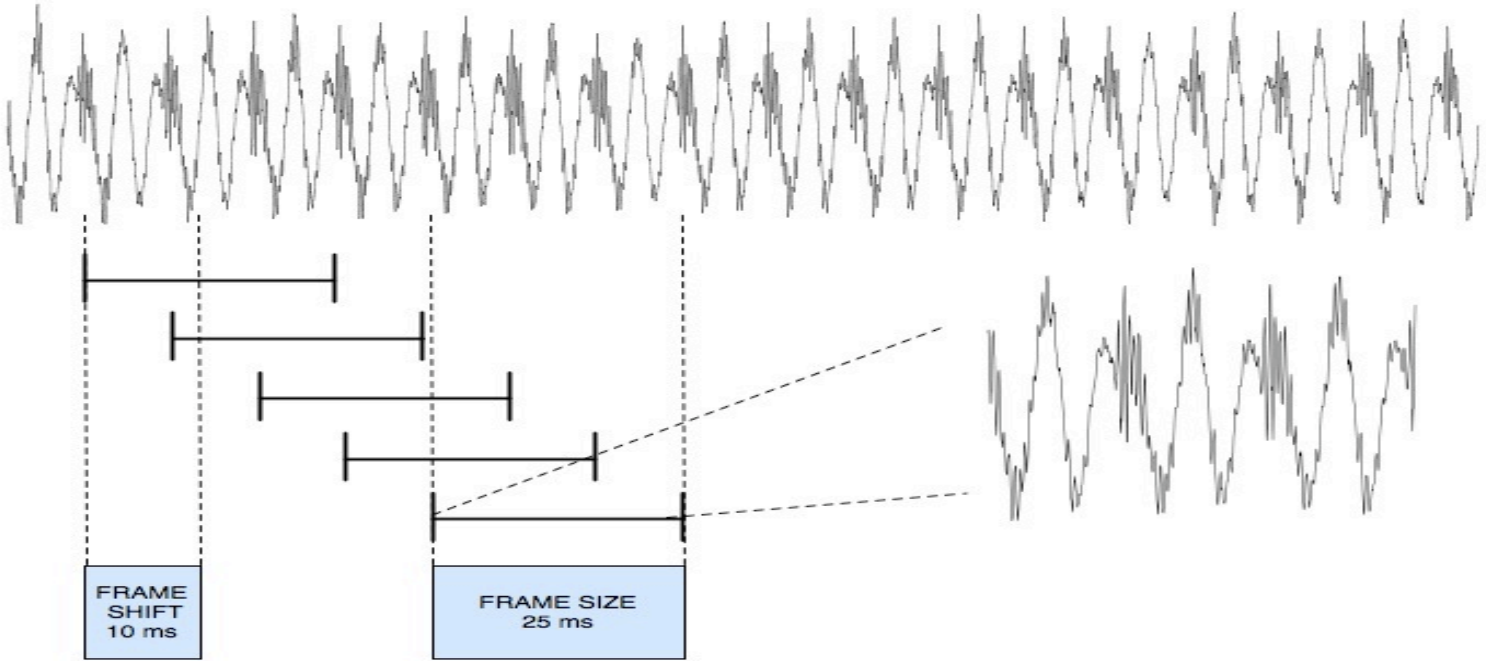


Σχήμα από τις διαφάνειες των Jurafsky & Martin (2008), προερχόμενο από τον B. Pellom.

- **Μέτρηση του αναλογικού** σήματος (**πίεση αέρα**) ανά **τακτά χρονικά διαστήματα** ($10Hz = 10$ φορές ανά sec).
  - Απαιτείται **συχνότητα δειγματοληψίας** τουλάχιστον **διπλάσια** από τη **μέγιστη συχνότητα** (συνιστώσα) του **σήματος**.
  - **Ομιλία**: $< \sim 10\ KHz$, άρα δειγματοληψία $\geq 20\ KHz$.
  - **Τηλεφωνία**: $< 4\ KHz$, άρα δειγματοληψία $\geq 8\ KHz$.
- Οι **μετρήσεις** αποθηκεύονται ως **ακέραιοι**.
  - Συνήθως των 8 bit (-128 ως 127) ή 16 bit (-32.768 ως 32.767).

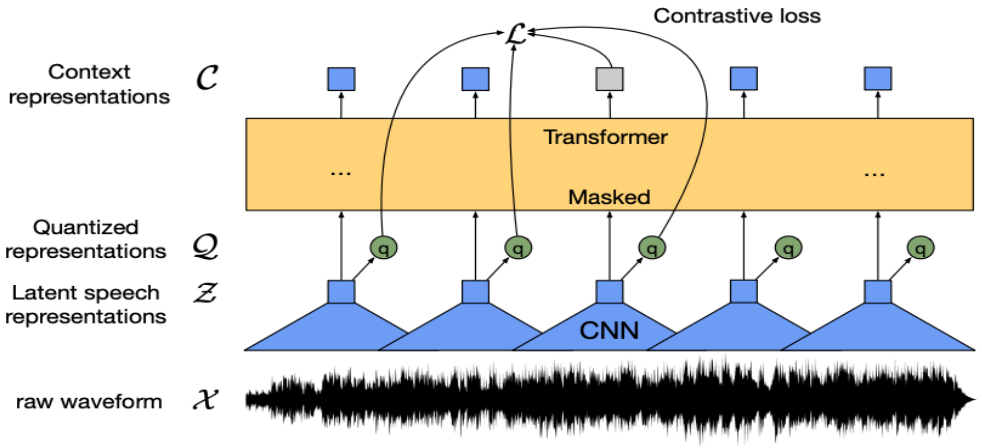
Τμήματα (frames)

Σχήμα από τις διαφάνειες των Jurafsky & Martin (2008).



- Εξάγουμε **επικαλυπτόμενα τμήματα (frames)** του σήματος.
  - **Σέρνουμε ένα «παράθυρο»** κατά μήκος του σήματος.
- **Κάθε τμήμα** συχνά παριστάνεται από ένα **διάνυσμα**.
  - Παραδοσιακά 39 **MFCC features** (βασισμένα σε μετασχηματισμό Fourier).
  - Πιο πρόσφατα **διανύσματα** που παράγονται με **προ-εκπαιδευμένους Transformers** (π.χ. **wav2vec**, **HuBERT**, βλ. παρακάτω).

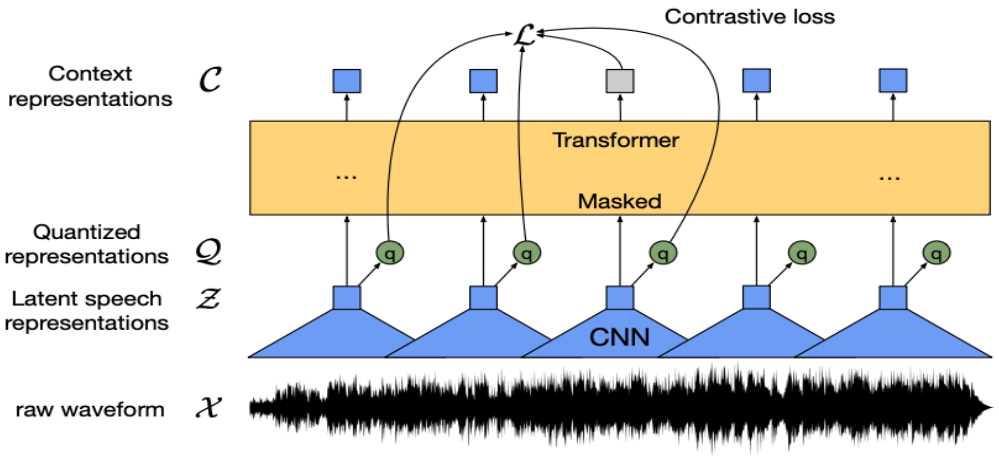
wav2vec



Σχήμα από το άρθρο των Baevsky κ.ά., «wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations», NeurIPS 2020 (https://arxiv.org/abs/2006.11477).

- Ένα **CNN** παράγει ένα **διάνυσμα** (z) **για κάθε τμήμα** (frame).
  - Η **είσοδος στο CNN** είναι οι (διακριτές) **τιμές ήχου ενός τμήματος** (διάνυσμα x). Σκεφτείτε τις σαν μια **μονοδιάστατη μικρή εικόνα** με ένα **κανάλι εισόδου**.
  - Με ***n* συνελικτικά φίλτρα → διάνυσμα *n* χαρακτηριστικών** (z) για κάθε τμήμα x.
- **Για κάθε διάνυσμα τμήματος** (z) που προκύπτει παίρνουμε και το **κοντινότερο διάνυσμα** (q) από ένα **codebook**.
  - Το **codebook** περιέχει **σταθερό αριθμό διανυσμάτων**, τα οποία **μαθαίνουμε**.
  - Ακριβέστερα χρησιμοποιούνται **πολλαπλά codebooks**. **Συνενώνουμε** τα διανύσματα που προκύπτουν (για το τμήμα) από κάθε codebook και τα περνάμε από ένα **dense layer** για να πάρουμε το q.

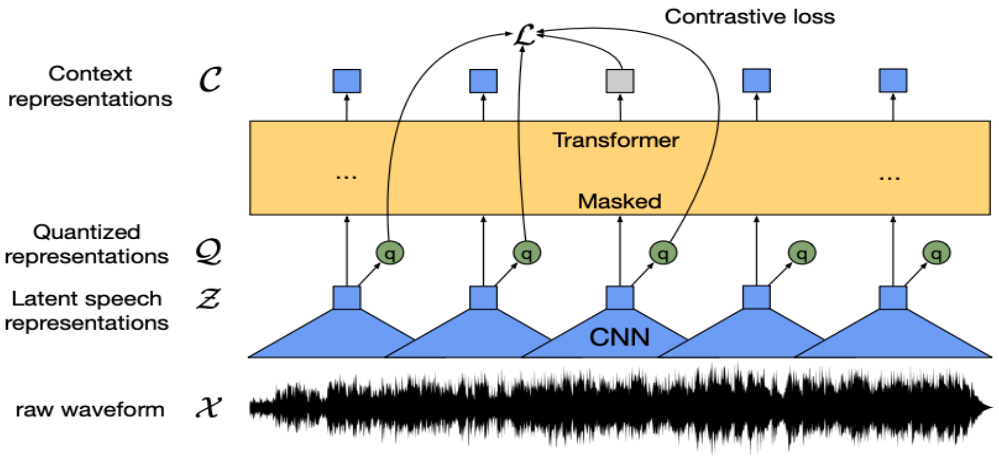
wav2vec – συνέχεια



Σχήμα από το άρθρο των Baevsky κ.ά., «wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations», NeurIPS 2020 (https://arxiv.org/abs/2006.11477).

- Τα **διανύσματα των τμημάτων** (z) που παράγει το CNN περνούν από στοιβαγμένους **κωδικοποιητές Transformer** (όπως στο BERT).
  - Έτσι παράγονται **νέα διανύσματα τμημάτων** (c) που είναι «γνωρίζουν» και τα υπόλοιπα τμήματα (**context-aware**).
- Κατά την **προ-εκπαίδευση**, **κρύβουμε τυχαία τμήματα** (διανύσματα z) και απαιτούμε **να «μαντέψει»** το wav2vec **τα διανύσματα *q*** τους **από το** αντίστοιχο **context-aware διάνυσμα *c***.
  - **Αντικαθιστούμε** στην είσοδο των Transformers τα **διανύσματα** (z) **για τα κρυμμένα τμήματα** με ένα **κοινό διάνυσμα** (σαν του [MASK] στο BERT).

wav2vec – συνέχεια



Σχήμα από το άρθρο των Baevsky κ.ά., «wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations», NeurIPS 2020 (https://arxiv.org/abs/2006.11477).

- Κατά την **προ-εκπαίδευση**, **κρύβουμε τυχαία τμήματα** (διανύσματα z) και απαιτούμε **να «μαντέψει»** το wav2vec **τα διανύσματα q** τους **από το** αντίστοιχο **context-aware διάνυσμα c**.
  - Ζητάμε από το wav2vec να επιλέξει το **σωστό q** διάνυσμα **μεταξύ των $\tilde{q}$** διανυσμάτων **όλων των κρυμμένων τμημάτων** (σφάλμα L_m).

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

  - Ένα **πρόσθετο σφάλμα** (loss, βασισμένο στην εντροπία) φροντίζει **να χρησιμοποιούνται όλα τα διανύσματα** του **κάθε codebook**.

HuBERT

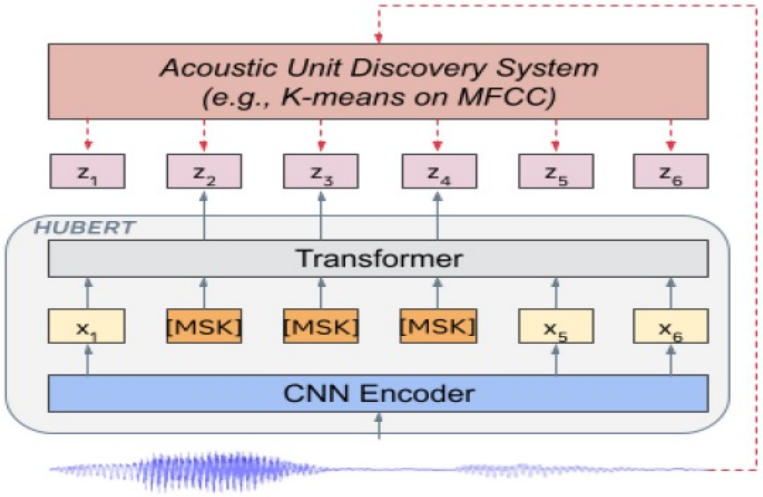


Fig. 1: The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.

Σχήμα και τύπος από το άρθρο των Hsu κ.ά., «HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units», 2021 (https://arxiv.org/abs/2106.07447).

- **Παρόμοιο με το wav2vec** αλλά τώρα **για κάθε κρυμμένο** (ή μη) **τμήμα** (frame) απαιτούμε κατά την **προ-εκπαίδευση** το μοντέλο να **μαντεύει τη συστάδα** (cluster) στην οποία **ανήκει το τμήμα**.
  - Οι αρχικές συστάδες παράγονται εφαρμόζοντας τον **k-means** στα **διανύσματα MFCC** των τμημάτων **όλων των δεδομένων** (ήχος μόνο) προ-εκπαίδευσης.
  - **Σε επόμενους κύκλους** παράγουμε νέες συστάδες-στόχους εφαρμόζοντας τον **k-means** στα **διανύσματα** των τμημάτων προ-εκπαίδευσης που παράγει το μοντέλο του **προηγούμενου κύκλου**.

Κωδικοποιητές/αποκωδικοποιητές

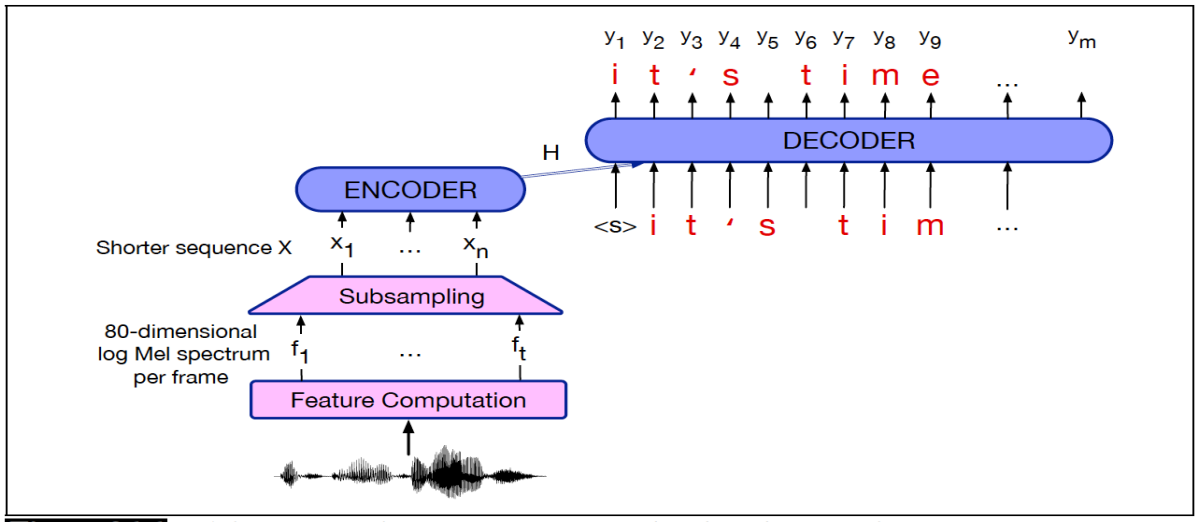


Figure 26.6 Schematic architecture for an encoder-decoder speech recognizer.

Σχήμα από το βιβλίο «Speech and Language Processing» των D. Jurafsky & J.H. Martin, 3η έκδοση (υπό προετοιμασία). http://web.stanford.edu/~jurafsky/slp3/

- Ο **κωδικοποιητής** και ο **αποκωδικοποιητής** μπορούν να είναι **RNNs**, όπως είδαμε στη **μηχανική μετάφραση**. (Συχνά είναι πια **Transformers**.)
  - Εκπαιδεύονται **μαζί**, σε **ζεύγη εισόδων-εξόδων**, όπως στην μηχανική μετάφραση.
- Ο **κωδικοποιητής διαβάζει** μια **ακολουθία διανυσμάτων**, δηλαδή ένα διάνυσμα **για κάθε τμήμα ήχου (frame)** (ή λιγότερα, αν κάνουμε υπο-δειγματοληψία τους).
  - Συχνά πια χρησιμοποιούνται τα **διανύσματα** που προκύπτουν από μοντέλα όπως τα **wav2vec** ή **HuBERT**. Παλιότερα χρησιμοποιούνταν διανύσματα **MFCC**.
  - Συνήθως υπάρχει και ένας **μηχανισμός προσοχής**, όπως στη μηχανική μετάφραση.

Κωδικοποιητές/αποκωδικοποιητές

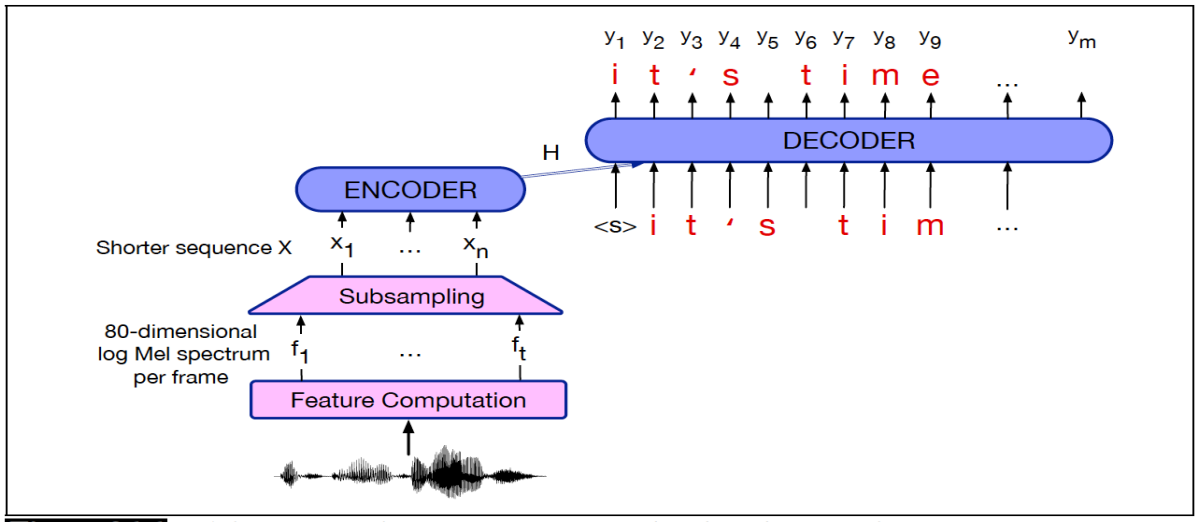


Figure 26.6 Schematic architecture for an encoder-decoder speech recognizer.

Σχήμα από το βιβλίο «Speech and Language Processing» των D. Jurafsky & J.H. Martin, 3η έκδοση (υπό προετοιμασία). http://web.stanford.edu/~jurafsky/slp3/

- Ο **αποκωδικοποιητής παράγει** γράμμα-γράμμα το **κείμενο**.
  - **Κατά την εκπαίδευση**, όταν υπολογίζουμε τη νέα κατάσταση του αποκωδικοποιητή, χρησιμοποιούμε ως **προηγούμενο γράμμα το σωστό** προηγούμενο (**teacher forcing**), ακόμα και αν ο αποκωδικοποιητής είχε επιλέξει άλλο (λάθος) προηγούμενο γράμμα.
  - **Σταδιακά** μπορούμε να χρησιμοποιούμε όλο και **συχνότερα** το **προηγούμενο γράμμα** που είχε επιλέξει ο ίδιος ο **αποκωδικοποιητής (scheduled sampling).**
  - Μετά την εκπαίδευση, **σε κάθε κατάσταση του αποκωδικοποιητή** επιλέγουμε **λαίμαργα το γράμμα στο οποίο δίνει μεγαλύτερη πιθανότητα** το μοντέλο. **Εναλλακτικά** ψάχνουμε τις πιθανότερες ακολουθίες γραμμάτων **με beam search**.

Κωδικοποιητές/αποκωδικοποιητές

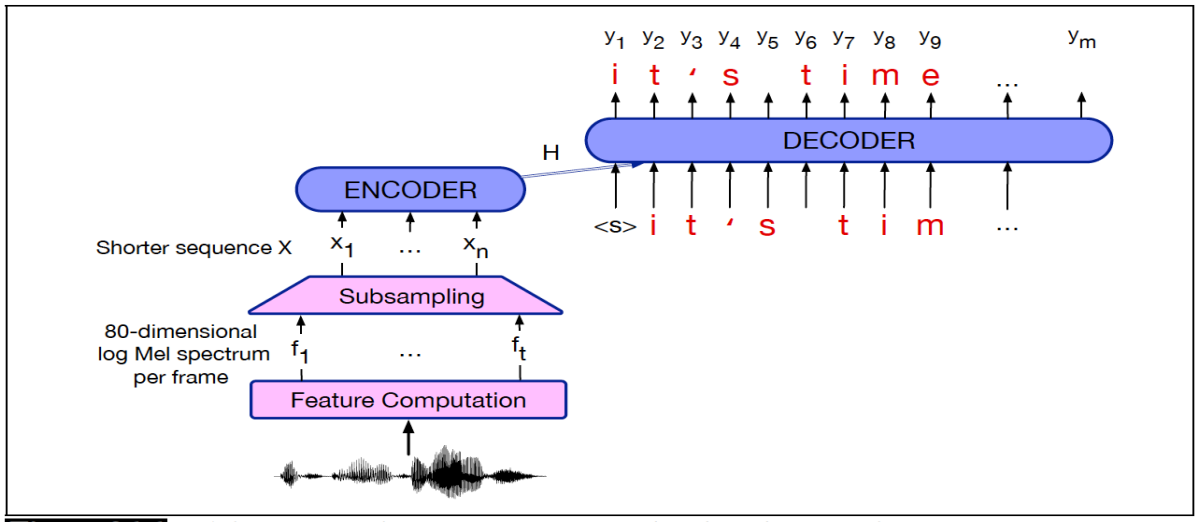


Figure 26.6 Schematic architecture for an encoder-decoder speech recognizer.

Σχήμα από το βιβλίο «Speech and Language Processing» των D. Jurafsky & J.H. Martin, 3η έκδοση (υπό προετοιμασία). http://web.stanford.edu/~jurafsky/slp3/

- Μπορούμε να παράγουμε **τις *n* πιθανότερες ακολουθίες γραμμάτων** (***n*-best list**) και μετά να τις **φιλτράρουμε** λαμβάνοντας υπόψιν τις **πιθανότητες** που τους δίνει ένα **γλωσσικό μοντέλο** εκπαιδευμένο σε πολύ μεγάλα σώματα κειμένων.
  - Ή λαμβάνουμε υπόψιν τις **πιθανότητες** και του **γλωσσικού μοντέλου κατά το beam search**.
  - Π.χ. **προσθέτουμε** την λογαριθμική **πιθανότητα** $p = \log(y_1) + \log(y_2) + \cdots$ που δίνει ο **αποκωδικοποιητής** σε ένα **μονοπάτι** $y_1, y_2, \ldots$ (που εξερευνούμε κατά το beam search) και την λογαριθμική **πιθανότητα** q που δίνει στο μονοπάτι $y_1, y_2, \ldots$ το **γλωσσικό μοντέλο** $(\lambda_1 p + \lambda_2 q)$.
  - Τα **γλωσσικά μοντέλα** «**προτιμούν**» **σύντομες ακολουθίες** (γιατί;), οπότε συνήθως χρησιμοποιούμε και έναν **παράγοντα που επιβραβεύει μακρύτερες** ακολουθίες γραμμάτων.

Μέτρα αξιολόγησης

- **Λόγος λαθών λέξεων** (Word Error Rate):
  - $WERR = \frac{\text{Insertions+Replacements+Deletions}}{\text{\#ReferenceWords}}$

Παράδειγμα από τις διαφάνειες των Jurafsky & Martin (2008).

Σωστή μεταγραφή (**reference**).

REF:	portable	****	PHONE	UPSTAIRS	last	night	so
HYP:	portable	FORM	OF	STORES	last	night	so
		I	R	R			

Έξοδος συστήματος (**υπόθεση**).

WER = (1+2+0)/6 = 50%

  - Υπολογίζεται **όπως η απόσταση Levenshtein**, αλλά με κόστος 1 και για R. Το WERR μπορεί να βγει και > 1.
- **Λόγος λαθών προτάσεων** (Sentence Error Rate):
  - Προτάσεις με ≥ 1 λάθος / πλήθος προτάσεων.

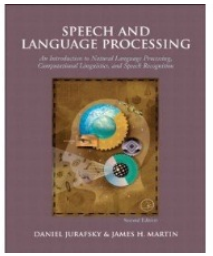
Ενδεικτικές επιδόσεις

English Tasks	**WER%**
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAL)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3
Chinese (Mandarin) Tasks	**CER%**
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

Figure 26.1 Rough Word Error Rates (WER = % of words misrecognized) reported around 2020 for ASR on various American English recognition tasks, and character error rates (CER) for two Chinese recognition tasks.

Πίνακας από το βιβλίο «Speech and Language Processing» των D. Jurafsky & J.H. Martin, 3η έκδοση (υπό προετοιμασία). http://web.stanford.edu/~jurafsky/slp3/

Διάβασμα



- Το μεγαλύτερο μέρος της ύλης αυτής της ενότητας καλύπτεται από το κεφάλαιο 26 του βιβλίου «Speech and Language Processing» των Jurafsky & Martin, 3η έκδοση (υπό προετοιμασία).
  - http://web.stanford.edu/~jurafsky/slp3/
  - Ενότητες 26.1–26.5. Για τις εξετάσεις, μόνο ό,τι περιλαμβάνεται στις διαφάνειες.
  - Όσοι ενδιαφέρεστε ιδιαίτερα, διαβάστε και το υπόλοιπο κεφάλαιο, που καλύπτει τη σύνθεση ομιλίας.
  - Τα MFCC features περιγράφονται εκτενέστερα στη 2η έκδοση (υπάρχει στη βιβλιοθήκη του ΟΠΑ).