Minimum Edit distance

Chapter 3 J&M'09

Building a spellchecker

Goal: given input w (e.g. 'thew'), guess the best alternate spelling candidate c ('the', 'thaw', ...)':

- Enumerate all words in a dictionary that within some range of spelling difference
- Order them in terms of their probability
  For example, by maximizing $P(c)P(w|c)$

  $P(c) =$ the probability of the candidate word

  $P(w|c) =$ the probability that w would be typed in a text when c was meant.

Ideally, condition c on all previously typed words, and introduce a model that has information about how likely specific errors are.
See for example the Birkbeck spelling error corpus.

MED

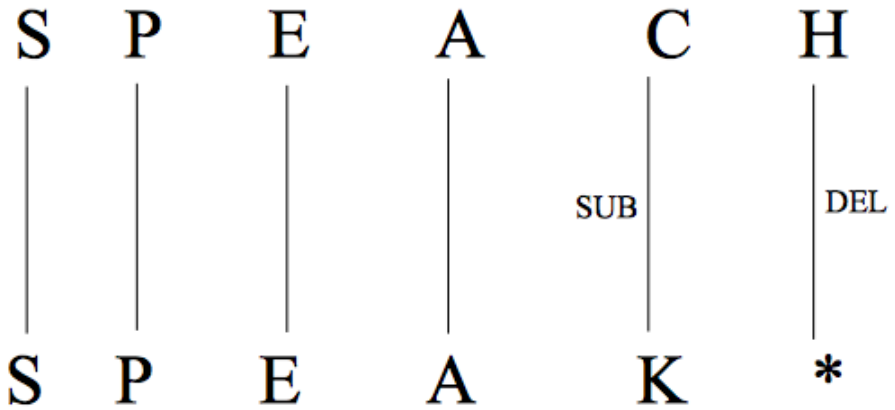
What kind of operations are needed to turn one string into another:

- Insertion
- Deletion
- Substitution

Minimum edit distance is the smallest number of operations needed for this transformation.

MED

In the following example, two operations suffice:



If each operation has cost of 1, then distance is 2.
If insertions and deletions cost 1, but substitutions cost 2 (Levenshtein edit distance), then the distance is 3.

MED

For source word s of length i, and target word t of length j:

- Base case:
  $D(0,0) = 0$
- Deletion cost 1
  $D(i,0) = D(i-1,0) + 1$
- Insertion cost 1
  $D(0,j) = D(0,j+1) + 1$
- Minimum Edit Distance:
  For every column 1 to i, for every row 1 to j:

$$D(i,j) = min \left(\begin{cases} D(i-1,j) + 1 \ (delete) \\ D(i,j-1) + 1 \ (insert) \\ D(i-1,j-1) + \begin{cases} 1 \text{ if } s_i \neq t_j \ (replace) \\ 0 \text{ o.w.} \end{cases} \end{cases} \right)$$

MED

(with substitution cost of 1)

e	7	6	5	4	3	2	**1**
f	6	5	4	3	2	1	2
f	5	4	3	2	1	2	3
a	4	3	2	1	2	3	4
r	3	2	1	2	3	4	5
i	2	1	1	2	3	4	5
g	1	0	1	2	3	4	5
-	0	1	2	3	4	5	6
	-	**g**	**r**	**a**	**f**	**f**	**e**

MED

(with substitution cost of 2)

e	7	6	5	4	3	2	**1**
f	6	5	4	3	2	1	2
f	5	4	3	2	1	2	3
a	4	3	2	1	2	3	4
r	3	2	1	2	3	4	5
i	2	1	2	3	4	5	6
g	1	0	1	2	3	4	5
-	0	1	2	3	4	5	6
	-	**g**	**r**	**a**	**f**	**f**	**e**

MED

(with substitution cost of 2)

y	9	8	7	6	5	4	5	4	5	**4**
m	8	7	6	5	4	3	4	3	4	5
o	7	6	5	4	3	2	3	2	3	4
n	6	5	4	3	2	1	2	3	4	5
o	5	4	3	2	1	0	1	2	3	4
r	4	3	2	1	0	1	2	3	4	5
t	3	2	1	0	1	2	3	4	5	6
s	2	1	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6	7	8
-	0	1	2	3	4	5	6	7	8	9
	-	**a**	**s**	**t**	**r**	**o**	**l**	**o**	**g**	**y**

MED

Useful for machine translation evaluation:

Human translation:
'The spokesman confirms that the senior government advisor was fired'

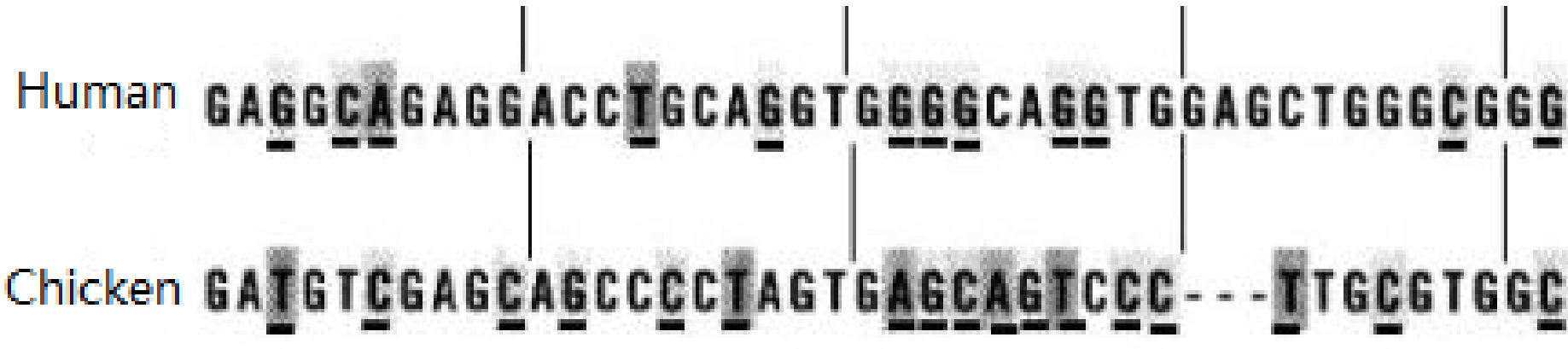
Machine translation:
'Spokesman said senior government advisor was dismissed'

MED estimates how good the machine translation was.

MED

Sequence alignment in computational biology:

- Comparing genes or gene regions across species
- Find important regions
- Determine function
- Look for mutations



MED

Edit distance defines a relation between strings, just like FSTs..
A weighed FST can provide the cost for each transition:

MED

Suppose that we are building a speech recognizer, and wish to use MED to determine candidate alternatives (or to measure how accurate the speech recognizer is)

We need:

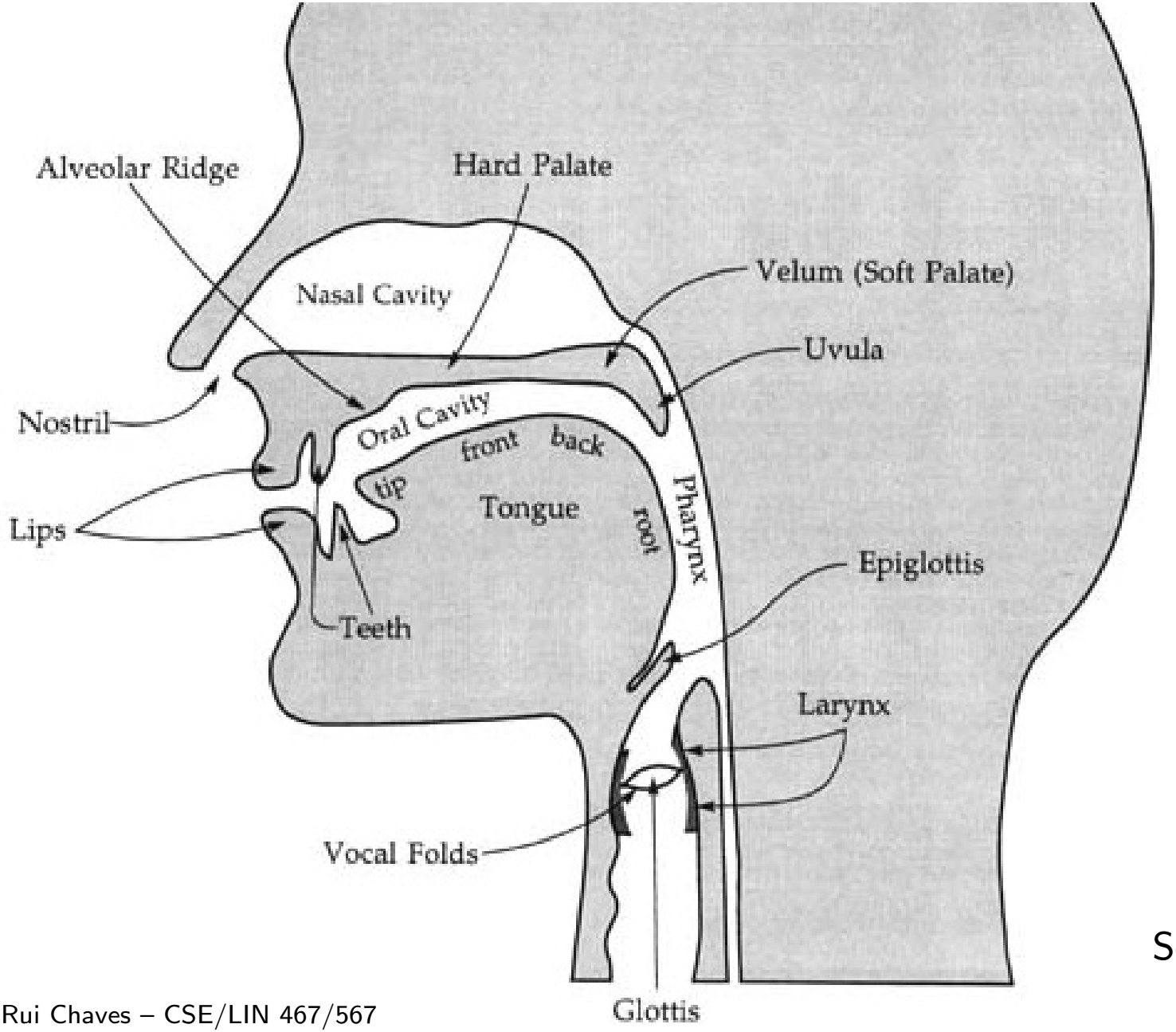
- A speech corpus (i.e. a corpus where words are transcribed into the sounds that are produced)
- A speech recognizer (complex! to be discussed later; see ch.9 of J&M'08)
- To adapt the MED for speech sounds

MED

Examples of speech corpora:

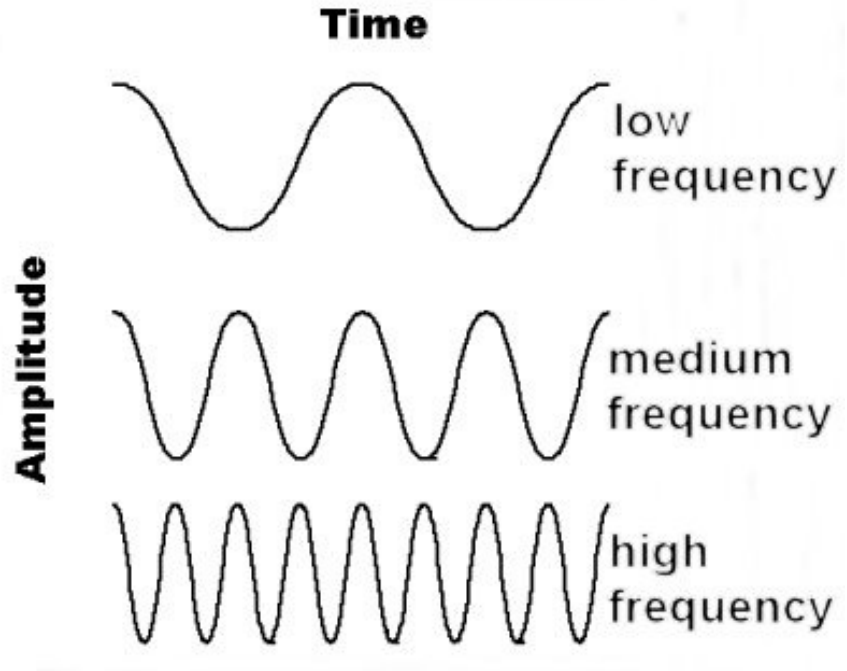
1. TIMIT Acoustic-Phonetic Continuous Speech Corpus
   (hand-verified speech from 630 people from 8 American dialects, time-aligned with transcripts (orthographic and phonetic))
2. CELEX2 (SUBTLEX)
   (orthography variations, phonological transcriptions, including syllables and stress)
3. TIDIGITS
   (326 speakers, each pronouncing 77 different sequences of digits in 1982)
4. Switchboard Transcription Project
   (several hundred informal speech dialogs recorded over the telephone, fairly representative of spontaneous discourse)

Airstream mechanics of speech



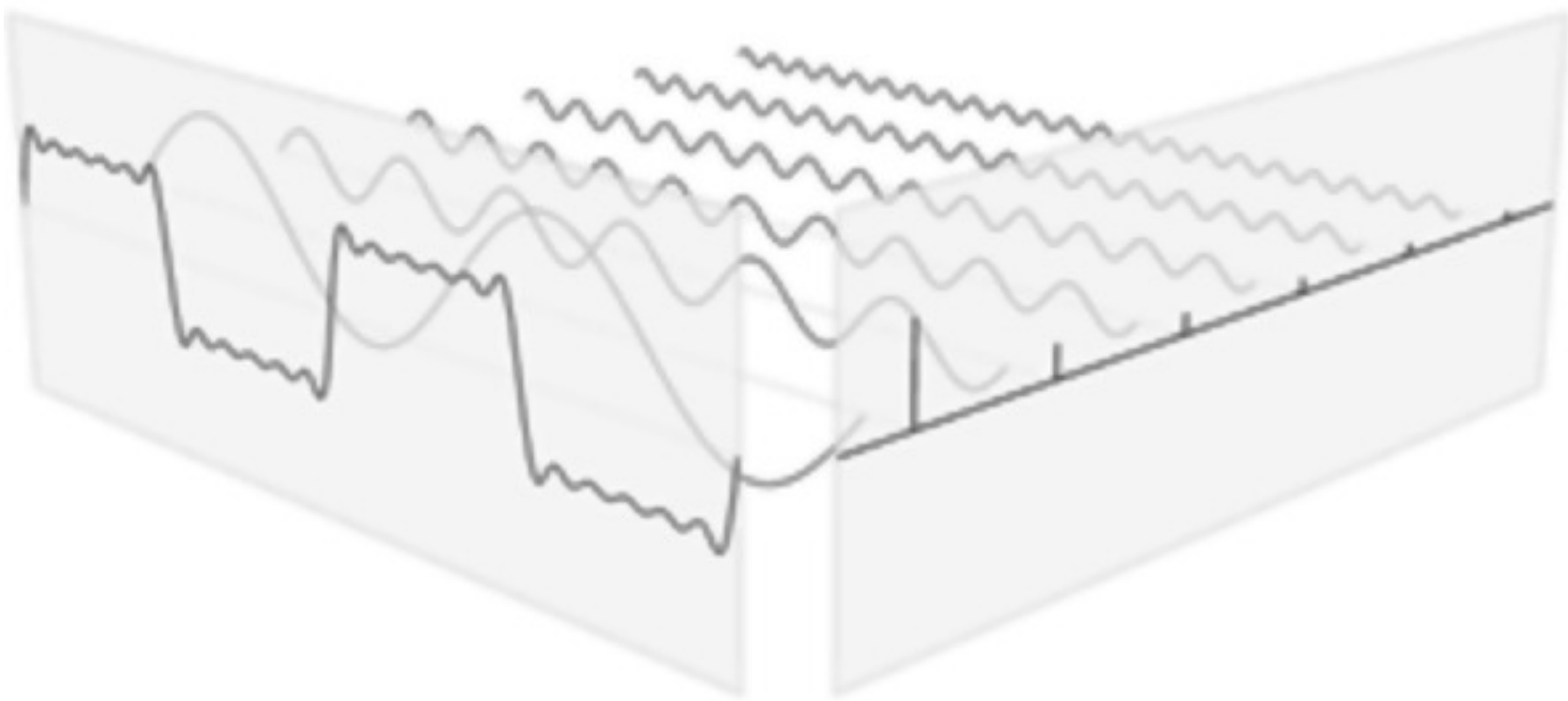
See an MRI scan.

Airstream mechanics



- **Sine wave**: the simplest kind of wave.
  - **amplitude (intensity)**: displacement of the vibrating medium from its rest position; Perceived as loudness.
  - **frequency**: number of complete vibration cycles per second. Perceived as pitch. 1 hertz = one cycle per second.
  - **duration**: length of a sound. Perceived as tempo.

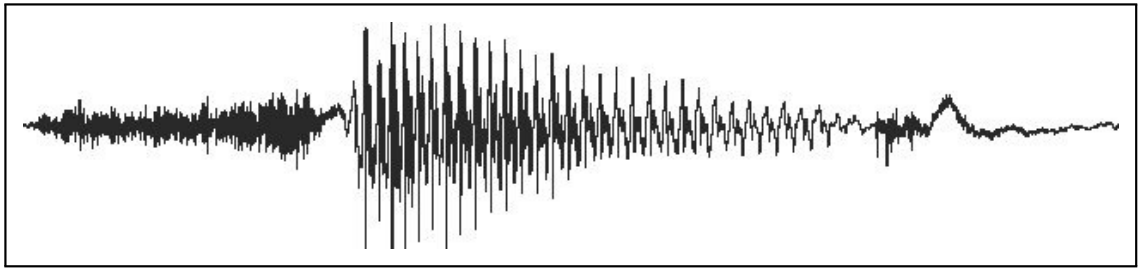
Spectrogram



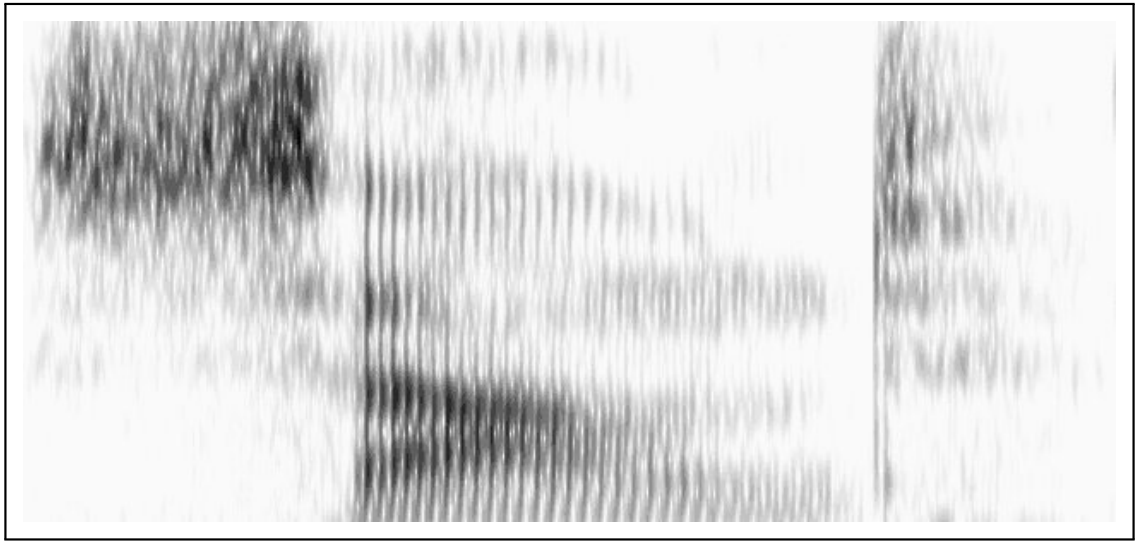
A **spectrogram** is a tridimensional representation of the sine waves that compose a complex wave:

Spectrogram

The word '*sound*'



soundwave

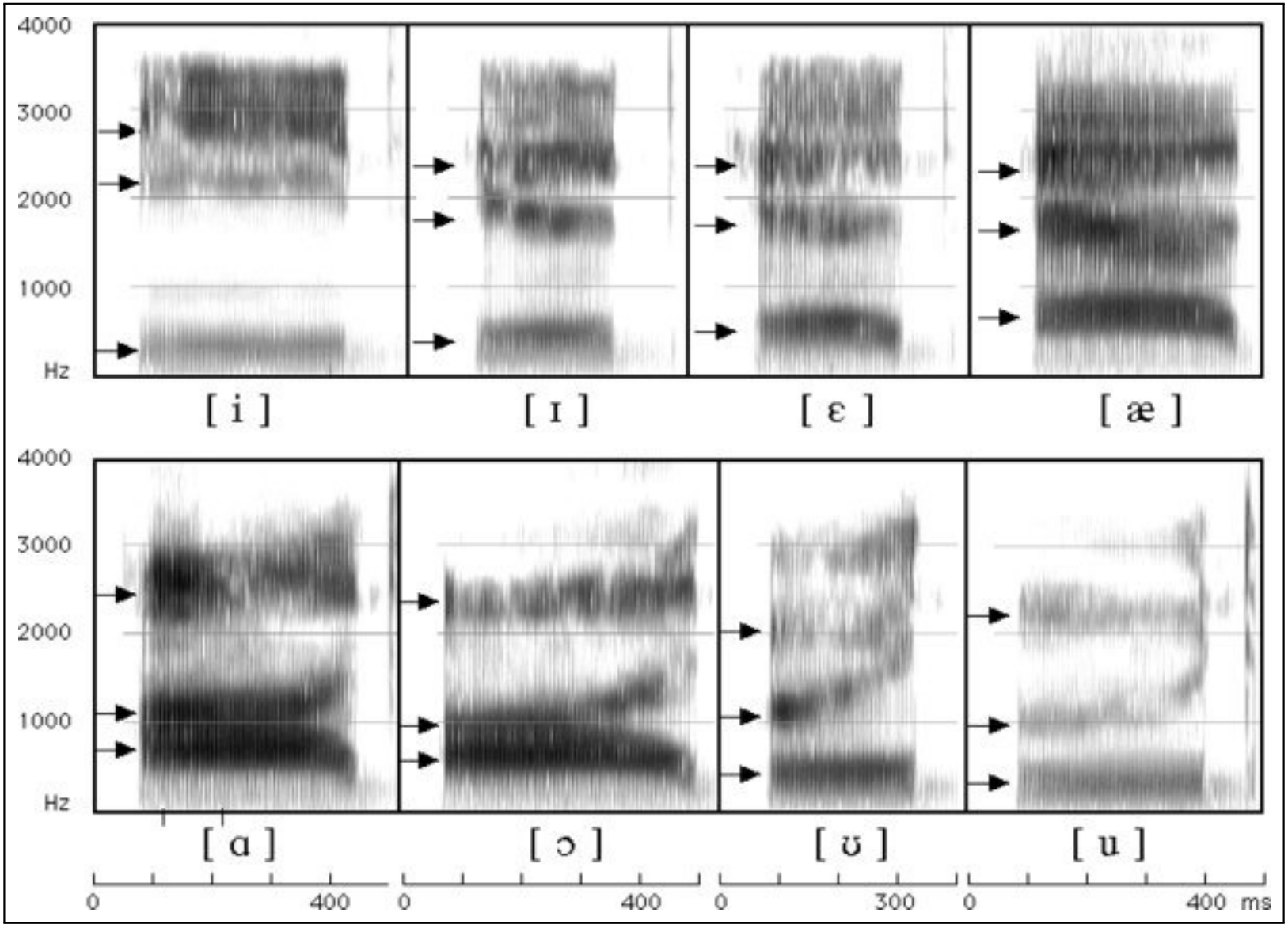


spectrogram

s ou n d

- frequency = **vertical** axis
- time = **horizontal** axis
- amplitude = **intensity** (darkness)

Spectrogram



Full interactive IPA chart.

MED

A phonetic MED:

- substitution distance between two segments is the number of features that are different.
  For example, [k] and [g] have a distance of 1.
- more generally, the segment distance is the sum of the non-shared features.
  For example, if a consonant with seven features shares only two features with a five-feature vowel, the minimum distance will be: $(7-2)+(5-2)=5+3=8$.
- Define Insertion and Deletion costs as half of the average substitution cost (i.e. the average cost of substituting every segment for every other segment).

Thus, substitutions costa about double of what insertions and deletions cost, as in Levenshtein.