CS 505: Introduction to Natural Language Processing

Wayne Snyder
Boston University

Lecture 23 – Automatic Speech Recognition (ASR)

Radio Rex from 1920s - The first speech recognition machine

Log MEL Spectrogram

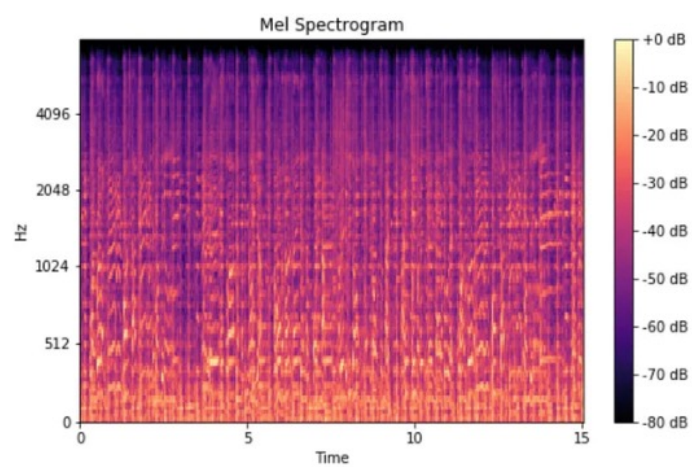
Therefore, to capture the human experience of sound, we typically use a Mel Spectrogram, where

- Pitch is given in Mels
- Loudness is given in Decibels:

Both of these are log scales

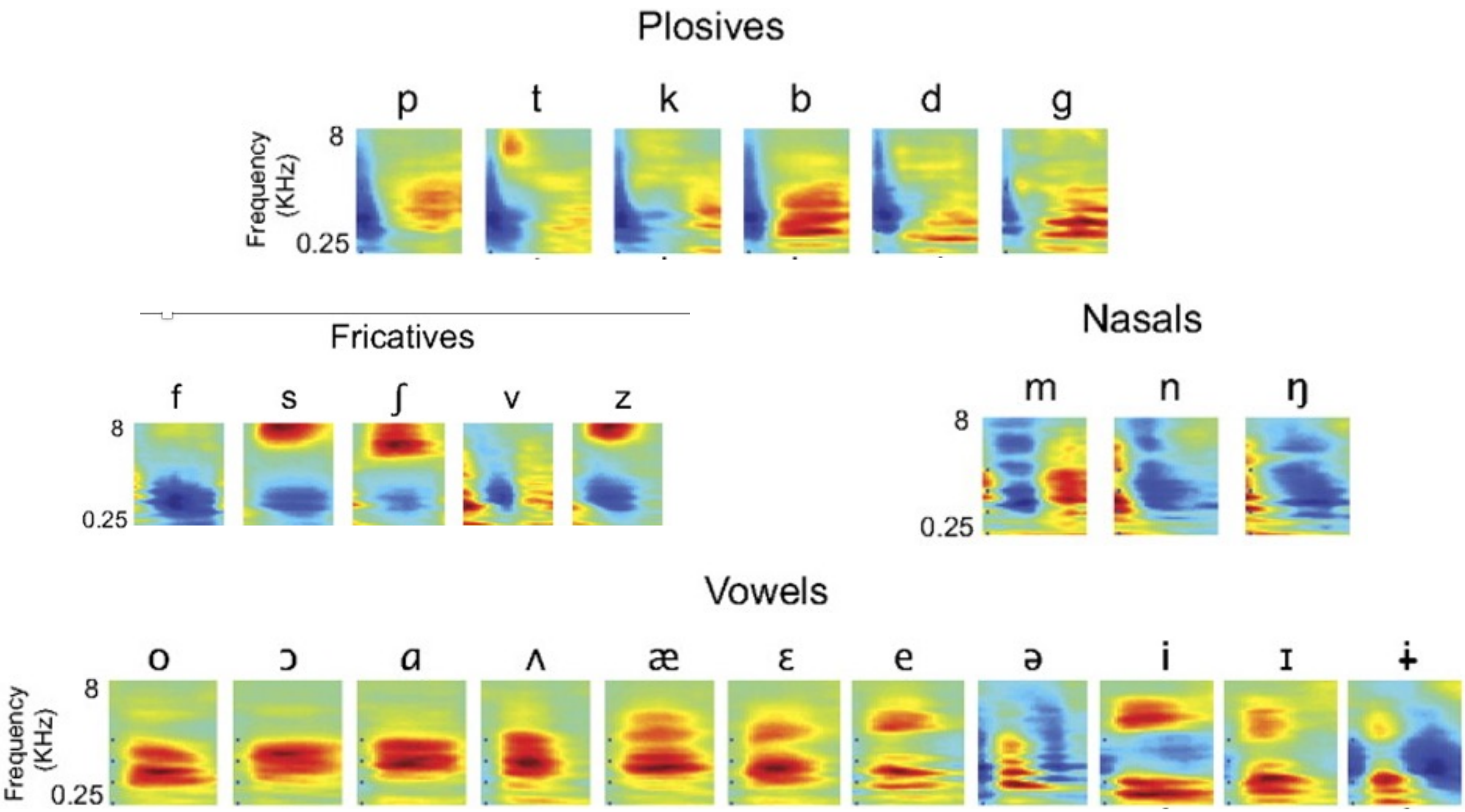
```
mel_spect = librosa.feature.melspectrogram(y=y, sr=sr, n_fft=2048,
hop_length=1024)
mel_spect = librosa.power_to_db(spect, ref=np.max)

librosa.display.specshow(mel_spect, y_axis='mel', fmax=8000,
x_axis='time');
plt.title('Mel Spectrogram');
plt.colorbar(format='%+2.0f dB');
```



Human Vocal Signals

Each phoneme in human language has a rather distinct spectrogram:



Phonemes and the International Phonetic Alphabet

Phonemes are smallest unit of sound in a particular language which convey meaning.

Each language has a distinct set of phonemes (English has 44) which describe the pronunciation of all words; the International Phonetic Alphabet (IPA) is a standard collection of phonemes for all the world's languages:

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

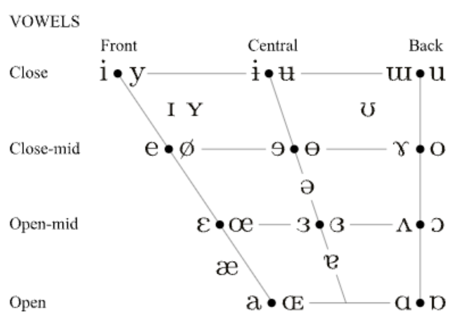
Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative
ɕ ʑ Alveolo-palatal fricatives
w Voiced labial-velar approximant
ɺ Voiced alveolar lateral flap



Where symbols appear in pairs, the one to the right represents a rounded vowel.

/p/	pit	/b/	bit
/t/	tin	/d/	din
/k/	cut	/g/	gut
/tʃ/	cheap	/dʒ/	jeep
/f/	fat	/v/	vat
/θ/	thigh	/ð/	thy
/s/	sap	/z/	zap
/ʃ/	Aleutian	/ʒ/	allusion
/x/	loch		
/h/	ham		

ARPAbet is an ASCII version of the IPA symbol set.

International Phonetic Alphabet

Translations into IPA from: https://tophonetics.com/

Here is a translation of English text into IPA. hir ɪz ə træn'zleɪʃən ʌv 'ɪŋglɪʃ tɛkst 'ɪntu aɪ-pi-eɪ.

Natural Language Processing 'nætʃərəl 'læŋgwədʒ 'prɑsɛsɪŋ

From dictionary.com:

language / ˈlæŋ gwɪdʒ / PHONETIC RESPELLING

See synonyms for *language* on Thesaurus.com

noun

1. a body of words and the systems for their use common to a people who are of the same community or nation, the same geographical area, or the same cultural tradition:
   the two languages of Belgium; a Bantu language; the French language; the Yiddish language.
2. communication by voice in the distinctively human manner, using arbitrary sounds in conventional ways with conventional meanings; speech.

SEE MORE

Spectra and Spectrograms for Vowels

Vowels are continuous sounds, formed by the shaping of the vocal cavity; therefore, each has a (instantaneous) spectrum.

These spectra have characteristic peaks, called formants, caused by the shape of the vocal cavity.

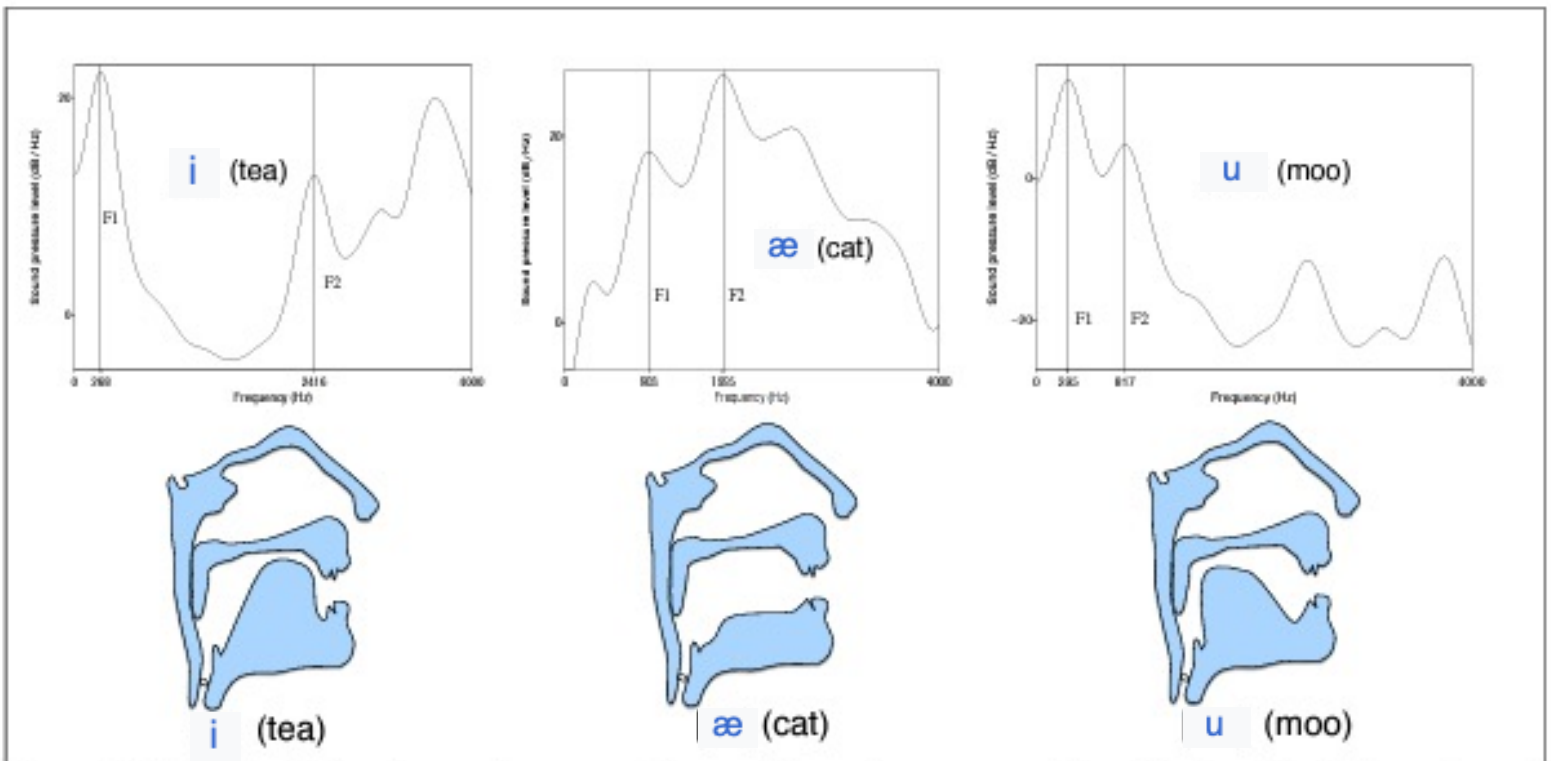


Figure 28.22 Visualizing the vocal tract position as a filter: the tongue positions for three English vowels and the resulting smoothed spectra showing F1 and F2.

Spectra and Spectrograms for Vowels

Vowels are continuous sounds, formed by the shaping of the vocal cavity; therefore, each has a (instantaneous) spectrum.

These spectra have characteristic peaks, called formants, caused by the shape of the vocal cavity.

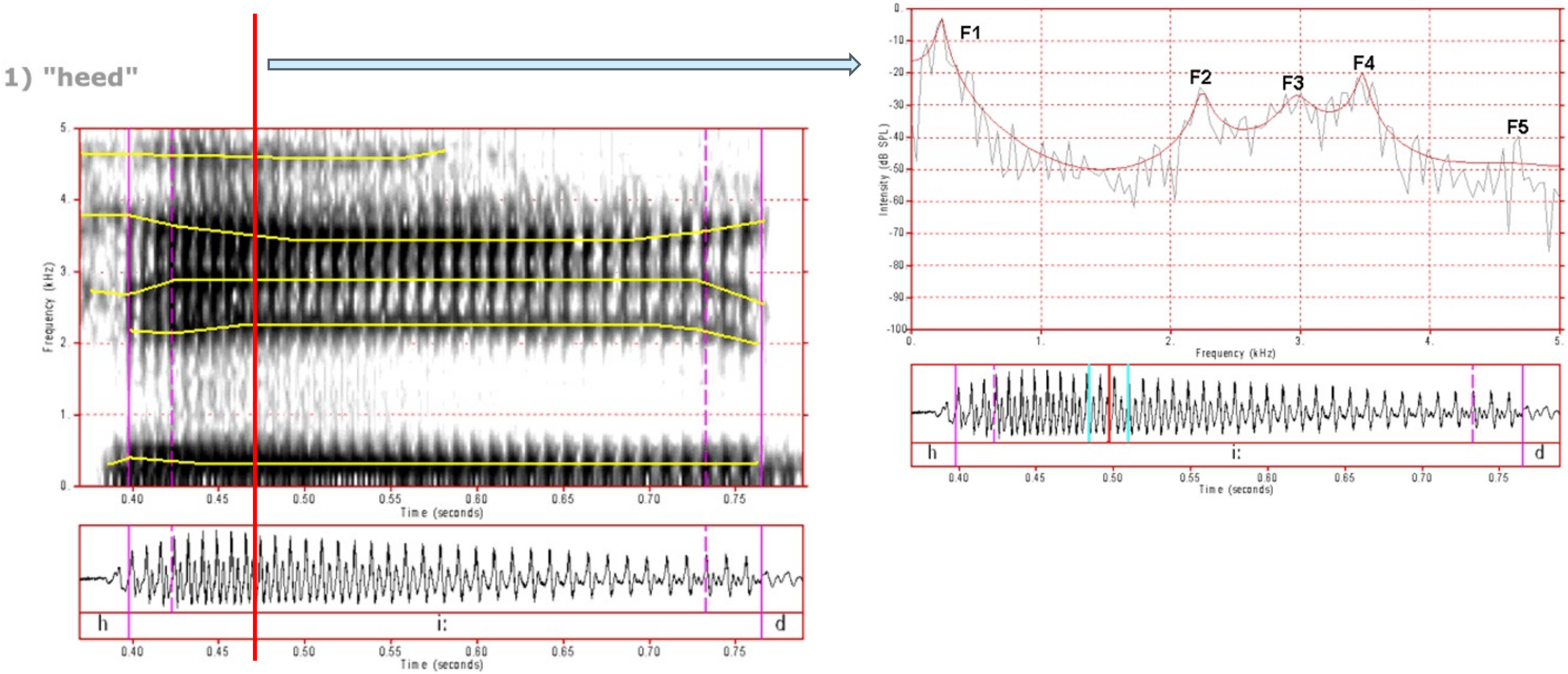
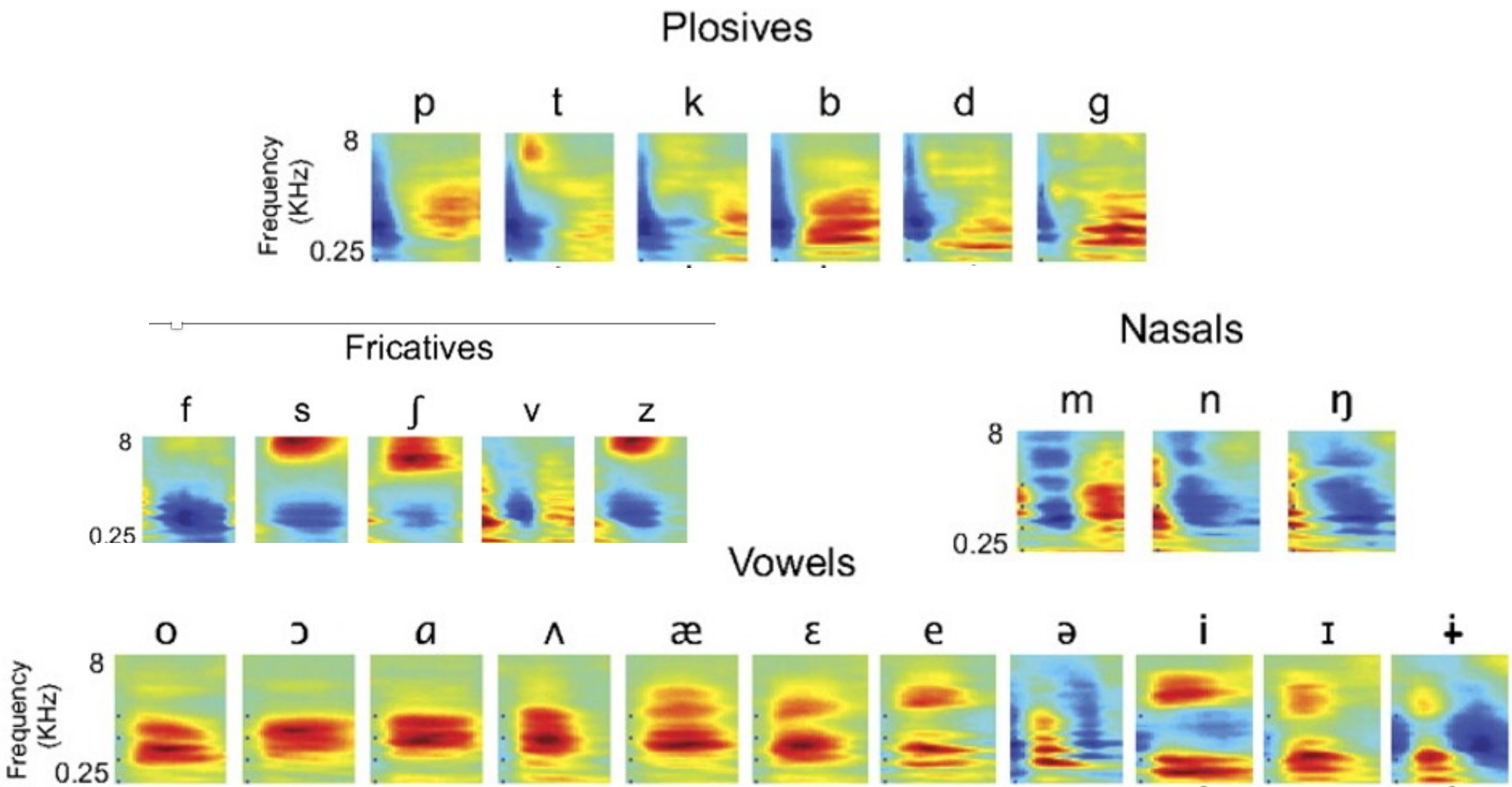


Figure 1: Broadband spectrogram of the vowel /i:/ from the token "heed".

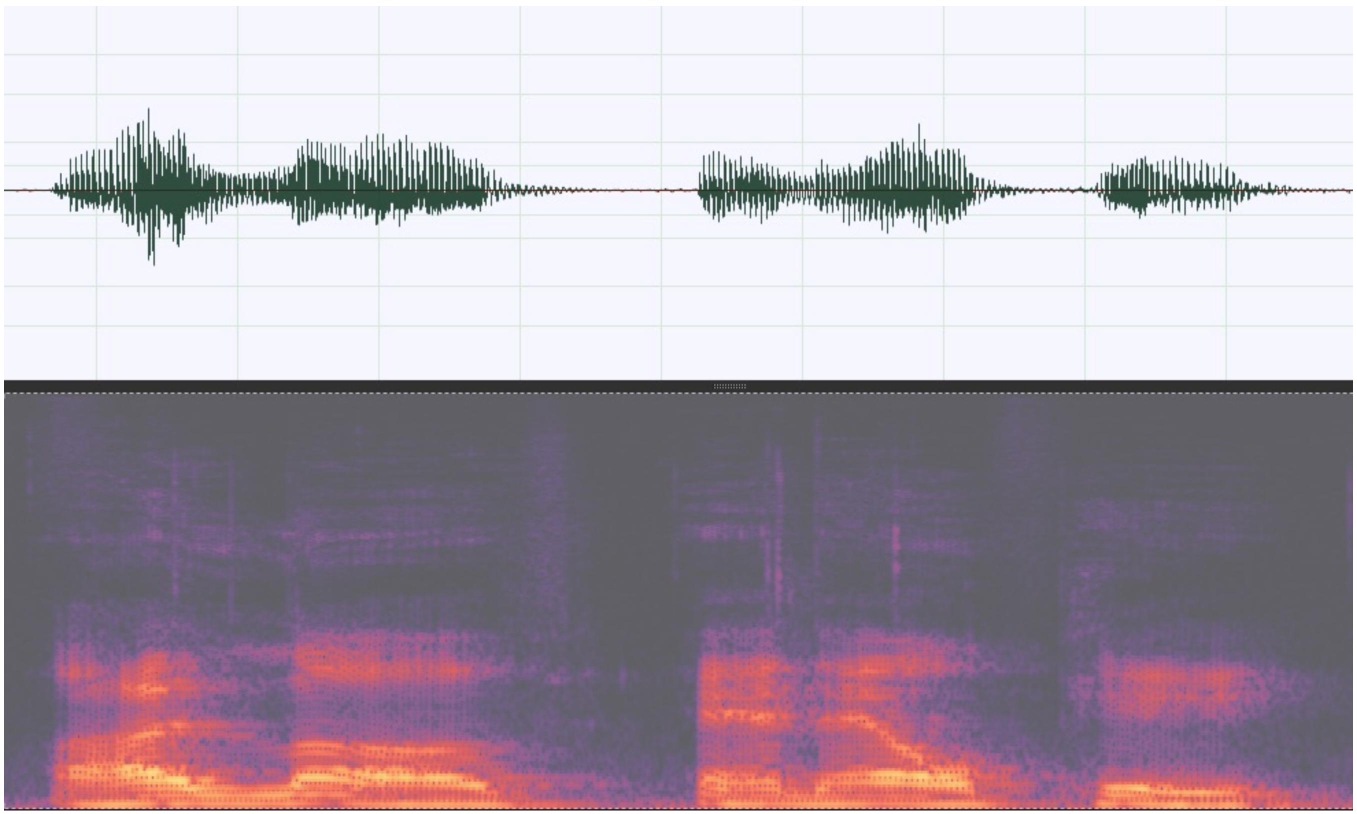
Spectrograms for Consonants and Vowels

Vowels can be recognized by their (instantaneous) spectra, but consonants and semi-vowels (such as w or y) have time-dependent characteristics, and are best represented by spectrograms:

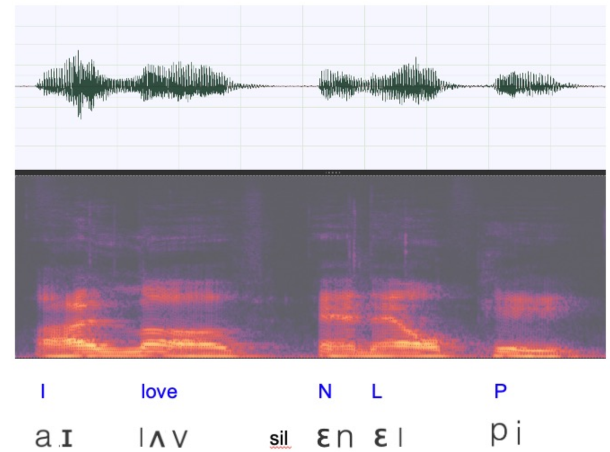


Continuous Speech

Continuous speech can be understood as a sequence of phonemes, possibly separated by periods of silence:



Continuous Speech



But continuous speech is complex!

In general, we must

- Identify individual phonemes
- Identify words
- Identify sentence structure and/or meaning
- Interpret prosodic features
- Deal with mistakes, different speakers, accents, self-corrections, etc.

Continuous Speech

Prosodic features are very important in deriving meaning from sequences of phonemes. Many of these have to do with lexical stress – what words or syllables are emphasized.

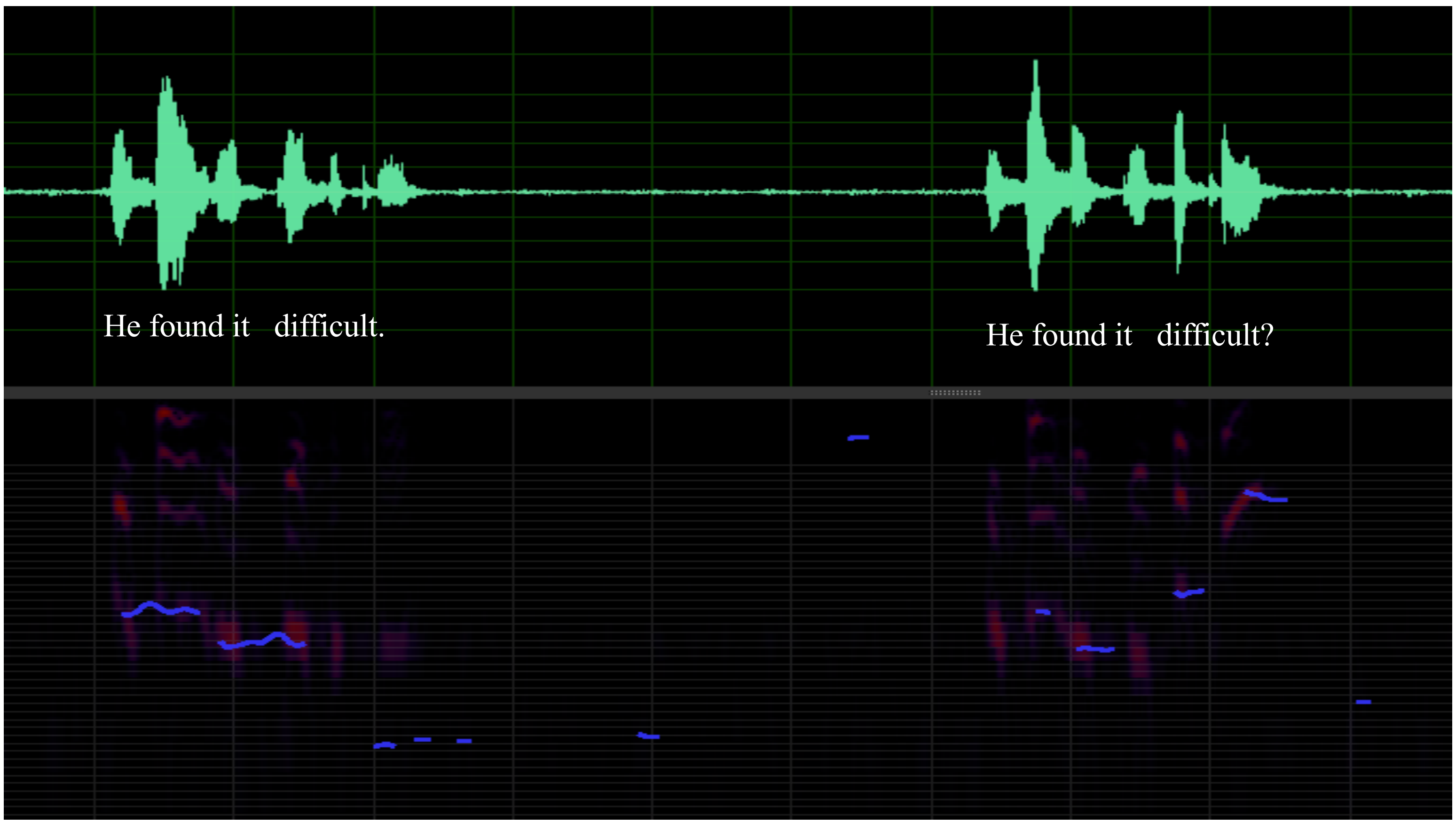
- Loudness
- Duration
- Pitch:
  - F0 (Fundamental Frequency)
  - Pitch Accent for lexical stress
  - Tune (Pitch over Time)

Lexical stress often involves loudness, duration, and pitch:

I'm surp<u>ris</u>ed that <u>some</u> people found HW 05 <u>diffi</u>cult.

Continuous Speech

Tune (pitch over time) is mostly observed in English with questions:



Automatic Speech Recognition (ASR)

ARS has a long and curious history...

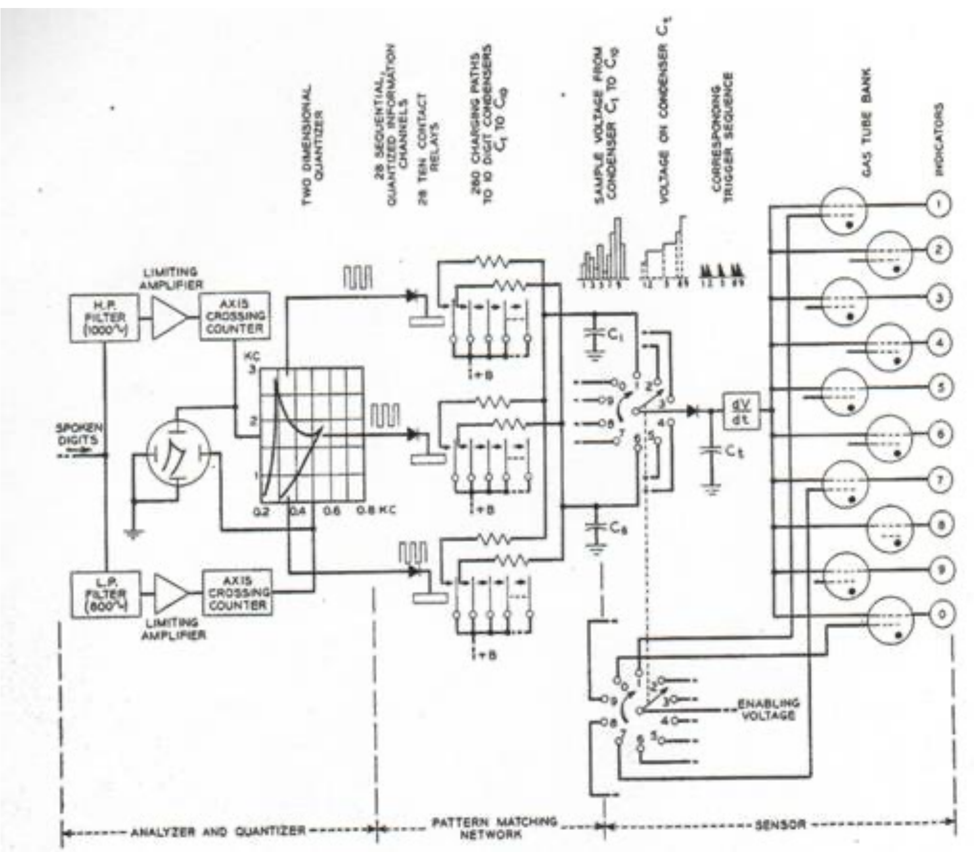
The first machine that recognized speech was probably a commercial toy named "Radio Rex" which was sold in the 1920's. Rex was a celluloid dog that moved (by means of a spring) when the spring was released by 500 Hz acoustic energy. Since 500 Hz is roughly the first formant of the vowel [eh] in "Rex", the dog seemed to come when he was called. (David, Jr. and Selfridge, 1962)

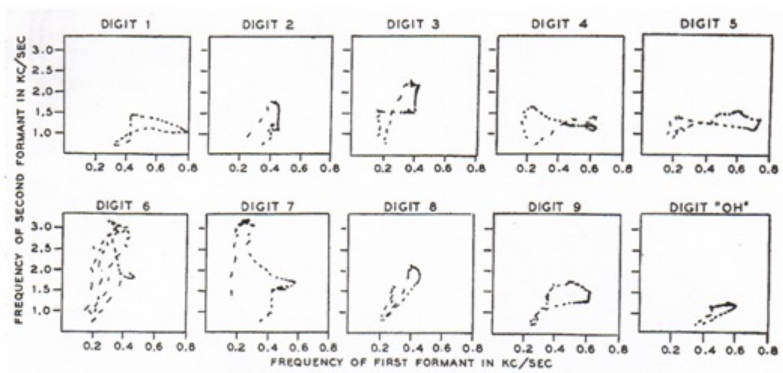
Radio Rex from 1920s - The first speech recognition machine

ASR: The Early Years

At first, ASR was an electrical engineering problem, e.g.,

- **Automatic Digit Recognition (AUDREY - 1952)**





ASR: Modern Era

In the modern era, ASR has benefitted from techniques from Electrical Engineering,Computer Science, and Linguistics:

- Large vocabulary
  - ~20,000-60,000 words or more…
- Speaker independent (vs. training on one speaker)
- Continuous speech (vs isolated-word)
- Multilingual, conversational
- World's best research systems:
  - Conversational speech: ~13-20% Word Error Rate (WER)
  - Human-machine or monologue speech: ~3-5% WER
- For much of the modern era, the best results were obtained by Hidden Markov Models (Viterbi Algorithm)

Recall: POS Tagging with Hidden Markov Models

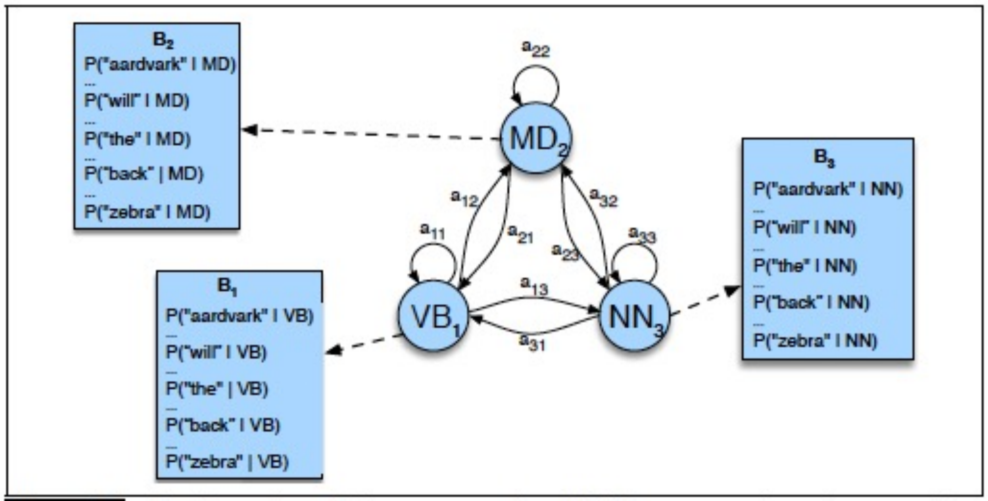


Figure 8.9 An illustration of the two parts of an HMM representation: the *A* transition probabilities used to compute the prior probability, and the *B* observation likelihoods that are associated with each state, one likelihood for each possible observation word.

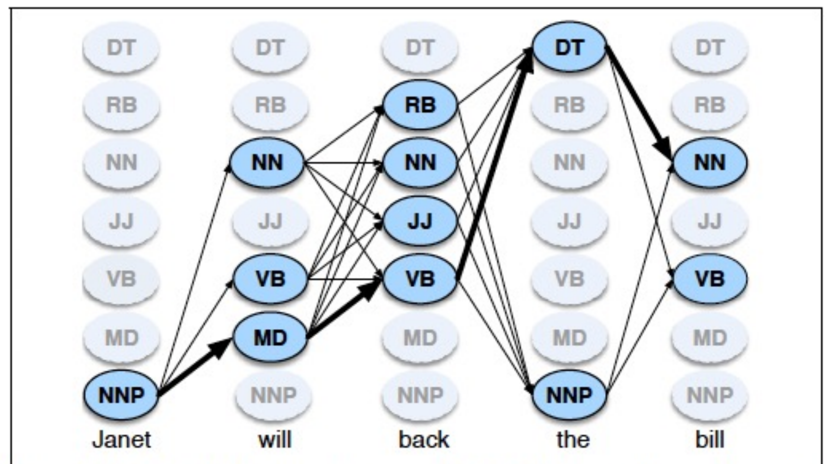
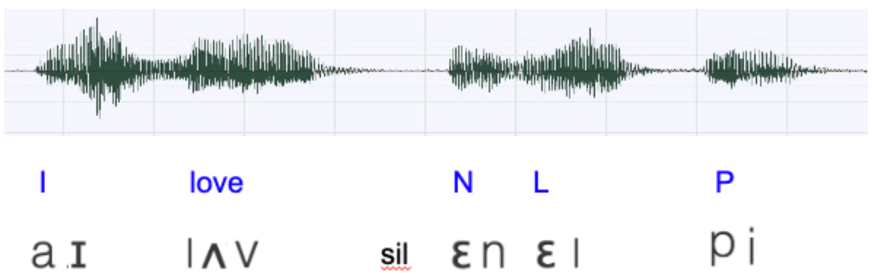


Figure 8.11 A sketch of the lattice for *Janet will back the bill*, showing the possible tags (q_i) for each word and highlighting the path corresponding to the correct tag sequence through the hidden states. States (parts of speech) which have a zero probability of generating a particular word according to the *B* matrix (such as the probability that a determiner DT will be realized as *Janet*) are greyed out.

ASR using Hidden Markov Models

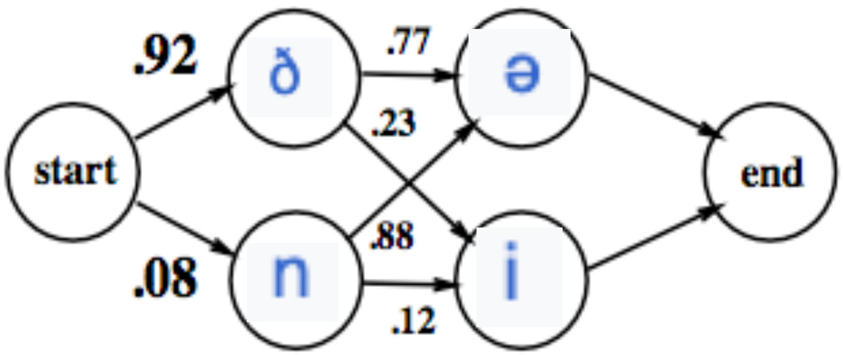
The basic approach starts out similarly to what you did in HW 05 by building a Viterbi model, but just for the words in the dataset:

1) Develop a training dataset from recordings, with annotations in IPA and English text:

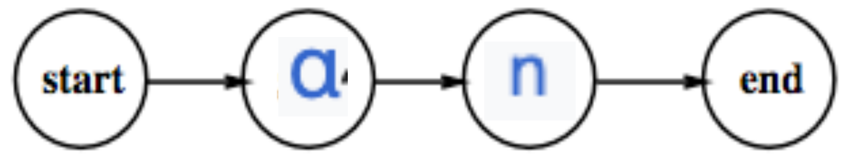


2) Build a Viterbi Word Model with phonemes (and sil) from the dataset:
   - Nodes are phonemes;
   - Start, Trans, and Emit dictionaries give probabilities of transitions among the nodes for words in vocabulary

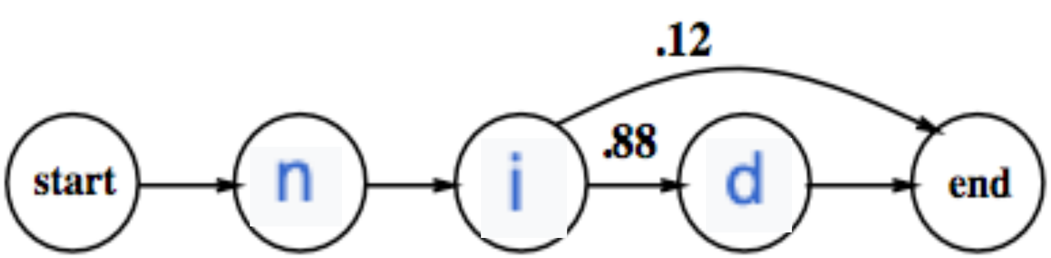
ASR using Hidden Markov Word Models



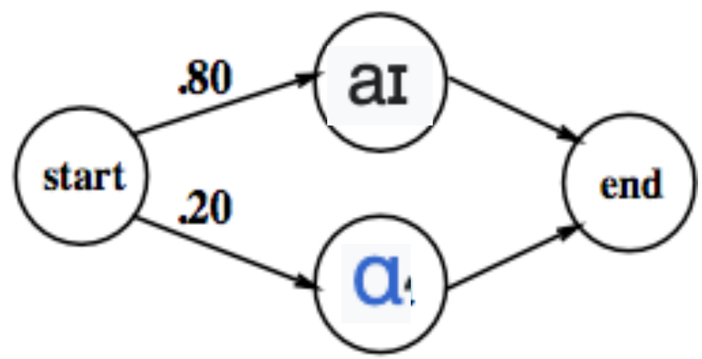
Word model for "the"



Word model for "on"



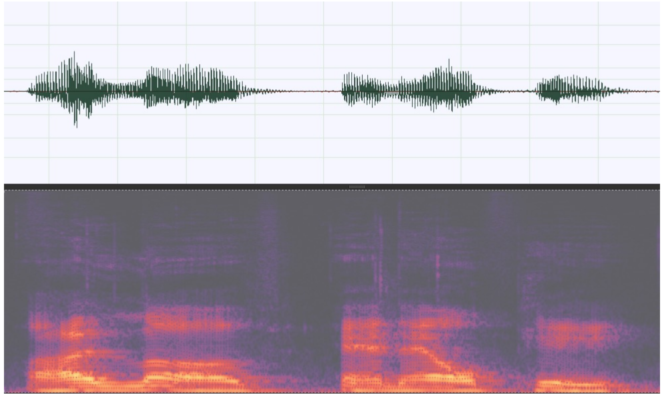
Word model for "need"



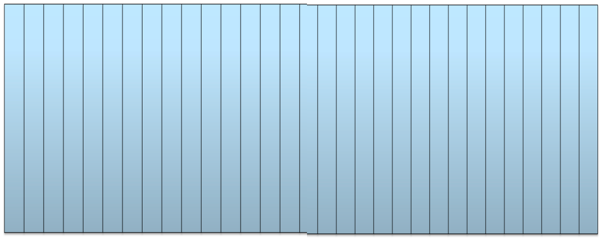
Word model for "I"

ASR using Hidden Markov Models

3) The test audio track is converted into a Log Mel Spectrogram

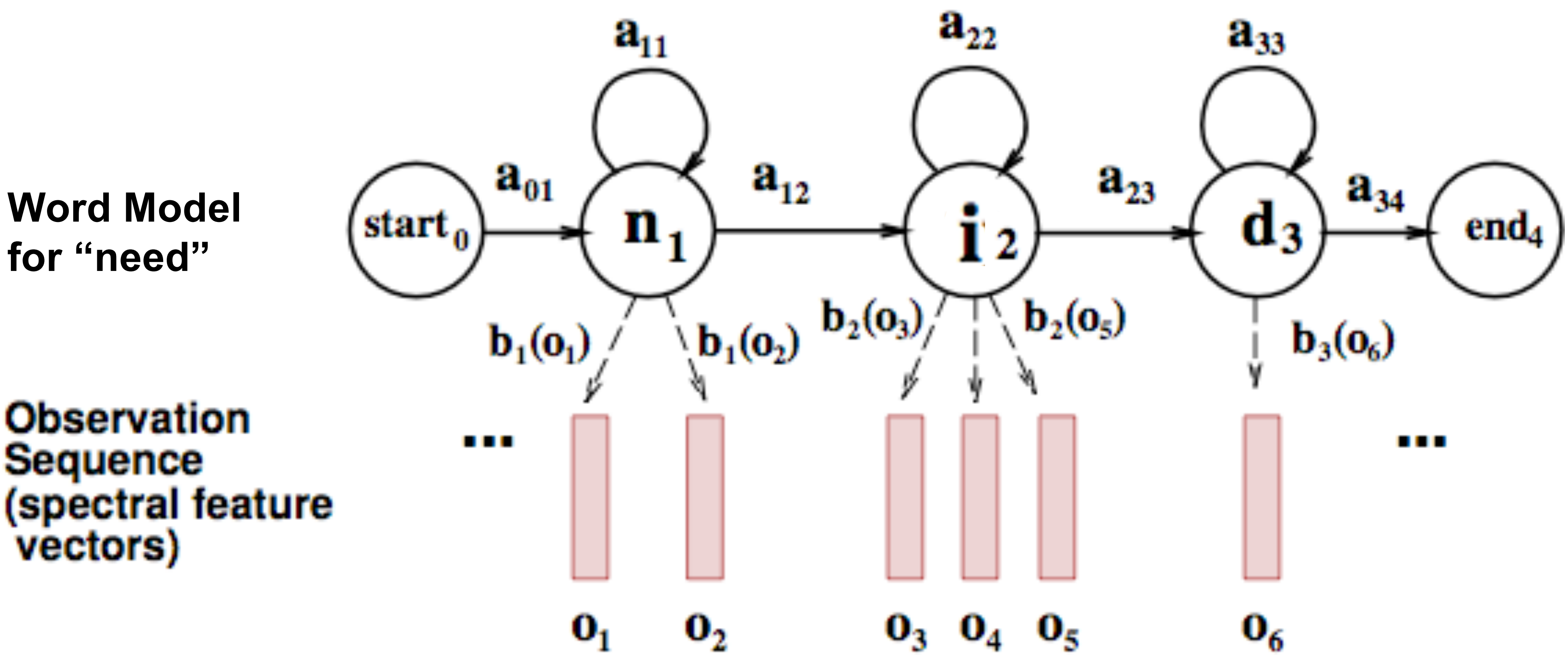


4) Then each spectrum (column in the spectrogram) is converted into an array of features (for a 50 msec section of the signal)
   - Mel Spectrogram Ceptrum Coefficients (MFCC)
   - Other statistical measures: spectral centroid, etc., etc.



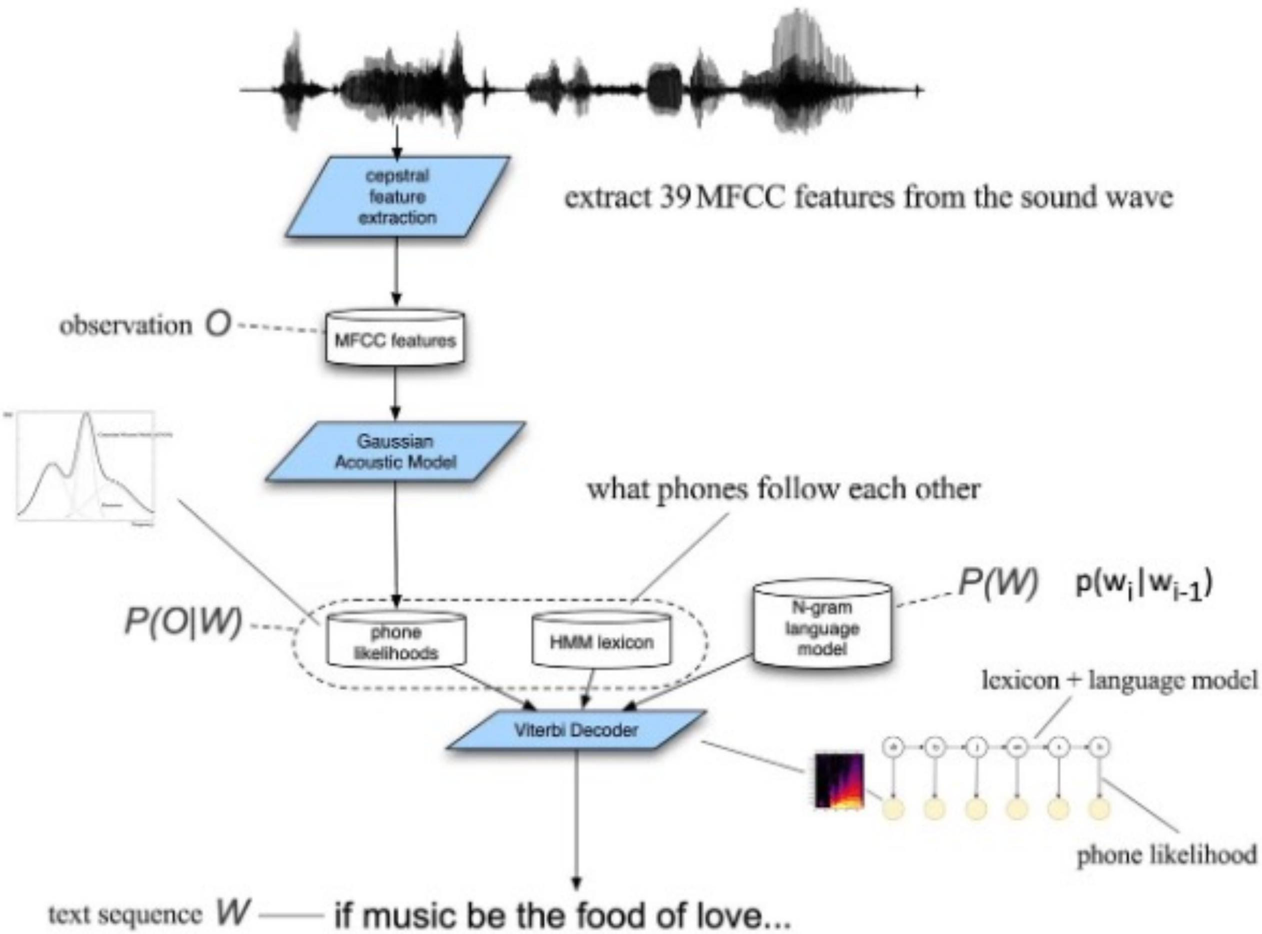
ASR using Hidden Markov Models

5) The feature vectors form the observation sequence input to the HMM:



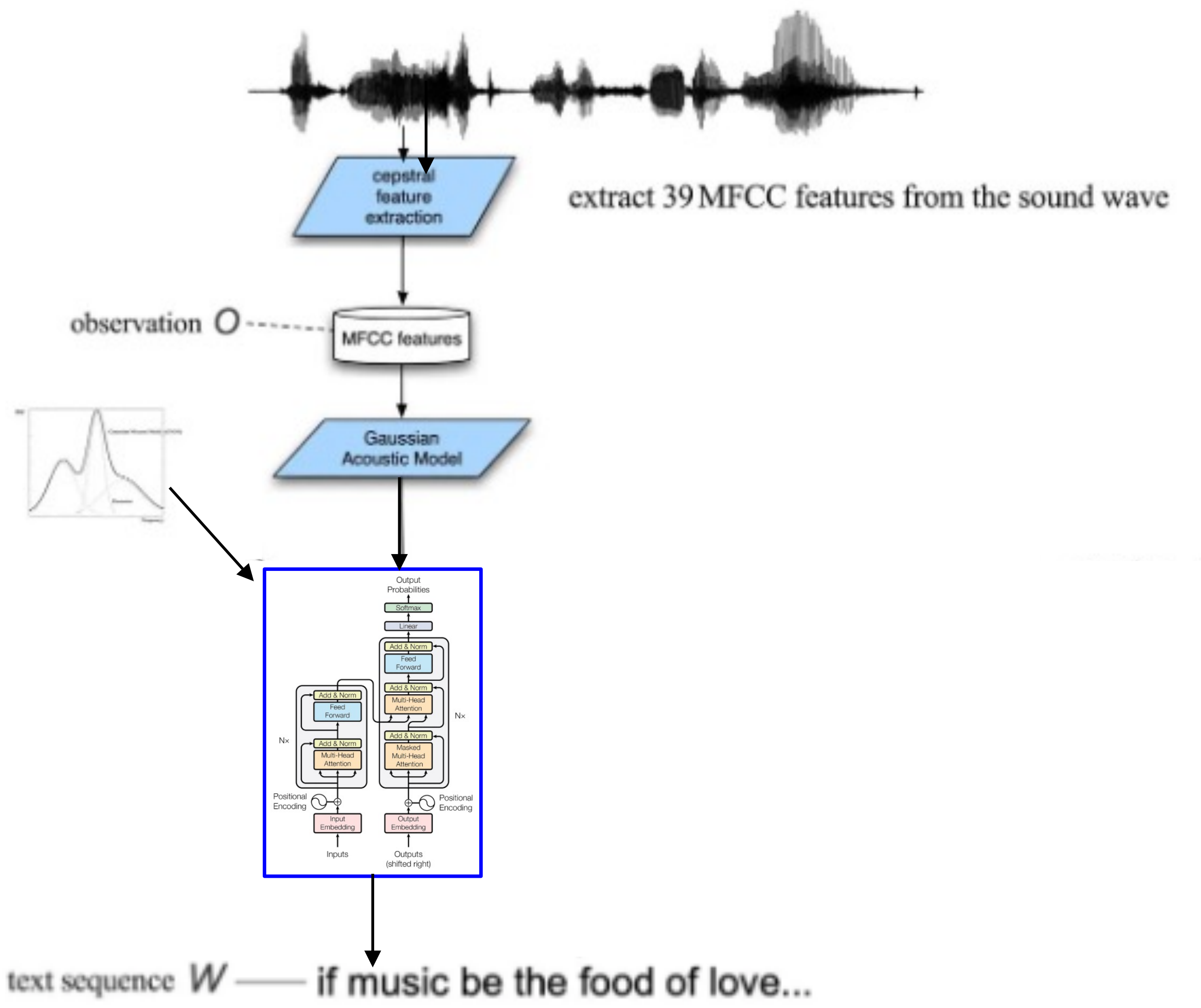
ASR using Hidden Markov Models

5) The decoding of the HMM is combined with an N-Gram language model to produce the most likely output text sequence:



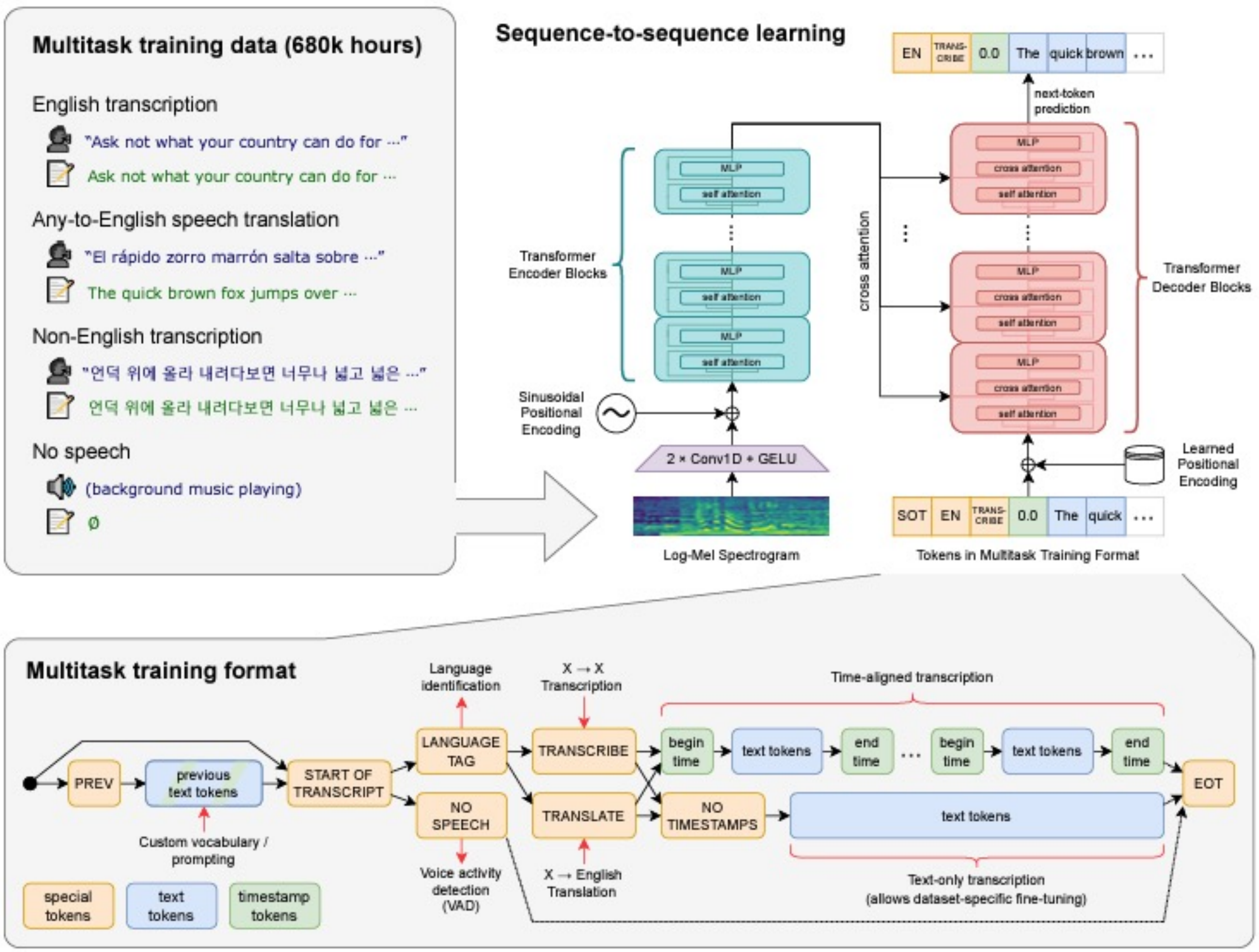
ASR in the Transformer Era

Since ASR is cast as a sequence to sequence task, it is not surprising that the most recent approaches use RNNs or Transformers:



ASR in the Transformer Era

The Whisper model from OpenAI is a good example of a transformer-based ASR system:



ASR in the Transformer Era

”Whisper is competitive with SOTA commercial and open-source ASR ystem in long-form transcription.”

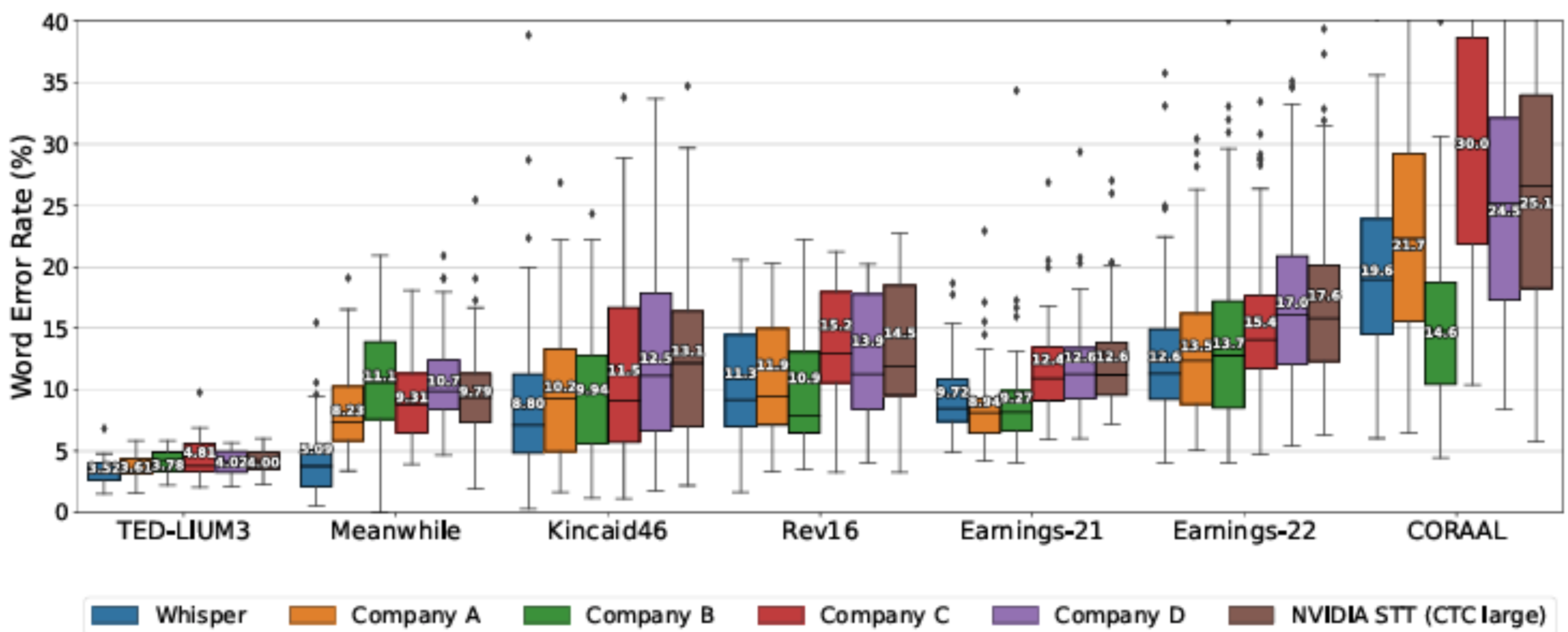


Figure 6. **Whisper is competitive with state-of-the-art commercial and open-source ASR systems in long-form transcription.** The distribution of word error rates from six ASR systems on seven long-form datasets are compared, where the input lengths range from a few minutes to a few hours. The boxes show the quartiles of per-example WERs, and the per-dataset aggregate WERs are annotated on each box. Our model outperforms the best open source model (NVIDIA STT) on all datasets, and in most cases, commercial ASR systems as well.