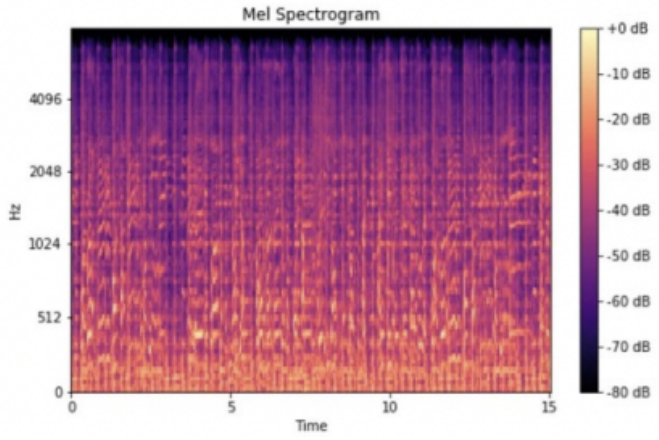
CAS CS 505
Lec 22

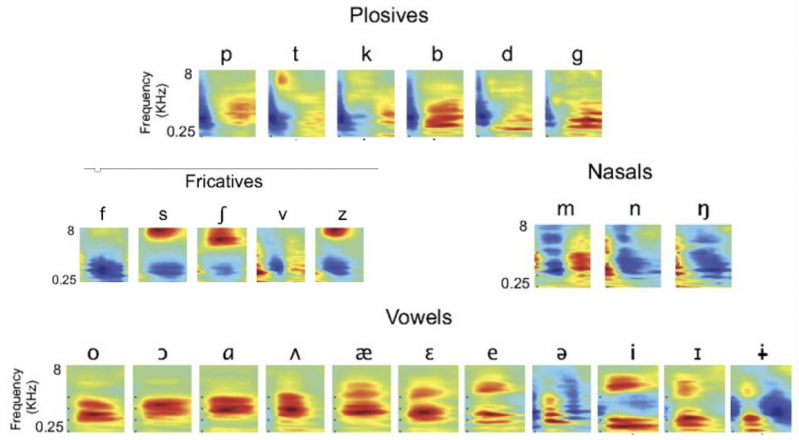
Automatic Speech Recognition (ASR)

1. Log MEL Spectrogram
    a. Therefore, to capture the human experience of sound, we typically use a Mel Spectrogram, where
        i. Pitch is given in Mels
        ii. Loudness is given in Decibels
    b. Both of these are log scales

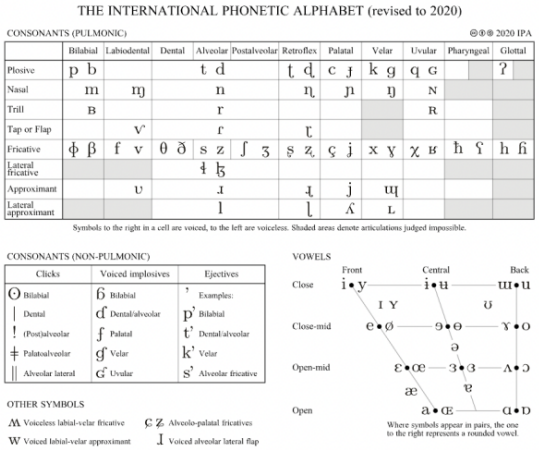
```
mel_spect = librosa.feature.melspectrogram(y=y, sr=sr, n_fft=2048,
hop_length=1024)
mel_spect = librosa.power_to_db(spect, ref=np.max)

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

    c. 

2. Human Vocal Signals
    a. Each phoneme in human language has a rather distinct spectrogram:



3. Phonemes and the International Phonetic Alphabet
    a. Phonemes are smallest unit of sound in a particular language which convey meaning.
    b. 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:



/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.

4. International Phonetic Alphabet
    a. 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

    b. 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

5. Spectra and Spectrograms for Vowels
    a. Vowels are continuous sounds, formed by the shaping of the vocal cavity; therefore, each has a (instantaneous) spectrum.
    b. These spectra have characteristic peaks, called formants, caused by the shape of the vocal cavity.

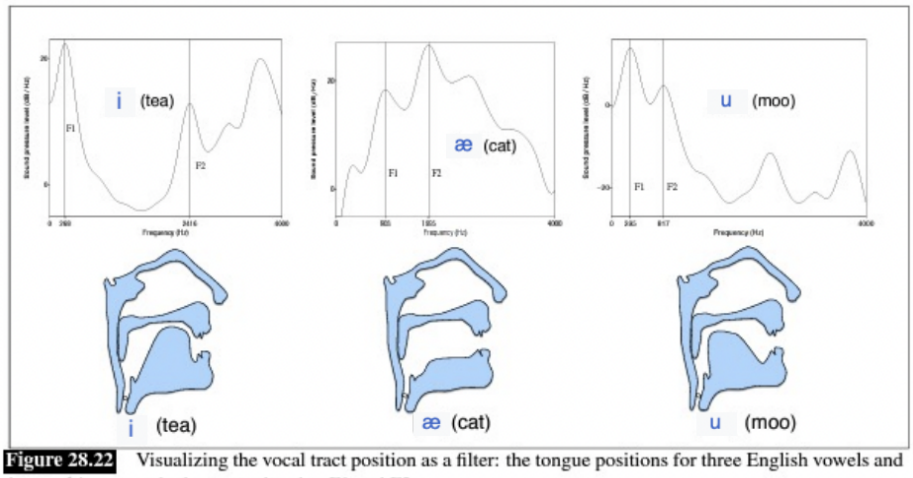


Figure 28.22 Visualizing the vocal tract position as a filter: the tongue positions for three English vowels and the resulting smoothed spectra showing F1 and F2.

c.

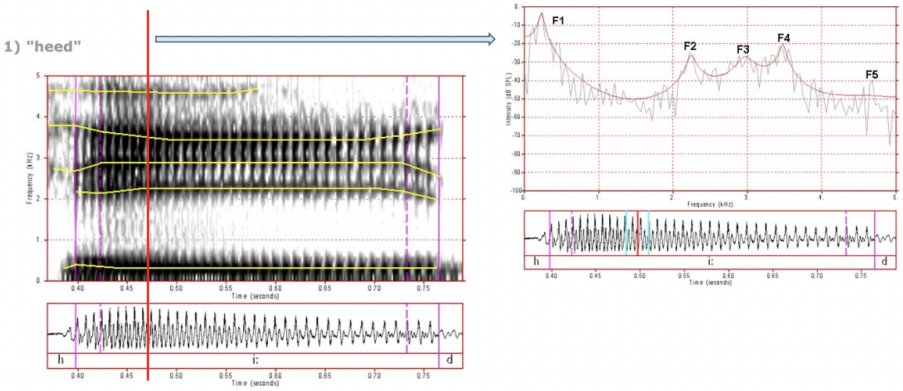
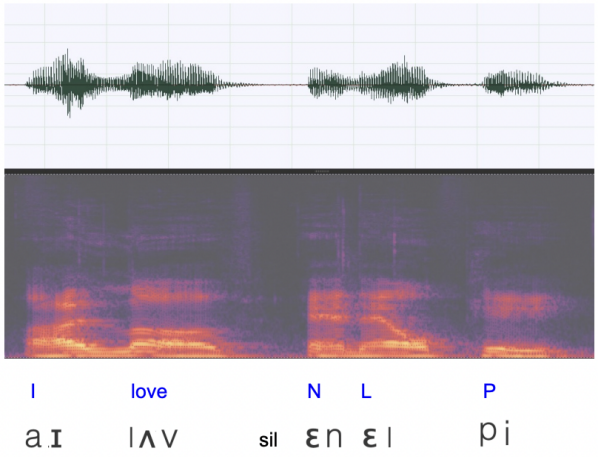


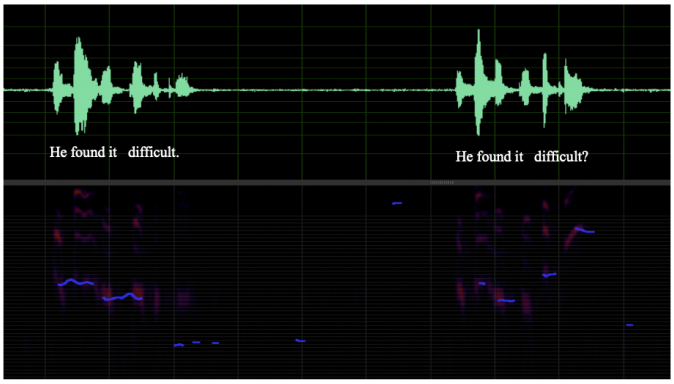
Figure 1: Broadband spectrogram of the vowel /i:/ from the token "heed".

d.

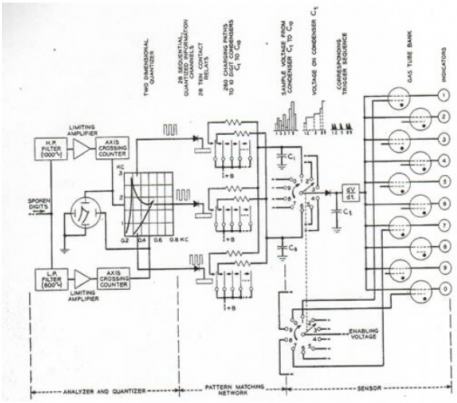
6. Spectrograms for Consonants and Vowels
    a. Vowels can be recognized by their (instantaneous) spectra, but consonants and semi-vowels (such as w or y) have time-dependent characteristics, and are best represented by spectrograms
7. Continuous Speech
    a. Continuous speech can be understood as a sequence of phonemes, possibly separated by periods of silence:

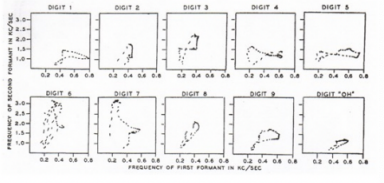


b. But continuous speech is complex!
c. In general, we must
    i. Identify individual phonemes
    ii. Identify words
    iii. Identify sentence structure and/or meaning
    iv. Interpret prosodic features
    v. Deal with mistakes, different speakers, accents, self-corrections, etc.
d. Prosodic features 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.
    i. Loudness
    ii. Duration
    iii. Pitch:
        1. F0 (Fundamental Frequency)
        2. Pitch Accent for lexical stress
        3. Tune (Pitch over Time)
e. Lexical stress often involves loudness, duration, and pitch:
f. Tune (pitch over time) is mostly observed in English with questions:



8. ASR: The Early Years
    a. At first, ASR was an electrical engineering problem, e.g.,
        i. Automatic Digit Recognition (AUDREY - 1952)





9. ASR: Modern Era
    a. In the modern era, ASR has benefitted from techniques from Electrical Engineering,Computer Science, and Linguistics:
        i. Large vocabulary
            1. ~20,000-60,000 words or more…
        ii. Speaker independent (vs. training on one speaker)
        iii. Continuous speech (vs isolated-word)
        iv. Multilingual, conversational
        v. World's best research systems:
            1. Conversational speech: ~13-20% Word Error Rate (WER)
            2. Human-machine or monologue speech: ~3-5% WER
        vi. For much of the modern era, the best results were obtained by Hidden Markov Models (Viterbi Algorithm)
10. Recall: POS Tagging with Hidden Markov Models

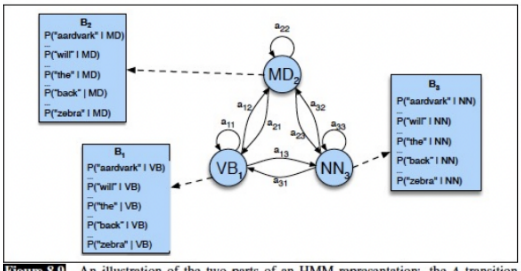


Figure 8.9 An illustration of the two parts of an HMM representation: the *A* transition probabilities used to compute the prior probability, and the *B* observation likelihoods that are associated with each state, one likelihood for each possible observation word.

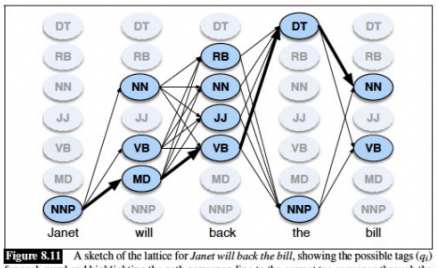
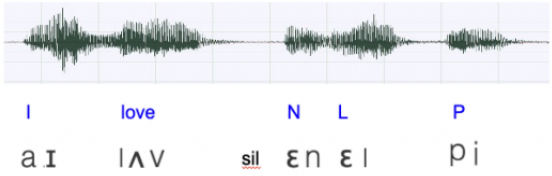


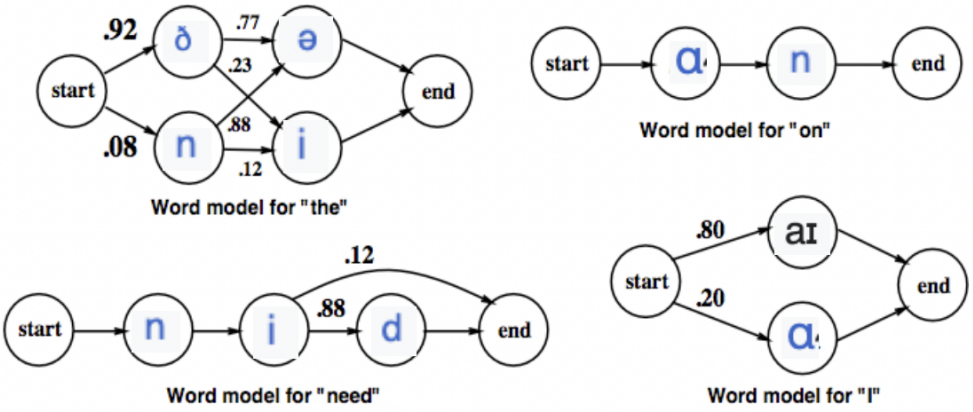
Figure 8.11 A sketch of the lattice for *Janet will back the bill*, showing the possible tags (q_i) for each word and highlighting the path corresponding to the correct tag sequence through the hidden states. States (parts of speech) which have a zero probability of generating a particular word according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

    a.
11. ASR Using Hidden Markov Models
    a. 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:
        i. 1) Develop a training dataset from recordings, with annotations in IPA and English text:

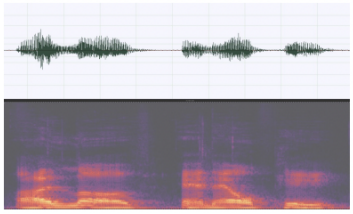


        ii. 2) Build a Viterbi Word Model with phonemes (and sil) from the dataset:
            1. Nodes are phonemes;

2. Start, Trans, and Emit dictionaries give probabilities of transitions among the nodes for words in vocabulary

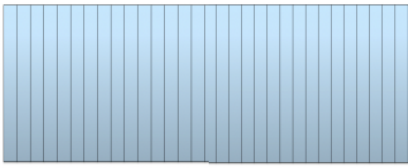


iii. 3) The test audio track is converted into a Log Mel Spectrogram

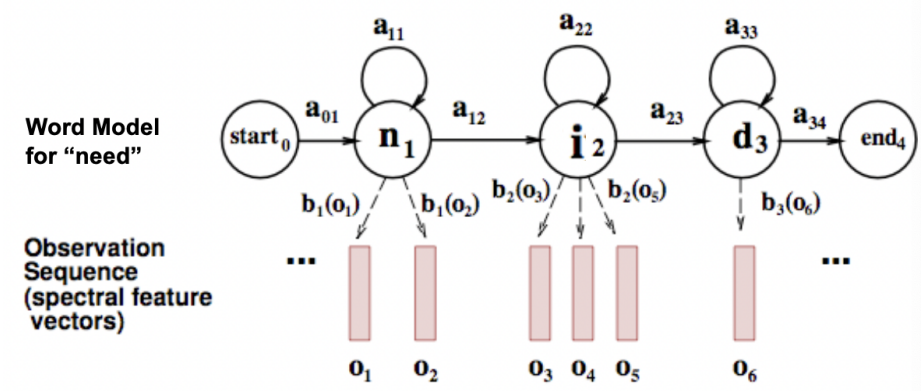


iv. 4) Then each spectrum (column in the spectrogram) is converted into an array of features (for a 50 msec section of the signal)

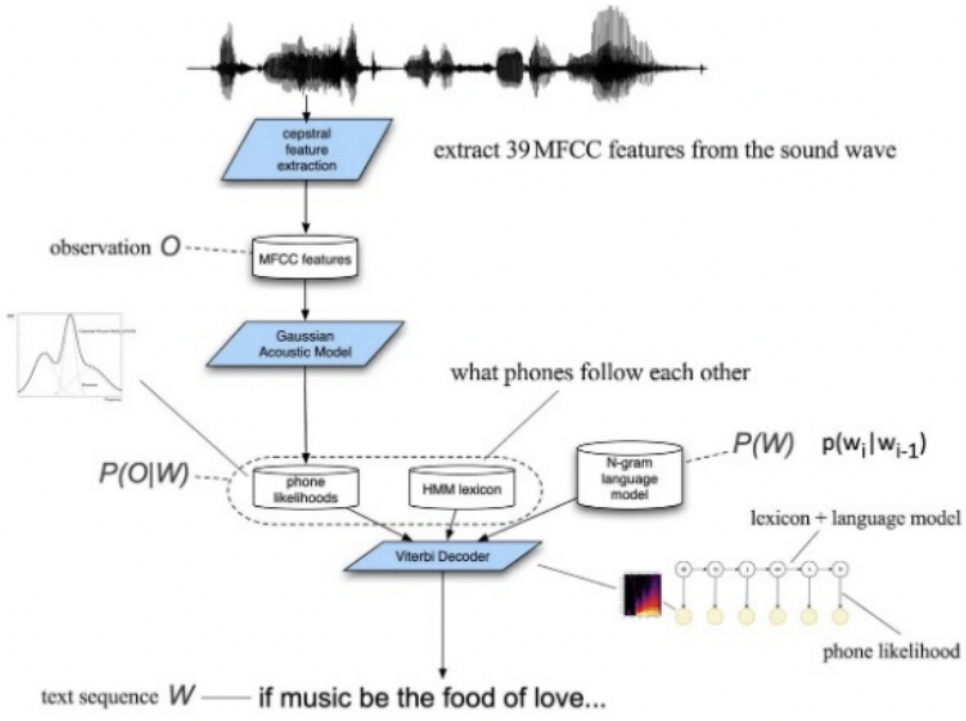
1. Mel Spectrogram Ceptrum Coefficients (MFCC)
2. Other statistical measures: spectral centroid, etc., etc.



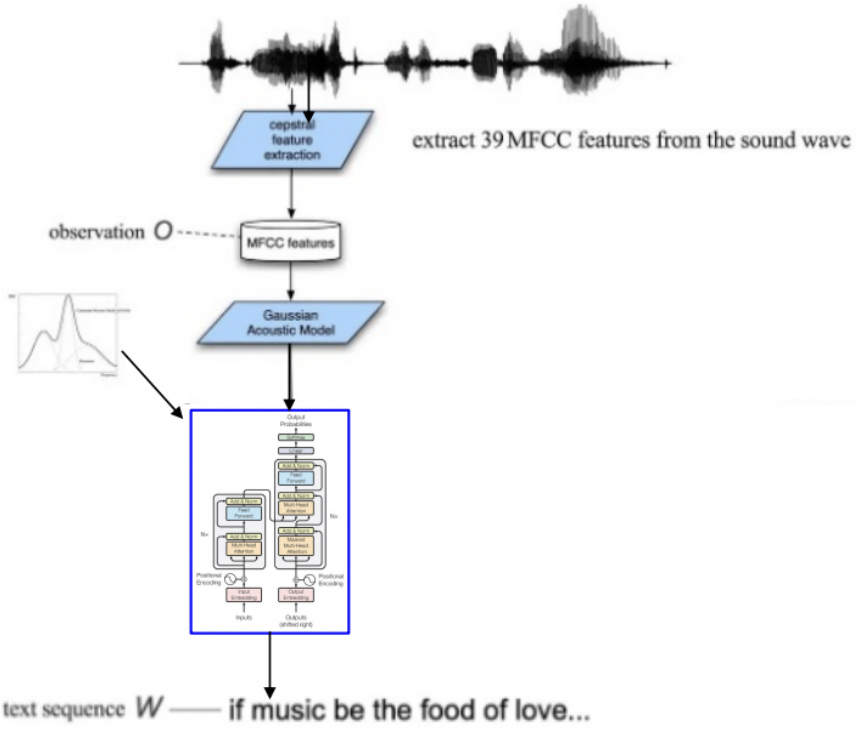
v. 5) The feature vectors form the observation sequence input to the HMM



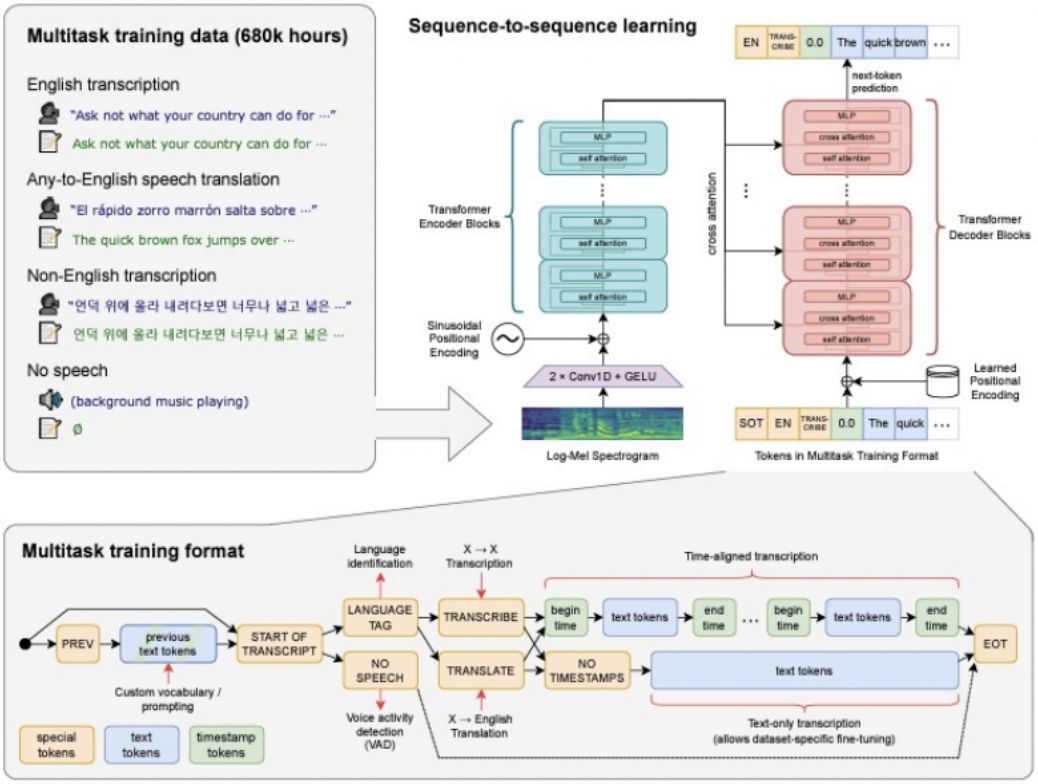
vi. 6) The decoding of the HMM is combined with an N-Gram language model to produce the most likely output text sequence:



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



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



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

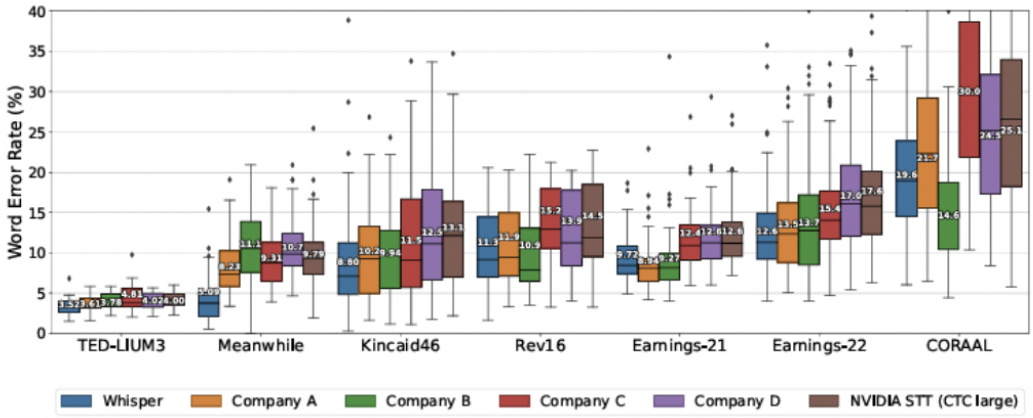


Figure 6. **Whisper is competitive with state-of-the-art commercial and open-source ASR systems in long-form transcription.** The distribution of word error rates from six ASR systems on seven long-form datasets are compared, where the input lengths range from a few minutes to a few hours. The boxes 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.