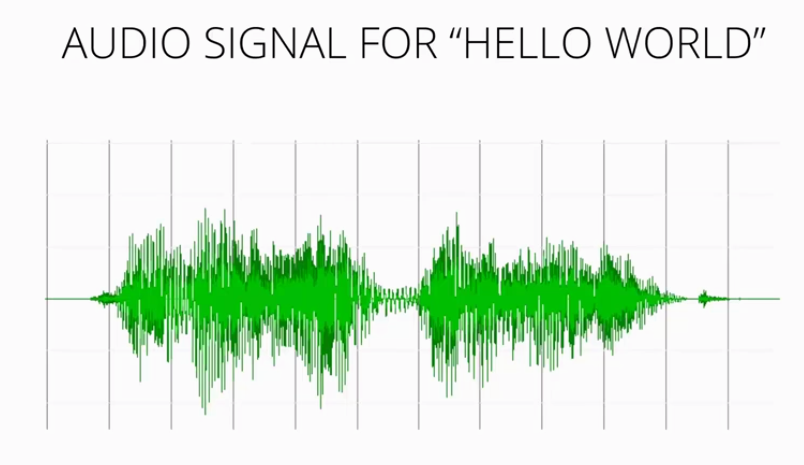
**References: Signal Analysis**

**Sound**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

Excellent explanations and definitions for vibration, frequency, sound waves, etc. can be found in [Wikipedia](https://en.wikipedia.org/wiki/Sound).

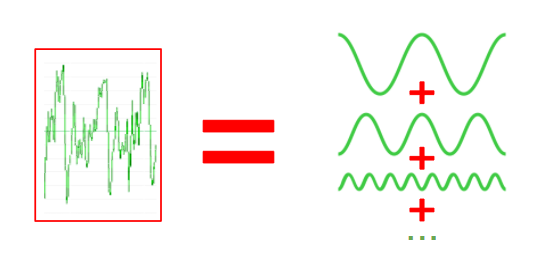
**Signal Analysis**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

Signal analysis in our context refers to the audio signal produced by speech. Sound vibrations cause pressure waves in the air that can be detected with a microphone and transduced into a signal. Detailed coverage of the topic as related to Speech Recognition can be found in the following:

[Cassidy, Steve. "Speech recognition." Sydney Australia (2002): Chapter 3.](http://web.science.mq.edu.au/~cassidy/comp449/html/ch03.html)

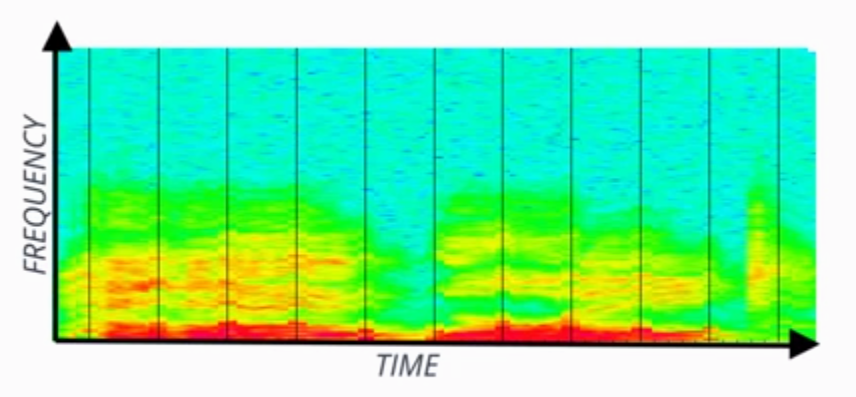
**Fourier Analysis**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

Fourier Analysis is the study decomposing mathematical functions into sums of simpler trigonometric functions. Since sound is comprised of oscillating vibrations, we can use Fourier analysis, and Fourier transforms to decompose an audio signal into component sinusoidal functions at varying frequencies. The following website explains the process:

[Fourier Transforms – the most important tool in mathematics?. (2014). IB Maths Resources from British International School Phuket.](https://ibmathsresources.com/2014/08/14/fourier-transforms-the-most-important-tool-in-mathematics/)

**Spectrograms**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

A spectrogram is the frequency domain representation of the audio signal through time. It's created by splitting the audio signal into component frequencies and plotting them with respect to time. The intensity of color in the spectrogram at any given point indicates the amplitude of the signal. The following reference includes interesting slides showing how sounds in spectrograms can be "read" by experts.

[Marcus, Mitch. "CIS 391 Artificial Intelligence." Philadelphia (2015). Seas.upenn.edu.](http://www.seas.upenn.edu/~cis391/Lectures/speech-rec.pdf)

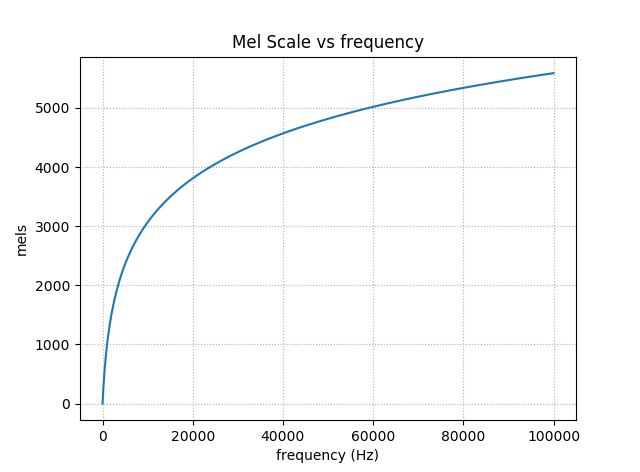
**References: Feature Extraction**

**Feature Extraction**

A summary of methods used in ASR:

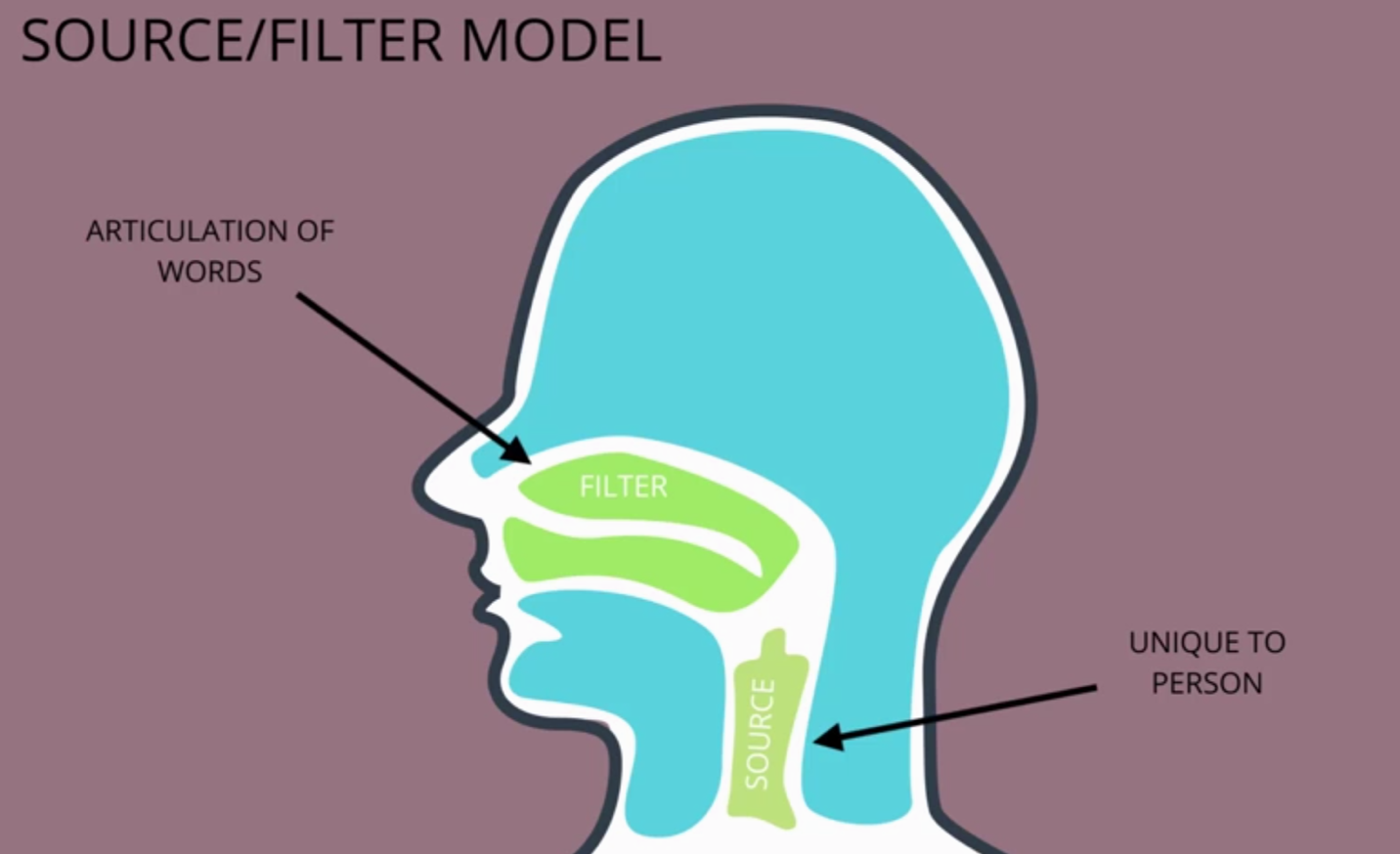
[Narang, Shreya, and Ms Divya Gupta. "Speech Feature Extraction Techniques: A Review." International Journal of Computer Science and Mobile Computing 4.3 (2015): 107-114.](http://www.ijcsmc.com/docs/papers/March2015/V4I3201545.pdf)

**Mel Scale**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

The Mel Scale was developed in 1937 and is based on human studies of pitch perception. At lower pitches (frequencies), humans can distinguish pitches better. Read more about it in [Wikipedia](https://en.wikipedia.org/wiki/Mel_scale)

**The Source/Filter Model**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

The source/filter model holds that the "source" of voices speech is dependent upon the vibrations initiated in the vocal box, and is unique to the speaker, while the "filter" is the articulation of the words in the forward part of the voice tract. The two can be separated through Cepstrum Analysis. A detailed explanation of the Source/Filter model for speech can be found at:

[Cassidy, Steve. "Speech recognition." Sydney Australia (2002): Chapter 7.](http://web.science.mq.edu.au/~cassidy/comp449/html/ch07.html#d0e1094)

**Cepstral Analysis**

The source/filter model motivates Cepstral Analysis. The intuition is that the "source" e(n)*e*(*n*) is multiplied by the "filter" h(n)*h*(*n*) to form the signal, s(n)*s*(*n*):

s(n)=e(n)×h(n)*s*(*n*)=*e*(*n*)×*h*(*n*)

This signal can be converted to the frequency domain through a discrete Fourier transform, or DFT (can use the FFT algorithm):

∣S(ω)∣=∣E(ω)∣⋅∣H(ω)∣∣*S*(*ω*)∣=∣*E*(*ω*)∣⋅∣*H*(*ω*)∣

Take the log and we can just add the source and filter instead of multiplying:

log∣S(ω)∣=log∣E(ω)∣+log∣H(ω)∣log∣*S*(*ω*)∣=log∣*E*(*ω*)∣+log∣*H*(*ω*)∣

Here's where it gets a bit tricky. By taking the inverse discrete Fourier transform, or IDFT, the signal can be split. This is the cepstrum  c(n)*c*(*n*) . Here's the final equation:

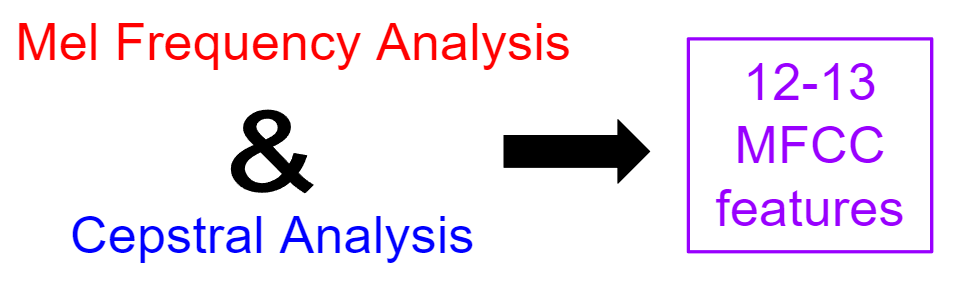
c(n)=IDFT(log∣S(ω)∣)=IDFT(log∣E(ω)∣+log∣H(ω)∣)*c*(*n*)=*IDFT*(log∣*S*(*ω*)∣)=*IDFT*(log∣*E*(*ω*)∣+log∣*H*(*ω*)∣)

Because we are splitting the logs of the frequencies, this is not the same as the original time domain, but rather now called the *quefrency* or *cepstral* domain. The vocal tract, or filter components that we want, can be extracted now because they vary slowly and are concentrated in the lower quefrency region.

Read more in the following thorough treatment complete with diagrams:

[Cepstral Analysis of Speech (Theory) : Speech Signal Processing Laboratory : Electronics & Communications : IIT GUWAHATI Virtual Lab](http://iitg.vlab.co.in/?sub=59&brch=164&sim=615&cnt=1)

**MFCC**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

Mel Frequency Cepstrum Coefficient Analysis is the reduction of an audio signal to essential speech component features using both mel frequency analysis and cepstral analysis. The range of frequencies are reduced and binned into groups of frequencies that humans can distinguish. The signal is further separated into source and filter so that variations between speakers unrelated to articulation can be filtered away. The following reference provides nice visualizations of the process of audio->spectrogram->MFCC:

[Prahallad, Kishore. "Speech Technology: A Practical Introduction, topic: Spectrogram, Cepstrum and Mel-Frequency Analysis." Carnegie Mellon University](http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf)

**MFCC Deltas and Delta-Deltas**

Intuitively, it makes sense that changes in frequencies, *deltas*, and changes in changes in frequencies, *delta-deltas*, might also be meaningful features in speech recognition. The following succinct tutorial for MFCC's includes a short discussion on deltas and delta-deltas:

[Mel Frequency Cepstral Coefficient (MFCC) tutorial. Practical Cryptography](http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/)

**References: Phonetics**

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

[Phonetics](https://en.wikipedia.org/wiki/Phonetics) is a branch of linguistics for the study of sounds of human speech: physical properties, production, acoustics, articulation, etc.

**Phoneme**

In any given language, a [phoneme](https://en.wikipedia.org/wiki/Phoneme) is the smallest sound segment that can be used to distinguish one word from another. For example "bat" and "chat" have only one sound different but this changes the word. The phonemes in question are "B" and "CH". What exactly these are and how many exist varies a bit and may be influenced by accents included. Generally, US English consists of 39 to 44 phonemes. See ARPAbet below for more phoneme examples.

**Grapheme**

The definition of a [grapheme](https://en.wikipedia.org/wiki/Grapheme) is somewhat inconsistent in the literature. In our context, a grapheme is the smallest symbol that distinguishes one written word from another. For example, "bat" and "chat" have a difference of two graphemes, even though "CH" is considered to be a single phoneme. In US English, 26 letters and a space combine for 27 possible graphemes.

**Lexicon**

A lexicon for speech recognition is a lookup file for converting speech parts to words. An example of this is [cmudict](http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/sphinxdict/cmudict_SPHINX_40" \t "_blank), the Carnegie Mellon tool for speech recognition compatible with the open source [Sphinx](https://cmusphinx.github.io/) project. Here's a short excerpt:

AARDVARK AA R D V AA R K

AARON EH R AH N

AARON'S EH R AH N Z

AARONS EH R AH N Z

AARONSON EH R AH N S AH N

AARONSON'S EH R AH N S AH N Z

AARONSON'S(2) AA R AH N S AH N Z

AARONSON(2) AA R AH N S AH N

...

**ARPAbet**

A set of phonemes developed by the Advanced Research Projects Agency(ARPA) for the Speech Understanding Project (1970's).

[ARPAnet on Wikipedia](https://en.wikipedia.org/wiki/Arpabet) [ARPAnet dictionary at CMU](http://www.speech.cs.cmu.edu/cgi-bin/cmudict" \t "_blank):

| **Phoneme** | **Example** | **Translation** |
| --- | --- | --- |
| AA | odd | AA D |
| AE | at | AE T |
| AH | hut | HH AH T |
| AO | ought | AO T |
| AW | cow | K AW |
| AY | hide | HH AY D |
| B | be | B IY |
| CH | cheese | CH IY Z |
| D | dee | D IY |
| DH | thee | DH IY |
| EH | Ed | EH D |
| ER | hurt | HH ER T |
| EY | ate | EY T |
| F | fee | F IY |
| G | green | G R IY N |
| HH | he | HH IY |
| IH | it | IH T |
| IY | eat | IY T |
| JH | gee | JH IY |
| K | key | K IY |
| L | lee | L IY |
| M | me | M IY |
| N | knee | N IY |
| NG | ping | P IH NG |
| OW | oat | OW T |
| OY | toy | T OY |
| P | pee | P IY |
| R | read | R IY D |
| S | sea | S IY |
| SH | she | SH IY |
| T | tea | T IY |
| TH | theta | TH EY T AH |
| UH | hood | HH UH D |
| UW | two | T UW |
| V | vee | V IY |
| W | we | W IY |
| Y | yield | Y IY L D |
| Z | zee | Z IY |
| ZH | seizure | S IY ZH ER |

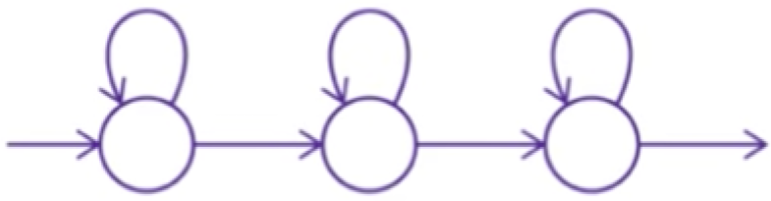
# References: Traditional ASR

### Traditional ASR

A bit of Computer History Museum nostalgia on Speech Recognition presents what we think of now as "Traditional" ASR:

[Kai-Fu Lee (Apple) in 1993. Computer History Museum video.](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjMtPu2vtfUAhWF6IMKHQIHDW0QtwIIJjAA&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DPJ_KCTsOCrs&usg=AFQjCNFVClgb-77HLUdZBhZjSDax7AYtAg)

### Acoustic Models with HMMs

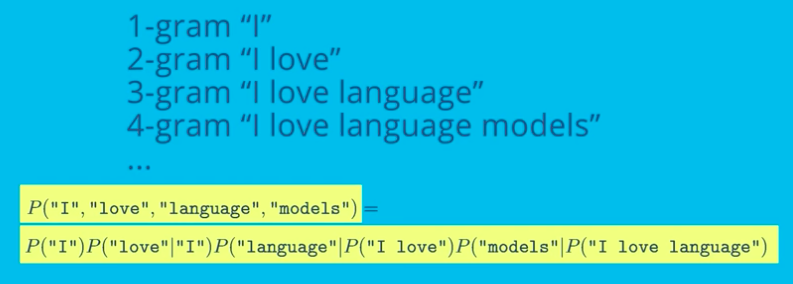
[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

HMMs are the primary probabalistic model in traditional ASR systems. The following slide decks from Carnegie Mellon include very helpful and detailed visualizations of HMM's, the Viterbi Trellis, State Tying, and more from the Carnegie Mellon:

**Raj, Bhiksha, and Rita Singh. "Design and implementation of speech recognition systems." Carnegie Mellon School of Computer Science (2011).**

* [slides - HMMs](http://www.cs.cmu.edu/~bhiksha/courses/11-756.asr/spring2014/lectures/class7-8.hmm.pdf)
* [slides - Continuous Speech](http://www.cs.cmu.edu/~bhiksha/courses/11-756.asr/spring2014/lectures/class9.continuousspeech.pdf)
* [slides - HMM tying](http://asr.cs.cmu.edu/spring2011/class21.6apr/class21.subwordunits.pdf)

### N-Grams

[[](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)](https://classroom.udacity.com/nanodegrees/nd892/parts/48e073f9-1f1d-441e-b98f-02549c679583/modules/03ef3f43-fa0b-4494-8350-897119d3dc6e/lessons/3c20697b-fd12-4ac8-afe3-fc8f80b932d4/concepts/bf895442-7584-420f-a9ac-324dc5af1160)

N-Grams provide a way to constrain a series of words by chaining the probabilities of the words that came before. For more on creating and using N-Grams, see the references below:

[Martin, James H., and Daniel Jurafsky. "Speech and language processing." International Edition 710 (2014). Chapter 4 Draft.](https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf)

[Jurafsky, Daniel. "CS124 - From Languages to Information". Stanford University.Language Modeling. Slides](http://web.stanford.edu/class/cs124/lec/languagemodeling2016.pdf)

**References: Deep Neural Network ASR**

**Deep Speech 2**

The following presentation, slides, and paper from Baidu on *DeepSpeech 2* were important resources for the development of this course and its capstone project:

* [Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." International Conference on Machine Learning. 2016.](https://arxiv.org/pdf/1512.02595v1.pdf)
* [Presentation](https://www.youtube.com/watch?v=g-sndkf7mCs)
* [Slides](https://cs.stanford.edu/~acoates/ba_dls_speech2016.pdf)

**Language modeling with CTC**

Gram-CTC from Baidu on integrating a language model into CTC for better performance:

* [Liu, Hairong, et al. "Gram-CTC: Automatic Unit Selection and Target Decomposition for Sequence Labelling." arXiv preprint arXiv:1703.00096 (2017).](https://arxiv.org/pdf/1703.00096.pdf)

Language modeling with CTC based on weighted finite-state transducers (WFSTs):

* [Miao, Yajie, Mohammad Gowayyed, and Florian Metze. "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding." Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015.](https://arxiv.org/pdf/1507.08240.pdf)
* [Slides](http://people.csail.mit.edu/jrg/meetings/CTC-Dec07.pdf)