Tutorial 2: Understanding Feature Development for Speech Recognition

Alok Debnath

11th January 2019

1 Introduction

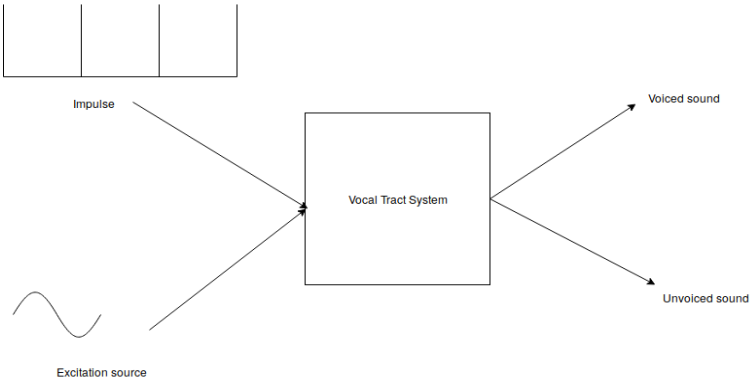

Figure 1: Basic Speech Processing Pipeline

The process of speech analysis is clear from the diagram above. The output of the system, which is the detection of voiced speech, unvoiced speech or silence, can be considered to be a spectrum (as is seen by a spectrograph). This output can be considered as a function of the input and the features extracted by the system. Abstractly, the input is a wave function (in the time or frequency domain, represented by $e(n)$) and the system feature extraction is also some mathematical function (represented by $v(n)$, discussed in the later sections of the tutorial). Therefore the spectrum ($S(n)$) is represented as follows:

$$S(t) = e(t) * v(t)$$

$$S(\omega) = e(\omega) \cdot v(\omega)$$

The operator in the first equation (in the time domain) is called convolution. By Fourier transformation on convolution, the latter equation can be obtained in the frequency domain, which is a simple multiplication operation. With this

idea in mind, the idea of this tutorial is to familiarize ourselves with the spectrum function $S(n)$ and the features extracted from this spectrum. Therefore, this tutorial is geared towards answering the following questions:

- What are features in speech analysis? What are some of the criteria for feature selection?
- What are MFCC and LPCC?
- How can we use prosodic features in automatic speech recognition?
- What are some of the applications of linguistic expertise in ASR?

2 What are features in Speech Analysis? What are some of the criteria for feature selection?

In speech analysis, a feature is an individual measurable property or characteristic of a phenomenon being observed. Features are a common term in machine learning and most AI applications rely on feature extraction, and speech analysis is no exception. While the definition and intuition of features is general across multiple subjects, the criteria for feature selection is extremely specific to the task at hand. For the system at hand, the features extracted have to satisfy the following properties:

- The feature should be identifiable and computable by a linear time invariant (LTI) system.
- It should be possible to uniformly model the feature regardless of the speech patterns in the input.
- The class of features extracted should have a computable number of members.

For reasons one and two above, we can not use features like pitch, amplitude, intonation and frequency (prosodic features) can not be used for short time periods. For reason three, the unit modelled is not a syllable or a word, but a phoneme. For sake of ease, the features are analysed in three time periods:

- Subsegmental (¡ 5 ms): These features are generally used for excitation analysis
- Segmental (10-20 ms): These features are spectral in nature.
- Suprasegmental (¿ 100 ms): Prosodic features and other volatile characteristics.

Prosodic features are used (reasons in a later section) for suprasegmental analysis. Features such as MFCC and LPCC are used segmentally. Features specific to certain sounds such as plosives and determining vowel onset point are subsegmental.

3 What are MFCC and LPCC?

3.1 MFCC: Mel Frequency Cepstrum Coeffecients

Mel-frequency cepstrum coefficients are coefficients, which are a representation of a short term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency of a pre-emphasized wave in the frequency domain.

Complex, right? Let's break it down:

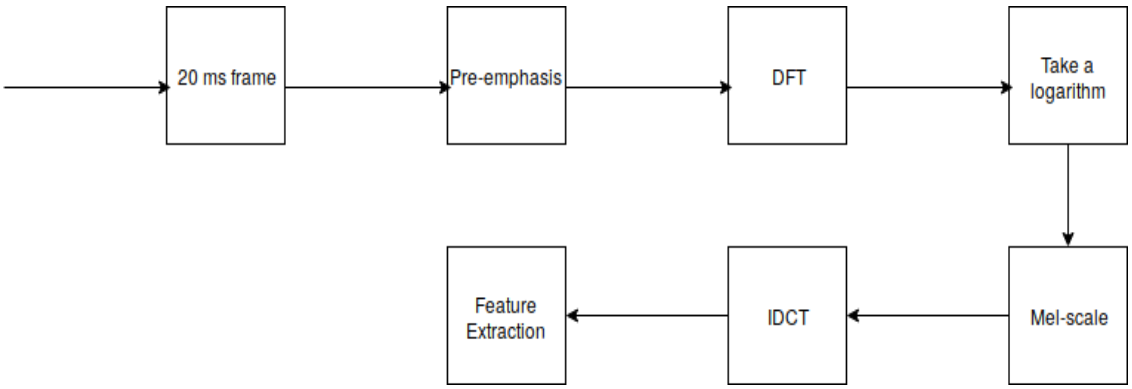

Figure 2: MFCC Extraction Pipeline

This diagram has the following steps:

- The signal is sampled, into frames of 20 ms (segmental). The frequency of sampling is usually 8 kHz (8000 samples per second), so 160 samples have been chosen for feature extraction.
- These samples are pre-emphasized, meaning that the high frequency (low energy) samples are normalized.
- A discrete Fourier transform (DFT) is applied on this set of samples $X_k = \Sigma_{n=0}^{N-1} x_n e^{\frac{-2\pi i k_n}{n}}$. The N chosen is usually 256, which is the final number if frequency features.
- A log is taken at each of the frequencies.
- The powers of the spectrum are then mapped onto a mel scale, with triangular overlapping windows.
- The spectrum now undergoes an inverse discrete cosine transformation (IDCT), which makes the resultant output a cosine function, therefore domain restricting and more computable.
- The resultant amplitudes are the MFCC characteristics, out of which the first 11 to 15 are usually the most used ones.

Mel Scale: The mel scale is the scale which approximates the reception of sounds at the ear drum. It is a scale of actual frequency to perceived frequencies, which is initially linear, but then plateaus out. Using the mel scale has proved to be a very useful feature, but also makes the frequency analysis very susceptible to noise.

Triangular window: In speech analysis, the analyzer has a choice of window size and shape. The window shape is important due to problems of boundary detection, amplitude and energy normalization and so on. For MFCC, the most common type of window is an overlapping triangular window, which looks like a wave, in order to preserve fuzzy boundaries and preventing loss of subsegmental information.

3.2 LPCC: Linear Prediction Cepstral Coefficients

The idea behind LPCC is similar to a mathematical concept known as Markov's assumption. Markov's assumption states that given the previous $n-1$ samples, the value (tag, class etc.) of the nth sample can be predicted. LPCC states that the current value of the spectral function is a linear weighted sum of the previous $n-1$ samples. This is written as $S_{LPCC}(n) = \Sigma_{k=1}^{p} a_k S(n-k)$. Note that this can only be applied to a 20 ms window, when the wave is still quasi-static. The value of a_k is found by performing cepstral analysis on the wave on the wave at that point.

LPCC, much like other machine learning algorithms, have an error function (calculated by deviation after reconstruction of the original wave), which can be simplistically expressed as $e(n) = (S(n) - S_{LPCC}(n))^2$

4 How can we use prosodic features for speech analysis?

Segmental spectral features like MFCC and LPCC are used for detecting voiced and unvoiced speech and for detecting excitation features. However, emotion detection, speaker detection and other speech analysis techniques also need suprasegmental features, which are not spectral in nature. As mentioned before, suprasegmental windows are not linear time invariant, so the features are not considered as it is. Rather, the prosodic features chosen are based on the average differences captured in analyzing multiple waveforms of multiple speakers and variations. Note that the prosodic features are time domain features, not frequency domain features like spectral features.

Prosodic features refer to certain properties of the speech signal such as intonation, duration and intensity in speech.

- **Intonation**: The dynamics of pitch or F_0 patterns over time is known as intonation contour. The patterns considered include:
  - Change in F_0

  - Distance of F_0 peak with respect to VOP
  - Amplitude tilt
  - Duration tilt

- **Duration**: The sequence of length of syllables is known as duration patterns. This is also called the rhythm features which includes:
  - Duration between successive VOP
  - Duration of voiced regions
  - Change in F_0

- **Intensity**: The dynamics of intensity patterns over time is known as intensity contour. Intensity features are also considered as stress features.
  - Change in log energy in voiced region
  - Change in F_0
  - Distance between successive VOP

4.1 What are the drawbacks of Prosodic features?

Drawbacks of using the prosodic features include the following:

- What is the fundamental frequency of the waveform?
- The detection of Voiced Onset Point based on detection algorithms is tedious.
- Identification of voiced speech and durational issues on alternating voiced and unvoiced sounds.

5 Applications of Linguistics in Speech Analysis

Speech analysis is a multi-domain field. The importance of linguistics in speech analysis can be seen widely in the following two applications.

5.1 Automatic Speech Recognition

Automatic Speech Recognition is one of the most important speech analysis applications. Linguistic analysis is used in almost all steps of the process. Take a look at the very simple ASR Pipeline shown. While the Acoustic Model is related to sound and wave physics in general, the language model (the prediction of the next phoneme given the previous phonemes in a given language) requires linguistic intuition based on data of that language.

Language models are a significant tool in NLP, most of which are statistical in nature. However, using and extracting certain linguistic features, the model can be tuned towards a particular language or domain.

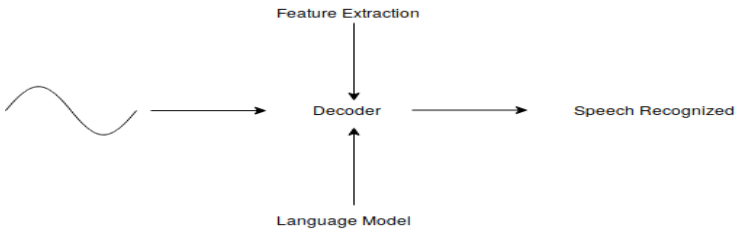

Figure 3: ASR Pipeline

5.2 Text to Speech

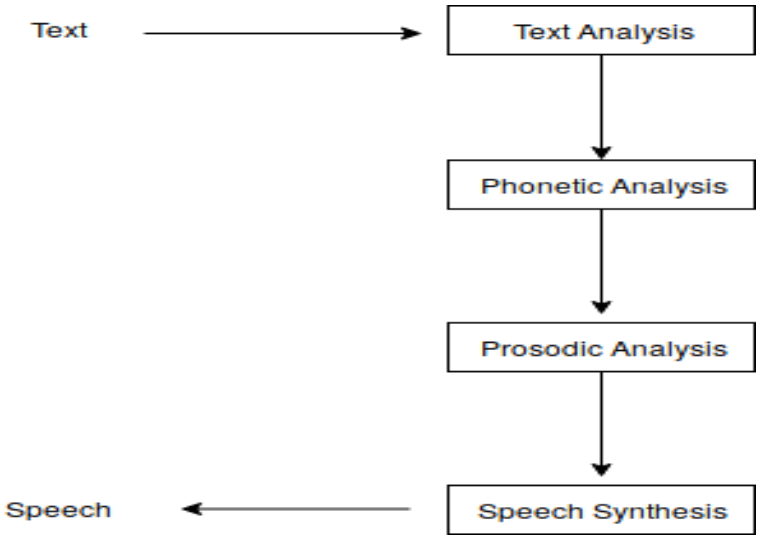

Figure 4: Text to Speech Pipeline

Self-explanatory