Tutorial 1: Introduction to Speech Analysis and Recognition

Alok Debnath

4 January 2019

1 Introduction

Speech analysis and processing is an important concept in computational linguistics, as speech is considered the most natural form of communication, noted in the fact that languages exist which have no writing systems, which are still phonetically, morphologically and syntactically rich and well developed. In this tutorial, we answer the questions:

1. What is speech? How is different from audio and voice?
2. What are the main goals of speech analysis?
3. Why is speech recognition a problem?
4. How do we model speech?
5. What are some basic tools, techniques and methods needed in speech analysis and recognition?

The last question will involve modeling of speech wave forms using a tool called *WaveSurfer*, which will be the lab component of this tutorial. Please ask if there are any doubts with respect to using the tool or analyses, as an assignment will be released on the same, shortly.

2 What is speech? How is different from audio and voice?

Speech is defined informally as a legal sequence of legal sounds produced by human beings. A legal sequence would be a sequence which corresponds to a meaningful utterance in the language intended for communication. Legal sounds are those which are in that subset of the IPA which the language "subscribes" to. Note that dialectal variations are what make this definition informal. On one hand some sequences are legal sequences, but the sounds may not be a part of the initial repertoire of sounds of the language. On the other hand, even with legal sounds, a particular sequence may not be the ”correct” pronunciation, but close enough to a legal sequence to be considered a dialectal variation. **Audio** is a cover term for speech and music (note that this is a rather layman distinction, audio seems to cover sounds which are not noise). **Voice** any sound produced by human beings.

3 What are the main goals of speech analysis?

The table below categorizes the features in a spoken sentence and the tasks in speech analysis corresponding to the same.

Information in Sentence	**Task in Speech Analysis**
Language	Language Identification
Accent and Dialect	Accent Detection, Dialect Detection
Message	*Automatic Speech Recognition*
Emotion	Emotion Detection

Another use of speech analysis is text-to-speech (TTS) generation system, which focuses on the generation aspect of speech as opposed to the analysis aspect.

4 Why is speech analysis a problem?

Multiple factors contribute to speech analysis and recognition being an ongoing puzzle in research. Apart from the obvious issues on the input end such as noise, dampening and capture of analog signals into a discreet signal procesing platform, the more linguistically oriented issues are as follows:

1. Natural variation in speech production is a common phenomenon. Speech characteristics can vary on an individual level (differences in the articulation patterns) or within social groups (dialectal and other variations).

2. Co-articulation, which is the production of two or more continuous speech sounds which affect one another due to consecutive production.

3. Recognition of continuous speech, dealing with lack of pauses, and context.

4. Analysis of tone, pressure, over aspiration and other prosodic features

5. Distance of the receiver from the source of articulation (distance capture)

5 How do we model speech?

A model of speech analysis needs to establish for speech to be analyzed. Speech analysis uses the basics of acoustics and wave physics in for primary analyses, and therefore basic terminology is borrowed from that domain. In this tutorial, some of the terms will simply be introduced on a need-to-use basis, not in too great a detail, as that is not necessary for an introductory course.

5.1 Terminology

1. **Phonemes**: A phoneme is the basic unit of sound that distinguish one word from another in a particular language.

2. **Voiced Activation Detection**: Algorithms used to differentiate the voiced phonemes from unvoiced phonemes is called voiced activation detection. Voiced phonemes produce stationary waveforms in a small time-slice, while unvoiced phonemes create non-stationary waves. Note that for all applications other than speech recognition, pauses and unvoiced sounds are ignored.

3. **Stationary waves**: Waves which have time-invariant properties such as amplitude, frequency and time period.

4. **Quasi-stationary waves**: Non-stationary waves which are composed of smaller, distinct stationary waves, resulting in the use of stationary waveform analysis algorithms from small time slices. Voice is a quasi stationary wave.

5. **Impulse**: Also known as glottal closure instance (GCI), instance of significant excitation (ISE) or epoch location, the impulse is that input to the vocal tract system which provides the times of high excitation of the vocal cords.

6. **Excitation Source**: The wave form which is given as an input to the vocal tract system.

7. **Pitch frequency**: The frequency of impulse.

8. **Speech sampling rate**: Voice is a purely analog signal. However, electronic devices represent and capture discreet value inputs. The speech sampling rate is that rate which the signal is sampled from the analog waves. The signal recorded is not an analog one, but a discrete one, which isn't digital, but the wave at small intervals.

9. **Encoding/quantization**: The number of bits used to represent the sample encoding is called the encoding. Higher encoding requires higher bandwidth for transfer.

10. **Bit rate**: Speech sampling rate x encoding = bit rate. For example: A clip which has been sampled at the rate of 8000 samples per second with an encoding of 8 bits for each sample will be transferred at 64 kbps (kilobits per second).

11. **Fourier Transformation**: In wave physics and the study of harmonics, the Fourier transformation is that transformation that relates the wave function of angular frequency and the wave function of time period. The function is written as: $H(\omega) = \mathcal{F}(h(t)) = \int_{\infty}^{-\infty} h(t)e^{-i\omega t}dt$.

12. **Spectral Energy Distribution**:The distribution of energy versus frequency is called the spectral energy distribution. The spectral energy distribution of sounds are an important sound feature.

13. **Window Size**: The size of the observation window, measured in milliseconds. A larger window size shows good frequency distribution, but bad time distribution. A small window size doesn't cover all the frequencies, but shows a good time distribution for the captured frequencies.

14. **Formant**: In the spectral energy distribution, the bands of concentrated energy (caused due to resonance of air in the vocal cavity) are called formants. The formants occur at particular resonant frequencies and are used to identify the sound produced.

5.2 Speech Processing Pipeline

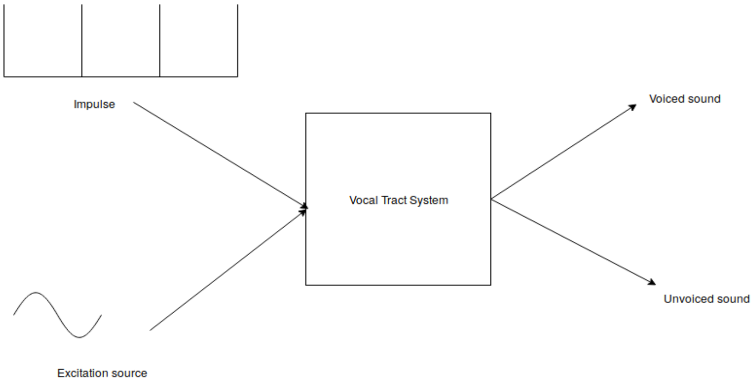

Figure 1: Basic Speech Processing Pipeline

So, the vocal tract system uses two inputs, a time-amplitude graph, called the excitation source, and a frequency-amplitude graph known as the impulse, or GCI. The latter can be derived from the former by applying a Fourier transformation. The vocal tract system then plots the spectral energy distribution of the sound form the graphs provided, and recognizes the formants. The formants are used to recognize the sound produced.

6 Basic Tools and Techniques

For the speech analysis part of this course, you will use a software called WaveSurfer. WaveSurfer is another free and open source software that is used to analyze the wave for sound visualization and manipulation, and is therefore used to analyze sound waves and speech patterns.

6.1 Lab Objectives

In this lab, we will:

1. Install the WaveSurfer software.
2. Record a speech signal for a sentence and display its waveform
3. Identify various regions such as voiced, unvoiced, plosives and silences

6.2 Installing WaveSurfer

Run the following command: `sudo apt-get install wavesurfer`.

6.3 Record a sentence

1. Run the following command: `wavesurfer`
2. On the wavesurfer platform, click the record button and record a short sentence.
3. From the pop-up, select the ”Waveform” representation.
4. Right-click and select ”Create Pane”. Choose the ”formant plot” and the ”spectrogram” options.

The three graphs now, from bottom to top, are waveform, formant plot and spectrogram. The axes on these graphs are time on the X axis and the frequency components are the Y axis. The spectrogram controls allow the choice of window length.

6.4 Identify various regions

Using the time domain waveform, the following characteristics can be used for identifying the regions of silence, voiced and unvoiced sounds.

1. Voiced sounds are those which are quasi-stationary, and they have high amplitude regions.
2. Unvoiced sounds are non-periodic, like noise.
3. Silence is very close to zero amplitude.

Using the formant plot (or the spectrogram) the characteristics used to identify the regions are:

1. Voiced: Regular formant structures, low frequency regions and pitch harmonics are used.

2. Unvoiced: Frequency at high energy, but no regular formant structure

3. Silence: No components