SPEECH PROCESSING

15 Sep 2020

SPEECH SOUNDS AND FEATURES :

The number of linguistically distinct speech sounds are variant to different people and are often subject to a matter of judgement.

However, phonetic symbols of American English can be generalized into some standard representations, which are as follows :-

- Vowels
- Diphthongs
- Semi Vowels
- Nasal Consonants
- Un voiced fricatives
- voiced fricatives
- Voiced and Unvoiced stops.

VOWELS :

Even though vowels are considered less important for the classification and representation of written text, most speech-recognition rely heavily on vowel recognition to achieve high performance.

for ex: This is an apple.

Th_s _s _n _ppl_

Avg reader can fill in vowels to decode written text.

__i_ i_ a_ a___e

The presence of vowels alone is not enough to decode the text.

Vowels are usually long in duration as compared to other types of sounds and are spectrally well-defined. Hence, they are usually easily and reliably recognized and therefore contribute a lot to our ability to recognize speech.

PRODUCTION OF VOWELS :

Vowels are produced by exciting the fixed vocal tract shape with quasi-periodic pulses of air caused by vibration of vocal cords.

The vowel sound produced is determined primarily by the following :-

- the position of tongue (primary point)
- the positions of jaw, lips, velum (slightly influence the resulting sound)

CHARACTERIZATION, CLASSIFICATION OF VOWELS :-

- Articulatory configuration : This refers to the tongue hump position (front, back, middle) tongue hump height (high, mid, low).
- Waveform plots : Front vowels show pronounced, high-frequency resonance, mid vowels show balance of energy over broad freq. range and back
- Spectogram plots : Front vowels show relatively high second, third formant frequency (resonance) whereas mid vowels show well-separated and balanced locations of formants. Back vowels show almost no energy beyond low frequency region with low first and second formant frequencies.

Vowels show mostly low-frequency spectral info

However, the concept of "typical" vowel sound is unreasonable because :-

- vowel pronunciation is variable among people with different regional accents

This leads to a wide range of variability in first and second formant frequencies for a given vowel sound.

In addition to this, different vowel sounds by different speakers may have overlapping formant frequencies.

Hence, measuring formant frequencies or spectral peaks is not enough to accurately classify vowel sounds.

Note - the primary frequency at which the vocal chords vibrate is also called the fundamental frequency (f_0)

Diphthongs : It is a gliding monosyllabic speech sound that starts at or near the articulatory position of one vowel and moves to or towards the position of another vowel.

There are 6 diphthongs in American English :-

- $a^{\jmath}$: Ex buy
- a^{w} : Ex down
- $e^{\jmath}$: Ex bait
- $o^{\jmath}$: Ex boy boy
- o : Ex boat
- ju : Ex you

PRODUCTION OF DIPHTHONG : Diphthongs are produced by varying the vocal tract smoothly between vowel configurations appropriate to the diphthong. This is highly prominent in $a^{\jmath}$, a^{w} etc but weaker for $e^{\jmath}$

because of closeness in the two vowel sounds.

Diphthong can also be thought of as ~~a~~ time-varying spectral characteristics, i.e. a plot of values of second formant versus first formant, as a func^n of time.

∴ diphthong can be characterized by a time-varying vocal tract area function that varies between two vowel configurations. (It can be seen in a spectogram : graphical plot with time on x axis and ~~amplitude~~ frequency of wave form on the y-axis)

SEMI-VOWELS : The sounds consisting of w, l, r, y are called semi-vowels because of their vowel-like nature. They are ~~gen~~ generally characterized by a gliding transition in vocal tract area function between two adjacent phonemes.

Semivowels are very difficult to characterize as the acoustic characteristics of w, l, r, y are strongly influenced by the content in which they occur.

Hence, semi-vowels are best described as vowel-like, transitional sounds, hence they are similar to both vowels and diphthongs.

NASAL CONSONANTS : The nasal consonants m, n, ŋ are produced with glottal excitation and the vocal tract is totally constricted at some point along the oral passageway.

PRODUCTION OF NASAL CONSONANTS :

- The velum is lowered so that air flows through the nasal tract, with sound being radiated at the nostrils.
- The oral cavity, though restricted, is still acoustically coupled to the pharynx.
- Mouth serves as resonant cavity that traps acoustic energy at natural frequencies.

For m : constriction is at lips
n : constriction is just behind the teeth
ŋ : constriction is just forward of the velum.

WAVEFORMS : The waveforms of m and n are very similar.

SPECTROGRAMS : Nasal consonants show a concentration of low-frequency energy with a mid-range of frequencies that contain no prominent peaks. This is because of the particular combination of resonances and anti-resonances that result from coupling of nasal and oral tracts.

UNVOICED FRICATIVES : The sounds f, θ, s, sh are produced by exciting the vocal tract by a steady air flow, which becomes turbulent in the region of constriction in the vocal tract.

The location of constriction determines the type of sound :-

f : constriction near the lips
θ : near the teeth
s : middle of oral tract
sh : back of the oral tract

Therefore, unvoiced fricatives consists of a source of noise at a constriction, which separates the vocal tract into two cavities.

- Sound is radiated from first cavity
- Back cavity (like in case of nasal consonants) traps energy and introduces anti resonances into vocal output

WAVEFORMS: Unvoiced fricatives have non-periodic waveforms