

# Objective Speech Analysis and Vowel Detection

Manuja Gokulan  
Department of Electronics  
and Telecommunication  
D.J. Sanghvi College of  
Engineering  
Mumbai, India

Maulik Gandhi  
Department of Electronics  
and Telecommunication  
D.J. Sanghvi College of  
Engineering  
Mumbai, India

Susmit Joshi  
Department of Electronics  
and Telecommunication  
D.J. Sanghvi College of  
Engineering  
Mumbai, India

Sunil Karamchandani,  
Ph.D  
Department of Electronics  
and Telecommunication  
D.J. Sanghvi College of  
Engineering  
Mumbai, India

## ABSTRACT

In considering application of digital signal processing techniques to speech communication problems, it is helpful to focus on three main topics: the representation of speech signal in digital form, the implementation of sophisticated processing techniques, and the classes of applications which rely heavily on digital processing. The objective analysis of the signal in terms of its various parameters is of primary concern. Pitch and Intensity are the most important parameters of a sound signal. Characteristics of a sound signal can be defined by observing the waveforms of pitch and intensity. A Spectrogram gives an efficient representation of the signal. Vowels and consonants can be separated by measuring the formant frequencies measured from the spectrogram.

## General Terms

Signal Processing, Characteristics of Sound, Pitch Measurement and Analysis, Intensity Measurement, Vowel Format Frequency Measurement

## Keywords

Intensity, spectrogram, pitch, speech, formant, vowels, frequency.

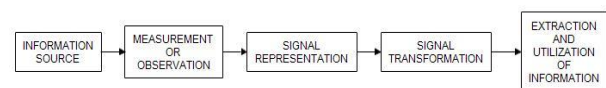
## 1. INTRODUCTION

The purpose of speech is communication. There are several ways of characterizing the communications potential of speech. One highly quantitative approach is in terms of information theory ideas as introduced by Shannon. According to information theory, speech can be represented in terms of its message content or information [1]. An alternative way of characterizing speech is in terms of the signal carrying the message information, i.e., the acoustic waveform.

In considering the process of speech communication, it is helpful to begin by thinking of a message represented in some abstract form in the brain of the speaker. Through the complex process of producing speech, the information in that message is ultimately converted into an acoustic signal. The message information can be thought of as being represented in a number of different ways in the process of speech production. The information that is communicated through speech is intrinsically of a discrete nature; i.e., it can be represented by a concatenation of elements from a finite set of symbols. The symbols from which every sound can be classified are called phonemes [5]. Each language has its own distinctive set of phonemes, typically numbering between 30 and 50.

## 2. INFORMATION MANIPULATION & PROCESSING

The general problem of information manipulation and processing is depicted in Figure 1.



**Fig 1: General view of information manipulation and processing**

In case of speech signals the human speaker is the information source. The measurement or observation is generally the acoustic waveform. Signal processing involves first obtaining a representation of the signal based on a given model and then the application of some higher level transformation in order to put the signal into a more convenient form [8]. The last step in the process is the extraction and utilization of message information. PRAAT is a software package that has been extensively used for analysis. Acoustic signals can be measured and observed and also can be processed using various transformation tools.

## 3. CONSIDERATIONS FOR ANALYSIS

Sound has got properties like pitch, intensity, etc. Pitch is related to frequency. It determines the nature of sound. Intensity is related to loudness. Therefore, measuring of sound signal can be achieved by measuring its pitch and intensity.

### 3.1 Pitch

It is the property of sound that varies with variation in the frequency of vibration [4]. Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale. Pitch may be quantified as a frequency. But pitch is not a purely objective physical property; it is a subjective psychoacoustical attribute of sound. Pitch is closely related to frequency, but the two are not equivalent. Frequency is an objective, scientific concept, whereas pitch is subjective. Sound waves themselves do not have a pitch, and their oscillations can be measured to obtain the frequency. Pitches are usually quantified as frequencies in cycles per second, or hertz, by comparing sounds with pure tones, which have periodic, sinusoidal waveforms.

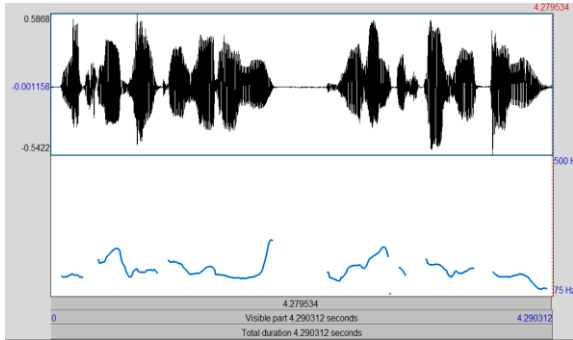
### 3.2 Intensity

Sound intensity or Acoustic intensity is defined as (1) the sound power per unit area [4].

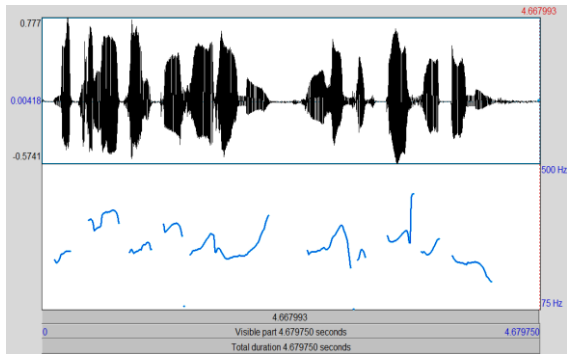
$$I = \frac{P}{4\pi R^2} \quad (1)$$

### 3.3 Pitch Measurement Results

Two same sentences read by two different people-one male and one female-were analyzed using PRAAT. Figure 2 and 3 show the variation in pitch [5] of a male and a female sound. Minimum noise interference from the environment was ensured.



**Fig 2: Male audio sample with its extracted pitch contour**



**Fig 3: Female audio sample with its extracted pitch contour**

As seen in Figure 2 and Figure 3, the pitch of female sound is higher than that of male sound. To show the difference objectively, readings were taken at different time instances and compiled in Table 1. Table 1 shows the Pitch of the two signals.

**Table 1. Pitch Observed for the audio samples**

Time Instance (s)	Pitch (Hz)	
	Male Voice	Female Voice
0.2	139.9	238.7
0.6	194.6	358.4
0.8	148.4	241.6
1.2	140.9	313.13
1.6	130.1	294.1
2.0	No Pulse	255.4
2.4	130	No Pulse
2.8	207.1	318.9
3.2	186.9	290.3

3.6	159.8	238.8
4.0	142.5	212

From the Table 1 it can be concluded that the pitch of a female voice is more than that of a male. This leads us to the fact that the female voice is shriller (sharp toned) than that of a male. A physical explanation of this difference is owing to biological variations. Men and women have different hormones. Testosterone in a man causes the vocal cords to thicken therefore vibrating at a deeper resonating sound. Women have less testosterone and more estrogen causing the vocal cords to be thinner and making a higher pitch sound.

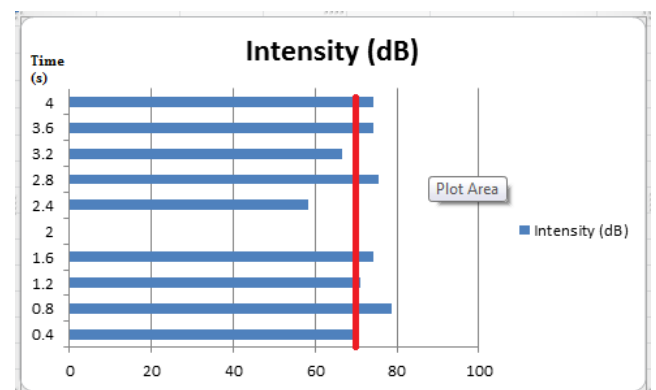
### 3.4 Intensity Measurement Results

Selecting only one sound signal, the intensity was measured. Intensity is related to the power of sound. Higher the power, higher is the intensity. Intensity is measured in dB(decibels). It is the measure of Loudness [5].

**Table 2. Observed Intensity for female audio sample**

Time instance (s)	Intensity (dB)
0.4	69.6
0.8	78.84
1.2	71.03
1.6	74.12
2.0	0
2.4	58.21
2.8	75.44
3.2	66.51
3.6	74.13
4.0	74.32

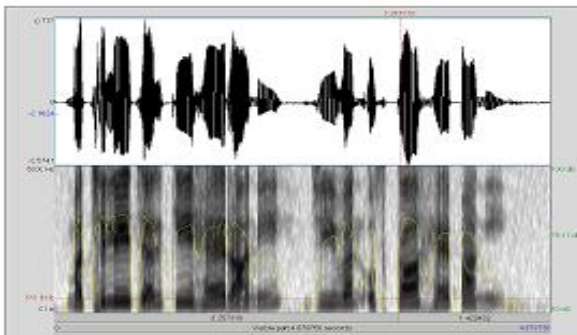
In Table 2, values of Intensity at different time instances were measured. It can be seen that the average sound Intensity is approximately 70. The graph in Figure 4 shows the average intensity.



**Fig 4: Intensity vs. Time**

Average normal human voice intensity is in the range 70-80 dB.

Figure 4 shows the female audio signal with spectrogram and intensity.



**Fig 5: Female audio signal with its spectrogram and intensity**

The yellow line traces intensity of the sound signal.

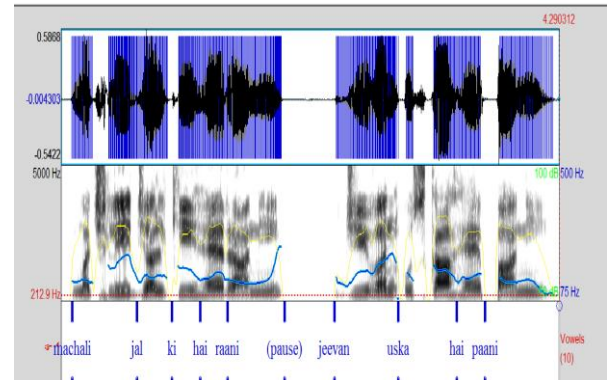
#### 4. VOWEL DETECTION

There are five vowels- *a*, *e*, *i*, *o*, and *u* -in the English language. Vowel is a speech sound made with the vocal tract open. Vowels are produced by exciting a fixed vocal tract with quasi-periodic pulses of air caused by vibration of the vocal chords. The way in which the cross-sectional area varies along the vocal tract determines the resonant frequencies of the tract(formants) and thus the sound that is produced. Formants are defined by Gunnar Fant [6] as - 'the spectral peaks of the sound spectrum of the voice'. It is often measured as an amplitude peak in the frequency spectrum of the sound, using a spectrogram (in the figure) or a spectrum analyzer, though in vowels spoken with a high fundamental frequency, as in a female or child voice, the frequency of the resonance may lie between the widely-spread harmonics and hence no peak is visible. In acoustics, it refers to a peak in the sound envelope and/or to a resonance in sound sources, notably musical instruments, as well as that of sound chambers [7].

By definition, the information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. In speech, these are the characteristic partials that identify vowels to the listener. Most of these formants are produced by tube and chamber resonance, but a few whistle tones derive from periodic collapse of Venturi effect low-pressure zones. The formant with the lowest frequency is called  $f_1$ , the second  $f_2$ , and the third  $f_3$ . Most often the two first formants,  $f_1$  and  $f_2$ , are enough to disambiguate the vowel. These two formants determine the quality of vowels in terms of the open/close and front/back dimensions (which have traditionally, though not entirely accurately, been associated with the position of the tongue). Thus the first formant  $f_1$  has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as [i] or [u]); and the second formant  $f_2$  has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as [u]). Vowels will almost always have four or more distinguishable formants; sometimes there are more than six. However, the first two formants are most important in determining vowel quality [7], and this is often displayed in terms of a plot of the first formant against the second formant, though this is not sufficient to capture some aspects of vowel quality, such as rounding. By listening to the sound

signal and identifying the vowel areas, the vowel formants were measured. The sound signal was 'machali jal ki hai









raani, jeevan uska hai paani machali.wav'. A representation of this sound signal is shown in Figure 5.









**Fig 6: Text Grid Annotated sample with distinct segmentation of words**

The measured formant frequencies from sound Spectrogram are shown in Table 3.

**Table 3. Vowel detection by Formant Measurement**

Vowel	Time instance (s)	Formant F1 (Hz)	Formant F2 (Hz)
 1.wav a	0.2	832	1418
 2.wav ee	0.5	2404	3558
 3.wav a	0.8	647	1752
 4.wav i	1	2572	3471
 5.wav a	1.3	672	2143
 6.wav aa	1.5	845	1690
 7.wav i	1.7	344	2671
 8.wav ee	2.6	2254	3373

 9.wav a	2.8	748	1926
 10.wav u	3	623	3325
 11.wav a	3.3	935	1551
 12.wav a	3.5	462	2274
 13.wav aa	3.85	935	1931
 14.wav i	4	310	2831

The universally accepted average vowel formant frequencies are given in Table 4[7].

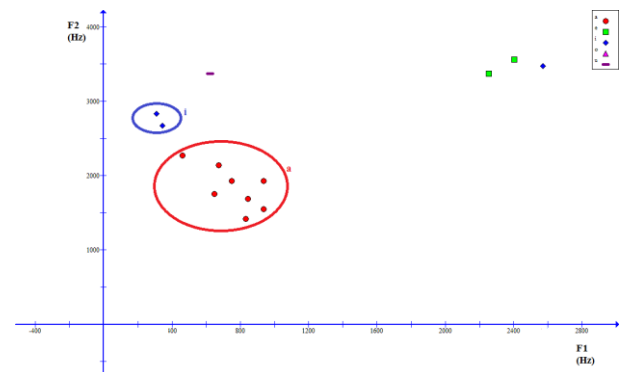
**Table 4. Average Vowel Formant Frequencies**

Vowel	Formant F1 (Hz)	Formant F2 (Hz)
i	240	2400
y	235	2100
e	390	2300
ɛ	370	1900
æ	610	1900
œ	585	1710
a	850	1610
æ	820	1530
ɑ	750	940
ɤ	460	1310
o	360	640
u	250	595

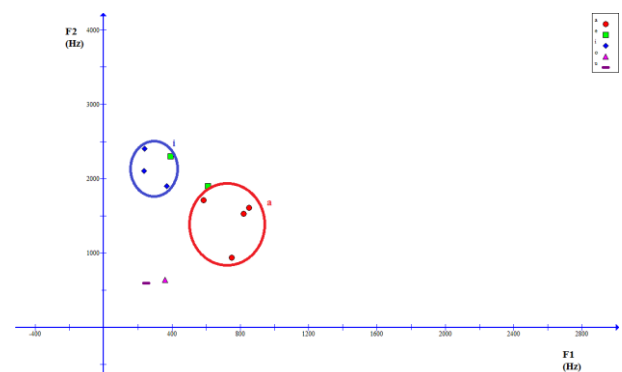
Comparing the readings in Table 3 and 4, it was observed that readings for the vowel sound ‘a’ and ‘i’ and some derived vowel sounds match to a certain extent with the average values. Other readings do not match. One reason may be the inherent differences in the vocal tracts of speakers. A great deal of variability is expected among speakers producing the same vowel.

A graphical representation of Table 3 and Table 4 is shown in Figure 6 and Figure 7. Encircled areas show the matching

between measured vowel formant frequencies and average vowel formant frequencies.



**Fig. 7: Measured Vowel Formant Frequencies (F1 vs. F2)**



**Fig. 8: Average Vowel Formant Frequencies (F1 vs. F2)**

Formant frequencies for vowels ‘a’ and ‘i’ match to a certain extent.

## 5. CONCLUSION AND FUTURE WORK

Pitch contour shapes clearly indicate that irrespective of the time of observation, the female voice possess a greater shrill than male as conformed by the readings. However, for intensity, it is purely dependent on the “force” with which the words are uttered. At different instances of time, there are fluctuations in the intensity for both male and female scales. The vowel detection mechanism conformed to the ideal frequency values as expected for all the words of the sentence except ‘e’ pronunciation in the word ‘chlee’. This deviation may be attributed to the external noise disturbances while recording.

Future work will target to automate the vowel detection system [3]. This feature of automated extraction can be put to use in real time service oriented sectors such as tele-booking. Moreover, a further enhancement could involve analysis of rapid continuous speech for content segmentation and system recognition such as study of prosody in native languages [2]. This task of evaluation becomes even more challenging as it demands more universal computational models.

## 6. ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Amit Deshmukh for giving us this opportunity to explore our areas of interest through this project and for guiding us throughout the project. His comments and suggestions have contributed greatly to an

extremely valuable learning experience. We would also like to thank and the EXTC faculty for their assistance and support.

## **7. REFERENCES**

- [1] C. E. Shannon, "A Mathematical Theory of Communication", Bell System Tech. J., vol.27, pp.623-656, October 1968.
- [2] L. R. Rabiner and R.W. Schafer, "Digital Techniques for Computer Voice Response: Implementations and Applications", Proc, IEEE, vol. 64, pp.416-433, April 1976.
- [3] D. R. Reddy, "Speech Recognition by Machine: A Review", Proc. IEEE, vol. 64, No. 4, pp. 501-531, April 1976.
- [4] J. L. Flanagan, Speech Analysis, Synthesis and Perception, 2<sup>nd</sup> Ed., Springer-Verlag, New York, 1972.
- [5] W. Koenig, H. K. Dunn, and L. Y. Lacy, "The Sound Spectrograph", J. Acoust. Soc. Am., Vol.17, pp. 19-49, July 1946.
- [6] R. Jakobson, C. G. M. Fant, and M. Halle, Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates, M.I.T. Press, Cambridge, Mass., 1963.
- [7] Datta, A.K., Mukherjee, B, "On the Role of Formants in Cognition of Vowels and Place of Articulation of Plosives", CMMR/FRSM 2011, LNCS 7172, pp.215-234, March 2012.
- [8] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Prentice-Hall Inc., Englewood Cliffs, N. J., 1975.