HUMANUS DISCOURSE Vol. 5. NO 1. 2024
ISSN 2787-0308 (ONLINE)

An Exploration of Machine and Digital/Computational Tools for Speech Analysis

Kofoworola A. Adedeji[1]
Department of English/Centre for Digital Humanities (CEDHUL)
University of Lagos, Akoka
Lagos, Nigeria

Abstract

A recent theoretical development in the field of Phonetics and Phonology is Stochastic phonology (Pierrehumbert 2022)[2]. Based on the relationship between speech and phonemes, this theoretical viewpoint strongly supports computational/digital analysis of speech rather than auditory transcriptions or even speech waveform and spectrograms majorly because of the inherent subjectivity and probabilistic nature of non-digital analysis. In support of this perspective, Harrington (2010)[3] adds that computational analysis is more exact because "transcription of speech can never get to the heart of how the vocal organs, acoustic signal, and hearing apparatus are used to transmit simultaneously many different kinds of information between a speaker and hearer." Two other advantages which digital tools offer is being able to provide more comprehensive information on the speaker and also the capacity to process huge amounts of data. Expansive information on the physical properties of speech sounds, sociophonetic details on the dialect, age group, educational level and socioeconomic class of the speaker can be more efficiently done using digital tools. Being able to abstract generalisations rather than idiosyncratic information is only possible through explorations of large data such as databases and this capacity gives digital tools an edge. This exploration is an in-depth examination of spectrographic analysis – the conventional instrumental analysis tool for acoustic analysis – as well a cursory look at two major digital tools for the examination of sound signals.

Keywords: digital analysis, computational tools, spectrograph, speech analysis

Introduction

Doing Experimental/Instrumental Phonetics

Experimental or instrumental phonetics makes it possible to describe sound production from the perspective of acoustic phonetics through the visual representation and objective description of sounds. Ordinarily, air is said to be made up of particles – known as air particles – which are stable, if not

[1]Email: kadedeji@unilag.edu.ng[1]
Phone: +234 802 335 3981
[2]Pierrehumbert, J.B. (2022) More than seventy years of probabilistic phonology. In B. Elan Dresher and Harry van der Hulst (Eds.) The Oxford History of Phonology. Oxford University Press.
[4]Harrington Jonathan (2010) Phonetic Analysis of Speech Corpora. Wiley-Blackwell.

ISSN 2787-0308 (ONLINE)

disturbed. In reality, there is regular movement of air resulting in disturbances in the surrounding air and such a movement can be caused by sounds, implying that sound generation is itself dependent on the slow or fast movement of air particles.

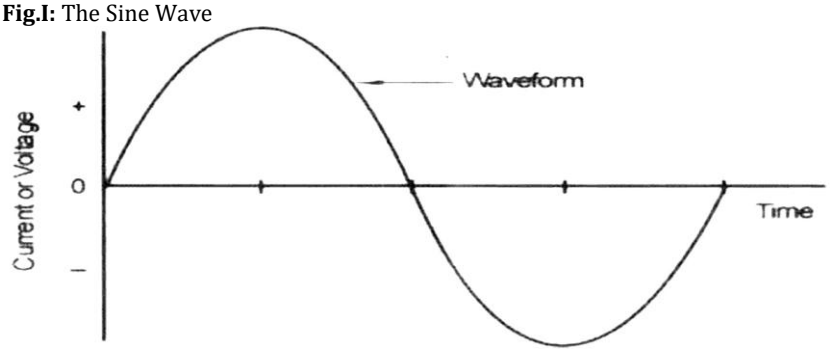
In theory and practice, for any object or particle to move, it must be displaced by some force which technically is energy. When energy or force is applied to a vibrating body, which may be air particle, the vibration of that body is transmitted to the particles of air surrounding it. The application of energy to a body is often explained by the simple experiment of a tuning fork used by music teachers, choir masters, musicians etc. When the fork is applied to an object, it is forced to show a blurry movement, often times forwards and backwards, generating sound This implies that as energy sets an air particle in motion, it hits and sets off another in motion etc. so that a pattern of motion travelling through space and time emerges until it reaches the ears of the hearer. Thus, a tuning fork right-hand prong moves from its position of rest, in an upright position, with maximum speed that decreases as it gets to the position of its maximum displacement and then gets back (a reverse journey) leftwards till it gets to its maximum leftwards position.

The principle of the tuning fork shows that air particle movement is forward and backward. As air moves, the air alongside it generates compression (when the air particles are crowded together) or rarefaction (when the air particles are further apart) thus showing that when compressed pressure is high and when rarefied, pressure is low. This is further explained by Hayward (2000:21)[5] that, "sound waves are difficult to visualize because the particles of air are moving back and forth in the same direction as the wave travels." This shows that sound description can be better understood although through the description of sound waves.

Sound Waves

The curve describing the motion of air in sound production is known as sinusoid or the sine curve. A diagrammatic vision of the wave form is shown in the two figures below which are reproduced from Hayward (2000) and Cox & Fletcher (2017)[6].

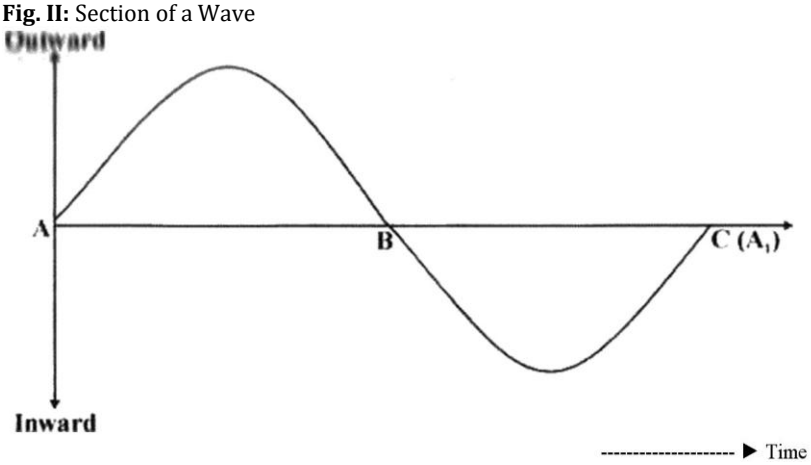
Fig.I: The Sine Wave



Source: Hayward (2000)

ISSN 2787-0308 (ONLINE)

Fig. II: Section of a Wave



Source: Cox & Fletcher (2017)

This figure shows movement from A to B and later to C. Since this in reality is a forward and backward movement, the movement from B to C as a backward movement is back to A; showing that C, is in reality A]. The movement from A to B and then back to C (A]) forms a *complete cycle.*

The same movement is shown in more technical details below as spatial variation in air pressure.

Fig. III: The wave and Air Pressure

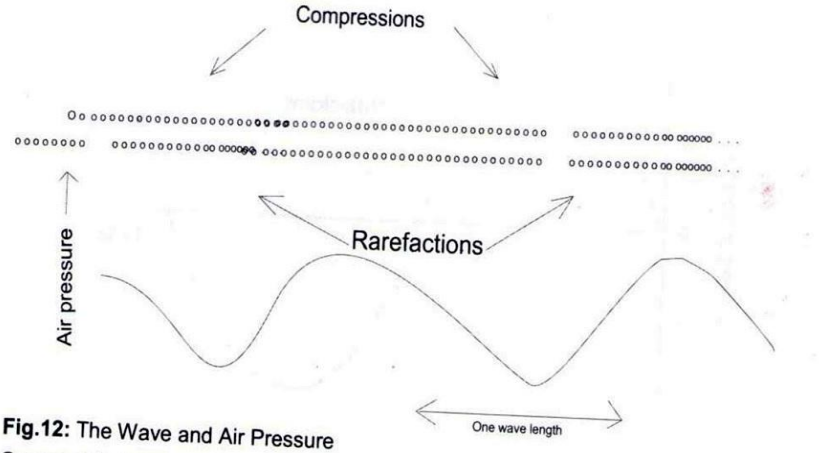


Fig.12: The Wave and Air Pressure
Source: Hayward (2000:25)

Source: Hayward (2000: 25)

The graphs show that pressure waveforms repeat themselves over and over to generate cycles. It is therefore important to understand the features of cycles or describe them scientifically or objectively.

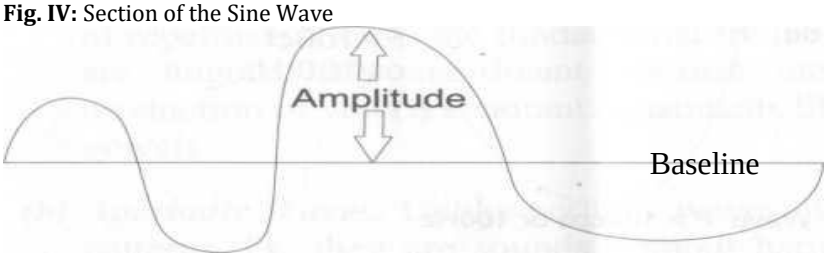
ISSN 2787-0308 (ONLINE)

In the attempt to understand the characteristic features of sound waves, there is the need to measure its distance, the time used in traversing a distance, the number of cycles in a sine wave etc. One of the commonest forms of measurement is frequency. Frequency is defined as the number of cycles produced or generated by a vibrating body in one unit of time or one second. Thus frequency (F) x T (time) = 1 (sec) or F = 1 /T while T = 1/F. Frequency is measured in cycles per second or c p s which is technically referred to as Hertz or Hz.

Closely related to frequency is duration or period which is time (T). Duration refers to the span of time during which sound is sustained. Duration is usually measured in the smallest possible unit of time which is milliseconds (msec or ms). When frequency and duration or time are crossed/paired, the relationship that shows is captured in the statement: *the shorter the period or T, the higher the frequency.*

Another form of measurement is known as amplitude, which refers to the maximum variation in pressure from normal position. Put in another way, it is the maximum height which a sine wave can achieve above a baseline (irrespective of frequency) or the maximum distance which a vibrating body moves as shown below.

Fig. IV: Section of the Sine Wave



Source: Hayward (2000)

Finally, there is the notion of wavelength which is the sine wave in space (not time) as it repeats itself over equal distances. Since sound is said to travel at constant velocity (c), the relationship between velocity and wavelength (h) is captured by the formula:
c = F x D N or F = C/N or N = C/F

It is important to understand basic concepts like loudness, pitch and quality and the relationship (if any) among them. Recall that amplitude as a maximum variation or distance from a baseline is related to loudness since a large movement of the source of sound produces loud sound. This implies that increase in amplitude makes sound to be very loud while decrease in amplitude indicates less loudness or the larger the amplitude, the louder the sound becomes (all other things being equal).

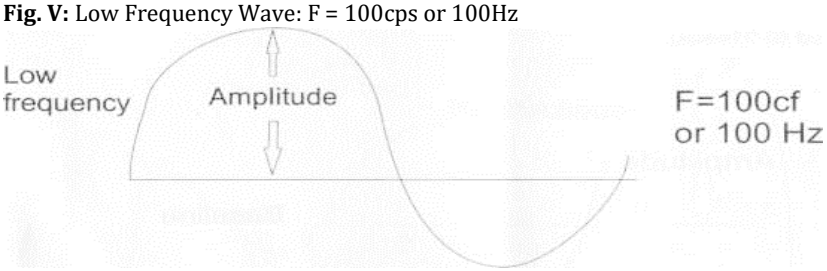
While amplitude is related to loudness of a sound, frequency is related to pitch. The faster the vocal folds vibrate in the supraglottal cavity, the higher the pitch. This implies that a higher pitch makes a greater number of vibrations per

ISSN 2787-0308 (ONLINE)

second than a lower pitch, i.e. the higher the pitch, the higher the frequency or high pitch, high F and low pitch, low F.

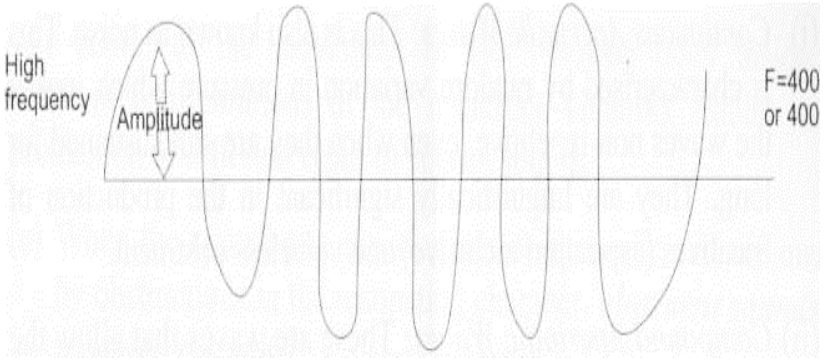
From the foregoing, it becomes clear that two sounds may have equal amplitude, while they have different pitch which depends on frequency as shown below.

Fig. V: Low Frequency Wave: F = 100cps or 100Hz



Source: Hayward (2000)

Fig. VI: High Frequency Wave: F = 400cps or 00 Hz



Source: Hayward (2000)

In practical terms, while loudness as a prosodic feature can be distinguished in its meaning as it functions as a component of stress, pitch can similarly be distinguished meaning at a suprasegmental level as a component of stress and also as a shaper of intonation, although it cannot change the function of an individual English sound. Where a change of function is signalled by pitch, what results is tone languages, i.e. pitch changes function in tone languages but not in intonation languages like English.

Finally, *quality* refers to the complexity in the structure of wave forms as a very complex wave form is said to differ in quality (wave structure) than a complex wave or a simple wave.

Types of Waves

As explained on pages 34 to 35, quality implies that there are different types of sound waves. Two types are often mentioned and they are *periodic* and *aperiodic*.

ISSN 2787-0308 (ONLINE)

a. Periodic Waves*:* These are cyclic in nature and with repetitive patterns as in the example of a simple sine wave. The frequency of repetition (FO) is the fundamental frequency.

b. Periodic waves are linguistically significant because they are used in the production of voiced sonorant consonants like /m, 1/ as well as in vowels.

c. *Aperiodic Waves*: Unlike periodic waves, they are not repetitive patterns, i.e. they are sounds without harmonic basis or those where the component frequencies are not related to each other in the series. Two sub-types are usually identified.

    i. *Continuous Aperiodic Waves:* This is also known as noise. This is characterised by random variation in pressure which makes the waves non-repetitive, even when they are still sustained' for long. They are linguistically significant in the production of fricatives (aspiration inclusive) and voiceless obstruent.
    ii. *Compound Aperiodic Waves:* These are waves that allow the super imposition of noise on (i) above or periodic waves. Examples include voiced fricatives and voiced obstruent.

The third type of wave that is not often referred to is known as *the transient wave*. Transient waves are of very brief duration and if they are periodic, they die off quickly. Waves can also be described as simple or complex. Most simple waves are periodic in nature while complex ones are oftentimes aperiodic.

Resonance and Secondary Vibration

When a body is set in motion by the vibrations of another body, the action-effect is known as resonance i.e. a body is said to resonate another. In a sense, therefore, resonance – oftentimes the resounding of an earlier/original sound – may be in a note played or in a vibration) is a form of secondary vibration. Resonance is therefore a secondary vibration of air particles.

Secondary vibration is generated by a resonating chamber, which implies that resonance depends on air particles and a cavity or chamber known as the resonating chamber or resonator. The box of a guitar is an example of a resonating chamber because when force is applied to the strings on it, the box as a chamber helps to resound the original sound - similar to echo effect when shouting in a deep forest. The vocal tract, where complex shapes of vibrating air particles pass through in complex waveforms, is also a good example of a resonating chamber which is used in the production of human speech sounds. The complexity of sizes or shapes of the chamber results in the different vibrations, frequencies, intensities, pitch etc. at a point in time, i.e. the way sound vibrates depends on the shape and size of the chamber.

The action of the movement of the vibrating body inside the different configurations of the resonance chamber results in the production of different types of speech sounds.

ISSN 2787-0308 (ONLINE)

a. When the waveforms are irregular or aperiodic and are marred by obstructions in the resonating chamber, obstruent sounds are produced e.g. stops.

b. When the vibrating air, as periodic wave, moves within the chamber without any form of obstruction (a form of open approximation) resonant sounds are produced, i.e. vowels.

c. Sometimes, the configuration of the resonance chamber is almost like that of a stop with its aperiodic waves, but at the same time air escapes through a narrow aperture. When this happens, sonorant sounds are produced. In this sense, while obstruent and sonorants are [+ consonantal], resonant are [- consonantal].

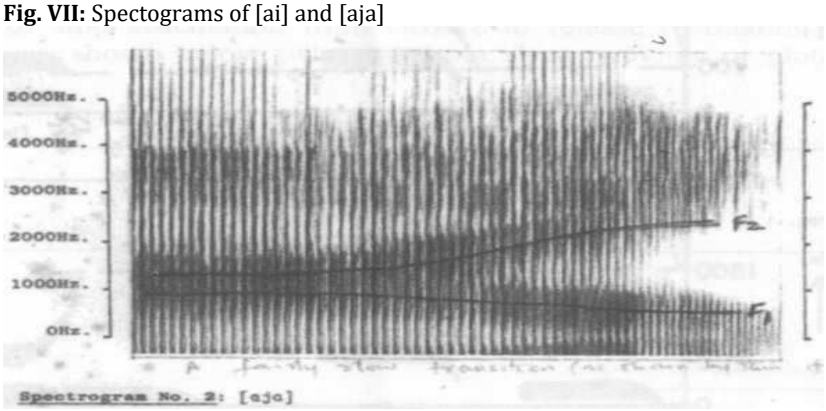
Spectrographic Analysis

Perhaps the most basic type of phonetic analysis is that which is done by the *spectrograph*. It measures the frequency, time or duration, forms of transitions between speech sounds. With the aid of the machine, a spectrogram is produced and printed on a paper or displayed on a computer screen. In a spectrogram, the x axis or left side scale is measured in Kilohertz (KHz) while in the *y* axis or bottom side, it is measured in milliseconds (ms). A spectrogram can show formant frequencies indicated by dark marks and when there is lack of acoustic activity it would, show there are no such marks. When there are regular vertical lines or striations, such indicate periodic vibrations of the vocal cords while irregular striations indicate aperiodic vibrations. For example, vowels (resonants) and sonorants are associated with periodic vibrations while fricatives and stops (obstruents) are associated with aperiodic vibrations, although the picture could be more complex when voiced and voiceless sounds are compared. Sometimes, waveforms also help to show pulses corresponding to regular or irregular vibrations. For example, voiceless sounds can be shown by straight lines while wiggly lines are indicative of voiced sounds.

In analysing voiced speech sounds, it is important to note the *fundamental frequency (F0*) at which the vocal cords are vibrating and it is also possible to identify certain frequencies above FO that correspond to resonances of the vocal tract or harmonics indicative of multiples of fundamental frequency. Thus, we are confronted with resonances above FO which are known as formant frequencies or formants. Phoneticians are interested in the emphasised frequencies above FO which is the fundamental frequency and this leads to the first (Fl), second (F2) and the third formant (F3). Fl to F3 display the frequency patterns of sounds associated with them. Sometimes F4 is said to be indicative of voice quality.

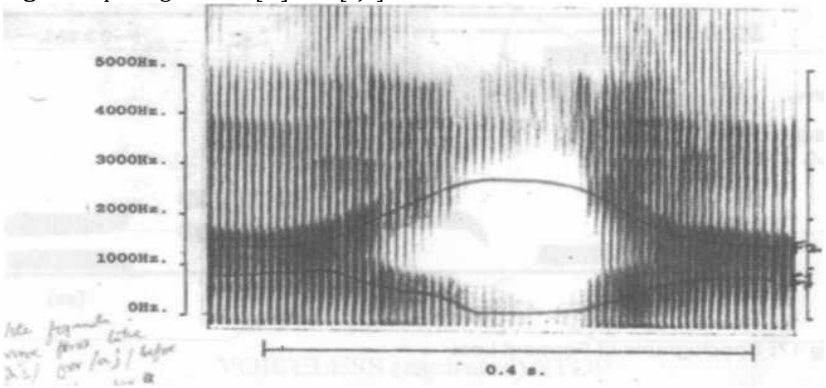
ISSN 2787-0308 (ONLINE)

Fig. VII: Spectograms of [ai] and [aja]



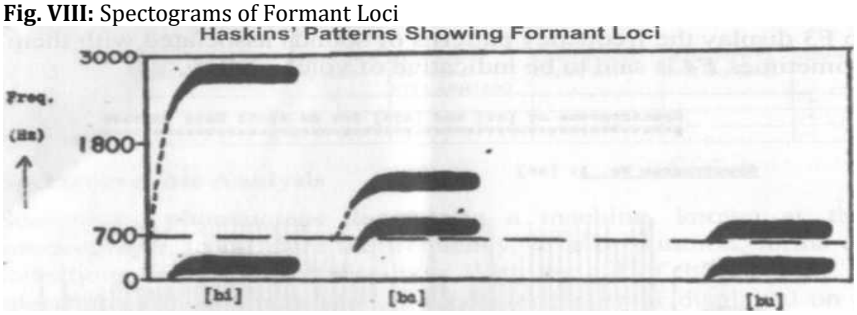
Source: Hayward (2000)

Fig. VII: Spectograms of [ai] and [aja]

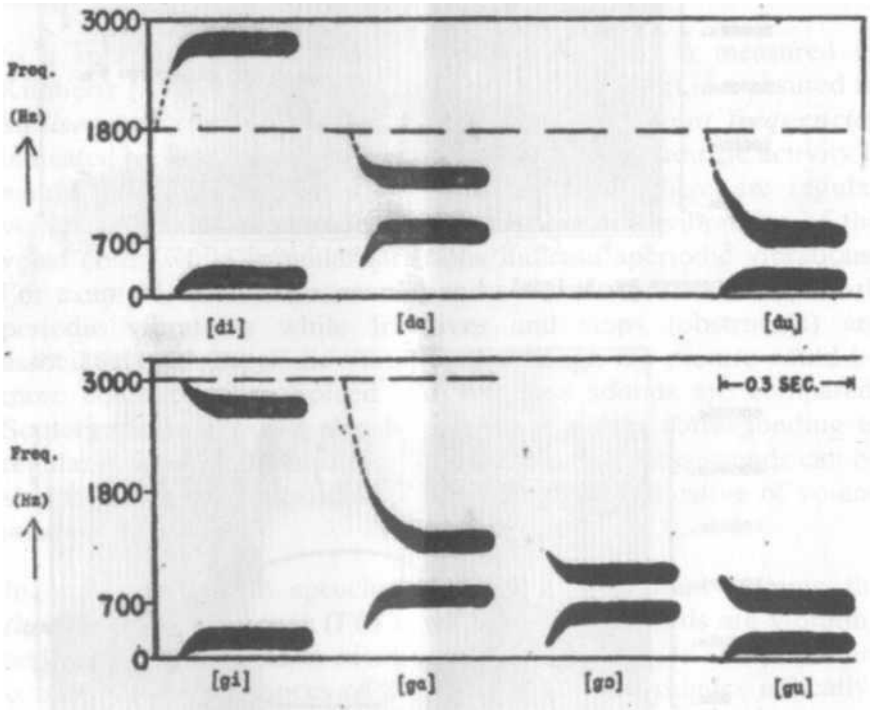


Source: Hayward (2000)

Fig. VIII: Spectograms of Formant Loci



ISSN 2787-0308 (ONLINE)



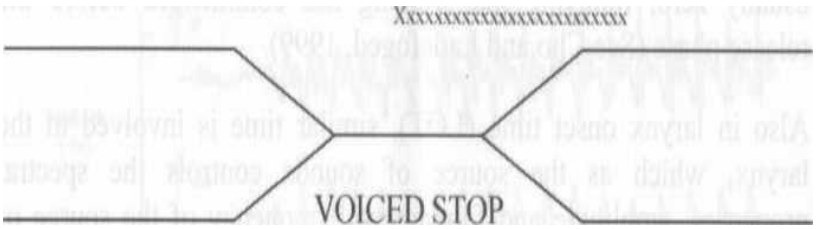
Source: Hayward (2000)

The non-presence of energy bust on the spectrogram also helps to understand pre-voicing or what is known as Larynx Onset Time(LOT). It can however be better seen in an oscillograph, displayed by the oscillogram machine, which may display waveforms and spectral features. Some of such features may include transition, pre-voicing and aspiration.

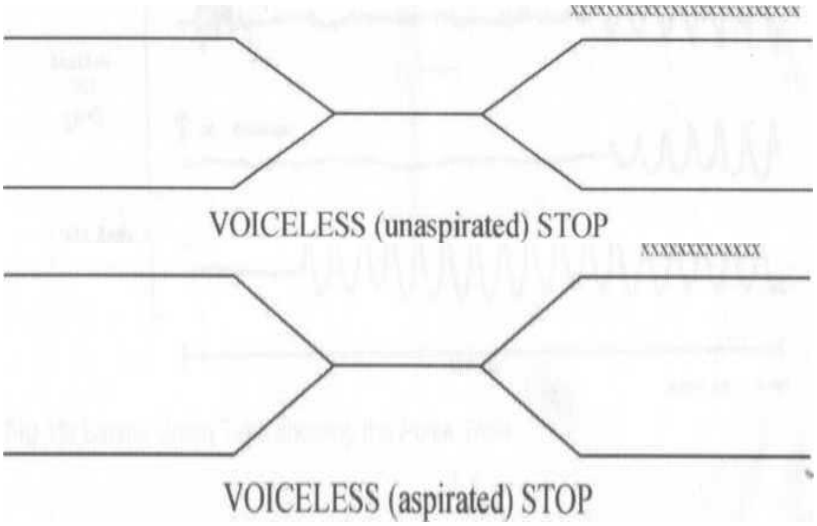
Voice Onset Timing

Another form of onset timing is Voice Onset Timing (VOT). It comes handy in the analysis of aspiration in stops. Ordinarily, the phases of stop articulation from closure to release m relation to voicing are shown below (where xxxxxx shows voicing or glottal excitation).

Fig. IX: VOT



ISSN 2787-0308 (ONLINE)



CLOSURE HOLD/STOPPAGE RELEASE PHASES
Source: Hayward (2000)

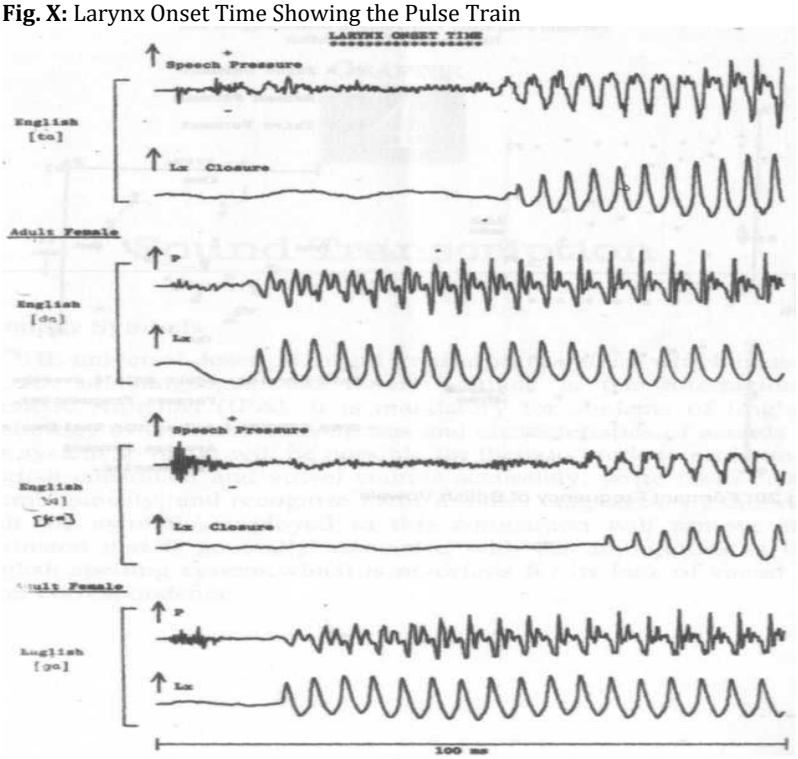
VOT which is also known as periodicity is the time or duration spent between the release phase of a stop and the commencement or onset of voicing or the glottal elicitation and vibration of vocal folds. In a pictogram, it is shown as the gap between release and voicing (as shown by the formant frequency). In unaspirated stops, VOT is always near zero while in aspirated stops, it could be up to 200ms. For example, [kh] is said to have 160ms as VOT. This implies that the longer the VOT, the longer the aspiration. Note however that because voiced stops are not aspirated, their VOT is usually zero, implying that voicing has commenced before the release phase (Ladefoged 2014)[8].

Also, in Larynx Onset Time (LOT), similar time is involved in the larynx, which as the source of sounds controls the spectral properties, amplitude and fundamental frequency of the source of voiced sounds, before sound production or glottal opening.

[8]Ladefoged Peter, Keith Johnson (2014) A Course in Phonetics. Cengage Learning. Patrick Svennson (2010) The Landscape of Digital Humanities. "Digital Humanities Quarterly, Vol 4:1

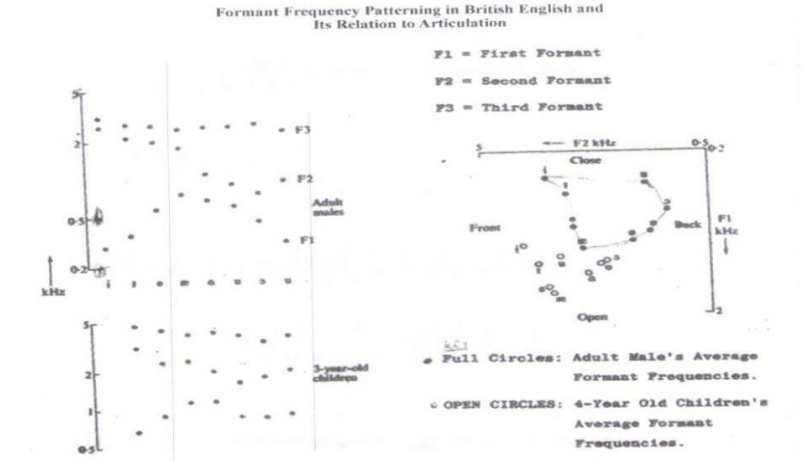
ISSN 2787-0308 (ONLINE)

Fig. X: Larynx Onset Time Showing the Pulse Train



Source: Hayward (2000)

Fig. XI: Formant Frequency of British Vowels



Source: Hayward (2000)

ISSN 2787-0308 (ONLINE)

Spectrographic analysis became less popular with the increase in access to desktop and laptop computers circa 2000. Many innovations in the area of computer software development aided the widespread use of many digital solutions. Emu-R and Forced Alignment Tool (FAT) are explored below:

Emu-R

Developed by the Institute of Phonetics and Speech Processing at Ludwig-Maximilians University of Munich in Germany, Emu-R has two components: Emu and R which run on Linux, Mac OS-X, and Windows platforms which can be downloaded at: http://emu.sourceforge.net and http://www.cran.r-project.org respectively. It can be used with existing corpora or with the researcher's designed corpus for speech signal processing such as:

a. Calculating, displaying and correcting formants

b. Querying annotation structures

c. Comprehensive analysis i.e. voice onset time, intragestural analysis &

d. differencing and velocity

e. Formant and formant transition analysis

f. Electropalatography

g. Spectral Analysis

h. Probability and theorem analysis

Emu-R was first developed as EMU (Electromagnetic Unit) in 1995. It now has six versions with the Emu-R 6.0 released in 2020.

Forced Alignment Tool (FAT)

FAT is a tool that automatically converts text into phonetic transcription thus 'forcing' the audio files and the transcripts into alignment. It creates accessible, annotated and time-stamped speech databases for both segmental and suprasegmental analysis involving consonants and vowels as well as syllable and phrase level analysis. It removes the tediousness and time-wasting that is associated with manual annotation of sound files. It also reduces the manpower involvement with multiple annonators and produces human-error free annotations (Wu et al.) Other features areCreates customized pronouncing dictionaries and pronunciation models:

a. Supports multiple languages with various rhythmic types

b. Data outputs can be in several formats: Praat, TextGrid etc.

c. Audio and text files are time-aligned

ISSN 2787-0308 (ONLINE)

Conclusion

The transition from the use of the spectrograph and similar machines started at the late 1990s and early 2000s, thus coinciding with the replacement of the term "humanities computing" with "digital humanities." Described by Svensson (2016)[9] as: "a meeting place for the humanities and the digital", it is within this DH space that tools such as Emu-R and FAT are built to provide technological solutions to sound analysis.

[9]Svensson Patrik (2016). Big digital humanities: imagining a meeting place for humanities and the digital. University of Michigan press.