Review Article

Speech acoustics: How much science?

Manjul Tiwari

Department of Oral Pathology and Microbiology, School of Dental Sciences, Sharda University, Greater Noida, Uttar Pradesh, India

Address for correspondence:
Dr. Manjul Tiwari, Department of Oral Pathology and Microbiology, D-97, Anupam Apartments, B/13, Vasundhara Enclave, New Delhi –110 096, India. E-mail: manjultiw@gmail.com

Abstract

Human vocalizations are sounds made exclusively by a human vocal tract. Among other vocalizations, for example, laughs or screams, speech is the most important. Speech is the primary medium of that supremely human symbolic communication system called language. One of the functions of a voice, perhaps the main one, is to realize language, by conveying some of the speaker's thoughts in linguistic form. Speech is language made audible. Moreover, when phoneticians compare and describe voices, they usually do so with respect to linguistic units, especially speech sounds, like vowels or consonants. It is therefore necessary to understand the structure as well as nature of speech sounds and how they are described. In order to understand and evaluate the speech, it is important to have at least a basic understanding of science of speech acoustics: how the acoustics of speech are produced, how they are described, and how differences, both between speakers and within speakers, arise in an acoustic output. One of the aims of this article is try to facilitate this understanding.

Key words: Speech, speech acoustic, speech science, speaker, speech sound

INTRODUCTION

Speech sounds

Speech sounds, just like any other sound, are rapid fluctuations in air pressure. Speech sounds are generated when air is made to move by the vocal organs.[1] While speaking, acoustic energy is radiated from the vocal tract. This acoustic disturbance, consisting of pressure fluctuations, causes the listener's eardrum to move rapidly in and out – in when the pressure is positive, out when negative.[1,2] Thus acoustic energy is transformed into mechanical energy at the eardrum. This mechanical energy, and the information it contains, go through several more transformations before arriving as patterns of neural energy at the listener's brain. The processing of the information in the listener's brain results in the percept of sound.[1,3]

The acoustic properties of the radiated speech wave are of great importance since they constitute the basis for both the phonetician's acoustic analysis and auditory transcription.[1,4,5]

Access this article online	
Quick Response Code:	Website: www.jnsbm.org
	DOI: 10.4103/0976-9668.95942

Speech waves

The speech wave is distributed, at any given instant, as a sound pressure wave in the air around the speaker, and can be looked at as pressure varying as a function of distance from the speaker.[5] Think of the distribution of the height of waves in the sea.[1]

At any particular instant one could describe the height of the wave as a function of the distance from, say, someone standing in the water. In the same way, it is possible to specify the pressure of the speech wave at a given instant as a function of distance, having such-and-such pressure at such-and-such point in space (including within the speaker's vocal tract). The mathematical equation that describes pressure fluctuations in this way is called a distance–pressure function.[1,3,5]

In speech acoustics, however, it is more common to consider the air pressure in a speech wave as varying not as a function of distance but as a function of time. This is equivalent to saying that at such-and-such point in space, the air pressure varies in such-and-such way over time.[4]

In speech, one such point in space is commonly the

microphone that transduces the acoustic signal. The mathematical equation that describes pressure fluctuations in this way is often referred to as a time–pressure function,[2] and variations in air pressure shown as a function of time are called time–pressure waves.[1,6]

Frequency

It can be seen that the magnified speech wave consists of rapid variations in air pressure as a function of time. Variations represent positive increases in pressure relative to atmospheric pressure and also negative.[7,8] The variations are periodic – they repeat and are obviously complex, in that the air pressure can be seen to be varying simultaneously at several different frequencies. These frequencies can be roughly estimated visually as follows.[1]

Fundamental frequency

First, the magnified portion of the wave apparently consists of a little less than three repetitions of a complex pattern. (One repetition is taken to occur between recurring events in the wave, for example, the peak values.) The first two repetitions can be seen to occur within 0.013 s. The elapsed time between recurring events is called the wave period, and so the average period of this wave is 0.013 s divided by 2, or 0.0065 s. The most common unit of acoustic duration is milliseconds (thousandths of a second, abbreviated msec or ms).[1,7,9] So the average period of this wave is 6.5 ms. Duration is also found quantified in centiseconds (hundredths of a second, abbreviated csec or cs), in which case the wave period is 0.65 cs.[1]

Frequency is expressed as number of times per second, or hertz. So if the wave has a period of 0.0065 s, in 1 second it will repeat 1/0.0065 = 153.8 times, or 154 Hz (rounded off to the nearest Hz). The frequency in Hz is thus the reciprocal of the period (1 divided by the period) in seconds (1/0.0065 s = 154 Hz).[1,6,10]

The rate of repetition of the complex wave is called its fundamental frequency (abbreviated *F*0, which is pronounced "eff-oh" or "eff sub-zero") and this wave therefore has an *F*0 of 154 Hz. Fundamental frequency is an extremely important measure in acoustic phonetics in general.[1,3,6,8]

From the point of view of speech production, the *F*0 corresponds to the rate at which the vocal folds vibrate. From the point of view of speech perception, had been in the vicinity of the waveform, an eardrum would have been going in and out 154 times per second as a response to these pressure fluctuations.[1,11]

Higher frequency components

Within each single repetition of the basic pattern, more fluctuations of pressure can be one pattern of fluctuation that recurs twice in each period. Between two peak values, for example, the pressure first goes down, then up, then down, and then up. These fluctuations in pressure are therefore occurring at about twice the *F*0 (i.e., 2 × 154 Hz = ca. 300 Hz).[9]

Finally, there is a second pattern of pressure fluctuations that recurs about 15–16 times per period. These can most easily been seen in the second repetition. Starting from the beginning of the second main peak of the waveform, for example, the pressure goes down a little, then up a little, then down a little, and then up a little about 15–16 times before arriving at the third main peak.[6,7] Since there are between 15 and 16 of these little fluctuations per period, this means a frequency of between ca. 2460 Hz (16 × 154 Hz) and ca. 2300 Hz (15 × 154 Hz).[1,7,8,12]

Fourier analysis

One way of looking at the speech wave is as complex fluctuations in air pressure as a function of time. Another is as a spectrum, which shows exactly what frequencies are present with what amplitudes.[4,9] But how does the spectrum relate to the time-domain speech wave? Fourier's theorem shows that any complex wave can be decomposed into, or represented as, a set of sine waves, also called sinusoids, each with its own frequency and amplitude.[13]

Let us assume the summation graphically. To do this, all waves – the complex wave and its sinusoidal components are – put together. The complex wave is made up of the three sinusoids, such that the amplitude of the complex wave at any point in time.[1,11,13]

The calculation of the pressure in the complex wave at t = 1.25 ms can be illustrated as follows. In each of the sinusoids, the value for the pressure at any given time is given by the following formula:

$P = A \times \sin(\omega t)$.

In this formula, P is the pressure to be determined, A is the maximum positive amplitude of the sinusoid wave, and t is the time, in seconds, at which the pressure is to be determined. ω, called the angular frequency, is 360 times the frequency of the wave, and is expressed as degrees per second. Thus if the frequency of the wave is 2 Hz, the angular frequency ω is 360 × 2 = 720 degrees per second. The number 360 and the use of degrees as a unit comes from the fact that the shape of a sinusoidal wave is related to circular motion, and 360° represents one revolution in a circle and one full cycle of a wave.[1,2,7,9,12] Sine (written as "sin" in formulae) is the name for one of the basic trigonometric functions relating the size of one angle in a right-angled triangle to the lengths of two of its sides, and

the sine of an angle can be found on a calculator or looked up in a table: for example, sin(90 degrees) is 1.[1]

It can be seen that the pressure is zero when $t = 0$ ms, when $t = 5$ ms, and so on; the maximum positive pressure is when $t = 2.5$ ms, $t = 12.5$ ms, etc., and the maximum negative pressure (−1000) is at $t = 7.5$ ms, $t = 17.5$ ms, and so on. The formula, pressure = amplitude × sin(ωt), can be used to work out the pressure in this wave when time is 1.25 ms, as follows. The frequency of this wave is 100 Hz, that is, it repeats 100 times per second, and therefore ω will be 100 × 360 degrees/s = 36,000 degrees/s. The maximum amplitude of the wave is 1000, and so when $t = 1.25$ ms, the value of the pressure will be the maximum amplitude (1000) times the sine of the product of the angular frequency ω and the time (36,000 × 0.00125). This product is 45, and the sine of 45 (degrees) is 0.707, and so the pressure at time (t) = 1.25 ms is 0.707 × 1000 = 707.[1,5,8,9]

The spectrum

The above section of the article showed how a complex wave can be represented as the sum of a set of sinusoids, each with its own amplitude and frequency. The concept of spectrum can now be introduced. The spectrum is plotted on a frequency axis and an amplitude axis. Each sinusoid is represented as a vertical line at its frequency with the length of the line corresponding to its amplitude. This spectrum thus shows the wave to consist of energy at three frequencies: 100 Hz, 200 Hz, and 1.6 kHz. It also shows the amount of energy present at each frequency: 1000 amplitude units at 100 Hz and 200 Hz, and 100 amplitude units at 1.6 kHz.[1,15]

The Fourier process works both ways: it can be understood in terms of both analysis and synthesis. Unlike the time-domain waveform, the spectrum no longer contains information about the variation of the wave over time.[1,15-17]

Harmonic spectrum

There are many different types of spectral representations. One type is called a fast Fourier transform or FFT. This is actually a spectral representation of the three periods of the [i] vowel and the axes are the same.[1,10-15]

The principle is the same for the spectral profile for the [i] and idealized line spectrum. The local spikes of energy correspond to the sinusoidal components, each with a given frequency and amplitude. Thus the sinusoidal component with the lowest frequency can be seen to have an amplitude of about 70 dB.[1,3,7]

These sinusoidal components are called harmonics, and the one with the lowest frequency is the fundamental frequency (the rate of repetition of the complex wave, and the main correlate of the pitch).[1,3,9] The fundamental frequency is the first harmonic (H1), the next higher in frequency is the second harmonic (H2), the next higher the third (H3), and so on. The harmonics occur at whole number multiples of the fundamental. That is, assuming $F0$ is 150 Hz, H2 will be at 300 Hz and H3 at 450 Hz, and so on.[1,18]

Smoothed spectrum

In addition to the fine harmonic structure in the spectrum, grosser structure can be detected. This can be best appreciated if the jagged structure of the harmonics is smoothed. The smoothing has been carried out by a rather complicated, but well-established signal processing technique called linear prediction (or LP) analysis.[1]

The smoothed spectral envelope from the LP analysis shows five major peaks; these are marked P1–P5. One peak (P1) is low in frequency, at about the frequency of the second harmonic, and the rest are above 2000 Hz.[1,15] The frequencies of the lowest three major peaks are the primary correlates of vowel quality.[1,10,11,13,15]

The acoustic theory of speech production

What gives rise to the radiated time–pressure variations of the speech wave? Where do its properties of fundamental frequency, harmonics, and spectral envelope crucial to the signaling of pitch and vowel quality come from? The theory that explains the radiated acoustics in terms of the vocal mechanism that produces them is called the acoustic theory of speech production, or source–filter theory.[1,19] It was developed by the Swedish speech scientist Gunnar Fant, and the first full account was given in his 1960 book *Acoustic Theory of Speech Production*. Part of the book tests how the theory works by first predicting, from the source–filter theory,[,1,2,4,19] the acoustic output of a Russian speaker using estimates of the size and shape of his supralaryngeal vocal tract derived from X-rays, and then comparing the predicted output with the speaker's actual output.[1,20]

It is worth pointing out that the source–filter theory is different from other theories in linguistics. Linguistics, as a hermeneutic science, that is a science of interpreting human behavior, is full of competing accounts of syntactic, semantic, and phonological phenomena (phonemics is one!). The source–filter theory, on the other hand, does not have any competitors and has not as yet been falsified. It is a received theory that is largely responsible for "raising the field of acoustic phonetics toward the level of a quantitative science."[1,17,19] The source–filter theory relates acoustics to production in terms of the interaction of two components: a source (or sources) of energy – the energy input into the system – and a filter, which modifies that energy. Let us consider the source first.[1,20]

Source

As the air flows through the glottis, vocal cord oscillation is started, whereby the cords come together, stay together for an instant, and then come apart. The cycle is then repeated as long as the aerodynamic and muscular tension conditions for phonation are met. The result of this is a periodic stream of high-velocity jets of air being shot into the supralaryngeal vocal tract. The volume velocity or glottal volume velocity which is volume of air flowing through the glottis .In terms of cubic centimeters (cm^3) of air /second it can be seen that the portion of the particular wave repeats twice in 20 milliseconds. This means its fundamental frequency is 100 Hz.[1,4,8,20,21]

The air-flow profile is a time domain which tells how the air-flow changes as a function of time. The precise energy content of a time-domain wave can be best understood in its spectral, or frequency-domain transformation. The spectrum of the wave shows the energy present in the volume velocity waveform (the energy input into the system).[1,19-21]

Filter

The energy content of the volume velocity wave at the glottis specifies the energy input into the system. This energy is then modified by its passage through the supralaryngeal vocal tract. The contribution of the supralaryngeal vocal tract is to act as an acoustic filter that suppresses energy at certain frequencies and amplifies it at others.[1,21]

During the production of a vowel, air is being expelled at a fairly constant rate through the vocal tract. (This is usually referred to as a pulmonic regressive airstream, because the movement of air is initiated by the lungs, and the direction of the movement of air is outward.)[1,22-25] This airstream is interrupted by the vibratory action of the vocal cords, so that a sequence of high-velocity jets of air is injected into the supralaryngeal vocal tract. The effect of these high-velocity jets is to cause the air present in the supralaryngeal vocal tract to vibrate.[1,23]

The way the air vibrates in particular, the frequencies at which it vibrates and the amplitude of those frequencies are determined by the shape of its container: the supralaryngeal vocal tract. The point in time at which the main response of the supralaryngeal air occurs corresponds to the most rapid change in the glottal flow rate, that is, when the cords are closing.[1,13,19,21,22,24,25-28]

The way in which the air in the supralaryngeal vocal tract will vibrate, given a particular supralaryngeal vocal tract shape, can again be conveniently observed by a frequency–amplitude spectrum, often called a transfer function. This spectrum, or transfer function, represents the acoustic response of the air in the supralaryngeal vocal tract for a schwa [m] – a vowel like that in the word "heard" – when said by a speaker with a supra laryngeal vocal tract 17.5 cm long. (The length of the supralaryngeal vocal tract means the distance between the glottis and the lips.)[1,2,5,8,9,14,16,19,26]

Interaction of the source and filter

The spectral envelope shows how the air would vibrate, given a supralaryngeal vocal tract in the shape of a schwa (with a uniform cross-sectional area). This is often called a filter or transfer function. Now we will find out what happens when some energy is actually provided from the laryngeal source. Recall that the energy input to the system, for vowels, is the spectrum of the volume velocity wave at the glottis.[1,5,11]

Talking about the time when the source is combined with the filter, visually, it appears as if the shape of the transfer function has been superposed on the harmonically rich spectrum of the source. It can be observed that energy contributed by the source is now present at the fundamental and harmonics. It can also be seen that the overall falling shape of the source spectrum has been modified by the envelope of the supralaryngeal vocal tract response.[1,8,9,14,26]

During speech, the fundamental frequency moves up and down and the harmonics concertina in and out as the speaker changes the pitch by changing the rate of vibration of the vocal cords.[1,27,28]

Spectrograms

One of the commonest ways of displaying speech data is using spectrograms, and so it is important to explain them, and enough speech acoustics has now been covered to do this.[1]

It has been shown how speech acoustics can be represented spectrally as a two-dimensional plot of amplitude against frequency. This shows how much energy is present at what frequencies at a particular instant, and it is a unique function of the vocal tract that produced it.[1,25,27,28]

Major peaks in the harmonic structure can be observed in the 2^{nd}, 8^{th}, 14^{th}, and 22^{nd} harmonics, and there are also some minor peaks at the 6^{th}, 19^{th}, and 27^{th} harmonics. There is also a clear zero, or antiresonance (i.e., absorption of energy), at about 4.5 kHz. The LP-smoothed spectrum can be seen to have resolved the major peaks in the harmonic spectrum, but not the minor ones. The first three major peaks in the harmonic structure coincide with the first three formants (marked F1, F2 and F3). The weak harmonic peak at 3.5 kHz actually reflects the fourth formant since it would expect it to be found at that frequency given the frequencies

of the first three formants, and the auditory quality of the vowel (the fact that it sounds like a schwa).[1,22,25,28,29]

A spectrogram allows us to infer quite a lot about the production of speech. To demonstrate this, some other acoustic features of the spectrogram that reflect aspects of articulation can be commented on at this point, as follows:[1]

- The closely spaced vertical striations in the vowel represent the energy pattern resulting from the individual glottal pulses. These are the points in time where the major excitation of the air in the supralaryngeal vocal tract occurs. The individual glottal pulses can also be seen in the vertical striations in the time-domain waveform.[1]

- The two vertical transients almost at the end of spectrogram indicate the energy from the release of the /d/ and the energy is concentrated into formant regions for a very short time after that. The double transient is not typical for alveolars.[1,3,9,15,16,20,23,28,30]

Between-speaker variation in vowel acoustics

The best illustration of how big between-speaker differences in vowel acoustics can be will be found between the vowels of a male (long vocal tract) and a child (short vocal tract). The next best demonstration will be between adult speakers of different sexes, and this is what is shown here.[1,30-36] In order to maximize the chances of getting large differences in the formant value assume a fairly tall male and a very short female for comparison. This is because, formant frequencies correlate with the vocal tract length, and the vocal tract length can be assumed to show at least some correlation with height. The data were maximally controlled for vowel quality.[1,4,6,8,9 31,35-37]

Formant frequencies not only reflect the overall length of the vocal tract that produced them, but also the particular vowel that is being produced.[1,38] Of course, from a investigative point of view, differences of such magnitude will never be the source of legal controversy because the voices that produced them sound so different that they will never be confused in the first place.[1,39,40-42]

The cepstrum

There are two kinds of acoustic parameters used in speaker recognition: traditional and automatic, the latter being used in commercial speaker identification. The undoubted algorithmic mainstay of automatic speaker recognition is the cepstrum, and this section gives a brief nontechnical idea of what it is like.[1,25,27,40,43-47]

From the mid-1960s to the mid-1970s was a very prolific period in the development of signal processing methods, and the cepstrum was first developed then,[1,41,42] as an analytical tool for automatically extracting the fundamental frequency from the speech wave.[1,21,28,43,44] The cepstrum very effectively decoupled the parts of the speech wave that were due to the glottal excitation from those that were due to the supra laryngeal response, and the former were used to estimate the *F*0.[1,34,45,46,48-51]

Its use in speaker and speech recognition rests primarily on its function as a spectral parameter, and not as a fundamental frequency estimator.[1,26,33,34,35,48,50]

Speech defects

Speech disorders belong to a broad category of disorders called communication disorders that also include language and hearing disorders. Speech disorders refer to difficulties in producing speech sounds or problems with voice quality. They may be characterized by an interruption in the flow or rhythm of speech such as stuttering, or by problems with the way sounds are formed, also called articulation or phonological disorders, or they may involve voice problems such as pitch, intensity, or quality. Often, there is a combination of several different problems.[1,49,50,52-54] Speech disorders can either be present at birth or acquired as a result of stroke, head injury, or illness. Major speech disorders include.[1]

Articulation disorders

Articulation is the production of speech sounds, and persons affected by articulation disorders experience difficulty in being understood because they produce incorrect speech sounds. As a result, their speech is not intelligible. They may substitute one sound for another or may distort the sound which results into incorrect sounds, even though still recognizable, or omit one or more sounds in a word.[1,49,53,55]

Phonological disorders

Phonology is the science of speech sounds and sound patterns and of the language rules that dictate how sounds may be combined to produce a language. Persons affected by phonological disorders do not use the conventional rules for their native language but substitute their own variants.[1,49,50]

Stuttering

Normal speech is fluent, in that it is spoken effortlessly and without hesitation. A break in fluent speech is called a dysfluency. Although some degree of dysfluency occurs in normal speech from time to time, stuttering has more dysfluencies than is considered average.[1,49,56,57]

Apraxia

This is a speech disorder in which voluntary muscle movement is impaired without muscle weakness. There are

two main types of apraxias: buccofacial apraxia and verbal apraxia. Buccofacial apraxia impairs the ability to move the muscles of the mouth for nonspeech purposes such as coughing, swallowing, and wiggling of the tongue. Verbal apraxia impairs the proper sequencing of speech sounds. Apraxias can either be acquired or developmental and have different degrees of severity, ranging from the inability to initiate speech to mild difficulties with the pronunciation of multisyllabic words.[1,50,51,58,59]

Dysarthria

This is a speech disorder that affects the muscles involved in the production of speech. As a result, speech is slow, weak, inaccurate, and hesitant. Dysarthria results from a weakness in any one of these elements or in the absence of proper coordination between them.[1,49-51,56]

Causes and symptoms

The causes of articulation and phonological disorders are unclear, although it has been observed that they tend to develop in children before the age of four and run in families. Articulation is considered a disorder when it is unintelligible or draws negative attention to the speaker. For example, the word "super" is pronounced as "thuper."[1,58,59]

The causes of stuttering are not very well understood. There is some evidence that stuttering has a genetic cause since it has been observed to run in some families.[1,49,52,57]

The main causes of organic voice disorders include neuromuscular disorder, cancer, vocal cord paralysis, endocrine changes, various benign tumors such as inflammatory growths (granulomas), or consisting of a mass of blood vessels (hemangiomas), or occurring on mucous membranes (papillomas).[1,50-52,54,58]

Diagnosis

Speech disorders are usually identified using a combination of hearing tests and physical exams. Physicians then recommend specialized evaluation by speech-language pathologists, who can best establish an accurate diagnosis.[1]

A stuttering diagnosis is established on the basis of the type, frequency, and duration of speech dysfluency. The number of dysfluencies occurring in 100 words is counted to determine the dysfluency percentage.[1,59] One half a stuttered word per minute is the usual criterion. Determining the type of stuttering behavior, either overt or covert, is the most important factor in diagnosing stuttering.[1,52]

Organic and functional voice disorders are diagnosed with the assistance of an ear, nose, and throat specialist, an otolaryngologist, who can identify the organic cause of the voice disorder, if present.[1,51-53,57,59]

Treatment

Speech pathologists have designed approaches for treating speech disorders with the type of treatment depending upon the type of impairment. A wide variety of treatment techniques are available for treating affected children, adolescents, and adults. A thorough assessment is normally conducted with the aim of determining the most effective and acceptable treatment approach for each disorder on an individual basis. A common treatment for many patients involves increasing sensory motor awareness of selected aspects of speech and systematically shaping the target speech behaviors.[1,49,51-53,56]

Speech-language pathologists use many different approaches to treat voice problems. Functional voice disorders can often be successfully treated by voice therapy. Voice therapy involves identifying voice abuses and misuses and designing a course of treatment aimed at eliminating them. Voice disorders may require surgery if cancer is present.[1,53,55]

The treatment of apraxia depends on the extent of the impairment. For individuals diagnosed with moderate to severe apraxia, therapy may be for them to start saying individual sounds and contrasting them, thinking about how the lips and tongue should be placed.

Treatment of dysarthria usually aims at maximizing the function of all speech systems with the use of compensatory strategies.[1,53]

Prognosis

The prognosis depends on the cause of the disorder; many speech disorders can be improved with speech therapy. In the case of childhood speech disorders, prognosis also significantly improves with early diagnosis and intervention. Children who do not receive speech therapy and do not outgrow their speech difficulties will continue to have the disorder as adults.[1,56-59]

Health care team roles

The treatment of speech disorders belongs to the field of speech-language pathology. Speech-language pathologists assist individuals who have speech disorders and collaborate with families, teachers, and physicians to design an appropriate course of treatment, which depends on the specific nature of the disorder. They also provide individual therapy to affected persons, consult with teachers about effective classroom strategies to help children with speech disorders, and work closely with families to develop effective therapies.[1,49,53,57,59,60]

CONCLUSION

This article tries to explain a little of the theory of how speech acoustics are produced and applied, speech defects, and two important acoustic parameters: vowel F-pattern and long-term fundamental frequency.

Speech acoustics reflect the vocal tract that produced them, and can be extracted and quantified with relative ease using currently available signal-processing software. Distributions of acoustic parameters like *F*0 can then be successfully modeled statistically to enable quantified comparison between speech samples, and spectral features can be modeled with the cepstrum.

Speech acoustics only reflect how a speaker is speaking on a particular occasion; however, just like any other parameter, they are not invariant and show both within-speaker as well as between-speaker variation. But for Successful comparison between speech samples requires well-controlled data from different occasions.

ACKNOWLEDGMENT

I want to give my sincere gratitude and acknowledgement to Philip Rose in preparing this article.

REFERENCES

1. Rose P. Forensic Speaker Identi?cation. 2002 Taylor & Francis, 1st Ed.pp 67-298.
2. Wolf JJ. 'Efficient acoustic parameters for speaker recognition'. Journal of the Acoustical Society of America (JASA) 1972;51: 2044-56.
3. Barlow M, Clermont F , Mokhtari P. 'A methodology for modelling and interactively visualizing the human vocal tract in 3D space'. Acoustics Australia 2001;29/1:5-8.
4. Atal BS. 'Effectiveness of linear predication characteristics of the speech wave for automatic speaker identification and verification'. Journal of the Acoustical Society of America (JASA) 1974;55:1304-12.
5. Davies SB, Mermelstein, P. 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences'. Proceedings of the Institute of Electrical and Electronics Engineers (IEEE), (SC) Speech Communication Transactions on Acoustics , Speech, and Signal Processing 1980;28:357-66.
6. AFTI (n.d.) Voice Print Identification, Applied Forensic Technologies International, Inc.HTTP://www.aftiinc.co./voice.htm (accessed 27/08/2001).
7. Deffenbacher KA. Relevance of voice identification research to criteria for evaluating reliability of an identification. Journal of Psychology 1989;123:109-19.
8. Diller AVN. Reflections on Tai diglossic mixing. Orbis 1987;32/1: 47-66.
9. Emmorey K, Van Lancker D, Kreiman J. Recognition of famous voices given vowels words and two-second texts. Working papers in phonetics, University of California at Los Angeles(UCLA WPP) 1984;26:120-4.
10. Baldwin J. Phonetics and speaker identification. Medicine, Science and the Law 1979;19:231-2.
11. Doddington GR. Speaker recognition - identifying people by their voices. Proc Proceedings of the Institute of Electrical and Electronics Engineers; Institute of Electrical and Electronics Engineers (IEEE), Speech Communication (SC) 1985;73/11:1651-64.
12. Broeders APA. The role of automatic speaker recognition techniques in forensic investigations . Proc. Intl Congress Phonetic Sciences 1995;3:154-61.
13. Foster KR, Bernstein DE, Huber PW.Science and the toxic tort. Science 1993; 261:1509-614.
14. Champod C, Evett I. Commentary on Broeders. Forensic Linguistics (FL) 2000;7/2:238-43.
15. Gish H, Schmidt M. Text-independent speaker identification. Institute of Electrical and Electronics Engineers; SC, Speech Communication (IEEE) Signal Processing Magazine 1994;11/4:18-32.
16. Goddard C. Can linguists help judges know what they mean? Linguistic semantics in the court-room. Forensic Linguistics (FL) 1996;3/2:250-72.
17. Kohler KJ. The future of phonetics. Journal of the International Phonetics Association (JIPA) 2000;30/1:1-24.
18. Osanai T, Tanimoto M, Kido H,Suzuki T. Text-dependent speaker verification using isolated word utterances based on dynamic programming. Reports of the National Research Institute of Police Science 1995;48:15-19.
19. LaRiviere C. Contributions of fundamental frequency and formant frequencies to speaker identification. Phonetica 1975;31:185-97.
20. Papçun G, Kreiman J, Davis A.Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America (JASA) 1989;85:913-25.
21. Wells JC. British English pronunciation preferences. Journal of the International Phonetics Association (JIPA) 1999;29/1:33-50.
22. Rose P. Differences and distinguish ability in the acoustic characteristics of Hello in voices of similar-sounding speakers. Australian Review of Applied Linguistics 1999a; 21/2:1-42.
23. Pruzansky S, Mathews MV. Talker-recognition procedure based on analysis of variance. Journal of the Acoustical Society of America (JASA) 1964;36:2041-7.
24. Künzel HJ. Beware of the telephone effect: the influence of transmission on the measurement of formant frequencies. Forensic Linguistics (FL) 2001;8/1:80-99.
25. McGehee F. The reliability of the identification of the human voice. Journal of General Psychology 1937;17:249-71.
26. Nolan F. Forensic phonetics. Journal of Linguistics 1991;27:483-93.
27. Goggin JP, Thompson CP, Strube G, Simental L R. The role of language familiarity in voice identification. Memory and Cognition 1991;19:448-58.
28. Furui S, Matsui T. Phoneme-level voice individuality used in speaker recognition. Proc. 3rd International Conference on Spoken Language Processing 1994;1:1463-6.
29. Noll AM. Short-time spectrum and cepstrum techniques for voiced pitch detection. Journal of the Acoustical Society of America (JASA) 1964;36:296-302.
30. Ladefoged J, Ladefoged P. The ability of listeners to identify voices. Working papers in phonetics, University of California at Los Angeles (UCLA WPP) 1980;49:43-51.
31. Clermont F, Itahashi. Monophthongal and diphthongal evidence of isomorphism between formant and cepstral spaces. Proc. Spring Meeting of the Acoustical Society of Japan, Meiji University Press 1999;56: 205-6.
32. Lisker L, Abramson AS. A cross-language study of voicing in initial stops: acoustical measurements .Word 1964;20:384-422.
33. Laver JMD. The nature of phonetics. Journal of the International Phonetics Association (JIPA) 2000;30/1: 31-6.
34. Goldstein AG, Knigh P, Bailis K, Connover J. Recognition memory for accented and unaccented voices . Bulletin of the Psychonomic Society 1981;17:217-20.
35. Harrington J, Cox F, Evans Z. An acoustic phonetic study of Broad, General and Cultivated Australian vowels. Australian Journal of Linguistics (AJL) 1997;17:155-84.
36. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. Proceedings of the Institute of Electrical and Electronics Engineers; SC, Speech Communication; (IEEE) Transactions on

Biomedical Engineering 2000;4:829-37.
37. Ingram J, Prandolini R, Ong S. Formant trajectories as indices of phonetic variation for speaker identification. Forensic Linguistics (FL) 1996;3:29-45.
38. Clermont F, Itahashi S. Static and dynamic vowels in a "cepstro-phonetic" sub- space'- acoustical letter. Journal of the Acoustical Society of Japan 2000;21/4:221-3.
39. Jassem W. Pitch and compass of the speaking voice. Journal of the International Phonetics Association (JIPA) 1971;1/2:59-68.
40. Kersta LG. Voiceprint identification. Nature 1962;196:1253-7.
41. Foulkes P, Barron A. Telephone speaker recognition amongst members of a close social network. Forensic Linguistics (FL) 2000;7/2:181-98.
42. Clermont F. Multi-speaker formant data on the Australian English vowels: a tribute to J. R. L. Bernard's pioneering research in McCormak and Russell (eds) 1996;88:145-50.
43. Butcher A. Getting the voice line-up right: analysis of a multiple auditory confrontation in McCormak and Russell (eds) 1996: 97-102.
44. Hombert JM. A model of tone systems. Working papers in phonetics, University of California at Los Angeles (UCLA WPP) 1977;36:20-32.
45. Clifford BR, Rathborn H, Bull R. The effects of delay on voice recognition accuracy. Law and Human Behaviour 1981;5:201-8.
46. Furui S, Itakura F, Saito S. Talker recognition by longtime averaged speech spectrum. Electronics and Communications in Japan 1972;55A/10: 54-61.
47. Collins B. Convergence of fundamental frequency in conversation. If it happens, does it matter? in Manell and Robert-Ribes (eds) 1998;3:579-82.
48. Rose P, Clermont F. A comparison of two acoustic methods for forensic discrimination. Acoustics Australia 2001;29/1:31-5.
49. Cox F. The Bernard data revisited. Australian Journal of Linguistics (AJL) 1998;18: 29-55.
50. Johnson NC, Sandy JR. Tooth Position and Speech-Is There a Relationship?. Angle Orthodentistry 1999;69: 306-10.
51. Kraus N, Cheour M. Speech Sound Representation in the Brain. Audiology and Neurotology 2000;5: 97-132.
52. Oller DK, Eilers RE, Neal AR, Schwartz HK. Precursors to Speech in Infancy: The Prediction of Speech and Language Disorders. Journal of Communication Disorders 1999;32: 223-45.
53. Postma A. Detection of Errors during Speech Production: A Review of Speech Monitoring Models. Cognition 2005;77: 97-132.
54. Rosen CA, Murry T. Nomenclature of Voice Disorders and Vocal Pathology. Otolaryngology Clinical North America 2007;33: 1035-46.
55. Yamazawa H, Hollien H. Speaking fundamental frequency patterns of Japanese women. Phonetica 2009:22:128-40.
56. Shipp T, Doherty E, Hollien H. Some fundamental considerations regarding voice identi?cation. Journal of the Acoustical Society of America (JASA) 2008;82: 687-8.
57. Tosi O, Oyer HJ, Lashbrook W, Pedney C, Nichol J, Nash, W. Experiment on voice identi?cation. Journal of the Acoustical Society of America (JASA) 2010;51: 2030-43.
58. Rietveld ACM, Broeders APA. Testing the fairness of voice identity parades:the similarity criterion. Proc. XIII International Congress of Phonetic Sciences 2007;5: 46-9.
59. McCawley JD. What is a tone language?. Working papers in phonetics, University of California at Los Angeles (UCLA WPP) 2009;49: 43-51.
60. Lyons J. Language and Linguistics: An Introduction. Journal of General Psychology 2010;18: 219-31.

How to cite this article: Tiwari M. Speech acoustics: How much science?. J Nat Sc Biol Med 2012;3:24-31.

Source of Support: Nil. **Conflict of Interest:** None declared.