

SPOKEN LANGUAGE : SOME SPECIFICS & A PERSPECTIVE*

ASOKE KUMAR DATTA

Abstract : The paper presents the formal composition of the *Prof. Suniti Kumar Chatterji Memorial Lecture-III*. Objective research on speech in India began in late sixties with the availability of instruments for objectively quantifying different aspects of speech. One of the pioneering research institutes in India engaged in this effort was Indian Statistical Institute (ISI). This paper presents the synopsis of the work done for about five decades there in relation to the dialect Standard Colloquial Bengali (referred to as Bangla hereinafter) among some other spoken languages. The presentation attempts to make it intelligible to students of spoken languages. It includes a discussion of the paradigm of research in spoken language. It is necessary because an intimate wholesome inquiry in spoken language needs to span two branches of knowledge gathering namely, subjective and objective. In the objective domain the paper discusses two different approaches for representation namely, spectral and time domain representations of speech sounds. The subjective categorization of some units of speech sounds is discussed. The objective cues for cognition of different phones are discussed in some details. Some new approaches, alternative to the traditional approach in spoken language research, which were experimented upon in ISI find some mention in the article. The paper also spells out some of the author's confusions related to certain units of spoken language and approaches traditionally used in language research. Finally the article advocates for the holistic approach in place of reductionist approach for a spoken language. A tentative model for spoken language research is also prescribed.

Key Words: artificial, spoken, perception, phones, speech, words

1.0. Introduction

The paper presents an overview of spoken language from the point of view of artificial intelligence (AI). A pure linguist's view, as I understand, seems to be an abstract ideational domain talking about formal things mostly focussed on subjective entities like semiotics including semantics and structures like grammar. These views also include the acoustics though primarily through cognitive images. Thus this domain can be thought of as a rationale based on subjectivity. Speech may also be viewed from the view point of hard sciences. This branch of science is primarily of western origin. We may trace its origin to the famous definition of reality by Democritus of Adbera. He said that reality is atom and void and thence the modern science meticulously adhered to this reality of strict material objectivity. The part of the natural world which represented the conscious wilful activity of the life

forms and was subjective was thrown out of the window for good. I have also heard of a school of thought which professes that we cannot know the reality as we live in a shadow world. As the source of light creates the shadow of an object, likewise our sensory organs, with their limitations, transforms nature into images of it and we live therein. We therefore can never have access to the true nature of reality as we have no idea of other objective parameters of reality that our sense organs cannot glean.

However, this knowledge does not resolve the confusion; it only transliterates it by adding an adjective 'shadow' before 'objective' or 'subjective'. Could there be another reality called public subjectivity? For example, let us take the statement 'the sun rises in the east'. This is no 'pointer reading'. This is a subjective reality common to many men. This public subjectivity bridges the two apparently disjoint sets, 'subjective' and 'objective'. It might transform some subjective realities into what we call objective reality. I believe, most of the so-called objective realities, we know of, belong to this class, unless they are axiomatic. So it makes sense to ponder over the universe even if we call our concept of it as a shadow. This is not to bore the audience but to enumerate the difficulties faced by hard technology to interact with an intelligent activity, one like language which is said to be the highest form of intelligent activity of human being. In my somewhat long association (Dutta Majumder et al. 1968) since late 60's with the spoken language as an AI student I oscillated between hard objectivity of acoustics and the subjectivity of cognition and ultimately settled (if I may call so) on what I may term as 'subjective-objectivity'. The following is a chronicle of this journey.

2.0. Spoken Language

I consider speech as the most common natural language of the human being despite the common practice of calling written language as the natural language. To me the written language is a highly reformed artificial language. Speech is almost unbridled. Like a sprinting gazelle it runs its own way. Being a subjective process and carrying the most knotty load of semantics, it really does not care much for the rules of grammar. Strangely, as my association with speech lengthened, my confusion grew with it. The confusion lies everywhere; in the search for objectivity of acoustics of speech, in understanding the subjectivity of cognition, in the search of their relationship, even in the doubt of the relevance of the chosen fundamental units of understanding speech, namely the phones/phonemes and the word, and finally, the reductionist approach of understanding speech. When I hear a speech, do I hear the phonemes first, then the words, then the message? I am not very sure.

2.1. Phones

Let me begin with the phonemes, rather the phones in the objective acoustic domain (Datta et al. 1978, 1985). Let us consider the nonsense syllable 'ɔkɔ'. Figure 1 presents both

spectrograph (upper half) and the signal (lower half) of 'əkə'. The three bars in between represent the states of the articulator and the glottis and a coloured bar represents the subjective hearing of the different segments. If we listen to the portion indicated by the double headed arrow tagged as 'k', the objective phone [k], which by definition consists of a silence followed by a burst, we do not hear [k], instead hear just a click. Similarly, if we listen to the objective portion representing the vowel [ə], which is defined to be the uninterrupted voice passing through the oral cavity, we hear the first red bar as [ək] the last red bar as [kə]. If you hear the signal assigned by the green bar below, you hear [əkə]. Up to this is elementary; we all know this from our school days, only nobody pointed out the vagueness of definition of phones which follows. If one takes two vowel signals representing [ə] and put the signal for [k] in between you don't hear [əkə], you hear only /ə/. One cannot just stitch together phones to make speech. In the continuum of natural speech there is no isolated vowel or isolated consonants, at least one cannot define boundaries. They are often in the mind of the listener. If one examines carefully a sentence taken from a normal conversation and microscopically examine them they will find lots of missing phones and non-speech sound items. Speech is a very complex and intricate interaction between different articulators, generated by neuro-motor commands controlled by the intention to convey a message. Coarticulatory (Ohman 1966) and anticipatory influences can go quite far. To me, phones do not represent fundamental robust objectively distinct units of speech in a real speech signal. It is a casual import from written language.

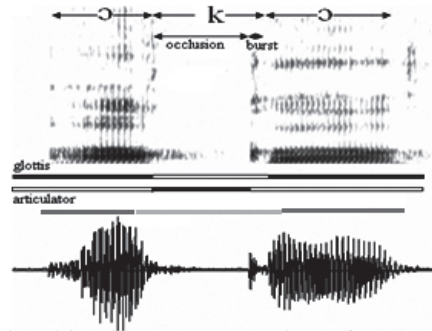


Figure 1: Objective representation of a VCV syllable

2.2. Words

My other confusion is about the objectivity of 'words' in spoken languages. I, along with many others in India and abroad, tried to find word boundaries objectively from speech signals. These searches became elusive in the sense that the recognition rate (87%) (Das Mandal et al. 2006) and the consistency was much below practical usability. I also doubt that anybody could subjectively detect even some words from continuous speech in a language unknown to him with a reasonable degree of confidence. I feel that units like

phones, words, phrases, grammar are needed to put the natural phenomena of speech in a reductionist perspective. It is likely that these concepts have been imported in speech from the written script without proper application of mind. I am not sure though. I shall await the opinion of scholars in this regard. While the relevance of such an approach has undoubtedly been useful for written languages, I cannot see its relevance in speech. The model which fits so well in artificial languages is, I am afraid, useless for natural speech. However it is possible, using signal processing to cut out portions consistently and robustly from speech which represent semantic chunks, often a group of words. To me these could be important elements in understanding speech. This will be elaborated later. I have reasons to believe that human brain, like many of its cognitive behaviour, does not use reductionist model in understanding speech. I think that the understanding of speech is a holistic process triggered by the association of certain key phrases without an order. Unfortunately I have no *proof* to offer. These are some of the problems an AI student faces with speech.

3.0. Objective representation of speech

It may be in order to refer several instrumental setups necessary for the extraction of the objective parameters directly from the speech sounds. Some of the instruments used in this field are electronic palatograph (EPG), elctroglottograph (EGG), spirometers, cine-fluorography, ultra-sonography, sound spectrographs, endoscopy and video graphs, etc. It is also possible to use sound spectrography to estimate some objective properties in the absence of a particular important instrumental set up. One needs to understand that these instrumentations are necessary to find the objective basis of categorization of units of speech as well as the mutual interactions of these units in continuous speech.

3.1. Spectral Representation of speech

Let us recall the traditional approach in the acoustic representation of speech first and talk about the confusion later. The traditional universally used approach for objectivity in speech acoustics is the spectral transform. Almost five decades ago Gunnur Fant (Fant 1970), often called the father of modern speech science, said that phones can be objectively defined by the behaviour of 'Formants'. For these five decades scientists considered formants to be solely responsible for cognition also. Numerous acoustic and psycho-perceptual experiments have been conducted worldwide to prove this. Neuro-physiological structures in the inner ear and the cochlear pathways have been found confirming spectral processing. Even neurons giving out 'auditory spectra' is detected. There has been no question about the all pervasiveness of formants in primary cognition related to speech. However in late 80's and early 90's some small doubts peeped in the mind of a group of Artificial Intelligence (AI) students in ISI (Datta 1993). The history of the pursuit of this doubt would be one of the mainstays of my talk today.

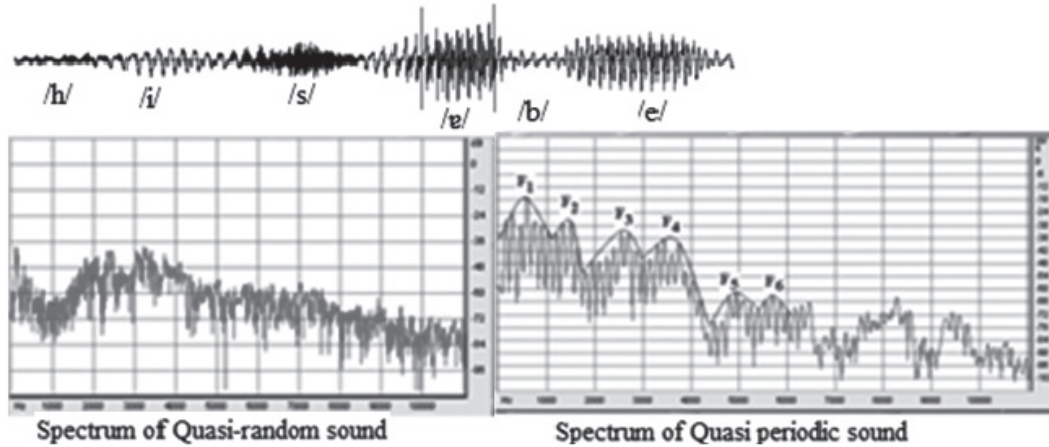


Figure 2: Illustration of spectral structure of speech sounds

Let me begin with some examples of the traditional objective representation of speech. Figure 2 illustrates the signal as well as the spectral structure of speech sounds. Upper section of the figure presents the signal representing the Bangla word ‘hisabe’ pronounced as */hifəbe/*. (An interesting point to note is that I found almost all /h/ in Bangla words I examined to be aspirated vowels). The lower section represents the spectrum of a quasi-random sound */f/* and the spectrum of the quasi-periodic sound */e/* showing formants. Formants are resonances in the spectral structures. One may notice that the spectral structure of quasi-random sounds shows no regularity which is characteristic of non-repetitive noise signals. On the other hand, quasi-periodic signals show regular harmonic structure (red line). If one draws the envelope (black line) one gets the resonances of the articulatory pathways. These are known as formants. The first two formants are held to be primarily responsible for categorisation of place of articulation of phonemes. It may be in order to explain quasi-periodicity (Sengupta et al. 2001, Teagre et al.1990) in the context of speech sounds. Contrary to the common approach of comparing voice production with reed vibrations in musical instruments, the vibration of vocal folds are quite complex. Figure 3 schematically illustrates the mechanism of voice production.

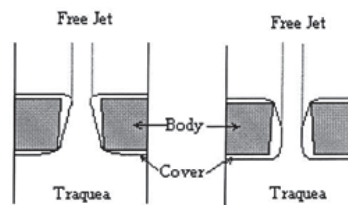


Figure 3 : Formation of free jet in supra-glottal tract

The vocal folds are mucous covered pliant tissues. The outer part usually referred to as ‘cover’ consists of the epithelium and the superficial layers overlaying the deep layers of

the vocal ligaments and muscles known as the body. Because of this inherent non-rigidity random variations occur in the oscillatory motion caused by the air streaming past them. The origin of these perturbations are to be sought either at the source or near the source in some sort of vertical flow created by the air streaming past the edge of vocal folds. The opening and closing of the tract is caused not by the motion of the rigid 'body' but by the semi-fluid motion of the 'cover'. During oscillations the cover of the vocal folds propagates a surface mucosal wave in the direction of the airflow. The body remains quasi-stationary. The glottal fold is considered as quasi steady and incompressible. The bulk of the flow is frictionless and laminar. Due to abrupt changes at the glottal exit, boundary layers separate from the wall and cause a free jet downstream. The air exits the glottis in a free jet and becomes turbulent. This is the source of random perturbations. These random perturbations are of three types: Jitter (variation in F_0), Shimmer (variation in amplitude) and Complexity Perturbation (variation in shape). Random perturbation cognitively manifest as a quality of voice. If it is too low, the voice appears to be mechanical and hornlike, if it is too large the voice appears harsh (Choudhury 2006-7).

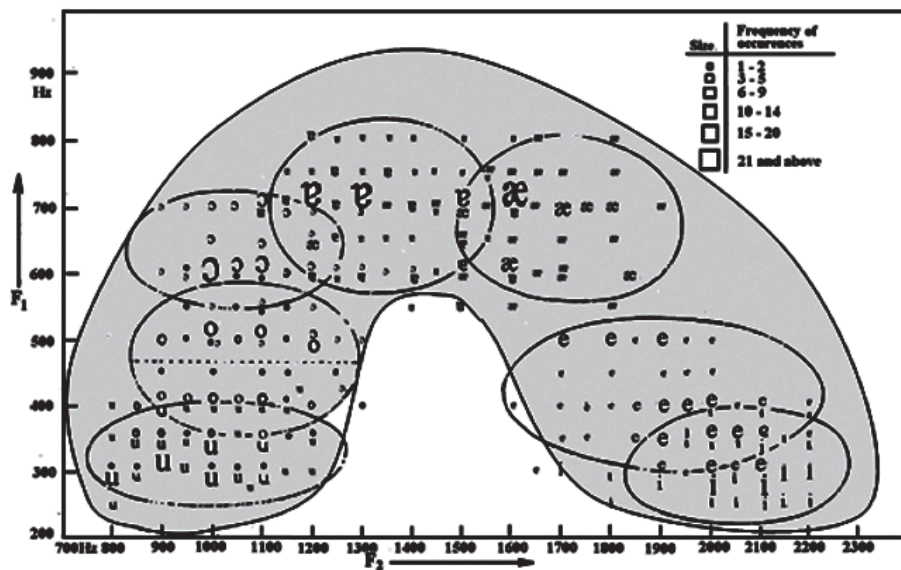


Figure 4 : Distribution of Bangla vowels in F_1 - F_2 plane (Datta 1988,1989)

Figure 4 shows the distribution of Bangla vowels in F_1 - F_2 plane. The vertical axis represents F_1 and the horizontal axis F_2 . F_1 increases with decreasing tongue height. The oval represents the spread of a vowel cluster, the axes being proportional to the standard deviations of the vowel clusters. The vowels are distributed in a space shaped like a boomerang. The almost free space between the two arms of the boomerang is occupied in

Bangla by lateral and rhotics. One may notice the overlaps between the vowels. The overlaps are generally between the adjoining vowels and on an average is 20%. The objectivity of the discrimination between vowels in formant plane is thus obvious.

It has been known for a long time that subjective categorization of vowels in terms of the height and the front back position of the tongue hump is strongly related to the first two formants. A recently reported technique (Pfitzinger 2003) enables one to represent formant data, together with fundamental frequency (F_0) values, into the traditional perceptual evaluation of the category of a vowel utterance in terms of height and backness of the tongue. The following equations have been used for converting the formant data into perceptual parameters of tongue height (h) and backness (b):

$$h = 2.621 \log(F_0) - 9.031 \log(F_1) + 47.873$$

$$b = -0.486 \log(F_0) + 1.743 \log(F_1) - 8.385 \log(F_2) + 59.214$$

Figure 5 below is drawn from the formant data of figure 4 along with the value of corresponding fundamental frequency F_0 using the aforesaid equations. One may note that this diagram represents the average of the normal pronunciations of SCB speakers taken from continuous speech, not those of erudite phoneticians imitating common SCB vowels either from single vowels and/or from those embedded in selected words. This possibly is the first report that gives the statistical estimates of the articulatory positions of Bangla vowels as spoken by common people in natural speech. It is also an example of obtaining subjective categories from large scale objective data compilation in an Indian context. A close perusal of it also indicates that there may be a need for a relook to fix the IPA symbols for Bangla vowels.

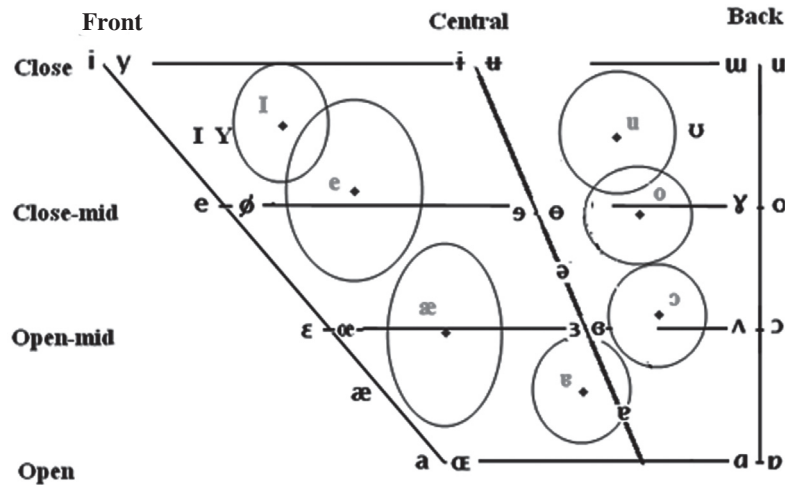


Figure 5: Perceptual vowel diagram for Bangla vowels drawn from objective data

Figure 6 presents formant transitions representing articulatory movements for four unvoiced unaspirated Bangla plosives (Datta 1980). It is a common knowledge that for the voiced and aspirated counterparts, the transitions do not significantly change. CV transitions are marked by white lines. One may notice from figure 5 that the transition of F1 and F2 are different for different places of articulation. While F1 transition is always downwards towards the burst, since at the point of release tongue always touches the palate and F1 is related to the height, F2 movement can both be upwards and downwards. The F2 transition varies normally according to the place of articulation of the consonant relative to the adjoining vowels. There is no F2 transition for [k] in [ɐkɐ] since it has almost the same place of articulation as that of the vowel.

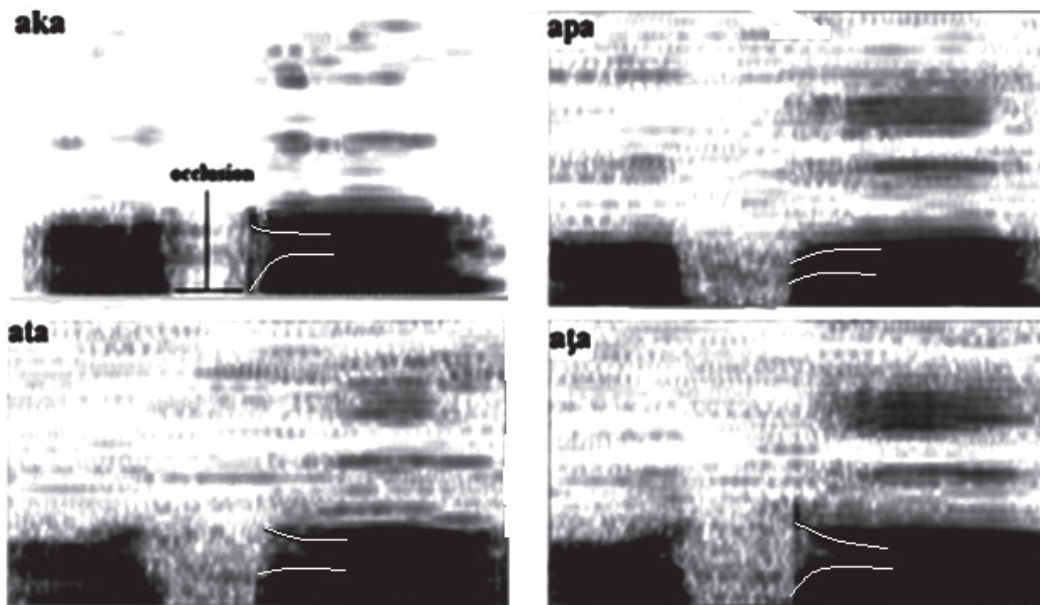


Figure 6: Formant transition representing articulatory movement for plosives

Aforesaid observation represents the flavour and the basics of the traditional model for both cognition and objective representation of speech sound. The vowels are represented by the formant structure at the quasi-steady state of the nucleus vowel, other vowel-like sounds like glides and diphthongs by both the steady states and transitory movements of the formants, place of articulation of consonants primarily by the transition of formants in VC/CV coarticulatory stages.

4.0 Alternative Hypothesis

As far as I know, nobody has ever questioned the traditional mode for cognition of speech. Some Jacks in ISI decided to probe this belief. The beginning was however, the objective representation. This entailed a prolonged engagement with speech signals on screen of oscilloscope, later on computer screens enabled them to recognise the signals in its time series form. We could recognise the vowels from their traces.

4.1 Time Domain Representation

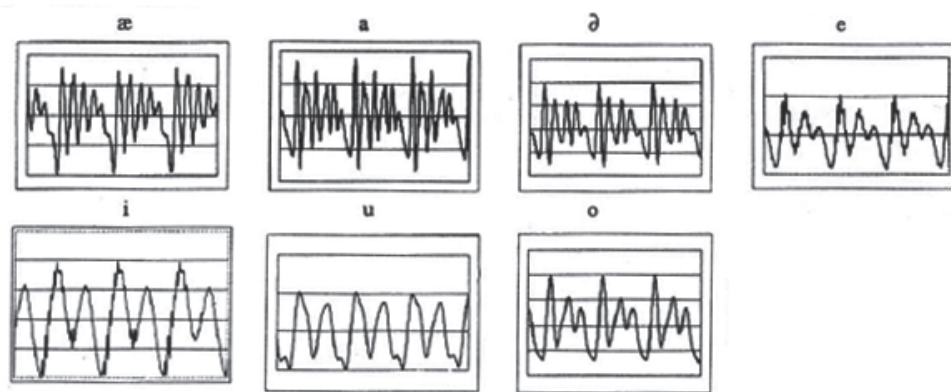


Figure 7: Representative signals of 7 Bangla vowels

Figure 7 shows the representative signals of 7 Bangla vowels. One can easily see the repetitive nature of the signals. One can also note that they are visibly distinguishable. This visual distinction led to the development of an alternative time-domain representation of speech both for Automatic Speech Recognition [12] and Text to speech synthesis (Dan et al. 1993) in ISI. It has been found that five parameters, namely (a) Peak to Peak Distance, (b) Number of Zero Crossings, (c) Perturbation Area (d) Extrema Rate, and (e) Intrinsic Intensity of the vowel, can distinctively determine the shape of the wave periods representing the vowels. As the definitions of these parameters is not felt required for the present lecture, it is not included here. These are described in detail elsewhere (Datta 1991, Das Mandal et al. 2003). It has been shown that the same parameters may be used to classify plosives according to their places of articulation [Das Mandal et al. 2007].

It has been noticed that while a pitch period can extend for more than 10 milliseconds, the perception of phonetic quality depends only on a small segment [13], about 1.5 msec. of the pitch-period measured from the epoch (Figure 8). This discovery led to the development of a new chapter in the domain of concatenative speech synthesis in early nineties again in ISI. This procedure has been used for singing synthesis too.

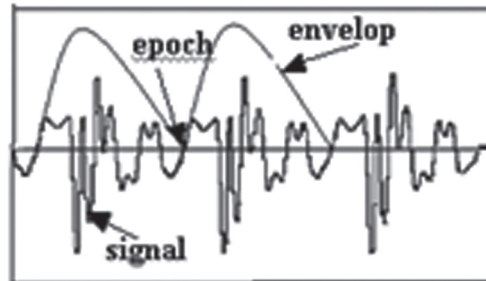


Figure 8 : Epoch of a pitch-period

4.2 Cognitive Behaviour

The aforesaid fact has been used to generate new signals to demonstrate that formant frequencies are neither necessary nor sufficient for cognition of vowels or plosives. Let us refer to figure 9. It shows three pitch periods. First one is taken from an utterance of vowel /ɔ/ and the second from vowel /e/. The portion A from /ɔ/ is copied and added in front of the portion B from /e/, pasted together and then passed through some selected filters to generate a new period shown as the last one in figure 8. If this new period is pasted several times consecutively to produce a signal of length say about 100 milliseconds, one would hear the vowel /ɔ/. If one extracts the formant frequencies of this signal it will be of vowel /e/. In this way 13 different vowel segments, each of length about 400 milliseconds were prepared representing seven vowels having contradictory formant indications [Datta et al. 2011].

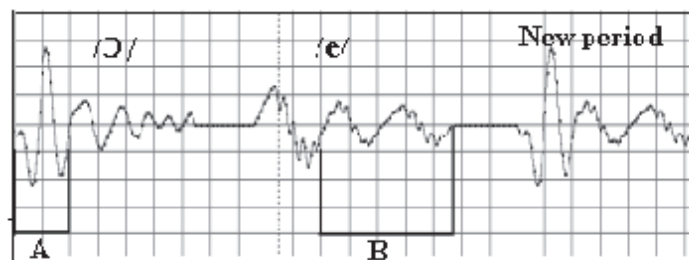


Figure 9 : Illustration of alternative vowel signal preparation

Figure 10 illustrates the process for straightening transitional movements of formants while retaining perceptual indicators for the place of articulation of the plosives. The red portions show what we believe to be the perceptual cue for tongue movement. The trick is to replace the rest of a period by the signal represented by the gray. This is the portion which represents the formant structure of the succeeding vowel. The whole signal is then treated with selected filters. The same process is repeated for the VC segment.

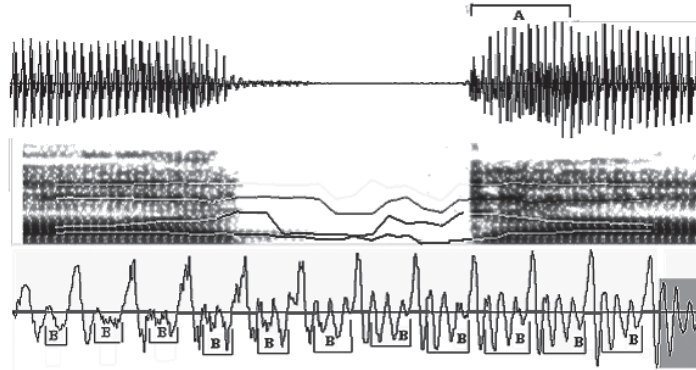


Figure 10: Illustration of signal preparation with VCV utterance /eʃe/

The transformed VCV signal is shown in figure 11.

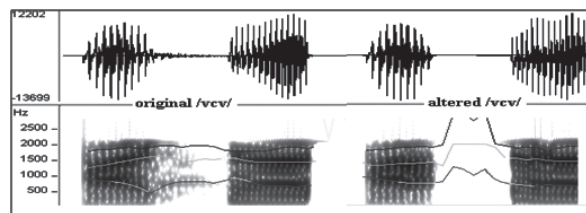


Figure 11: Example illustrating straightening of the formant transitions

Altogether 22 VCV utterances are used to prepare 22 separate altered sound files. Those VCV are chosen which normally reveal large transitions. As stated earlier the altering consisted of flattening the formant transitions.

All these signals (13+22) were used for listening tests by 30 native SCB speakers. The results conclusively demonstrated (Datta et al. 2011) that formants are neither necessary nor sufficient for cognition of vowel category or the place of articulation of plosives. I should point out, we did not prove that human cognition does not use formants for cognition of speech sounds.

5.0 Summary

I failed to fully comprehend the usefulness of a word, particularly as a fundamental semantic unit, in speech. When I look at a book I can identify the words because they are separated by a blank space. As reported earlier, one can find word boundaries objectively with only a little over 85% success rate. If I look at the speech signals of sentences consisting of words like /rɛdzpuʃrɔ/, /dinɔbondhu/, etc. I hear each of the aforesaid words broken into

two by stress as indicated by the accent mark in the above examples. May be it is so, because in Bangla the first syllable of a word is said to be stressed.

It is well known that a listener can detect only about 40% of the phonemes when they listen to an unknown language. I am not aware of experiments on the perceptual capability of listeners with regard to the detection of words in an unknown language. I surmise this would not be great either. Thus a student of AI is faced with a paradoxical situation. One cannot decipher a message unless one knows the language. One cannot know a language unless one can extract the cognitive elements, like words necessary for deciphering the message.

Up to this are statements on fact findings. What do they import?

It often pays to question even the well established notions.

Time domain parameters provide representation of speech sounds which is as good as spectral domain representation.

Spectra are not the only objectivity responsible phenomena for cognition of speech. Human cognition is difficult to understand because the brain may be using more than one algorithm for a particular task. (I may be allowed to report one of my experiences in visual cognition. It was sometimes in mid eighties one of my associate students was studying illusory contours. When he used a low pass spatial filter, the illusory triangle came out objectively. Around the same time an article was published in the American Journal of Psychology (if I remember correctly) that an area in the striate cortex of Rhesus monkey exists which is involved in spatial filtering. It was exciting. About a year later another article was published in the same journal which reported that in human the brain rarely uses this area in cognition. It was doubly exciting. Probably the brain found a better algorithm for the task.)

The spoken language is direct man-to-man communication; it does not require any further transformation for getting the message. Artificial language on the other hand, transmits artificial codes which have to be interpreted to obtain the semantics of different messages. Interestingly this interpretation, I believe, has to be through verbalisation.

6.0 A Skeleton Model for Spoken Language

It appears to me that speech being fundamentally different from artificial languages and the existence of so called smallest crisply defined fundamental units like phonemes, words etc., being vague, a reductionist approach is not the proper paradigm. Moreover, the semantic information is highly contextual. In such a situation a holistic approach seems to be more appropriate. Let us begin with a set Θ of key element 's' having a binary semantic value, 'TRUE' or 'FALSE'. These key elements are perceptually and objectively segmented portions of the continuum of speech signals. These elements, the disjunct chunks of signals have

consistent existence. Let me illustrate this with an example of the spoken sentence /dinobondhubhorbælykəʃirnoditetʃhipniyefikeregelen/. Signal processing gives out the following chunks /dinobondhubhorbæly/, /kəʃirnodite/, /tʃhipniye/ and /ʃikeregelen/. All possible permutations, I_k , of these 's's will give the same message though the texts are different. Let us consider a commutative and transitive operation $\oplus$ on elements of θ such that for $s = s1 \oplus s2$, all the $s, s1, s2 \in \theta$. $\oplus$ is basically a concatenation. The θ is a spoken language (Figure 12) and θ is closed with respect to the operation $\oplus$.

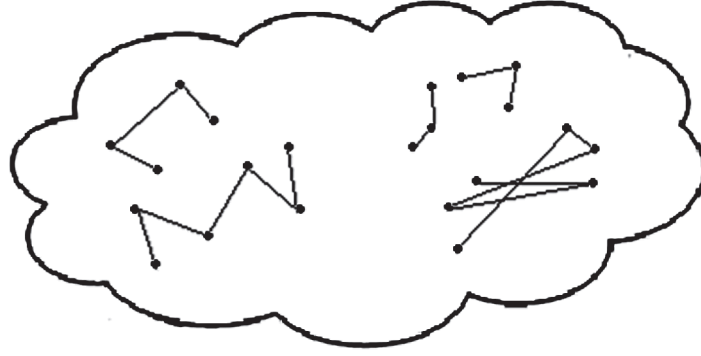


Figure 12: The schematic representation of a tentative model for spoken language

The operation $\oplus$ is a mapping of θ into itself. $\oplus$ on a specific cohesive subset θ_k produces the image subset $I_k \in \theta$ which relates to a specific message. For example let $\theta_k = [\text{emi}, \text{kəl}, \text{kolkətə}, \text{d'əbo}]$. Then a four-fold operation $\oplus$ will produce I_k with cardinality of ${}^4p_4 = 24$. All the 24 elements will have equivalent messages. Some of them are /emikəlkolkətəd'əbo/, /emikəld'əbokolkətə/, /emikolkətəkəld'əbo/, /emikolkətəd'əbokəl/, etc. θ_k is a cohesive set. The cohesiveness of θ_k is important. Figure 13 shows all elements of I_k corresponding to $\theta_k = [\text{emi}, \text{kəl}, \text{kolkətə}, \text{d'əbo}]$.

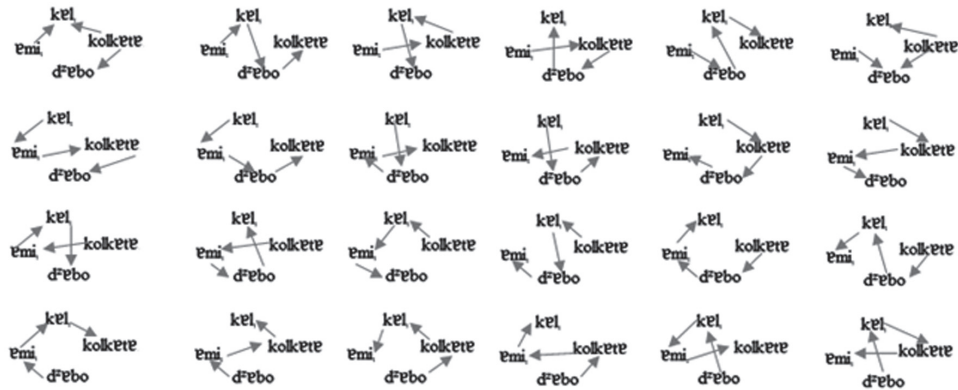


Figure 13: Example of message construction in spoken language

This cohesiveness is fundamental. Let us consider the subset $\theta_l = [\text{v̥mi, goru, poʈi}]$. The 3-fold operation would produce I_l the semantic value of the elements of which in Bangla would be 'FALSE'. This subset is not cohesive.

One may say that the semantic value in normal linguistic sense of an element s_k is the subset I_k generated by all possible iterative operations of $\oplus$ on the cohesive subset ${}^c s_k$, all elements of I_k being 'TRUE'.

The aforesaid is the skeleton of a postulated model for spoken language. If this has any potential I am sure scholars will take up rigorous formalisation and extension to explain various properties of spoken languages.

It is interesting to note that all of the elements of I_K above carry the same overall semantics. This reveals a distinctive property of a natural language, speech. Spoken language transmits basic meanings directly, not mere messages. Let me clarify what I imply by basic meaning through an analogy. While working on emotion Hevner listed 66 English adjectives signifying differentiable emotions. Human-Machine Interaction Network on Emotion (HUMAINE) classifies all of them into 10 basic groups.

The aforesaid formalisation of a spoken language seems to be attractively simple. One may note that θ also contains all possible linguistically defined words, at least embedded in chunks, if not occurring as a single element. I am not sure if it is possible to extract them mechanically from θ . However, it is possible to extract a subset θ_w which consists of all shortest chunks which occur in other chunks. The elements of θ_w may be considered to play a role equivalent to the 'words' in written language.

I believe that in Bangla a fixed order does not exist between three types of phrases namely, subject (S), object (O) and verb (V). In the same vein, this classification for chunks has no relevance here. In such a scenario, I feel, the concepts of 'parts of speech' and 'karaka' relations will also be irrelevant. For other languages, one has to figure out whether this holds any relevance. I shall give an example. Let us consider the spoken message / b̥t̥er̥mmat̥ʰerdz̥ʰoldies̥d̥əb̥h̥et̥k̥h̥et̥b̥h̥el̥ob̥əse/. Chunking of the speech signal provides the following chunks, /b̥t̥er̥m/, /mat̥ʰerdz̥ʰoldie/, /s̥d̥əb̥h̥et̥/ and /k̥h̥et̥b̥h̥el̥ob̥əse/. It is obvious that chunking removes the need of both 'parts of speech' and karaka relations. One may argue however that the chunks may bear the label of 'parts of speech'. But because all

the possible permutations of these chunks convey the same meaning, S, O and V become meaningless and so is the need of 'parts of speech'. What is the grammar here? Whatever I said in this paragraph is my view and may be taken only as an idiosyncrasy. I shall wait till they are disproved.

References

- Chowdhury, S. 2006-7. Concatenative Text-To-Speech Synthesis: A Study on Standard Colloquial Bengali. Ph.D. Thesis. Kolkata: Indian Statistical Institute.
- Dan, T., B. Mukherjee and A. K. Datta. 1993. 'Temporal Approach for Synthesis of Singing (Soprano 1)', Proceedings of Stockholm Music Acoustic Conference (SMAC93), pp.282-287.
- Das Mandal, S., B. Gupta and A. K. Datta. 2006. 'Word Boundary Detection based on Suprasegmental Features: A case study on Bangla Speech', *International Journal on Speech Technology*, Sept. 2006.
- Das Mandal S., B. Gupta and A. K. Datta. 2003. 'Study of Time Domain Shape Parameters and Comparison with Mel Frequency Cepstral Coefficients for Recognition of Bangla Vowels', *Young Horizon – Computing and Informatics*, 1(1): 15-21.
- Das Mandal, S. K. 2007. 'Role Of Shape Parameters In Speech Recognition: A Study On Standard Colloquial Bengali (SCB)'. Ph.D. thesis. Kolkata: Jadavpur University.
- Datta, A. K. 1993. 'Do Ear Perceive Vowels through Formants?', Proceedings of 3rd European Conference on Speech Communication and Technology, Genova, Italy, 21-23 Sept, 1993.
- Datta, A. K. 1991. 'Time domain analysis of speech sounds', XIIeme Cong. Int. Des. Sc. Phone, en-Provence France, vol 3, pp 522-525, 19-24 August, 1991.
- Datta, A. K. 1988. 'Acoustic Phonetics of Non-Nasal Standard Bengali Vowels: A Spectrographic Study', *JIETE*, 34.
- Datta, A. K. and B. Mukherjee, 2011. On the Role of Formants in the Cognition of Vowels and Place of Articulation of Plosives. In: Ystad et al (eds.) *Speech, Sound and Music Processing: Embracing Research in India*. pp. 215-234, LNCS 7172, Springer.
- Datta, A. K. and N. R. Ganguly 1985. 'Behaviour of Terminal frequencies in VC Combination', *Acustica*, 67: 26-33.
- Datta, A. K., N. R. Ganguly and D. Dutta Majumder. 1978. 'Some Studies on Acoustic Features of Telugu Vowels', *Acustica*, 41: 55-64.
- Datta, A. K., N. R. Ganguly and S. Ray. 1980. 'Recognition of Unaspirated Plosives-A Statistical Approach', *IEEE Trans. Acous. Speech and Sig. Process*, ASSP-28 (1): 85-91.
- Dutta Majumder, D. and A. K. Datta. 1968. 'Some Studies on Automatic Speech Coding and Recognition Procedure', *Indian Journal of Physics*, 42.
- Fant, G. 1970. *Acoustic Theory of Speech Production*. Mouton De Gruyter.

- Ohman, S. E. G. 1966. 'Coarticulation in VCV utterances- Spectrographic measurements', *JASA*, 39: 151-168.
- Pfützing Hartmut, R. 2003. Acoustic Correlates of the IPA Vowel Diagram. Paper presented at 15th ICPHS, Barcelona, Spain. August 3-9.
- Sengupta, R., N. Dey, D. Nag and A. K. Datta. 2001. 'Comparative Study of Fractal Behaviour in Quasi-random and Quasi-periodic Speech Wave Map', *Fractals*, 9 (4) 403-414.
- Teagre, H. M. and S. M. Teagre. 1990. A phenomenological model for vowel production in the vocal tract. In: R. G. Daniloff (ed.) *Speech sciences: Recent Advances*. pp 73-109. San Diego: College Hill.

* The lecture was delivered in the event of 'Prof. Suniti Kumar Chatterji Memorial Lecture –III' organized by Kolkata Society for Asian Studies held on May 23, 2016.

Bio-Profile of the Author:

Former Professor
Indian Statistical Institute, Kolkata
Email: dattashoke@gmail.com