Interdisciplinary Approaches to Language Processing

Editors

Denis Burnham, Sudaporn Luksaneeyanawin,
Chris Davis, and Mathieu Lafourcade

Speech as Encoded Language

Arthur S. Abramson

Abstract

I address here several topics that focus on speech as the output of the human vocal tract constrained by a language. Human language apparently differs from all other means of animal communication through three properties: (1) It is unlimited in its scope of reference. (2) It is free from control by identifiable stimuli. (3) It can be shifted into other perceptuomotor modalities. Language generativity has particulate origins in that discrete units are drawn from a set of primitive elements. These units are permuted and combined to yield larger units. The primitive units are meaningless as are some combinations of them. Above a certain level we have meaningful rule-governed combinations of the units. A language is never completely static. A drift from earlier states leads to the rise of regional and social dialects. In our research we must take cognizance of variation and clearly identify the variety that is the object of our study. The human vocal tract can produce a large gamut of sounds for exploitation by languages. Its supraglottal cavities resonate to periodic and aperiodic excitation sources. The spectral resonances (formants) shift with changing tract-configurations, including the possibility of coupling between the oral and nasal cavities. One goal of speech-perception research is to find the information-bearing elements in the signal, the acoustic "cues" to phonemic distinctions. Another is a better understanding of the links between relevant acoustic dimensions and the parameters of physiological control mechanisms.

1. Introduction

As practitioners of rather diverse disciplines that have not always worked together toward common goals, we face the problem of understanding each other well while laying out important questions and some possible answers in the matter of human and machine processing of language and speech. Our worthy organizers, Sudaporn Luksaneeyanawin and Denis Burnham, seemed to think that from my vantage point I could give a broad overview of the object of our concern to the satisfaction of our kind of group. My vantage point, I should explain, is that of a linguist who has devoted his career to experimental phonetics. My research on the production and perception of speech has been done for the past 40 years in the heady, richly interdisciplinary atmosphere of Haskins Laboratories, where linguists, phoneticians, speech and hearing people, experimental psychologists, engineers, physiologists, and physicians collaborate productively.

In this paper my plan, under the influence of my phonetic bias, is to address sever topics that emphasize speech as the output of the human vocal tract under tl constraints of the language being spoken. This will include the particulate nature language structure, variation within language, human vocal capacity, and tl systematic use of acoustic dimensions in production and perception.

2. Linguistic Constraints

When we talk about speech, we normally mean spoken language. That is, utteranc are emitted from a human vocal tract - or a simulated human vocal tract - under tl constraints of a language. There is a linguistic code, e.g., English or Thai, from whic messages are encoded into speech signals. But is there anything special about huma language? After all, much research has been done on the communication systems such higher non-human animals as chimpanzees and dolphins. Perhaps the communication systems and ours are essentially the same on an evolutionar continuum, with ours at the top with its richer, more elaborate nature.

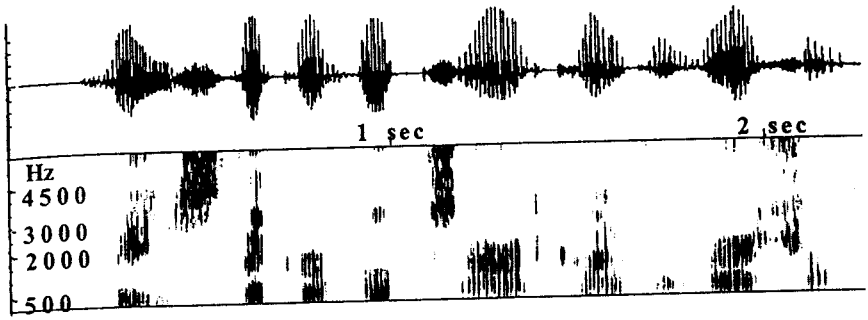
Language as a system of animal communication has three unique propertie (Chomsky, 1966, *passim*; Studdert-Kennedy, 1998). (1) It is unlimited in its scope reference. (2) It is free from control by identifiable stimuli. (3) It is amenable t transduction into alternative perceptuomotor modalities.

Unlimited scope means that language can be used to discuss anything. As far as w know, other animal systems are limited to some dozens of calls for specific purpose Messages generated from a language are not just responses to stimuli or needs in th immediate environment. Thus, one can talk about philosophy while feeling hung pangs and flit easily from one topic to another, with reference to any place, in th past, present, or future. Also, linguistic messages can be encoded not only in speec but also in such other modalities as writing and finger-spelling.

This leads us to the particulate principle in language. Let us begin our consideratio of this topic by examining the utterance depicted in Figure 1. If the reader will be with me and pretend not to have seen the printed matter at the bottom of the figur we can focus for the moment on the waveform at the top. One is probably n surprised to learn that it is the graphic transform of the pressure disturbance caused b a speech signal. Indeed, the pattern of amplitude rises and falls is rather suggestive the results of the opening and closing articulatory gestures that form syllables. A experienced speech researcher can look at the pattern and infer the high probability the kinds of opening and closing gestures, namely, vocalic and consonantal. On could even say something, perhaps more easily with greater magnification of th waveform, about the probable phonetic classes of some of the presumed consonant That would be more or less the limit for most of us.

If we go to the lower display, a sound spectrogram[1] made by fast Fourier transform one can do somewhat better even without being told what the language is. That is, th results of decades of work in speech acoustics enable students of the subject to infe much more about the nature of the articulatory movements and sound sources th yielded this spectrogram. For example, even in this display, which is limited to abou 5000 Hz, we can not only detect alternations between quasi-periodic and aperiodi sound sources, as in the waveform, but also see spectral patterns that are characterist of certain classes of vocalic and consonantal gestures. Indeed, some speech scientist through intensive practice, have become rather adept at "reading" sour

spectrograms.
Beneath the spectrogram is a fairly broad transcription of the utterance in the symbols of the International Phonetic Alphabet. Below that is the conventional orthographic representation of what turns out to be an English sentence.



ðizθɪkəbuksɑɹnt ˌkʰantɹə'vəʃəl

"These thicker books aren't controversial."

Figure 1. Waveform, wide-band sound spectrogram, and broad phonetic transcription of an English sentence spoken by a native of the United States northeastern seaboard.

There is a more revealing way to think about what we have seen in Figure 1. Inspired by the arguments of William Abler (1997) from chemistry, biological inheritance, and language, Michael Studdert-Kennedy (1998; in press) discusses the "particulate origins of language generativity." He explains that in self-diversifying systems, including language, discrete units are drawn from a finite set of primitive elements. These units are permuted and combined to yield larger units above them in a hierarchy. The units thus combined are not lost; they can be recovered by appropriate analysis. For language such an analysis is through the mechanisms of human speech perception.

Let us go back to Figure 1. An old conventional view is that the primitive elements underlying such an utterance are the "sounds of speech" or, perhaps, for the linguist, abstractions of these sounds called phonemes. Some phonologists, linguists who specialize in the study of the sound patterns of language, break the phonemes down into even more primitive units called distinctive features, such as nasality, voicing, and place of articulation. A more recent approach, espoused by some of the Haskins Laboratories researchers, takes the articulatory gesture to be the basic unit in both production (Browman & Goldstein, 1986) and perception (Liberman & Mattingly, 1989; Fowler, 1994). As for the perception of speech, the matter remains controversial (e.g., Lindblom, 1996). Indeed, even among the proponents of the role of the gesture in perception, there are rather different theoretical stances.[2] This, however, need not detain us here. Rather, let us say that my broad phonetic transcription in Figure 1 may be taken to point to the underlying primitive elements, be they phonemes, features, or gestures. Whichever outlook is convincing to you, it has to be said that these units are meaningless. Thus, for example, the [z] at the end of

these is by itself meaningless, whether it is viewed primarily as a manifestation of an abstract phoneme /z/, a movement of the front of the tongue to form a narrow constriction in the alveolopalatal region together with phonatory approximation of the vocal folds, or an acoustic disturbance caused aerodynamically by turbulent air in a narrow orifice accompanied by pulses from the laryngeal source.

Combinations of the basic units can be meaningless too. In Figure 1 the sequences [θɪ], [bʊ], and [ks] are meaningless. Since this is, after all, an utterance of a possible English sentence, there are certainly meaningful combinations of basic units in it. This is because meaning has been assigned to many combinations in any language by social convention. Such meaningful combinations become the morphemes and words of the language. In our sentence, *thicker* is recognised as a word but a word composed of two morphemes (minimal meaningful units), *thick*, which can also stand by itself as a word, and the ending *-er*, which carries the meaning of 'more so than something else' and cannot stand by itself. Further complexity is seen in interactions between morphology and phonology. In our sentence a more deliberate utterance would have yielded [nɑt] instead of [n̪t]; the two of them are rule-governed variants of the morpheme *not*. Another example of such rule-governed behavior is the primary stress marked on the fourth syllable of *controversial*. For the speaker, indeed for American English by and large, in the uninflected noun *controversy*, the primary stress is on the first syllable.[3]

Above the level of the morpheme and word we enter the domain of syntax, that part of the grammar that controls how words are combined into sentences. If the speaker in Figure 1 had meant only one book, he would have started the sentence with the word *this*. Such a choice would have entailed two other differences, yielding "This thicker book isn't controversial."

From the foregoing examples we can see that these larger units in the hierarchy have structures and functions beyond, and more diverse than, those of their constituents (Studdert-Kennedy, 1998). Of course, if I had composed the message of the sentence in Figure 1 in Thai, important cross-language differences would be striking; nevertheless, for anyone who cares to try it, the particulate principle will still be apparent.

3. Dialects and Related Languages

Over the course of its history, a language is not likely to remain uniform in time and space. Linguists have long agreed that language is always in a state of flux. That is, gradual change is always going on, even though at any given moment it may not be very noticeable. Even within a small, closely knit speech community there is a drift away from earlier states. The drift is likely to include slow shifts over the generations in articulatory habits, changes in vocabulary, deviations in the meanings of words, the coining of new words and the loss of some old ones, and alterations in syntactic patterns.

The causes of these changes are more often than not hard to identify. Some may have to do with fad and fashion, as in the imitation of the linguistic idiosyncrasies of a popular person. Now if the language is spread over a large territory with poor communication between parts of that territory, such changes will probably not be completely the same in all the subgroups. This gives rise to regional dialects. In addition, if some of the dialects are in close contact with other languages along the

borders of the territory, those dialects are likely to undergo further changes through interference from the languages in contact with them by way of bilingual speakers. This kind of diachronic splitting into regional dialects may go on long enough to lead to the recognition of some of the resulting varieties as separate but related languages descended from a common ancestor, a proto-language. Thus, Thai (Siamese), Lao, Shan, Phuan, and a number of other languages are members of the Tai family. Each of its members, in turn, has dialects.

To all the foregoing factors should be added the social variable. In addition to spatial and temporal separation of populations of speakers, there can also be social separation. If a community is marked by the existence of social classes, the barriers between them, just as with a wide river or a high mountain range, can lead to the emergence of dialects, in this case social dialects.

For those of us who try to approach the study of language and speech scientifically, the question arises as to how to decide when two or more varieties of a language are no longer merely dialects of that language but separate related languages. Alas, such a distinction appears to defy any kind of formal quantitative analysis that has been tried. The question is largely decided on sociocultural or nationalistic grounds. Southern Thai of Thailand and Central Thai, on which Standard Thai is based, are treated as regional dialects of "Thai," yet they may be as different, if we could but apply the right metric, as Central Thai and Lao, which we all regard as separate languages.

In our study of language and speech we must take into account the existence of variation. If you are working on English, which of the great number of "Englishes" is it? The English of England, Scotland, Wales, Ireland, the United States, Canada, New Zealand, Australia? And with the choice of the English of any one of them, we cannot ignore its own lack of uniformity. In our work we must recognize the fact of variation and then make clear what variety, what regional or social dialect, is the object of our study. Such a decision may be a matter of convenience or practicality, but it should be expressed. Of course, if we are lucky or clever enough, we may ultimately be able to extract a common core of all the dialects and show its relevance to our research goals. This is a problem not only in speech research but also in linguistic analysis.

4. Speech Production

Let us turn now to the sound-making capacity of the human vocal tract, which has stimulated not only descriptive work but many provocative attempts at developing satisfactory theories (see Lõfqvist, 1997). No matter how outlandish it may sound, even when we do not understand it, an utterance strikes the ear as the vocal output of a person speaking a language. Of course, we may be wrong on two counts. It could be the output of an excellent speech synthesis program that passes for a human being talking. Also, whether a natural or simulated talker, it could be pronouncing gibberish. (When investigators record nonsense syllables for experiments, they do so according to the phonological norms of the language of concern.) A brief outline of the production of speech may be helpful here.

Speech is normally formed on outgoing air from the lungs.[4] If the vocal folds of the larynx are drawn close enough together with appropriate muscular contractions, the subglottal pressure of the air stream, coupled with the Bernoulli effect of air passing through a narrow channel, will set the folds into periodic or quasiperiodic motion. The rather regularly emitted pulses from the glottis are heard as voice. Our very fine control of laryngeal mechanisms and aerodynamic forces includes the ability to vary

both the repetition rate of these pulses to yield the impression of changing pitch ar
the amplitude of the voice source to contribute to the impression of changing loudne
or prominence (Dickson & Maue-Dickson, 1982, pp. 130–136). Indeed, fundament
frequency and amplitude may interact to affect both percepts, pitch and loudness.
Aside from glottal pulsing, speech also makes use of aperiodic sound sources. One
them is turbulence formed aerodynamically either at the somewhat open glottis, as
aspiration, or in an articulatory constriction somewhere in the supraglottal tract, as
voiceless fricative consonants. Another is a transient, a noisy burst caused by t
sudden release of air under pressure behind a complete closure, as in plosi
consonants. Various combinations of these sound sources occur in speech, e.g., glott
pulsing and local turbulence in voiced fricatives.
By and large, the sounds of speech that we call vowels and consonants are formed
supraglottal articulatory states and movements regulated by coordinated contractio
and relaxations of the muscles of the pharynx, tongue, jaw, and lips (Gracco, 199
These articulatory gestures are themselves coordinated with the switching on and
of the possible sound sources. Examples of consonantal gestures formed at t
boundary between the supraglottal and subglottal tracts are the medial compression
the vocal folds for a glottal stop and lowering of the larynx for imploded stops.
A cautionary note on the matter of the segmentation of speech may not be amiss he
For the linguist dealing with abstract units that in successively more complicat
sequences and sets are seen as furnishing the building blocks of language, it is usef
and perhaps necessary to view the sounds of speech as beads on a string. The mo
control of speech production, however, is very different; it is dynamic. Two and ev
more "sounds" in such a string are normally coarticulated or coproduced. The to
has been much explored and discussed (e.g., Recasens, 1987; Fowler & Saltzm
1993; Gandour, Potisuk, & Dechongkit, 1994). That is, a traditional lin
segmentation may have its uses as a kind of "spelling" of morphemes and words,
it seems to have little to do with the production and perception of speech.[5]
All of the foregoing neuromotor activity, of course, leads to the radiation of
acoustic speech signal from the mouth of the speaker. The cavities of the supraglot
vocal tract, varying in shape with articulation, resonate to the excitation sources.
illustrated in the spectrogram of Figure 1, these resonances form the spectra of
sound waves travelling through the air or some other medium. This is the essence
the source-filter theory of speech production (Fant, 1960; Flanagan, 1965) that
motivated research in speech acoustics for decades.
A revealing way of thinking about the acoustic output of speech is through
concept of the carrier nature of speech presented by Homer Dudley, a Bell Telepho
Laboratories engineer (Dudley, 1940). He says (p. 495), "Speech is like a radio w
in that information is transmitted over a suitably chosen carrier. In fact the mod
radio broadcast system is but an electrical analogue of man's acoustic broadc
system supplied by nature." The speaker's message appears as "muscular w
motions in the vocal tract." These motions, however, cannot be heard, because t
occur at too slow a rate, but the speaker's voiced and voiceless carriers are q
audible. The carrier chosen is modulated, through selective transmission, by
gestural information containing the message, yielding audible speech signals. Thi
likened to the modulating of the distinctive carrier wave of a radio station with
content of a radio program—speech, music—for transmission to radio receivers

the ears of listeners.
Dudley's reasoning led to his invention of the vocoder, "so named because it handles the speech in a coded form" (p. 505). With a bank of filters at the input and corresponding modulators at the output, as well as buzz and hiss excitation sources, one could analyse an utterance acoustically and selectively transmit spectral properties to the modulators for imprinting on the carriers to yield a resynthesized version of the original speech. The vocoder, originally meant for bandwidth compression in long-distance telecommunications, became a powerful tool in speech research. For example, my own early research on the five phonemically distinct tones of Standard Thai (Abramson, 1962) included perceptual experiments with variants of fundamental-frequency (F_0) contours that might suffice for identification of words minimally differentiated by tone. I did this by controlling the F_0 of the simulated voice (buzz) of the vocoder and modulating the resulting carrier with the spectral information retained from the original input of natural speech. Native speakers identified randomised sequences of the stimuli as Thai words, demonstrating the adequacy of five "ideal" contours for the tones. Let this example be a point at which to turn to a brief look at studies of speech perception.

5. Speech Perception

The search for the information-bearing elements of speech, the acoustic cues (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Delattre, Liberman, Cooper, & Gerstman, 1952) has traditionally started with hypotheses derived from acoustic analysis of phonemic distinctions of interest to the investigator (Fry, 1979). The indispensable acoustic analysis of speech was greatly facilitated when the sound spectrograph, newly invented at Bell Telephone Laboratories, became available for non-military use after the Second World War (Koenig, Dunn, & Lacy, 1946; Joos, 1948). Many of us grew up in the field of experimental phonetics watching a stylus burn patterns into "teledeltos" paper wrapped around the rapidly spinning drum of the spectrograph. Now, of course, we generally use computer programs for the purpose.
Acoustic analysis typically yields a welter of details about the acoustic phonetics of utterances differentiated by the distinction in question. Much of that information turns out to have little or no bearing on the distinction, although it may tell us other things, such as the sex and approximate age of the speaker or, for an acquaintance, the identity of the speaker. From the point of view of both the speech scientist and the communications engineer there is the problem of normalisation across acoustic differences in the "same" utterance caused by differences in the dimensions of the vocal tract from person to person within the same sex and age range and even more so between the sexes and between adults and children. Somehow the ear-brain of the human being does this normalising with little or no trouble.
For perceptual research it is necessary to form testable hypotheses as to what properties of the signal provide acoustic cues to the distinction of interest. Typically, especially as the field developed, such a hypothesis would seize upon an acoustic dimension that could plausibly be a function of a crucial articulatory state or movement. Testing is then done by making incremental changes along that dimension and playing the resulting stimuli to native speakers of the language for labelling or discrimination. In the early decades of such experimentation, it was difficult or impossible to create such stimuli with natural speech while not allowing anything to change but the parameter of interest, so terminal analog synthesisers were used under

the control of carefully drawn schematic spectrograms or, later, programme instructions to the parameters of the synthesiser. Thus, one specified excitation type and the times of switching between them, the F_0 and amplitude contours of the voic source, the formant frequencies and amplitudes for resonances of vowels, formar transitions, and all other spectral and temporal properties that seemed relevant.

In recent years investigators have been turning to the newer articulatory synthesiser (Mermelstein, 1973). Although still in need of improvement, they provide a parameters a set of simulated articulators such as tongue body, tongue tip, velum, lips and mandible. Thus, one can base hypotheses on the role of articulatory gesture rather than aspects of the acoustic signal. Of course, the articulatory synthesise makes use of algorithms for calculating area functions of varying configurations o the simulated vocal tract and deriving acoustic outputs therefrom.

Nowadays, instead of using pure synthesis, it is possible to make carefully controlled spectral and temporal changes in natural speech to make stimuli for experiments in speech perception. A current such line of research of mine, to be outlined briefly here, is a hunt for the acoustic cues that distinguish word-initial distinctively short and long consonants in Pattani Malay. The language is extraordinary in that this distinction is relevant in word-initial position for all classes of consonants. Measurements revealed (Abramson, 1987) that on average the closures of long consonants were three times longer than those of short consonants. An example of a minimal pair is shown in Figure 2. In it the /b:/- closure of the second word is 2.6 times longer than the /b/- closure of the first word.

In a control test (Abramson, 1986) I demonstrated the robustness of the distinction in isolated words by showing that when either closure excitation or, for voiceless stops, intervocalic position makes the relative closure durations audible to native speakers, they perceive the difference and identify the words correctly. Even so, that alone did not prove that duration itself is an important cue to the distinction. I used closure duration as the variable in two experiments with voice-excited stop closures in the word pairs /labɔ/ 'profit' vs. /l:abɔ/ 'spider' and /gamɔʔ/ 'approximately' vs. /g:amɔʔ/ 'to be shy.' Working with the digitized files, I shortened the original long closure of each pair in 10-msec steps to yield as many shortenings as would make the last one in the series a bit shorter than the original short consonant. The subjects readily separated the randomized stimuli into the two length categories with an increase in ambiguity around the middle of the range.

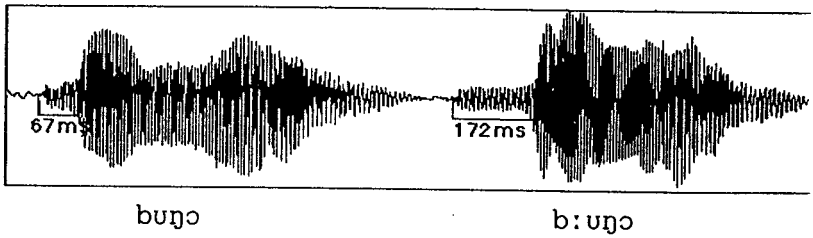


Figure 2. Waveforms of /buŋɔ/ 'flower' and /b:uŋɔ/ 'to bloom,' showing the durations of the voiced closures. Recorded separately but put together in the figure. Adapted from Abramson (in press) Fig. 1.

For voiceless stops the task is more complicated. Although the control tests had shown that subjects could identify words beginning with them at rates far better than

chance, in utterance-initial position there is no way—at least in the acoustic domain—of finding the beginnings of these silent spans and making changes in them. Instead, I had my speaker embed the words /paka/ 'to use' and /paka/ 'usable' after a word ending in a vowel in a carrier sentence that was recorded fluently with no pause within it. The closure durations could then be marked off in spectrograms and waveforms. For this pair I ran two experiments. In one I shortened the closure of the long consonant in its carrier sentence as already described. In the other I lengthened the closure of the short consonant by inserting a series of 10-msec stretches of silence, making the longest variant a bit longer than the original short closure.

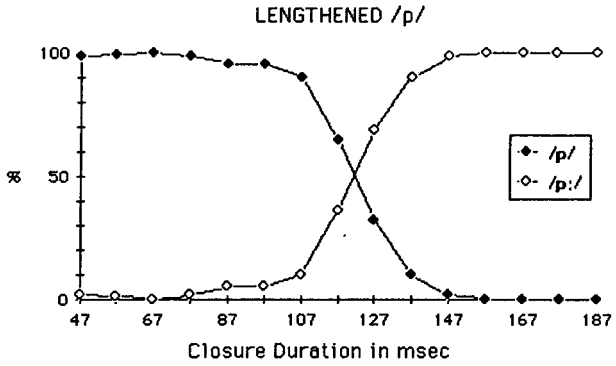


Figure 3. Identification functions for original short /p/ and lengthened variants. (Adapted from Abramson (1986), Figure 4.

The results of the experiment with the lengthened intervocalic voiceless closures are shown in Figure 3, which resembles the identification functions for all the experiments described. That is, in all of them relative duration proved to be a sufficient and powerful cue to the distinction. In the two experiments with voiceless stops, however, a provocative ancillary result appeared. The 50% crossover point in the identification function for the lengthened variants is 16 msec later than for the shortened variants. This highly significant difference [$F(1, 20) = 27.5$, $p < 0.0001$] implies that even in intervocalic position some property other than relative duration contributes somewhat to the distinction. Could that be the sole property that enables speakers to hear the distinction even when voiceless stops are in utterance-initial position?

Recalling that the linguist Christopher Court, who has done a great deal of field work with Pattani Malay, has wondered whether a prosodic distinction is not gradually becoming concomitant with the length distinction, I listened carefully to many recordings of such words in the speech of several informants. I, too, had the impression of greater salience in the first syllable of disyllabic words beginning in a long consonant, perhaps more so for stops than for other categories.

Figure 4 illustrates the results of analysis of a corpus of utterances for phonetic properties that could underlie an impression of an accentual difference between

disyllabic words beginning with long and short consonants Let us limit our attention here to amplitude (Abramson, 1987) and fundamental frequency (Abramson, in press), both of which are significantly greater in the first syllables of words beginning with long consonants, especially for voiceless stops, as exemplified in Figure 4.

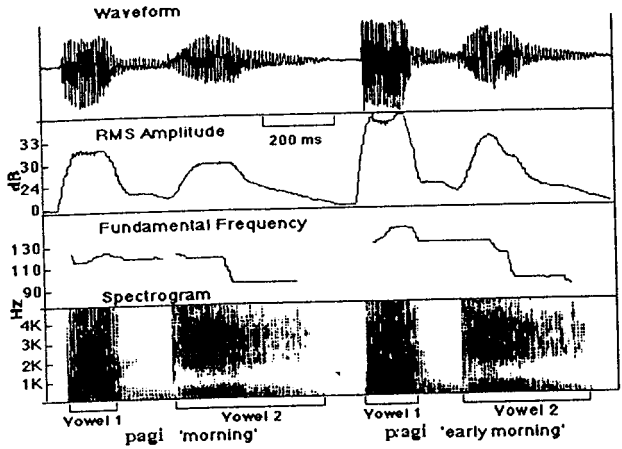


Figure 4. Acoustic analysis. The space between Vowel 1 and Vowel 2 in each spectrogram is the voiced /g/ closure. Recorded separately but put together for analysis and display. Adapted from Abramson (in press), Figure 2.

Having obtained my root-mean-square amplitude values first, I turned once again to natural speech to begin this phase of my research with tests of the efficacy of relative amplitude in the identification of voiceless stops (Abramson, 1991). In one set of experiments I pitted relative amplitude against relative closure duration in three minimal pairs of words with short and long labial, dental, and velar stops embedded after a vowel in a carrier sentence. The closure durations were varied incrementally as before. The first syllable of each variant of the original long stop was decreased in amplitude in five 2-dB steps; the first syllable of each variant of the original short stop was increased in five 2-dB steps. The results of the perception tests showed relative duration to be the clearly dominant cue, although the crossover points between the labeling categories were significantly different for the two conditions, indicating that amplitude makes at least a small contribution. In another pair of experiments I tested the perceptual efficacy of amplitude alone in isolated words, the pair /pagi/:/p:agi/, thus without the help of closure-duration variants. Here, for increments of amplitude on tokens of the word with a short stop and decrements on tokens of the word with a long stop, the effect was small. The identification curves in both tests converged but never crossed each other. Clearly it is not a strong cue by itself.

It is likely that in nature relative amplitude works together with other stronger cues to give the impression of prominence. The property with the most promise shown in Figure 4 seemed to be fundamental frequency (F_0). In new stimuli I have raised the F_0 on the first syllable of words with short stops in half-octave steps through a range of three octaves and lowered it in the same way on the first syllable of words beginning

with long stops. I have played these variants to a large group of Pattani Malay speakers for word identification. So far, a quick look at the data suggests greater efficacy for F_0 than for amplitude. A detailed report awaits full compilation and analysis of the data.

There is also a considerable literature of the perceptual role of relative duration in languages with vowel-length distinctions. My own involvement in such research has been largely limited to Thai (Abramson, 1962, 1974), although, in this domain too, I have been intrigued by the possible efficacy of a concomitant property, spectral differences, in distinguishing short and long vowels in Swedish (Hadding-Koch & Abramson, 1964) and Thai (Abramson & Ren, 1990).

6. Conclusion

This paper tries to give a rather broad outlook on speech, the subject of so much rich interdisciplinary research in our times. Speech, complex as it is physiologically, aerodynamically, acoustically, auditorily, and perceptually, is itself the output of a language with its internal structure. Listening to an utterance, the listener makes use of "bottom-up" information from the basic elements combined hierarchically into larger and larger units, as well as "top-down" information from the higher levels of the grammar, the lexicon, and semantics. Of course, the situational context can help too. The transmission of linguistic information can be burdened by the existence of a range of regional and social dialects of a language. In the extreme case, a listener's inability to understand an utterance may lead to the conclusion that a foreign language is being spoken, rather than a very different and unfamiliar dialect of his own language.

Language as observed in all cultures has characteristics that, as far as we know, make it uniquely human. It is unlimited in its scope of reference, free from control by identifiable stimuli, and amenable to transduction into perceptuomotor modalities other than speech.

The study of the production and perception of speech has depended heavily upon the close cooperation of linguists, phoneticians, psychologists, engineers, computer scientists, physicians, and physiologists. More work remains to be done on the mechanisms of production and perception. New instrumentation used in the light of newer ways of thinking about motor control and coordination will lead us to a better understanding of the planning and execution of an utterance. The quest for the acoustic cues for speech perception has gone very far, but there are still some that we do not have under full control. Such research will be aided by improvement in our knowledge of articulatory gestures and their acoustic correlates. Some of us have participated in extensive work on the differential use by a variety of languages of the same articulatory and acoustic dimensions as befitting their separate phonologies. A very stimulating and revealing overview of the development of research into the perception of speech is given in a recent book by Alvin M. Liberman (1996). It traces the evolution of thinking by this pioneer and leader in the field to his current gestural theory of speech perception, which, controversial though it may be, merits close attention. Indeed, much of the interest of experimental psychologists in speech research stems from their concern with higher-order aspects of human cognition and behavior.

[1] A sound spectrogram gives a graphic display of the acoustic spectrum over tin That is, patterns of energy are shown at their frequency locations (vertical axis) they change or remain steady over time (horizontal axis). The darkness of a ba grossly represents its relative intensity.

[2] An excellent and very clear explanation of the direct realist account and the mot theory is given by Catherine T. Best (1995). She compares these two gestur approaches with the traditional psychoacoustic approach.

[3] In other dialects of English the rule is different. In *controversy* the major stress is o the second syllable, and the vowel in it is of the type /ɑ/.

[4] Some languages make phonological use of momentary shifts to in-going air, as i the imploded stops of Sindhi.

[5] Just what bearing all this has on the question of how morphemes and words ar represented in the mental lexicon is not yet clear. Some will claim that the ready application of the alphabetic principle to so many and such varied languages argues for the psychological reality of the phoneme or a unit like it.

Acknowledgement

Much of the research described in this paper has been supported by two NIH grants to Haskins Laboratories, HD01994 and DC 02717. The University of Connecticut Research Foundation provided funding for the author's travel to Thailand to participate in the International Workshop on Human and Computer Processing of Language and Speech and to do related research there for three months.

References

Abler, W. L. (1997). Gene, language, number: The particulate principle in nature. *Evolutionary Theory, 11,* 237-248.

Abramson, A. S. (1962). *The vowels and tones of standard Thai: Acoustical Measurements and Experiments.* Indiana University Research Center in Anthropology, Folklore and Linguistics *International Journal of American Linguistics, 28, No. 2, Part III*], Pub. 20, Bloomington.

Abramson, A. S. (1986). The perception of word-initial consonant length: Pattani Malay. *Journal of the International Phonetic Association, 16,* 8–16.

Abramson, A.S. (1987). Word-initial consonant length in Pattani Malay. *Proceedings of the XIth International Congress of Phonetic Sciences, Vol. 6* (pp. 68–70). Tallinn: Academy of Sciences of the Estonian S.S.R.

Abramson, A.S. (1991). Amplitude as a cue to word-initial consonant length: Pattani Malay. In *Proceedings of the XIIth International Congress of Phonetic Sciences, Vol. 3* (pp. 98–101). Aix-en-Provence: Université de Provence.

Abramson, A. S. (in press). The complex acoustic output of a single articulatory gesture: Pattani Malay word-initial consonant length. In *Proceedings of the 4th International Conference on Southeast Asian Linguistics.* Tempe,

Arizona: Southeast Asian Studies Publishing Program, Arizona State University.

Abramson, A.S., & Ren, N. (1990). Distinctive vowel length: Duration vs. spectrum in Thai. *Journal of Phonetics, 18,* 79–92.

Best, C.T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*(pp. 171–204). Baltimore: York Press.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. In C. Ewan, & J. Anderson (Eds.), *Phonology yearbook 3* (pp. 219–252). Cambridge: Cambridge University Press.

Chomsky, N. (1966). *Cartesian linguistics.* New York: Harper & Row.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., & Gerstman, L.J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America, 24,* 597–606.

Delattre, P., Liberman, A.M., Cooper, F.S., & Gerstman L.J. (1952). An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word, 8,* 195–210.

Dickson, D.R., & Maue-Dickson, W. (1982). *Anatomical and physiological bases of speech.* Boston: Little, Brown and Co.

Dudley, H. (1940). The carrier nature of speech. *Bell System Technical Journal, 19,* 495–515.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton. [2nd ed., 1970].

Flanagan, J. L. (1965). *Speech analysis, synthesis and perception.* New York: Academic.

Fowler, C. A. (1994). Speech perception: Direct realist theory. In R. E. Asher (ed.), *Encyclopedia of language and linguistics* (pp. 4199-4203). New York: Pergamon.

Fowler, C.A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36,* 171–195.

Fry, D.B. (1979). *The physics of speech.* Cambridge: Cambridge University Press.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics, 22,* 477-492.

Gracco, V. (1990). Characteristics of speech as a motor system. In G. Hammond (Ed.), *Cerebral control of speech and limb movements* (pp. 3–28). Amsterdam: Elsevier.

Hadding-Koch, K., & Abramson, A.S. (1964). Duration versus spectrum in Swedish vowels: Some perceptual experiments. *Studia Linguistica, 18,* 94–107.

Joos, M. (1948). *Acoustic phonetics* (Language Monograph No. 23). Baltimore: Linguistic Society of America.

Koenig, W., Dunn, H.K., & Lacy, L.Y. (1946). The sound spectrograph. *Journal of the Acoustical Society of America, 18,* 19–49.

Liberman, A. M. (1996). *Speech: A special code.* Cambridge, Mass. & London: MIT Press.

Liberman, A., & Mattingly, I. (1989). A specialization for speech perception. *Science, 243,* 489-494.

Lindblom, B. (1996). Role of articulation in speech perception: Clue production. *Journal of the Acoustical Society of America, 99,* 1683-1(

Löfqvist, A. (1997). Theories and models of speech production. In W. Hardca Laver (Eds.), *The handbook of phonetic science* (pp. 405–426). Blackwell.

Mermelstein, P. (1973). Articulatory model for the study of speech pro *Journal of the Acoustical Society of America, 53,* 1070-1082.

Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulator in Catalan and Spanish VCV sequences. *Journal of Phonetics, 15,* 29

Studdert-Kennedy, M. (1998). The particulate origins of language generativi syllable to gesture. in J. Hurford, M. Studdert-Kennedy, & C. Knigh *Approaches to the evolution of language: Social and cognitive ba* 202-221). Cambridge: Cambridge University Press.

Studdert-Kennedy, M. (in press). Evolutionary implications of the pa principle: Imitation and the dissociation of phonetic form from s function. In C. Knight, M. Studdert-Kennedy, & J. Hurford, *Evoltionary emergence of language: Social function and then or linguistic form.* Cambridge: Cambridge University Press.