AN ATTACK ON THE PROBLEMS OF SPEECH ANALYSIS AND SYNTHESIS

WITH THE POWER OF AN ON-LINE SYSTEM

Glen J. Culler
University of California, Santa Barbara
Santa Barbara, California

Summary

How hard a problem can you solve? The answer to this question is no longer fixed by the solver's education and intelligence alone. We are already in an era where very substantial and maneuverable on-line systems can greatly enhance our capabilities. Consequently, the difficulty of a problem must now be measured by its distance from the combined powers of a man-machine problem-solving union. Furthermore, we wish to allow the starting man-machine combination to be used experimentally to develop an extended machine of increasing power in the direction required for the solution of a given hard problem. This is by no means just hand waving, for such an arrangement has already been used in a fundamental attack on the problems of speech analysis and synthesis. The substance of this writing consists of a simple description of the developing system with motivation for its development shown by direct illustrations of the structure of sounds in speech deduced through its use.

Introduction

The development of interactive software and hardware to permit convenient man-machine communication has been underway for approximately ten years. Hardware which provides the channels for this communication necessarily relates to man's I/O, his senses. The software supporting this communication must consist of two parts: one which extracts the programmatic implications from the man's inputs to the computer system and the other which generates outputs perceivable by man's senses. The nature of these outputs depends upon the computer's programmatic perception and response to its own internal data, program, and status configurations. The use of speech as the mode of communication has become urgently attractive to many of us engaged in building interactive systems. There are many reasons for this concern: man's ease in chattering, the freedom to do manual things in parallel with speaking or listening, and the universal character of our existing communication systems. The required hardware is trivial; it need only consist of a modest microphone and speaker system with an ordinary audio-amplifier which is attached to the A/D and D/A input-output of the computer to be used. The problem lies in the two part software. Here is an outline of our approach to this problem:

1. ASCON WAVE FUNCTIONS
   Determine a mathematical formulation natural for the representation of wave forms observed as microphone outputs during speech.

2. ASCON TRANSFORM
   Develop algorithms required to transform the digital representatio' of the microphone signal into the parameters of the mathematical formulation.

3. ASCON INVERTERS
   Develop algorithms required to transform lists of mathematical parameters into the digital form required by the b/A which drives the speaker system.

4. PHONE RECOGNIZER
   Classify the parameters actually occuring in speech output as functions of the elements of speech.

5. PHONE SYNTHESIZER
   Generate parameter strings to drive the output system to produce high quality program controlled speech.

6. LINGUISTIC FORMATTER
   Classify phone sequences to determine phonetic, phonemic, and prosodic content and list an encoded representation as a linguistic string.

7. VOICE SYNTHESIZER
   From linguistic strings, construct phone sequences to drive the phone synthesizer.

8. LINGUISTIC ANALYZER
   Extract meaning from linguistic strings.

9. LINGUISTIC SYNTHESIZER
   Construct linguistic strings from programmatic meaning.

At the present time we can announce excellent success in (l.), because we already have sufficiently good success in (2.) and (3.) which is, of course, the first criterion for the adequacy of (l.). We have completed successful experiments in the pair *(4.)* and (5.) for a restricted set of speech elements and are currently engaged in completing this part of our effort. We can give a

very tentative and premature definition of 'Voice" and "linguistic strings" in relation to (6.) and (7.). Only after (4.) and (5.) are complete can we make firm specifications for these definitions. At present, consideration of (8.) and (9.) are beyond the scope of our development.

The Speech On-Line gystem

The initial hardware complex started with an inter-connection of an early form of the U.C.S.B. on-line system with a process control computer. The configuration consisted of:

1. RW-400 polymorphic system
2. IBM-1800 mod II process control computer

The RW-400 system was contributed to U.C. S.B. by the Bunker-Ramo Corporation for further investigation of on-line systems.

The IBM-1800 system was made available on our ARPA project, Contract # AF 19(628)-600l+ by the Department of Electrical Engineering at U.C. S.B.

3. Tektronix 611 , 11 storage tube
4. Microswitch operator-operand keyboard

which were combined as shown in Figure 1.

The first form of our speech station was reported in [l.] at "The Conference for Interactive Systems", August 1967. The principal hardware change since that time has been the addition of Bill Proctor's "Nanohumper"*, which is the right way to tackle the problem of broad band A/D conversion. Our major task has been the experimental

"The Tektronix 611 was donated to U.C.S.B. by the Tektronix Corporation to provide early experience in its use.

The operator-operand keyboard was developed at U.C.S.B. under ARPA contract # AF 19(628)-600U as the input portion of a classroom system.

Records time and voltage of fixed voltage level transitions. Thus the volume of data output is proportional to frequency rather than time.

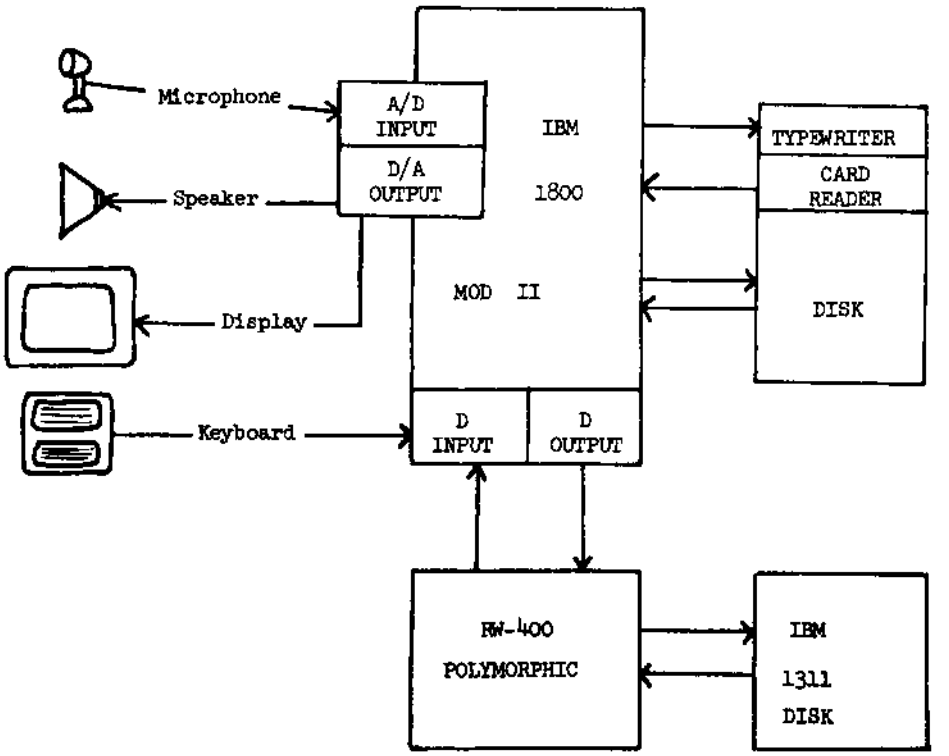


Figure 1. System layout for an experimental on-line speech system.

construction of software required for the investigation of (1.) and its validification by (2.) and (3.)- Roughly speaking, it consists of a set of operations which process a long data list and a set of others which process a long parameter list in the process control computer which is supervised by the early U.C.S.B. on-line system running in the RW-4OO system. For a detailed description of this system and its operations, the reader is referred to the Final Report ARPA Contract # AF 19(628)-6004.

Elementary Sounds

After some initial investigation of the inadequacies of both Fourier series and transforms as mathematical frames for representing the voltage signal output by a microphone during the speaking process, we turned to wave function representations of limited time duration. When one of these wave functions is given as the specified voltage used to drive a voice coil on an audio speaker, the result is an "elementary sound". Sounds of arbitrary complexity can be constructed by superposition of these elementary sounds. The extent to which all speaking voices can be duplicated in this manner is a measure of the completeness of these sound elements as a basis for speech. The simplicity of the resulting representations of live speaking voices in terms of elementary sounds is a measure of the wisdom of selection of that particular basis, for simplicity of representation carries continuing value as we move up the ladder of efforts outlined above. The parameters used to characterize the elementary sounds can be motivated by considering the pressure wave function. This pressure wave is a disturbance propagating through the medium surrounding the speaker and can be recorded by a standing observer with a microphone. Assuming the speaker and microphone are of excellect fidelity, we can thus obtain an experimental recording of an elementary sound. In the coordinate system of the observer, the disturbance will have an amplitude A, will last for a certain time span S, and will have associated with its occurrence a center time C half way through the disturbance. These are called the global parameters of the elementary sound, since they depend only upon its traveling envelope. Now, within this traveling envelope, we characterize the pressure oscillations by N and 0, where N is the number of oscillations recorded during the time interval S and 0 is the phase of the internal oscillation at t = C. As a means to

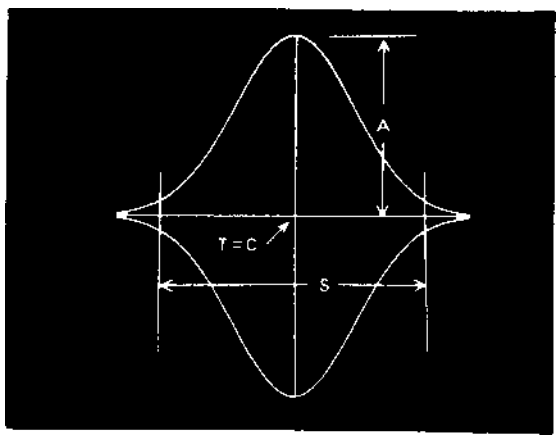
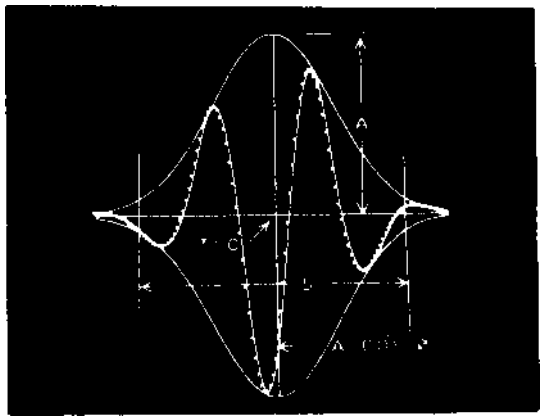


Figure 2. Envelope of an elementary sound.



A = amplitude

S = time span

C = traveling center time

0 = oscillation phase

N = oscillation order

Figure 3. Graph of pressure wave of an elementary sound with its ASC0N parameters.

familiarize the reader with the visual aspects of our wave functions and their ASC0N parameters, we invite you to try estimating all parameters for those wave functions shown below. Correct answers are given at the end of the paper, A in volts, S in milliseconds and 0 in radians.

Elementary Strings: treble, mid, and bass

The speaking vcices of man are so personalized and rich in frequeny range that one despairs early of hoping to c aracterize speech in terms of statical averages o- frequences, extremes, zero crossings, etc. Instead, we elect to carry out a detailed reduction of each speech expression to the elementary sounds of which it is composed, thus deferring sound recognition and interpretation processes as shown in the outline above. This is the purpose of the ASC0N-transform of (2.). With classical transforms that are defined as mathematical processes, one can always state a transformation formula in explicit terms. Doing

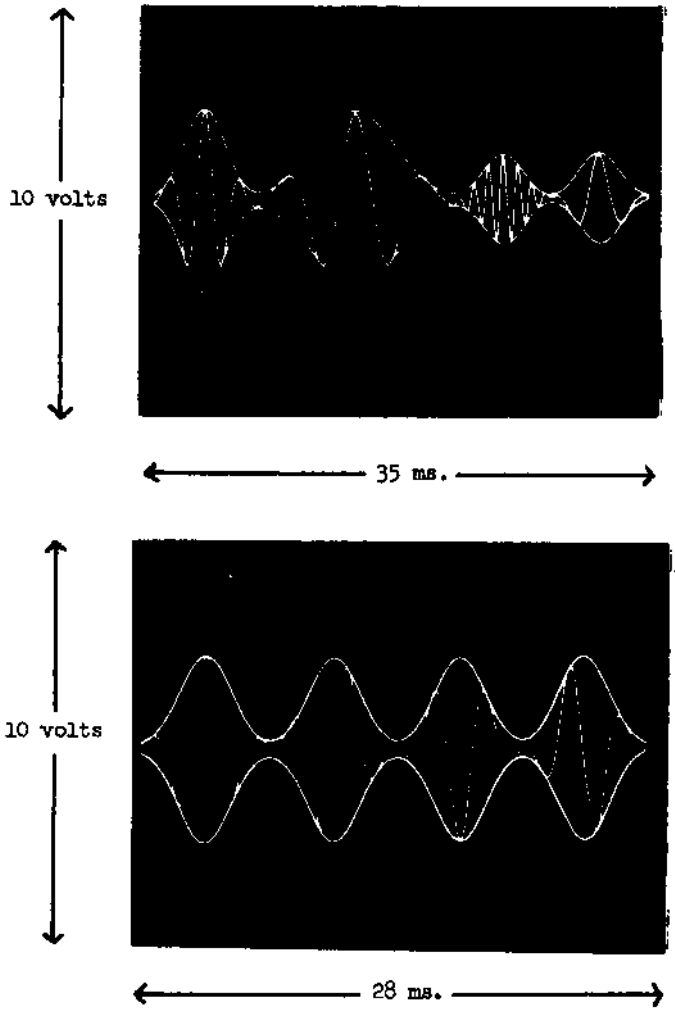


Figure 4. Wave functions of some typical elementary sounds in their envelopes.

so, however, does not remove the need for an algorithmic computational method, since these formulae frequently are not formulated for computational requirements but rather for descriptive mathematical expression. In the case of our ASCON transform, we can state an algorithmic method for computing the transform, although no mathematical formula is known. The computer effort required to carry out this algorithm is greatly benefited by some initial reduction of complexity in speech data. For us, this reduction consists of decomposing the speech data into strings of elementary sounds of separated frequencies. These strings are in rough agreement with the classical selection of formants, although they occasionally differ. We distinguish three strings: treble, mid, and bass. In some sounds one or another of these strings may have very low amplitude, and with different sounds one or another may assume major importance. In general, the treble string variation indicates most sharply transitions in sound; thus it is often the most distinctive of the three. The mid string is the most familiar of the three in terms of voice; when combined with the treble, it makes a voice much like that recorded in an anechoic chamber. The bass is not often of distinctive importance; it represents much of the acoustic interaction of the room, the voice, and the recording system. In our on-line speech system we have included the facility for programmatically constructing digital filters. Since we are still in the study phase, we have the freedom to carefully observe a sample word and experimentally select a means of separating out the strings. In crude terms, the treble is everything above 1800 Hertz, the bass is everything below 450 Hertz, and the mid is everything between. A good test for completeness of separation into strings is that the second derivative of each of the three must remain in the same category as the original string, or the separation is incomplete.

Each string is then decomposed into its constituent elementary sounds by the method of evacuj ation, i.e., first remove those of primary influence from the string by insisting that after removal of a wave function, the residue string is of minimum amplitude in the neighborhood of the one removed. After each evacuation the residue is a similar string of smaller amplitude, and the process can be repeated. In Figure 5. we show the three strings, in the decomposition of the signal produced by a male voice saying "Corby". As is apparent from this figure, the frequency is not constant across a string, but when properly measured for individual elementary sounds will provide an important parameter for sorting these sounds. The evacuation method is shown in Figure 6. for the mid string of "Corby" during the first 210 ms. Figure 7. shows the sum of the first two evacuations in comparison with the original mid string.

Each elementary string f(t) can be expressed as a sum of elementary sounds through the evacuation process. This sum takes the form

$$4.1) \quad f(t) \approx \sum_{j=1}^{k} A_j \, E(t-C_j, S_j, \phi_j, N_j)$$

where E is computed from the wave equation:

$$4.2) \quad \frac{S_j^2}{4\pi^2} \frac{d^2E}{dt^2} + (t-C_j) \frac{dE}{dt} + (N_j^2 - \tfrac{1}{2})E = 0$$

$$E(0) = \cos \phi_j \qquad E'(0) = \omega_j \sin \phi_j$$

$$\omega_j = \frac{2\pi N_j}{S_j} \quad \text{radians/millisecond.}$$

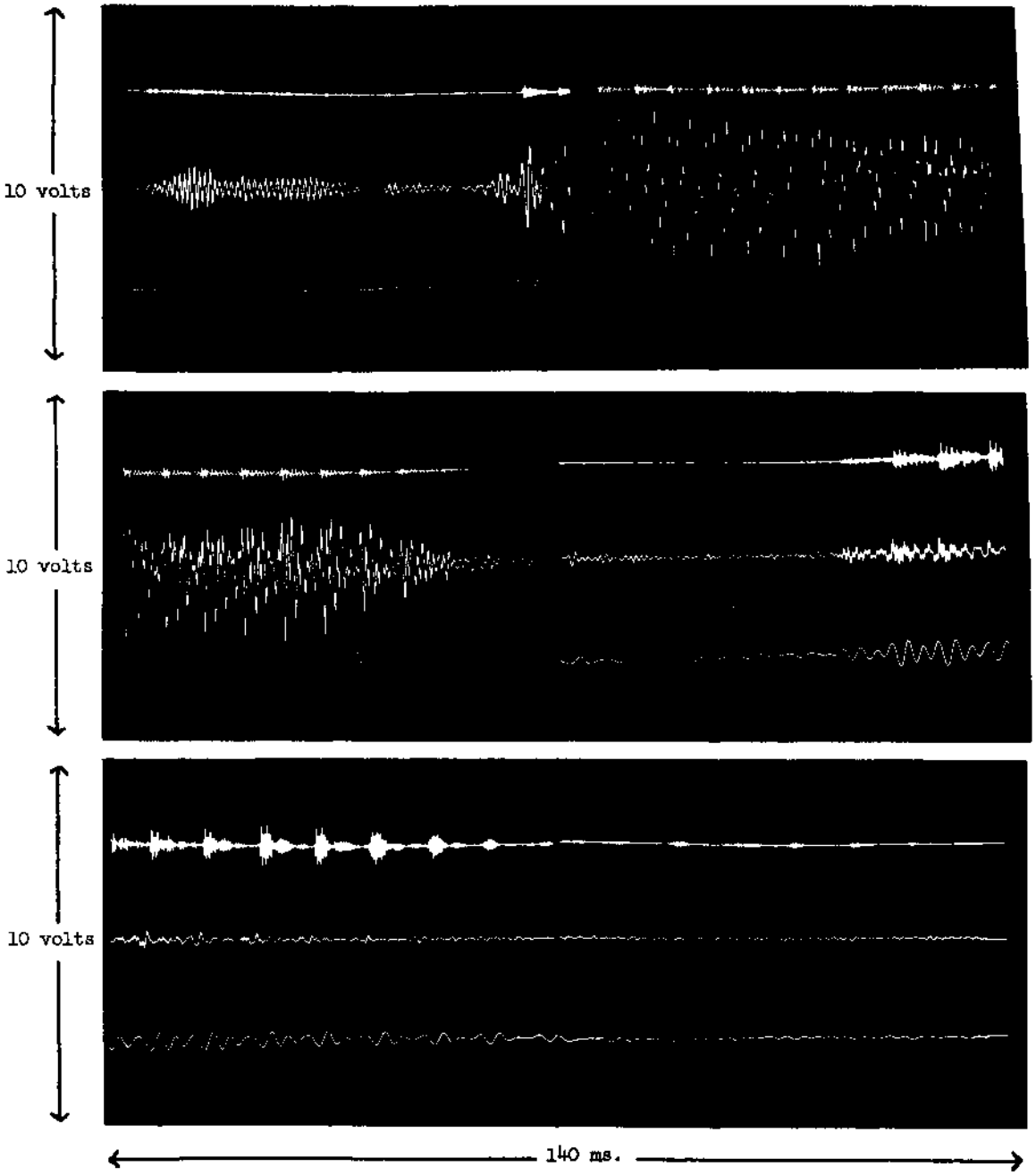


Figure 5. Separation of strings in the word "Corby".

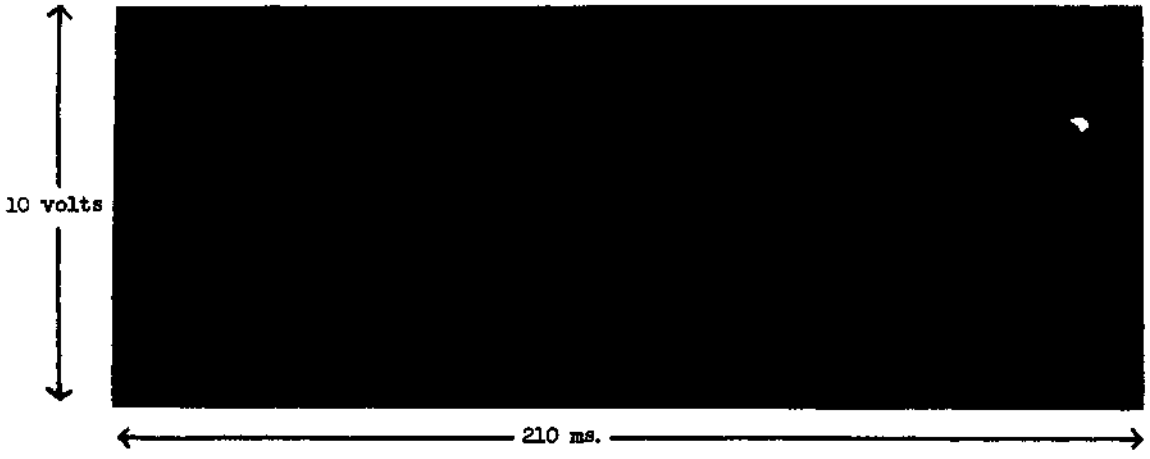


Figure 6. Mid string of "Corby" and first evacuation.

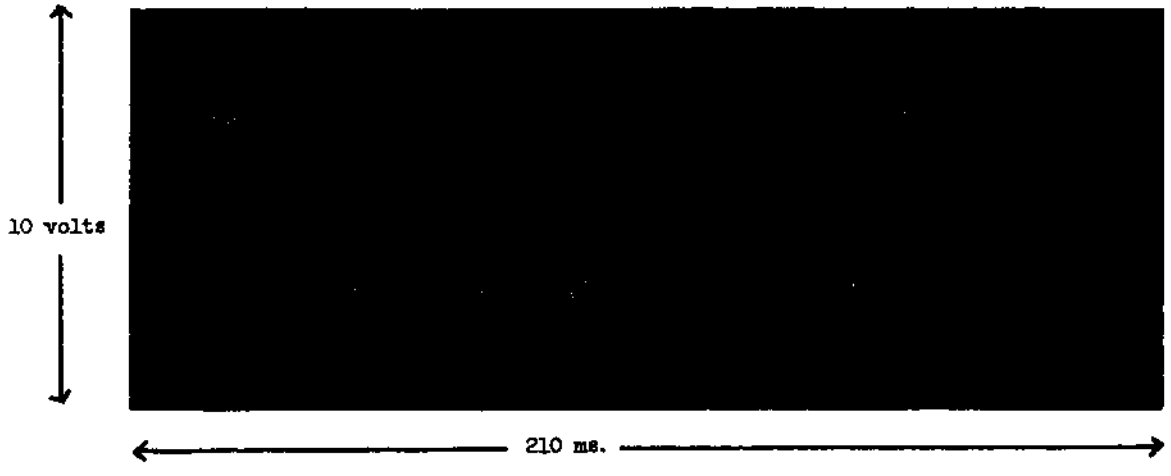


Figure 7. Comparison of "Corby" and its two-evacuation approximation.

Part (3.) of the outline is thus a computational analysis of how to solve 4.2) in real time without introducing undesirable computational anomalies. One can experiment with many methods and observe their shortcomings by listening to 4.1). This way we can frequently hear troubles which are difficult to detect by ordinary means. We are still working to improve our methods here and produce a hardware specification for a "string" instrument, i.e., a special-purpose module which can read a string of ASCON parameters and produce the real time sum 4.1). Three speech strings are capable of producing arbitrarily good speech.

Phones, The Simplest Structural Elements in Speech

The transformation of digital sample data to elementary strings and finally to lists of elementary sounds represented by lists of ASCON parameters puts us in a position to study these lists and determine which subsets have sufficient inner cohesion to be distinguished as speaking sounds. In Figure 8. we show A, S, AC, 0, and cu plotted against the index j for a single mid string recording of the word "odd". Clearly, what catches our eye (and our ear as well) is the modularity of the above data and an apparent regularity of the evo-

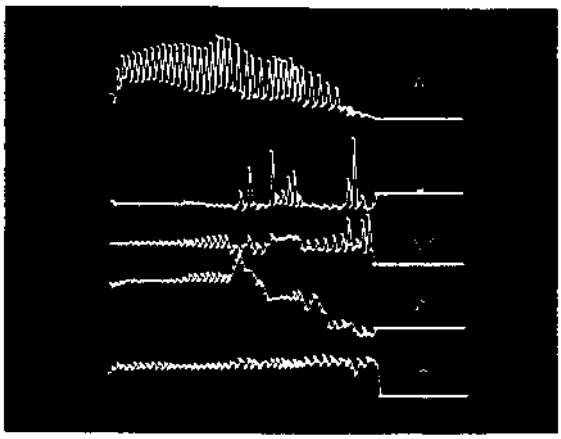


Figure 8. ASCON parameters for one sample of the word "odd".

lution of the parameters up to ragged changes, at which time other trends set in. Thus, to isolate the simplest substructures of these parameters, we base the selection on the smoothness of variation of the parameters. We then define:

A phone is a set of non-overlapping elementary sounds with smoothly varying ASCØN parameters.

a. The pitch function for a phone is the ordered set of time displacement inverses, $\{(\Delta C_j)^{-1}, j = 1, \ldots, \ell\}$; the pitch is the mean value of the pitch function.

b. The frequency function for a phone is the ordered set $\{(^{N}j/S_j) \cdot 1000\}$ (in Hertz); the frequency is the mean value of the frequency function.

c. The span function for a phone is the ordered set $\{S_j\}$; the span is the mean value.

d. The phase function for a phone is the ordered set $\{\emptyset_j\}$; the phase is the mean value of the phase function.

e. The amplitude function for a phone is the ordered set $\{A_j\}$; the amplitude is maximum value of the amplitude function.

The condition of non-overlapping elementary sounds in phones is too stringent for some of the consonants whose elementary sounds have broad spans in the bass string and is too lenient for vowels, as we do not want to allow more than one elementary sound in the same phone per pitch period. The pitch period of an elementary string is determined by the pitch functions of the principal phones, those of largest amplitude in the string. The principal phones of the first 210 ms. of the mid string of "Corby" are shown in Figure 9.

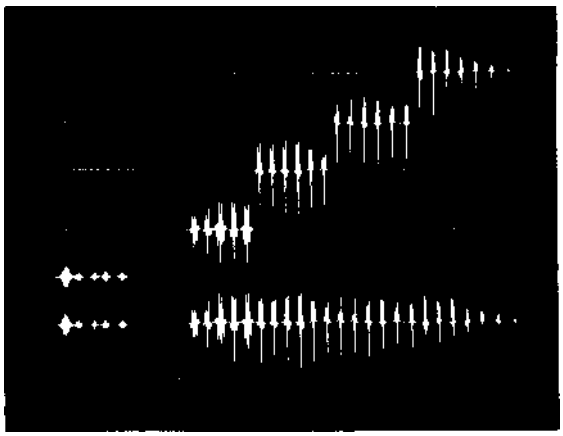


Figure 9. Principal phones from the mid string of "Corby".

Voice; The Comprehensive Structure of Sound in Speech

The concept of a "phoneme", that which minimally distinguishes between two words, leads us to a study of complexes of phones. Strictly speaking, a complex of phones that in a given utterance operates according to the concept of a phoneme comprises a phonetic realization of the phoneme in question.

Our tentative definition of voice is a set of prescriptions of phone complexes that give rise to phonetic realizations of the speaker's phonemes. The expression of voice, sometimes called prosodic control is a transformation of the amplitude and pitch of these phone complexes just prior to final construction of the phonetic realizations. That is, the voice consists of prescriptions for standard phones (unit amplitude and unit pitch) which are modified by the prosodic controls. To extract the voice from input data, the speaker may recite some well-chosen words containing the presumed phonetic realizations of different phonemes (certainly with cross-checking to diminish confusion). The resulting phone complexes must then be simplified by smoothing and stored in the voice table in a way that is advantageous for comparisons. This is a two-way table associating phonetic symbols with phone complexes. One way provides a basis for analysis, the other way for synthesis.

In summary, while there still is much difficult and challenging work to be done, we have succeded in constructing thoroughly tested foundation elements which permit us to record accurate and simple mathematical parameters of speech sounds

which provide high-fidelity reproduction and which open the way to future, perhaps general, computational analysis and synthesis of speech.

Acknowledgements

The work reported here was primarily supported by the Advanced Research Projects Agency, Contract # AF 19(628)-6004. A number of dedicated co-workers have applied themselves in an extraordinary way and have contributed broadly to our present success. We are especially indebted to Gordon Buck and Helen Smith for their programming success; John Greaves, Gary Nelson, and Bill Proctor as graduate research assistants; Ray Bjorkman, Gordon Buck and Dennis Grubbs for their hardware design and construction; Professors Jim Howard, Roger Wood and their students for their continuing experimentation with our underlying engineering processes. Finally, I would like to thank the ARPA personnel, formerly Ivan Sutherland and presently Larry Roberts and Robert Taylor for their encouragement and stimulations to undertake this research.

Bibliography

1. Culler, Glen J., Proceedings of the Symposium on Interactive Systems for Experimental Applied Mathematics, Associate for Computing Machinery Washington D.C., August 26-28, 1967.

2. Howard, J.A., Wood, R.C., Hybrid simulation of speech waveforms utilizing a Gaussian wave function generation, Simulation, Vol. 11, No. 3, September 1968.

ASCØN parameters for wave functions in Figure 4.

A = 2.25, 2.25, 1.125, 1.125	volts
S = 7, 11, 7, 7	milliseconds
Ø = 0, 0, 0, 0	radians
N = 2.75, 2.75, 6.125, 2.1	
A = 2.25, 2.25, 2.25, 2.25	volts
S = 7, 7, 7, 7	milliseconds
Ø = 0, π/2, π, 3π/2	radians
N = 2.35, 2.35, 2.35, 2.35	