CR 86024

SINGLE EQUIVALENT FORMANT EXTRACTOR SYSTEM

By Louis R. Focht

November 1967

Distribution of this report is provided in the interest of information exchange and should not be construed as endorsement by NASA of the material presented. Responsibility for the contents resides in the organization that prepared it.

N68-15768

Prepared under Contract No. NAS 12-582 by
PHILCO-FORD CORPORATION
Blue Bell, Pennsylvania

Marshall Space Flight Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Mr. Wayne E. Lea
Technical Monitor
NAS 12-582
Electronics Research Center
565 Technology Square
Cambridge, Massachusetts, 02139

Requests for copies of this report should be referred to:

NASA Scientific and Technical Information Facility
P.O. Box 33, College Park, Maryland 20740

SINGLE EQUIVALENT FORMANT EXTRACTOR SYSTEM

By Louis R. Focht

November 1967

Prepared under Contract No. NAS 12-582 by
PHILCO-FORD CORPORATION
Blue Bell, Pennsylvania

Marshall Space Flight Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

TABLE OF CONTENTS

FIGURES

SINGLE EQUIVALENT FORMANT EXTRACTOR SYSTEM

by Louis R. Focht

Philco-Ford Corporation
Blue Bell, Pennsylvania

SUMMARY

The Single Equivalent Formant (SEF) is a transformation by which the information bearing parameters of speech are represented with only three parameters, the SEF, the amplitude and the state-of-voicing. The equipment built to extract these three parameters is an approximation of the theoretical transformation and as a result possesses certain performance limitations. However the equipment's simple implementation and small number of parameters considerably reduces the size of the total recognition logic.

The utilization of the SEF parameters for recognition purposes requires an apriori knowledge of the perceptually significant features found in the SEF parameters. A total of five such features have been found to date and techniques for their utilization are suggested.

The comparison of the usefulness of the SEF parameters with other analyzer techniques is considered. It is pointed out that any such evaluation should ideally be carried out as a comparison of the ultimate potential of each analyzer system. However, in practice this is very difficult in as much as it requires the solution to the general recognition problem for each analyzer system. As an alternate, a simpler technique is suggested that is felt will provide a reasonable and fair performance comparison of different analyzer systems for specific recognition tasks. Specific areas of the recognition task are also pointed out in which superior performance might be expected by the use of SEF techniques as well as those areas in which difficulty might be encountered.

INTRODUCTION

The concept of the Single Equivalent Formant is based upon the theory that phoneme perception is a Gestalt response to a time series of "elemental speech sounds" and their relative changes in amplitude. These strings of elemental speech sounds have perceptual values for which the whole is not equal to the sum of the parts. Strings of elemental speech sounds are not perceived as a sequence of individual elementary sounds but rather each string produces its own unique response in the human dependent upon the sequence, duration, and relative amplitude of each elemental sound within the string.

The term elemental speech sound, as used here, refers to those few phonemic values that may be absolutely classified by the human in the absence of all contextual and speaker identity cues. The elementary speech sound may thus be thought of as the sounds perceived by listening to a word through a time window narrow enough to eliminate all contextual cues and listened to with enough time between samples to eliminate speaker identity. Figure 1 shows the results of such analysis on the vowel combination i-a. It will be noted that while only two vowel sounds are perceived in the continuum a total of five vowel sounds are perceived in the context free state. Thus, elemental speech sounds only assume a perceptual significance in this context free state. Its perceptual relation to the entire phoneme, syllable, word or sentence within which it is found may only be defined by its environment. Herein lies the problem of speech recognition.

The instantaneous value of the Single Equivalent Formant parameters provides the information necessary to recognize the context free elemental sounds of speech while the task of recognizing the perceptual value of the strings of these elemental speech sound is accomplished by a knowledge of the perceptual modifications imposed by context.

It is the purpose of this report to identify the known factors (both contextual and context free) which contribute to the creation of a perceptually significant event. It is only with such information that one may effectively utilize the SEF parameters.

PERCEPTUALLY SIGNIFICANT FEATURES OF THE SEF PARAMETERS

Absolute Values of the SEF Parameter

Listening tests have shown that when the human is deprived of time varying elements, contexual cues, and speaker identity, he is only capable of accurately recognizing six vowel, three fricative, and three voiced fricative sounds or a total of twelve " elemental speech sounds" (Ref. 1). In particular, these tests dealt with the accuracy of the human's phoneme recognition ability when he is forced to base his decision solely upon the steady state spectral distribution of the phoneme and in the absence of all of the above mentioned contextual clues. An examination of the results from these tests shows that the recognition errors are not just random but rather groups of phonemes tend to be confused. Each of these groups of phonemes are called elemental speech sounds because phonemes within such groups are not resolvable by the human under the context free conditions of the experiment. As stated, the total number of elemental speech sounds was found to be twelve (in normal speech, some 20 phonemes would be perceived in these twelve phonetic categories). An elemental speech sound is thus defined as the smallest reliably recognized speech element in the absence of all time varying, contextual, and speaker identity cues.

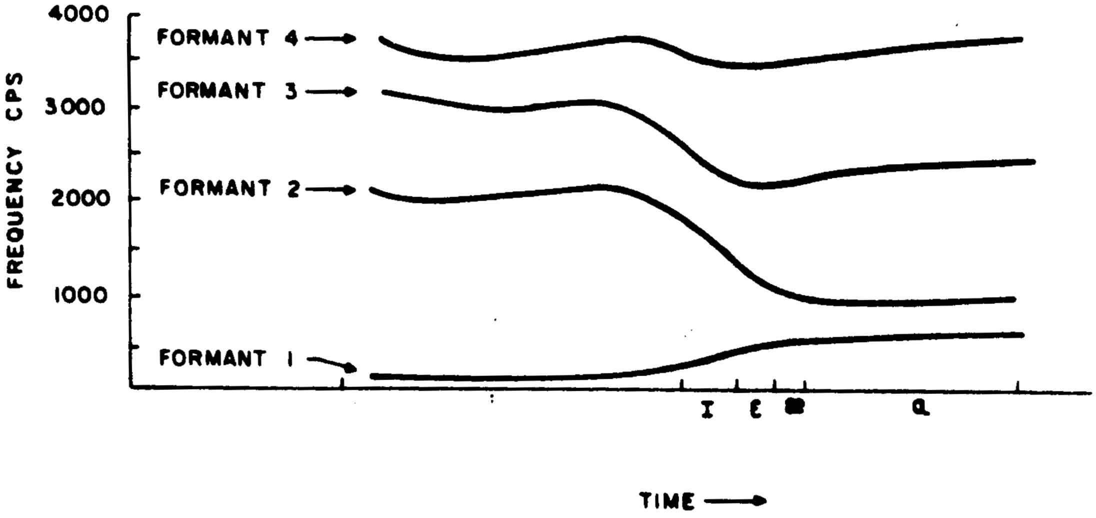

Figure 1. Perceived Phonemes in the Vowel Combination i-a

With this information in mind, it is only natural to assume that under similar context free conditions the SEF parameter should also be limited to six voiced, three unvoiced and three voiced fricative values. This is obvious when one considers the context free nature of the SEF concept presented in the introduction (the theoretical SEF parameter is defined as the context free phonetic value of speech sounds).

The SEF parameter extractors developed for the contract display the expected degree of phonetic confusions. For example, when the absolute level of the SEF parameter for various voiced phonemes is tabulated for a group of speakers, it is found that only six quantum levels for voiced sounds may be readiably established. These are in general i-I, ɛ-æ, ɝ, ʌa, ɔUul, mŋn. This means that when experiments deal with multiple speaker input data (and no means is provided in the recognizer to identify each speaker and then optimize the SEF quantum levels to the speaker) there is no point in quantizing the SEF parameter with a resolution greater than six levels for voiced sounds. This also applies to unvoiced and voiced-fricative sounds with the exception that they must be limited to three quantum levels each i.e., for the fricatives f-θ-h, s, ʃ, and for the voiced fricatives v-ð, z, ʒ.

If of course, speaker adaption is provided for by the recognition logic then additional quantization levels will provide additional perceptually significant information.

Phonetic Environment Modifications of the SEF Parameters

If the preceding conclusions are valid, it would seem logical that the contextual information is the mechanism which enables the human to refine his decision from one of the phoneme categories just described to one of the thirty or more actual phonemes defined by linguistics. Each of the contextual clues appears to be important for the recognition of a different type or combination of phonemes. The stop consonants, for example, are identified principally by their connective vocal transient. On the other hand, diphthongs and vowels appear to be resolved by contrasting them with adjacent phonemes. In fact, significant shifts have been observed in the perceived value of the accoustically identical stimulus simply as a result of adjacent phonetic information. The results of an adjacent vowel interaction study are shown in Figure 2. This study investigated the shifts in perception occurring in Single Equivalent Formant Speech for adjacent vowels "A" and "B". Single formant speech was used for the experiment to avoid the problems of shifting dominance assignment that might occur in multiformant speech.

The " Isophonetic Chart" illustrated in Figure 2, is a plot of the formant frequency of the first sound heard, " A" , versus the formant frequency of the second sound heard, " B" . The vertical lines represent the boundaries of the perceived " B" sounds. It can be readily observed from the chart that the changes in perception grow more pronounced when the spectral content of neighboring stimuli resembles the sound under study. The chart shows that the interpretation that a listener places on the second sound is strongly influenced by the preceding sound. The phonemena is analogous to the shift in color perception experienced by an observer who has been looking previously at a different color (sequential contrast).

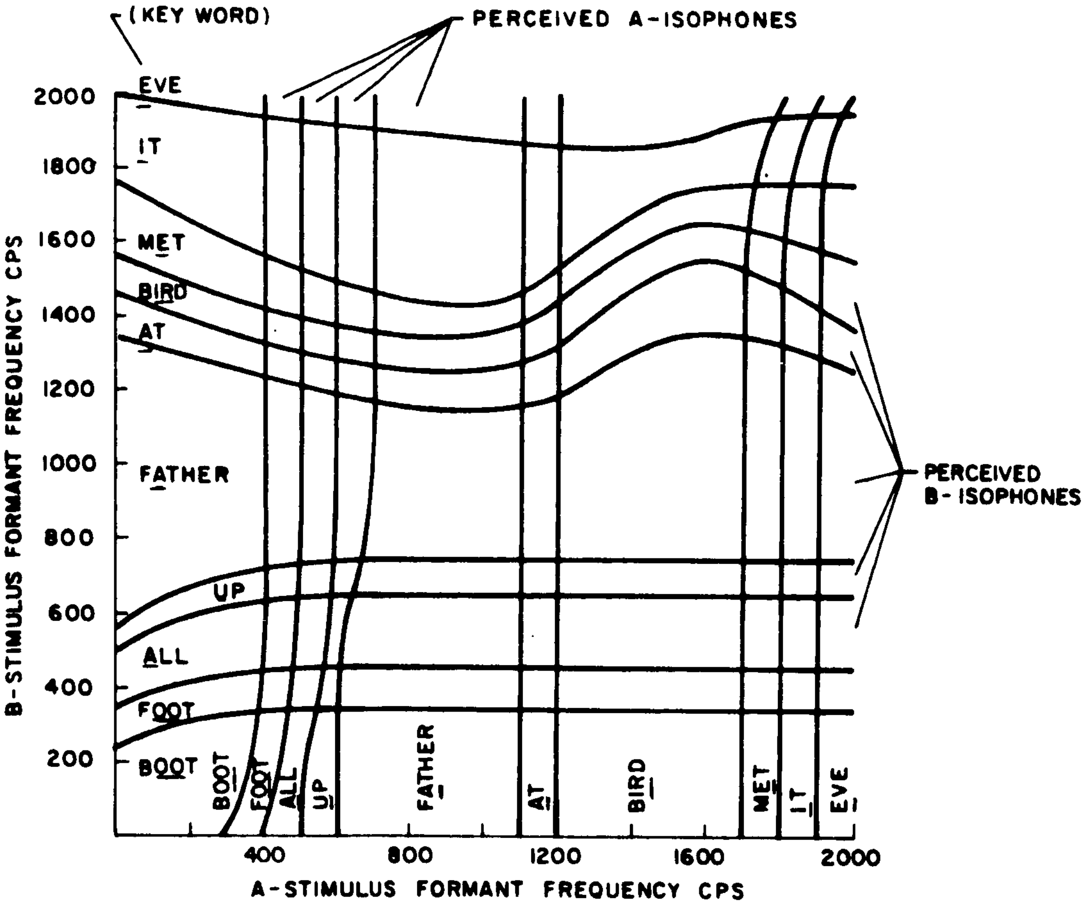

Figure 2. Isophonetic Chart for Sequential Contrast Effects

This phonemenon may explain many of the ambiguities reported by designers of automatic recognition devices. In order to verify these perception shifts, further experiments were performed using two phoneme combinations uttered by human speakers (I-ε and æ-I). Sound spectrograms were made from these utterances and the locations of the first and second formants were extracted and plotted against the perceived sound. This information is shown in Figure 3. It can be seen that the confusion between I and ε exists in the area of F_1 = 450 cps and F_2 = 1600 cps. This confusion is predicted from the data of Figure 2. This can be seen by noting that for the sound æ-I, the I region covers the same range that both I and ε cover for the sound I-ε.

This sequential contrast information most certainly plays a rol in the humans ability to resolve phonemes in the presence of context (both phonetic and speaker).

Energy Environment Modifications of the SEF Parameters

Another perceptual modification of the SEF parameter that has been observed involves rapid changes in the amplitude parameter and their effect upon the ability to perceive simultaneous changes in the SEF parameter. This effect occurs only when the amplitude falls rapidly. Under such conditions the value of the SEF parameter is masked for a duration of time that is directly proportional to the magnitude of the decrease in amplitude. Thus, the larger the decrease in amplitude the longer the duration of the succeeding SEF palues must be to insure perception.

Such effects are most pronounced at the end of a word. When the amplitude drops quite rapidly at the end of a word, the SEF parameter usually changes value or tends to move in a rondom fashion. These changes in the SEF parameter are not usually perceived because the change-duration requirements just described are not satisfied.

An exact quantative measure of this effect has not been established, however an approximate rule that has been used with success states that for every 6 db of amplitude drop the following interval must display a sustained SEF level at the new and lower amplitude for at least 20 ms to be perceived. Thus, 18 db of amplitude drop requires approximately a 60 ms SEF interval at the new amplitude level to be perceived.

Single Parameter Time Modifications of the SEF Parameters

A total of five perceptually significant events have been observed that are the result of independent changes in one of the three SEF parameters. Each of these events produce a different perceptual value. The SEF and amplitude parameters each account for two of these events while the voicing parameter is responsible for the fifth.

The two SEF parameter events differ primarily in the time interval over which they occur. Similarities do, however, occur in their perceptual

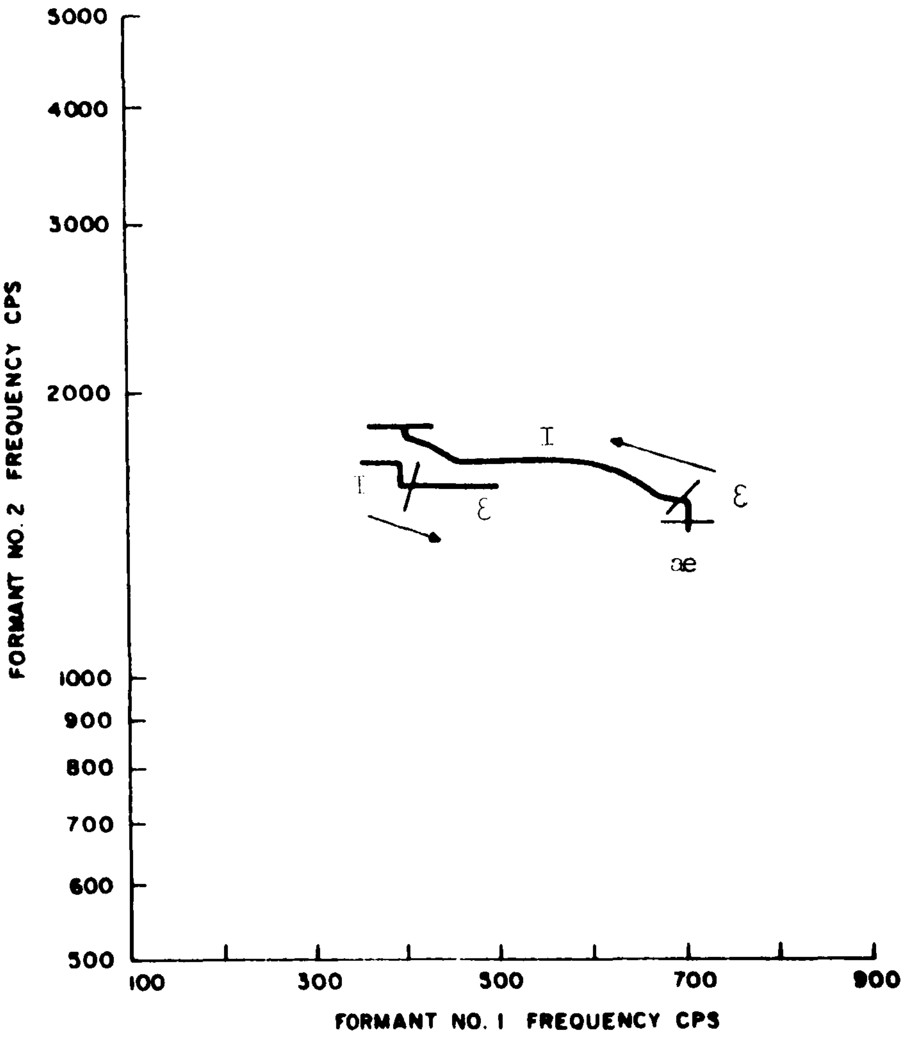

Figure 3. Formant No. 1 and 2 Locus for the Vowel Combinations I- ϵ and ɑe - I

characteristics in as much as phonemic values are not assigned by the human to the isolated events.

The occurrence of the first of these two events results in the perception of a noise component within an elemental speech sound and as such should be processed in such a manner to eliminate its effects. This type of event is the result of both the human speech process and errors in the SEF parameter extractors. The technique that has been developed to eliminate its effects in subsequent recognition logic is termed quantization before integration. This means that the presence of absolute values of the SEF parameter are detected or quantized before any smoothing is applied to the SEF parameter waveform. The output of the SEF level-quantizers are smoothed in a manner to eliminate the effects of the noise like feature. The task of smoothing is designed with two considerations in mind. The first assumes that a noise pulse causes the SEF parameter to momentarily jump to a different SEF value and then return. This produces what might be described as a hole in the ouput waveform of the SEF level-quantizer in question. The smoothing circuits fill in this hole provided it does not exceed approximately 15 ms in duration. Thus, the effects of noise may be eliminated from the output of the quantizer.

The second way of considering the effects of a noise pulse is to assume that the SEF value jumps momentarily into the range of the quantizer in question during the noise pulse. In this case the function of the smoothing circuits is to reject all pulses that are shorter than 15 ms in duration.

The second type of SEF parameter event is the result of connective transients produced by joining two phonemes. These transient events are differentiated from noise events by their longer duration and the fact that they always occur during a voiced interval. Events which remain within a particular SEF quantum range for more than 15 ms but less than 50 to 80 ms are potentially a connective transient or a portion of a connective transient. (Events remaining longer than 70 ms within a particular SEF quantum level, whether voiced or unvoiced, indicates the presence of an elemental speech sound.) Unfortunately more specific rules for determining the occurrence of a connective transient can not be given at this time because of a lack of information concerning their detection and perception in terms of the SEF parameters.

Two amplitude parameter events are associated with the articulation of phonemes that change the rate of air flow in the vocal tract. This change of flow can either be complete, such as produced by the articulation of a stop consonant, or partial as would be produced by a nasal.

The first of these amplitude features is termed an onset. It is defined as the fall in the amplitude parameter accompanying the articulation of a stop consonant or nasal. In general, it has been found that this feature may be detected by noting the presence of a negative slope in the amplitude waveform exceeding an empirically determined value. The detection criteria for the feature is not effected by the phonetic environment , however criteria differences do exist between an onset produced by a stop and an onset produced by a nasal.

The second amplitude feature is called a release. This feature is defined as a rapid rise in the amplitude parameter attending the articulation of a stop consonant or nasal. The basic detection criteria of the release feature is dependent upon both the position of the stop within a word and its phonetic environment. This necessitates dividing releases into several groups, the minimum of which are releases occurring at an initial position within a word, a mid position within a word, and a midposition within a word following a nasal (the latter case is of course for the occurrence of a stop consonant). There can be no final position release by its definition. The detection logic for the stop feature consists of a threshold detector operating upon the positive portion of the differentiated amplitude parameter. Different threshold detector settings are used for each of the three positional and environment categories.

A detailed description of the circuits actually used to implement onset and release detectors may be found in the ''Voice Sound Recognition'' RADC Technical Document Report (Ref. 2).

The fifth and last perceptually significant event is produced by the voicing parameter changing state. While this is an obvious feature it is also very important. To emphasize this point, it is pointed out that the SEF parameter values for an ʃ and i are identical and only the state-of-voicing provides a distinction between these two very different phonemes.

Multiple Parameter Time Modifications of the SEF Parameters

The last type of perceptually significant feature in the SEF parameter is the simultaneous or sequential occurrence of changes in more than one parameter at a time. These features describe the occurrence of a class of transient phonemes, i.e., the voiced stops and unvoiced stops. These phonemes cannot be defined as just a sequence of elemental speech sounds but require additional information concerning changes in amplitude and voicing.

The voiced and unvoiced stops are defined by the time sequences of onsets, releases and the state-of-voicing. The basic feature sequence distinguishing the voiced stop from the unvoiced stop is simply the condition of voicing during the release or onset detection. This very simple accoustic definition of voiced and unvoiced stops does not satisfy the perceptual distinction in all cases. The exception being the case of a short duration aspiration that may proceed the articulation of voiced stop consonants. This special condition voiced stop may be recognized by the fact that the aspiration may be no longer than 50 ms for a mid position voiced stop and no longer than 30 ms for an initial position voiced stop. Longer values of aspiration (or unvoicing) indicate the occurrence of an unvoiced stop.

While the stops represent the most important class of multiple parameter events, several other classes have been observed but not as yet studied sufficient to provide guidelines for their utilization. These are the nasals and the initial ''ℓ''. Both types at phonemes display simultaneous changes in SEF and amplitude parameters that appear to be a perceptually significant event related to the articulation of these phonemes.

RECOMMENDATION FOR EVALUATING SPEECH ANALYZER SYSTEMS

It would be most desirable to carry out a comparative evaluation of speech analyzer techniques in such a manner as to show which system provides the best general solution to the problem of speech recognition. This goal however is most difficult to achieve for many reasons. It must be assumed that any analyzer system to be tested in the near future will not be perfect. Rather they will be characterized by varying strengths and weaknesses in their ability to extract the various information bearing elements of speech. The relative performance of each analyzer system will, of course, be different for each information bearing element necessitating a decision as to the relative importance to the overall recognizer of these various strength and weaknesses. To further complicate the problem, each analyzer system results in raw parameters that are conceptually different. Performance evaluation must thus be deferred to a later point in the logical process where identities in the information bearing elements begin to occur i.e., phonemes. This means that the perfection of these intermediate recognition logical processes must be high or at least equal to insure a fair comparison. Thus, it is felt that a comparison of the ultimate performance capability is very desirable but also very difficult. In fact to carry out such a comparison, a complete general solution of the speech recognition problem would have to be made for each type of analyzer to be evaluated.

In consideration of these problems, it is felt that a far more reasonable approach would be the comparison of analyzer systems in the environment of limited or restricted recognition tasks. Furthermore, to simplify the problem of creating an equal information bearing base, upon which to judge relative performance, it is suggested that phonemic categories be utilized. These two suggestions greatly simplify the task of comparative evaluation.

The restriction of the recognition task should for example, initially, limit the vocabulary size, use discrete speech, and utilize a single speaker. This sufficiently reduces the number of variables created by the different analyzer systems to permit evaluation. Later comparisons may increase the number of speakers or words, etc., one at a time to evaluate individual aspects of each analyzer system.

The use of phonemic categories (groups of acoustically similiar phonemes), as a common base for evaluation, significantly reduces the problems associated with differences in the power or perfection of the required higher level logic. Furthermore, phonemic categories provide a better match, in terms of logic sophistication and complexity, to the suggested limited recognition tasks. Imperfection in the phonemic category logic that may exist in the logics developed for the various analyzers being evaluated should be minimized or at least made less critical. It is also pointed out that any recognition logic must ultimately be evaluated in terms of its cost effectiveness for a particular task. The use of the above recommended procedure will provide a common ground upon which such evaluations may be made.

In evaluating the SEF parameters as a system of analysis, it is important to be aware of the strength and weakness of the SEF technique in relation to other analyzer systems. This is of course just as true for other systems to be compared. The most significant strength of the SEF parameters are their inherent lack of parameter rate of change limiting (other than the speech production process itself). The value of the SEF, for example, may change in one pitch interval from a maximum to a minimum value and be exactly represented in the output of the SEF parameter. The amplitude parameter is also so designed in such a manner as to allow a rise time equal to the fastest frequency component found in speech. The ability to resolve such changes is particularly obvious when considering the recognition of stop consonants and other transient reinforced phonemes. It is pointed out that both the Analogue Ear and Filter Bank analyzer approach do not have such advantages. This is because of the discrete bandwidth associated with the analyzer filters and the methods used for detecting and low pass filtering of the output parameters. However as was pointed out in the previous section of this report, this wide bandwidth of the SEF parameters must be used with some cautions.

It is a characteristic of the SEF concept that many speech sounds are not sustainable when isolated. For example the nasal has been found to be indistinguishable from the u when sustained and isolated from connective cues. The same is true for certain vowel sounds when speaker identity is not preserved. Thus, the SEF parameters must be evaluated on speech sounds in context.

A final point should be made regarding the operation of the voicing extractor built for the contract. This extractor is not as accurate in its decision (both timing and value) as is desirable. The voicing extractor normally used with the SEF parameter utilizes a separate throat microphone and achieves very accurate results. However, because of limitations imposed by the desire to provide a tape recorder input to the SEF extractor the throat microphone type of voicing decision was not practical. An alternate voicing detector that is compatable with the input requirements was supplied but unfortunately as stated this detector does possess some performance limitations that will effect the evaluation of the SEF parameters.

CONCLUSIONS

The successful utilization of the SEF parameters is dependent upon the degree of utilization of the perceptually significant features in the parameters. The extraction of these factors should therefore be of prim concern in the evaluation study.

The comparison of results from system to system should be made on some common ground such as the phoneme or phonemic category. Phonemic categories would appear to be perferred because it tends to minimize differences that might occur in the degree of logic perfection.

In comparing the SEF parameters it should be noted that the SEF parameters are most useful when contexual information is utilized. Furthermore, careful attention should be paid to the full utilization of the fast response nature of the SEF parameter extractors.

REFERENCES

1. "The Single Equivalent Formant", Louis R. Focht, 1966 IEEE International Convention Digest of Technical Papers, p. 108.

2. "Voice Sound Recognition", Louis R. Focht, RADC Technical Documentary Report for the period April 1966 to April 1967, Project No. 4027, AD 802-997.

NEW TECHNOLOGY APPENDIX

After a diligent review of the work performed under this contract, no new innovations, discovery, improvement or invention was made.