17.

A FUZZY LOGICAL MODEL OF SPEECH PERCEPTION

Dominic W. Massaro
Program in Experimental Psychology
University of California, Santa Cruz
Santa Cruz, California 95064

ABSTRACT

The thesis of this paper is that humans achieve robustness in speech perception by evaluating and integrating multiple sources of information. The sources of information are assumed to be continuous rather than discrete (categorical) and represent both bottom-up and top-down contributions. The bottom-up sources take the form of acoustic features, which are the perceptually relevant physical properties of the speech signal. The top-down sources are constraints given by the phonological, lexical, syntactic, and semantic context of the message. All sources of information are represented by fuzzy truth values, which indicate the degree to which each source supports each possible alternative. The multiple sources of information are integrated together in such a manner that the least ambiguous sources have the most impact on recognition.

These assumptions about speech perception have been formalized within the framework of a fuzzy logical model of speech perception. Three operations are assumed. First, feature evaluation involves the derivation of various perceptual features. The features are assumed to be continuous rather than discrete. The outcome of featural evaluation is a truth value, t(x), representing the degree to which each relevant feature is present in the speech stimulus. The second operation is prototype matching which involves the integration of the features. The featural information is compared with perceptual unit definitions, or prototypes, to determine to what degree each prototype is realized in the speech sound. Prototypes define a perceptual unit in terms of arbitrarily complex fuzzy logical propositions. The third operation is pattern classification. During this stage, the merit of each potential prototype is evaluated relative to the summed merits of the other potential prototypes. The relative goodness of a perceptual unit gives the proportion of times it would be selected as a response or its judged magnitude.

The model is tested against human perceptual results derived from identification experiments using synthetic speech. The experiments are designed to manipulate independently various acoustic characteristics of the speech signal and higher-order constraints. The acoustic characteristics that have been studied include voice-onset time, aspiration intensity, fundamental frequency, formant patterns of stop consonants, and consonant and vowel duration of stops and fricatives. The higher-order constraints that have been varied include phonological, lexical, and syntactic/semantic context. The results of these experiments support the model and also allow rejection of alternative models of speech perception.

17 .1 SPEECH PERCEPTION: HEARING MORE OR LESS THAN THERE IS

From a pattern-recognition perspective, speech perception is an amazing skill. There does not seem to be an exact relationship between the acoustic signal and the perceived patterns in the message. As an example, the units of recognition do not seem to coincide with units of the speech signal. In some cases, we hear word boundaries where there is no or little silence and hear complete words with significant silent periods. In the statement "That you may see," there is usually more silence during the /aet/ portion of "that" than between the words "may" and "see." This example shows that successive segments of sound can produce a coherent or unitary percept and that discrete percepts can result from a relatively continuous sound segment.

The acoustic properties characterizing a speech segment seem to vary depending on the placement of that segment in the speech message. As an example, certain properties of stop consonants depend on their position in words. Voice onset time (VOT) is an important property for the voicing of stops in word-initial and medial position, whereas this property is less common in word-final position. Perceptually, VOT appears to be important for the identification of stops in initial and medial position, whereas preceding vowel duration is important for voicing of stops in word-final position. Thus, to recognize stops, the listener must allow for their position-dependent properties.

Another obstacle in the recognition of small speech segments is that the acoustic signal specifying a particular linguistic unit is context sensitive. That is, the acoustic properties of a unit found in one context are significantly modified in another. Consider the classic example of the syllables /di/ and /du/ (Liberman, Cooper, Shankweiler, & Studdert-Kennedy 1967). The acoustic signal corresponding to the initial /d/ sound is significantly different in the two syllables. The following vowel context modifies the properties of the preceding stop. This example shows that perceptual recognition of some speech sounds must take into account the contribution of the surrounding context.

One reason that speech perception might be accomplished, even in the presence of the aforementioned difficulties, is the contribution of linguistic context. It is generally agreed that the listener normally achieves good recognition by supplementing the information from the acoustic signal with contextual information generated through the utilization of knowledge in long-term memory. There is considerable debate concerning how informative the acoustic signal actually is (Blumstein & Stevens 1979; Cole & Scott 1974; Liberman, et al. 1967; & Massaro 1975b). Even if the acoustic signal proved to be sufficient for speech recognition under ideal conditions, however, few researchers would believe that the listener relies on only the acoustic signal.

17.2 INFORMATION-PROCESSING MODEL

Our study of pattern recognition in speech has been carried out within a general information-processing model (Massaro 1975a, 1975b, 1979). A schematic representation of the stages of processing in the model is presented in Figure 17.1. At each stage of processing, memory and process components are represented. A particular memory component (indicated by a rectangle) corresponds to the information available at that stage, whereas the corresponding process component (indicated by a circle) represents the operations applied to the information in the memory component.

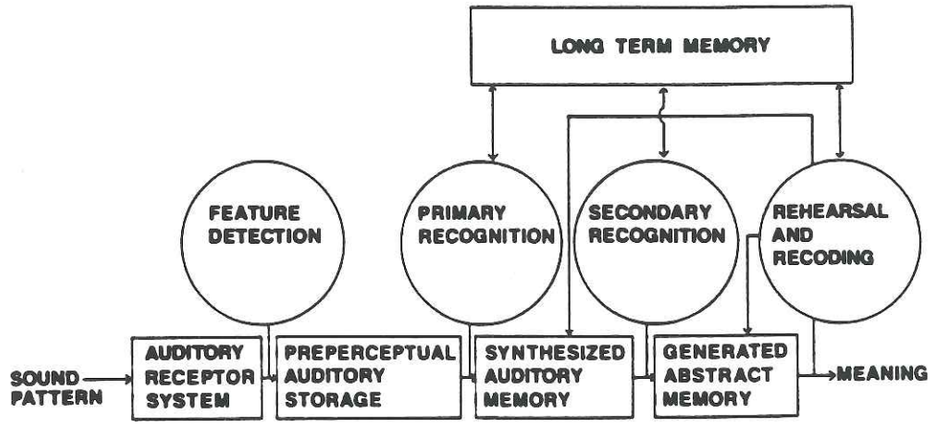

FIGURE 17.1 SCHEMATIC DIAGRAM OF INFORMATION-PROCESSING MODEL.

The feature detection process transforms the energy pattern created by the language stimulus and transduced by the appropriate receptor system into a set of features held in preperceptual storage. The changes in sound pressure set the eardrums in motion, and these mechanical vibrations are transduced into a set of neural impulses. It is assumed that the signal in the form of continuous changes in vibration pattern is transformed into a set of relatively independent features. Features are not limited to primitive attributes but can be relatively complex. In speech, for example, the amount of energy in a particular frequency band would be a relatively simple feature, whereas a complex feature might include information about the direction and rate of frequency change of the sound (Stevens 1981). It would be possible, for example, to have a feature detector that responds to the rising first formant transition that is characteristic of the class of voiced stop consonants. Primary recognition evaluates and integrates these features into a percept which is held in synthesized memory.

Secondary recognition transforms synthesized percepts into meaningful forms in generated abstract memory. In speech perception, it is assumed that the input is analyzed syllable by syllable for meaning. The secondary recognition process makes the transformation from percept to meaning by finding the best match between the perceptual information and the lexicon in long-term memory. Each word in the lexicon contains perceptual and conceptual codes. The concept recognized is a function of at least two independent sources of information: the perceptual information in synthesized memory and the semantic/syntactic context in the message. Rehearsal and recoding processes operate at generated abstract memory to maintain and build semantic/syntactic structures. There is good evidence that this memory has a limited capacity, holding about five, plus or minus two, chunks of information.

The following example may help clarify the differences among these three levels of processing; presented with a tone, the listener can detect or sense the presence of sound, hear and remember a tone of a particular quality, and even identify it as a particular note on the musical scale. Figure 24.2 illustrates the outcome of these three stages as detection, perception, and conception, respectively. Each of these stages makes information available to a response selection and programming process. Accordingly, some response execution can be initiated by any of the three stages of language processing. Although the boundaries between these stages

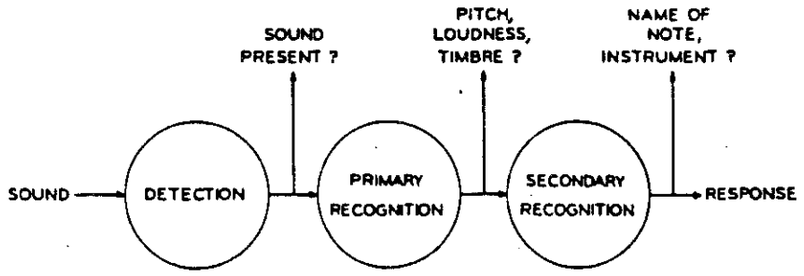

FIGURE 17.2 THREE STAGES OF PROCESSING A SOUND IN MUSIC.

are sometimes fuzzy, we have found it helpful to maintain these distinctions in our experimental and theoretical research.

17.2.1 Pattern Recognition

This information-processing model has served as the basis for a model of pattern recognition in speech perception. Central to the model is the idea of prototype descriptions stored in the listener's long-term memory. These descriptions are propositions specifying the featural properties of speech sounds of roughly syllabic size (Oden & Massaro 1978).

According to the model, well-learned patterns are recognized in accordance with a general algorithm regardless of the modality or particular nature of the patterns (Massaro, 1979; Oden & Massaro, 1978). The model postulates three operations in perceptual (primary) recognition: feature evaluation, prototype matching, and pattern classification. Continuously valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. The model is called a fuzzy logical model of perception (abbreviated FLMP), and it will be helpful to discuss the concept of fuzzy logic and how it is used in the model.

17.2.2 Fuzzy Logic

In fuzzy logic, propositions are neither entirely true nor false but rather take on continuous truth values. For example, we might say that a team is having a "good" season or that a meal is somewhat spicy. Ordinary logical quantification would require that the team be performing well or not and that the meal is either spicy or it isn't. Fuzzy logic theory (Zadeh, 1965; Goguen, 1969), on the other hand, allows us to represent the continuous nature of things. In fuzzy logic, we can construct a membership function: for example, short(x) which is true to the extent that item x is a member of the set short. It should be noted that fuzzy truth is different from probability. If we say that a whale is a fish to degree .2, that does not mean that there is a .2 probability that a particular whale is a fish. Rather, it is true that the whale is a fish to degree .2.

An important part of fuzzy logic theory concerns the realization of the standard logical operations of conjunction, negation, and disjunction. The range of truth values, t(x), in fuzzy logic goes from Ø for perfectly false to 1 for perfectly true. Thus, a reasonable definition for negation is the additive complement:

$$t(\tilde{}x) = 1 - t(x). \qquad (1)$$

where $t(\tilde{}x)$ is the truth of not x. Goguen (1969) has suggested two possiblities for the conjunction (^) of two events a and b:

$$t(a\hat{}b) = t(a) \times t(b), \qquad (2)$$

and

$$t(a\hat{}b) = \min(t(a), t(b)). \qquad (3)$$

Massaro and Cohen (1976) tested an additive definition for the conjunction of two events

$$t(a\hat{}b) = t(a) + t(b). \qquad (4)$$

Research in a number of domains has consistently supported the multiplicative over the other two forms of conjunction. Massaro and Cohen (1976) the conjunction of voice-onset-time and fundamental frequency as perceptual cues to the /si/-/zi/ distinction. A multiplicative combination of the cues values described the results about four times more accurately than did an additive combination. Oden (1977) investigated which set of definitions of fuzzy logical conjunction best fit judgments about logical combinations of pairs of statements about class membership functions (e.g., a bat is a bird, and a refrigerator is furniture). The data from the experiment were better explained by the multiplication rule of Equation 2 than by the minimization rule of Equation 3.

We do not assume that humans actually carry out the process of multiplication, just that multiplication closely represents the processes involved in conjoining or integrating different sources of information. A model mimics the behavior of the phonemenon of interest. A mathematical model doesn't necessarily attribute the mathematical processes to the entities involved; it simply describes the outcome of the processes. By assuming that the feature values are multiplied by the listener, we do not mean that the listener actually multiplies feature values. We simply mean that the features are combined in such a way to produce an outcome that is equivalent to a multiplication of the feature values. As acknowledged by Rumelhart and Norman (1983), modeling a bouncing ball with differential equations does not imply that the ball itself understands or uses these equations. Modeling human performance with various formalisms should not be taken to mean the human understands or uses these formalisms (Lopes 1981).

17.3 FUZZY LOGICAL MODEL OF PERCEPTION (FLMP)

According to the fuzzy logical model of perception (FLMP), primary recognition is carried out in three stages. The first stage is feature evaluation, during which the features transduced by the sensory systems are assigned truth values. The features are assumed to be continuous rather than discrete, and thus featural evaluation provides truth values, t(x), representing the degree to which each relevant feature is present in the speech stimulus.

The second stage of recognition is prototype matching which involves the integration of the features. During this stage, the featural information is compared with perceptual unit definitions, or prototypes, to determine to what degree each prototype is realized in the speech sound. Prototypes define a perceptual unit in terms of arbitrarily complex fuzzy logical propositions.

The third stage of recognition processing is pattern classification. During this stage, the merit of each potential prototype is evaluated relative to the summed merits of the other potential prototypes. The relative goodness of a perceptual unit gives the proportion of times it would be selected as a response or its judged magnitude. This is similar to Luce's (1959) choice rule. In pandemonium-like terms, we might say that it is not how loud some demon is shouting but rather the relative loudness of that demon in the relevant crowd of demons. An important prediction of the model is that one cue has its greatest effect when the second is at its most ambiguous level. Thus, the most informative cue has the greatest impact on the judgment.

The FLMP defines a representational system for the processing of speech. Following the illuminating analyses of Palmer (1978), we define five attributes that must be specified for any representational system. These attributes and their instantiations are given in Table 17.1.

Table 17.1 Five attributes of the representation system assumed by the fuzzy logical model of speech perception

1. the represented world -- speech

2. the representing world -- listener

3. aspects of the represented world being modeled -- acoustic characteristics of the speech, speech sounds (syllables, words), memory of speech sounds, integration of characteristics, decision, and classification processes.

4. aspects of the representing world doing the modeling -- features, truth values, integration operations, prototype descriptions, and classification algorithms.

5. correspondences between represented and representing worlds --

represented	representing
1. acoustic characteristics	1. features
2. speech sounds	2. perceptual units
3. memory for speech sounds	3. prototypes
4. integration of characteristics	4. prototype matching
5. decision/classification	5. pattern classification

The represented world is speech, and the representing world is the human listener. The aspects of the represented world being modeled, the aspects of the representing world doing the modeling, and their correspondences are listed in attributes 3, 4, and 5, respectively. The value of presenting the model in these terms is that it makes explicit attributes of speech that would have to be included in almost any reasonable representational system. The attributes listed in the representing world are those assumed by the FLMP are subject to falsification in experimental and theoretical tests.

17.4 INTEGRATING ACOUSTIC FEATURES

In order to illustrate how the FLMP is applied and tested within the domain of pattern recognition, consider an experiment carried out by Massaro and Oden (1980a). Seven levels of voice onset time (VOT) were crossed with seven levels of the onsets of the F_2-F_3 transitions in the synthesis of stop consonant-vowel syllables. The VOTs ranged from a completely voiced to a completely voiceless sound. The values were 10, 15, 20, 25, 30, 35, and 40 msec. The seven levels of the F_2-F_3 onset frequencies ranged from 1345 to 1796 Hz for F_2 and 2397 to 3200 Hz for F_3 to give a continuum of sounds going from a labial to an alveolar place of articulation. Subjects made repeated identifications of random presentations of the 49 unique syllables from the alternatives /bae/, /dae/, /pae/, and /tae/.

The four panels of Figure 17.3 present the percentage of /bae/, /pae/, /dae/, and /tae/ identifications, respectively, as a function of the two independent variables. The levels along the abscissa are not equally spaced but rather have been adjusted to be proportional to the differences between the marginal means across the levels of the F_2-F_3 transitions. These differences were computed separately for each of the four response alternatives and then averaged over response types so that all four of the panels have the same spacing along the abscissa.

In the FLMP, the following propositions specify the prototype descriptions for the four response alternatives in the experiment.

/bae/: (short VOT) and (low F_2-F_3 onsets) (5)

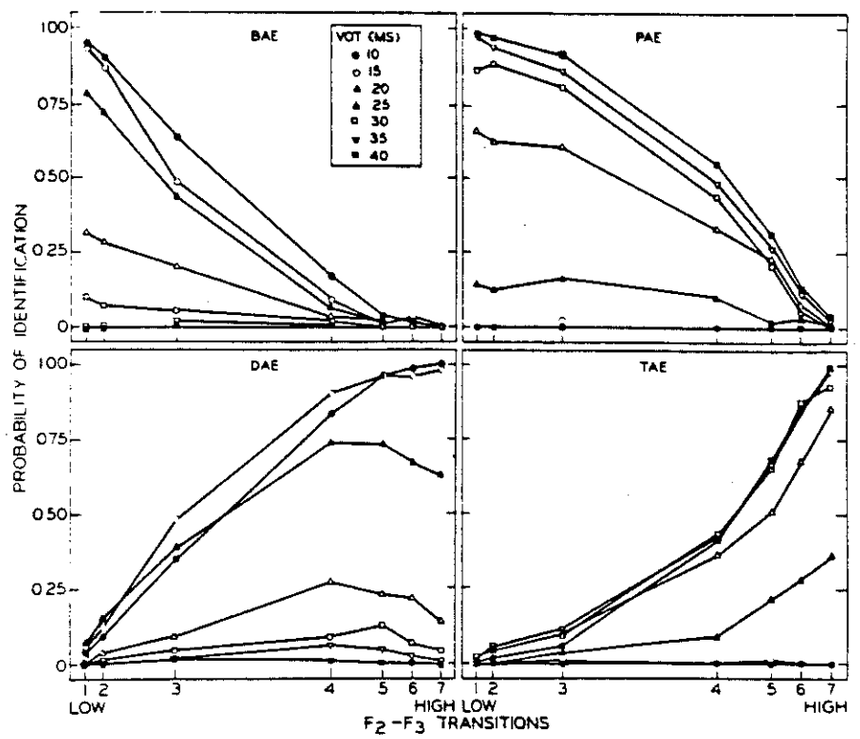

FIGURE 17.3 PERCENTAGE OF BAE, PAE,DAE, AND TAE IDENTIFICATIONS AS A FUNCTION OF VOT AND F2-F3 TRANSITIONS.

/pae/: (long VOT) and (low F_2-F_3 onsets) (6)

/dae/: (short VOT) and (high F_2-F_3 onsets) (7)

/tae/: (long VOT) and (high F_2-F_3 onsets) (8)

The prototype propositions specify the ideal values of each of the acoustic features for the particular speech sound. The properties for the prototypes would also include other acoustic features characterizing stop consonants and vowel /ae/. These properties are not included in the propositions, since they are present in all of the four alternatives. The mathematical form of the fuzzy logical model necessitates consideration of only those properties which differentiate the relevant prototypes.

Upon presentation of a speech sound, the feature evaluation process produces a fuzzy truth value specifying the degree to which it is true that the sound has the relevant acoustic feature. For example,

$$t[\text{short VOT}(S_{ij})] = .60 \qquad (9)$$

represents that it is .60 true that the speech sound S_{ij}, from the ith row and jth column of the factorial stimulus design, has a short VOT. To simplify the notation, let SV_i and LV_i correspond to short and long VOTs, respectively. The subscript i signifies that the values change only with changes in the variable VOT. Similarly, LO_j and HO_j correspond to low and high F_2-F_3 onsets, respectively. The values change only with changes in the column j variable of the F_2-F_3 onset frequencies. Also, we will henceforth use the expression "short VOT" to represent its truth value t(short VOT). The truth of the negation of a feature is defined as one minus the truth value of the feature. In this case, if .6 specifies the truth value of a short VOT, then 1 - .6 = .4 would specify the truth value of a long VOT. In general, the value $LV_i = 1 - SV_i$. A similar complementary relationship is assumed between high and low F_2-F_3 onsets; the value LO_j is equal to $1 - HO_j$.

At the prototype matching stage, a determination is made regarding the degree to which the conjunction of features in each prototype definition has been realized in the speech signal. The multiplicative rule for conjunction gives the matching function

$$bae(S_{ij}) = SV_i \times LO_j. \qquad (10)$$

and so on for the other three prototypes.

Given the matching functions for each of the alternative prototypes, the speech sound is identified on the basis of the relative degree of match. Following the rationale of Luce's (1959) choice model, it is assumed that the probability of identifying a stimulus to be a particular syllable is equal to the relative degree to which that syllable matches the stimulus compared to the degree of match of the other syllables under consideration. Given that the speech sound must be identified as either /bae/, /pae/, /dae/, or /tae/, the probability of a bae identification given stimulus S_{ij} is given by

$$P(bae{:}S_{ij}) = \frac{bae(S_{ij})}{bae(S_{ij}) + pae(S_{ij}) + dae(S_{ij}) + tae(S_{ij})} \qquad (11)$$

where the variables in the ratio represent the matching functions for the four alternative speech sounds.

The simple model described above failed to capture the fine detail of the results. A more complex model assumes that features defining the prototypes include modifiers to provide more specific information about the ideal feature values. For example, the locus of F_2 at the instant of release is usually higher for /t/ than it is for /d/ (Fant, 1973, Chapter 11). Thus, the prototype for /tae/ might be defined as

/tae/: (long VOT) and very (high F_2-F_3 Onsets) (12)

where "very" is implemented as an exponent on the feature value. If the exponent were 2, for example, a truth value of .8 would be only .64 when it is squared. This captures the possibility that F_2-F_3 are normally higher for /t/ than /d/ and listeners require a higher value of F_2-F_3 for /tae/ than for /dae/. The predictions of the complex model with a single prototype modifier gave a reasonably good description of the 147 independent response probabilities with just 15 free parameters.

17.5 INTEGRATING HIGHER-ORDER CONTEXT

One important test of the FLMP involves the contribution of higher-order context to speech perception. In the model, the bottom-up information remains independent of the top-down information and the two sources are combined in the same manner as two bottom-up sources. In terms of the model, one basic limitation in previous research is that it has been primarily directed at showing a positive contribution of linguistic context rather than at how it is integrated with information from the acoustic signal (Cole & Jakimik, 1978; Marslen-Wilson & Welsh, 1978; Pollack & Pickett, 1964). Recent research in our laboratory, along with other current studies, overcome these limitations in previous research and provide quantitative tests of the FLMP.

17.5.1 Phonological Context

Massaro and Cohen (in press) assessed how the information from the acoustic signal is combined or integrated with information from phonological constraints in English. Phonological constraints refer to the fact that languages are redundant in terms of the possible sequences of speech sounds. There are constraints on the ordering of speech sounds within English words such as /r/ but not /l/ following word-initial /t/. Listeners were asked to identify sounds along a continuum between /li/ and /ri/, which was made by varying the starting frequency of the third formant (F_3) transition. These sounds were placed after each of four initial consonants. When the sounds are placed after the initial consonant /s/, /l/ is phonologically admissible in English, but /r/ is not. If phonological constraints influence perception, listeners should tend to hear /l/ following the sound /s/. Given an initial consonant /t/, however, listeners might be more likely to hear /r/ than /l/. In English, /l/ does not follow initial /t/. In addition to these two conditions, the contexts /p/ and /v/ are also included. Both /l/ and /r/ are phonologically admissible following /t/, but neither is admissible following initial /v/. The results not only provide a test of whether phonological constraints contribute to speech perception, the experimental design allows quantitative tests of how context and an acoustic feature are integrated together in speech perception.

Each speech sound was a syllable beginning with one of the four

consonants, /p/, /t/, /s/, or /v/, followed by a liquid consonant ranging in seven levels from /l/ to /r/ followed by the vowel /i/ (Massaro & Cohen, in press). On each trial, the syllable was randomly selected without replacement from the set of 28 syllables generated by the factorial combination of the four initial consonants and the seven F_3 levels of the following liquid. Subjects identified the sound by pressing one of eight buttons, labeled PLE, PRE, TLE, TRE, SLE, SRE, VLE, and VRE. Subjects were told that their task was to identify the syllable on the basis of what they heard. They were told that there was no correct response and simply to make the best judgment they could. Subjects were tested for two days with 28 practice trials and two sessions of 280 experimental trials each on each day.

The recognition of the initial context consonant was very good, averaging about 95 percent correct. Figure 17.4 gives the probability of /r/ identifications for each of the two days of the experiment. The points in Figure 17.4 show that the identification of the liquid sound was an orderly function of both F_3 of the liquid and the initial consonant. As expected, the proportion of /r/ identifications increased with decreases in the starting value of F_3 of the liquid. More importantly, an /r/ identification was more likely in the context of /t/ than in the context of /p/ or /v/ and least likely in the context of /s/. The effects of the initial context consonant were largest at the more ambiguous values of F_3. In addition, the context effects did not appear to decrease with experience in the experiment. In summary, the results of the experiment show large effects of both the acoustic featural information of F_3 and the phonological context of the initial consonant. The significant interaction of these two variables reveals that the magnitude of the context effect was largest at the more ambiguous levels of the acoustic featural information.

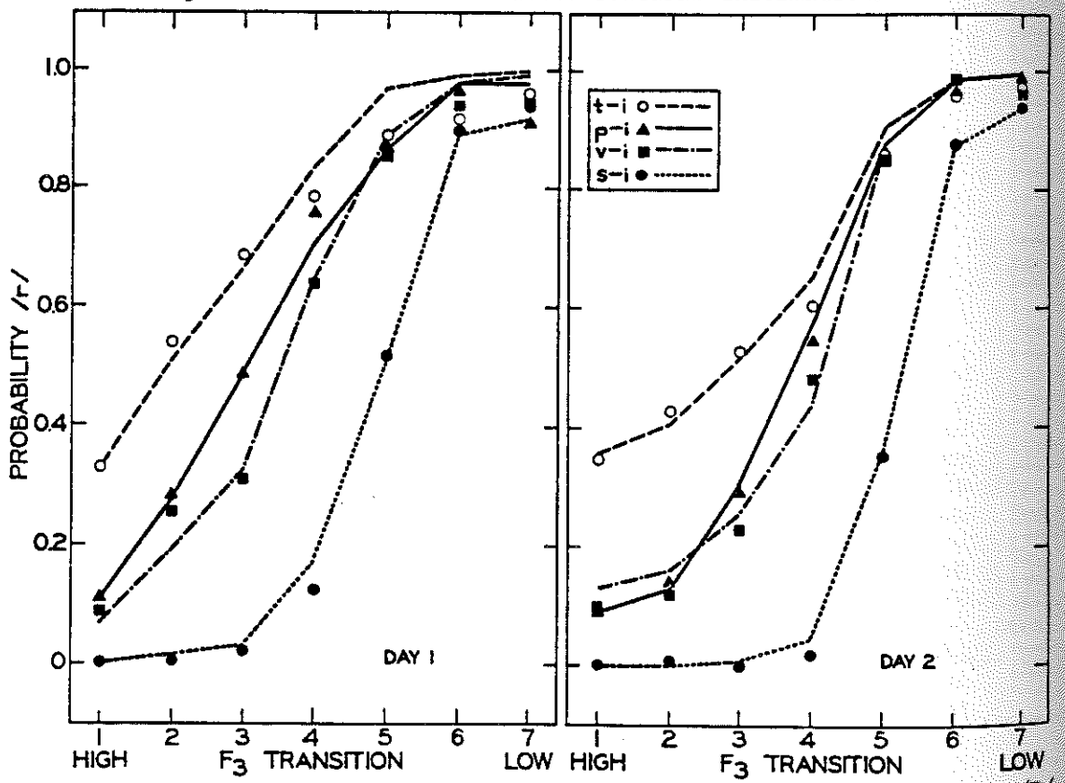

FIGURE 17.4 THE OBSERVED (POINTS) AND PREDITED (LINES) PROBABILITY OF AN /R/ IDENTIFICATION AS A FUNCTION OF THE F3 TRANSITION AND PHONOLOGICAL CONTEXT.

The lines in Figure 17.4 present the predictions of the fuzzy model of speech perception. The critical assumption of the model is that the featural information from the liquid and the phonological context provide independent sources of information. The featural information representing the liquid can be represented by the truth value T_i where the subscript i indicates that T_i changes only with the F_3 transition. For the /l/-/r/ identification, T_i specifies how much R-ness is given by the critical F_3 transition feature. This value is expected to increase as the onset frequency of the F_3 transition is decreased. With just two alternatives along the continuum, it can further be assumed that the amount of L-ness given by the featural information is simply one minus the amount of R-ness given by that same source. Therefore, if T_i specifies the amount of R-ness given by the F_3 transition, then $(1-T_i)$ specifies the amount of L-ness given by that same transition.

The phonological context also provides independent evidence for R and L. The value C_j represents how much the context supports the consonant R. The subscript j indicates that C_j changes only with changes in phonological context. The value of C_j should be large when R is admissible and small when R is not admissible. Analogous to the treatment of the featural information, the degree to which the phonological context supports the consonant L is indexed by $(1-C_j)$.

The listener is assumed to have two independent sources of information. The amount of R-ness and L-ness is determined by integrating these two sources. The amount of R-ness and L-ness for a given syllable can be represented by the conjunction of the two independent sources of information:

$$\text{R-ness} = (T_i \wedge C_j) \tag{13}$$

$$\text{L-ness} = [(1-T_i) \wedge (1-C_j)] \tag{14}$$

In this case, for pattern classification, the probability of an R-response is predicted to be

$$P(R) = \frac{T_i C_j}{T_i C_j + (1-T_i)(1-C_j)} \tag{15}$$

The model was fit to the proportion of /r/ identifications as a function of the F_3 of the liquid and the initial consonant context. Seven values of T_i are required for the seven levels of the F_3 transition of the liquid. Unique C_j values are required for each of the four different initial consonant contexts. Fitting the model to the observed data, therefore, requires the estimation of 11 parameters. The predictions of the model were obtained by minimizing the squared deviations between predicted and observed values, using the routine STEPIT (Chandler, 1969).

Figure 17.4 shows that the model provides a good description of the results, with an average squared deviation of less than 5 percent. In addition, the parameter estimates of the model are meaningful. The T_i values, representing the degree of R-ness, increase systematically with decreases in the starting frequency of F_3. The C_j values change systematically with phonological context; the degree of R-ness given by context is much larger for initial /t/ than for initial /s/. Relative to the context /v/, the context /p/ is somewhat more supportive of /r/ than of /l/. This could be due to the fact that, in natural English, initial /p/ is more likely to be followed by /r/ than by /l/ (Roberts, 1965).

17.5.2 Lexical Context

Ganong (1980) assessed the contribution of lexical context to the perception of stop consonants. The voice-onset time of the initial stop consonant was varied to create a continuum from a voiced to voiceless sound. The following context was varied so that either the voiced or the voiceless stop would make a word. For example, subjects identified the initial stop as /d/ or /t/ with the following context ash where (where /d/ makes a word and /t/ does not). Voiced (/d/) responses were more frequent when /d/ made a word than when /t/ made a word. The contribution of lexical context was largest at the more ambiguous levels of voice-onset time. These results have been described quantitatively by the FLMP with the basic assumption that acoustic featural information and lexical context make independent contributions to perceptual recognition (Massaro & Oden, 1980b).

17.5.3 Sentential Context

In a study of sentential context effects, Isenberg, Walker, and Ryder (1980) created a speech continuum between the function words the and to. The continuum was created by beginning with a natural utterance of the word to and attenuating the onset energy between 14 and 36 dB in steps of 2 dB. With little attenuation, the word is heard consistently as to; with a lot of attenuation, the word is heard as the. Intermediate levels of attenuation give more ambiguous percepts. The test word was placed as the initial word in one of two sentence contexts

To/the go is essential.

To/the gold is essential.

The only difference between the sentences is whether go or gold follows the test word. The appropriate syntactic constructions are "To go" and "The gold," and the experimental question was whether the syntactic context would influence perceptual recognition of the test word.

Subjects were presented with the twelve versions of the test word in the two sentence contexts and were asked to identify the test words as the or to. Figure 17.5 gives the observed results. As expected, the percentage of the identifications increased systematically with increases in the attentuation of the onset of the test word. Sentential context also had an effect, especially at the intermediate levels of attenuation of the test word. In terms of our model, the acoustic features in the test word and the syntactic constraints given by the sentence provide independent sources of information for identification of the test word. The predictions of the model are identical in form to those given in the study of phonological constraints. If F_i is the featural information supporting the percept the, and C_j is the syntactic information supporting this same percept, the probability of a the identification given the contexts gold and go are

$$P(\text{the:gold}) = \frac{F_i C_j}{F_i C_j + (1-F_i)(1-C_j)} \qquad (16)$$

$$P(\text{the:go}) = \frac{F_i (1-C_j)}{F_i (1-C_j) + (1-F_i) C_j} \qquad (17)$$

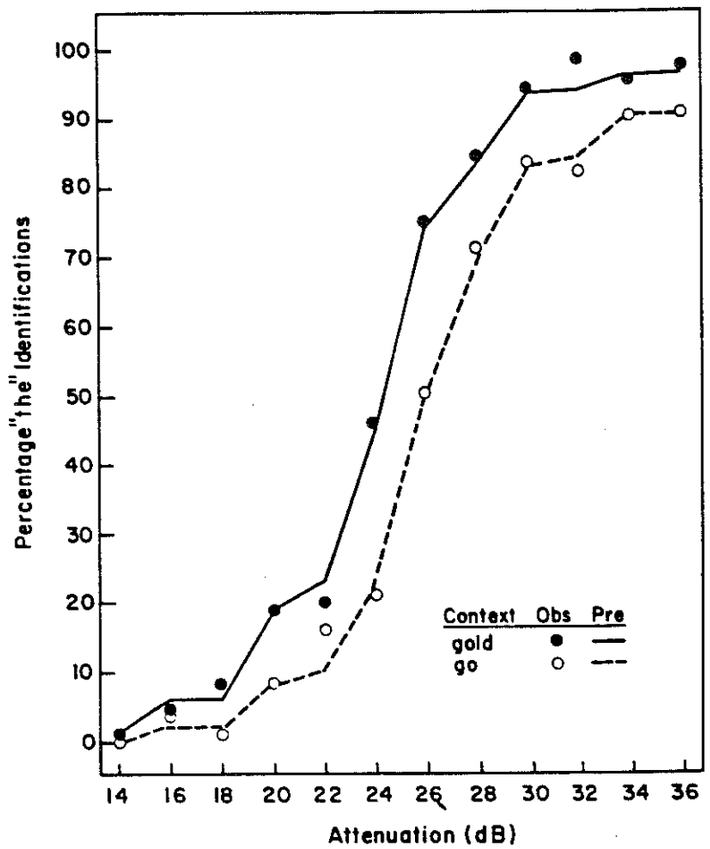

FIGURE 17.5 OBSERVED (POINTS) AND PREDICTED (LINES) PERCENTAGE OF "THE" IDENTIFICATIONS AS A FUNCTION OF ONSET ATTENUATION AND SENTENTIAL CONTEXT.

Figure 17.5 also gives the predictions of the model. The model provides a good description of the results with an average square deviation of less than 2 percent.

The role of context in speech perception is relevant to contemporary issues in psychology, phonology, and artificial intelligence. One persistent issue in psychological theory is whether or not the context effects modify lower level feature analysis processes (Broadbent, 1967; Morton, 1969). The description of the results given here and research in other domains provide strong evidence that context effects operate independently of lower level featural processing (Massaro, 1979). Recent theories of phonology (Chomsky & Halle, 1968; Ladefoged, 1975) and syntax have begun to give more weight to actual psychological performance and the present results indicate that phonological and syntactic semantic constraints are psychologically significant. Finally, with respect to artificial intelligence, it is now generally agreed that automatic speech recognition cannot be completely bottom-up but must involve the utilization of linguistic constraints in perception and recognition of the message (Klatt, 1977). The present results suggest that phonological constraints might be successfully utilized in automatic speech recognition by machine.

17.6 FUZZY TRUTH VALUES VERSUS PROBABILITIES

Fuzzy truth values are clearly not probabilities; yet there appears to be a close mathematical correspondence between models based on these concepts. In this section, a general decision theory based on probabilities is contrasted with the FLMP. The heart of the probability model is Bayes Theorem, which is an optimal decision rule for obtaining and revising probabilities. Bayes Theorem states that

$$P(H_1{:}E) = \frac{P(E{:}H_1) \times P(H_1)}{\sum_i P(E{:}H_i) \times P(H_i)} \qquad (18)$$

where $P(H_i{:}E)$ is the probability that some hypothesis H_i is true given that some evidence E is observed, $P(E{:}H_i)$ is the probability of the evidence E given the hypothesis H_i is true, and $P(H_i)$ is the a priori probability of the hypothesis H_i. The likelihood of hypothesis H_1 given some evidence E is equal to the likelihood of the evidence given the hypothesis times the a priori likelihood of the hypothesis divided by the sum of analogous likelihoods for all possible hypotheses. If the a priori probabilities of all possible hypotheses are equal, Bayes Theorem reduces to:

$$P(H_1{:}E) = \frac{P(E{:}H_1)}{\sum_i P(E{:}H_i)} \qquad (19)$$

It can be seen that if each hypothesis corresponds to a particular response alternative, the equation is similar in form to the pattern classification operation in the FLMP. The important question that remains is how different sources of evidence are combined according to Bayes Theorem. Given two pieces of evidence E_1 and E_2, the likelihood of a hypothesis H_1 is equal to

$$P(H_1{:}E_1 \text{ and } E_2) = \frac{P(E_1 \text{ and } E_2{:}H_1)}{\sum_i P(E_1 \text{ and } E_2{:}H_i)} = \frac{P(E_1{:}H_1) \times P(E_2{:}H_2)}{\sum_i P(E_1{:}H_i)\ P(E_2{:}H_i)} \qquad (20)$$

Equation 20 follows from probability theory in which the likelihood of the joint occurrence of two independent events is the multiplicative combination of the likelihoods of the separate events. The likelihood of two heads in two tosses of a coin is the multiplicative combination of the likelihood of a head on each toss. Given a multiplicative combination of independent sources of evidence in the FLMP, it is identical in form to a probability model based on Bayes Theorem. In the FLMP, a parameter is estimated for each level of each source of evidence. The same would be true with respect to the probabilities assumed by Bayes Theorem.

The notions behind Bayes Theorem seem to have a better justification for the subject's internalization of probability values, since previous experience would determine the probability values. The same might be said of the fuzzy truth values, but there has been no formalization of how

exactly the values are changed with experience. At some level, the process will be similar in form to the Bayesean analysis; the fuzzy logical value for some evidence E_1 will be related to the likelihood that the evidence came from some particular alternative. Bayesean analysis also has a mechanism for a priori probability. Fuzzy logic could provide such a mechanism by simply treating a priori probability as an additional source of information (Massaro, 1979).

17.7 AN ALTERNATIVE VIEW: CATEGORICAL PERCEPTION

An alternative view of speech perception has been proposed. People might be biologically disposed to discriminate easily those distinctions important to language and not to discriminate differences that are not informative. For example, it has been proposed that voice onset time (VOT) is the primary acoustic cue distinguishing voiced and voiceless stop consonants. Infants and adults are claimed to discriminate voiced and voiceless values of this cue, but apparently not a similar physical difference within either the voiced or voiceless category. According to this theory of categorical perception, we perceive a continuum of acoustic differences in terms of only two categories, and, ideally, the categories are represented by having very noticeable differences between categories and few, if any, noticeable differences within a category.

Speech perception, given such a sensory interface, would be an easy matter. Infants and children would learn to form speech categories on the basis of these categorical differences (Blumstein & Stevens, 1981). Stevens (1981) offers a number of distinctions in speech that might be conveyed categorically rather than continuously. These distinctions include a rapid amplitude change versus a slow amplitude change such as that which distinguishes the initial consonants of chop and shop. However, there is now good evidence that this distinction is not perceived categorically nor is the nonspeech analog of stimulus rise-time (Cutting & Rosner, 1974; Hary & Massaro, 1982; Rosen & Howell, 1981). Although Stevens (1981) provides some nice insights into some likely acoustic features in speech perception, there is no convincing evidence that these features are perceived categorically (Hary & Massaro, 1982); Massaro & Cohen, 1983).

Although categorical perception is often viewed as a very appealing explanation of speech perception (Gleitman & Wanner 1982), it is unlikely to be correct. One justification for the explanation is its simplification of the speech recognition problem. As noted by Gleitman & Wanner (1982), categorical perception of speech would provide the infant with the relevant linguistic categories. However, categorical perception clearly is not the case for all acoustic distinctions functional in speech perception. Vowel quality, segment duration, and frication quality are clearly noncategorically perceived, and yet these distinctions have been shown to be functional in speech perception. Thus, categorical perception cannot explain all of speech perception, and there is no reason that it should explain some of it. The theory that explains the noncategorical perception of speech contrasts might also explain those few cases that appear to be somewhat categorical.

Other reasons to reject the categorical perception of speech come from cross-language, learning, and context studies. Williams (1977) found different category boundaries along a voice-onset-time continuum for Spanish and English monolingual subjects. Did the different subjects have different

natural categories as would be necessary if perception were categorical, or do the results reflect simply the influence of language experience. That is, subjects must be able to learn to categorize graded sensory events depending on how they are used to represent categories in the language. As demonstrated by Pisoni, Aslin, Perey, and Hennessey (1982), English adults can easily learn to form a new category across a voice-onset-time continuum, even though the new category is not functional in their language. Given that these subjects were able to learn a new category in just a short training session, we have good evidence that the voice-onset-time continuum is perceived continuously. Finally, the differences in the voice-onset-time boundaries for the voicing of stop consonants as a function of place of articulation cannot be explained by simple psychoacoustic factors (Summerfield 1982). Thus, it is unlikely that a natural boundary separates voicing categories, and it is most reasonable to assume that the listener has continuous featural information in speech perception.

17.8 REFERENCES

Blumstein, S.E. & K.N. Stevens (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants, Journal of the Acoustical Society of America, 66, 1001-1017.

Broadbent, D.E. (1967). Word frequency effect and response, Psychological Review, 74, 1-15.

Chandler, J.P. (1969). Subroutine STEPIT - Finds local minima of a smooth function of several parameters, Behavioral Science, 14, 81-82.

Chomsky, N., & M. Halle (1968). The sound pattern of English. New York: Harper and Row.

Cole, R.A., & J. Jakimik (1978). Understanding speech: How words are heard. In, G. Underwood (Ed.) Strategies of information-processing, London: Academic press.

Cole, R.A., & B. Scott (1974). The phantom in the phoneme: Invariant cues for stop consonants, Perception and Psychophysics, 15, 101-107.

Cutting, J.E., & B.S. Rosner (1974). Categories and boundaries in speech and music, Perception & Psychophysics, 16, 564-570.

Fant, G. (1973). Speech Sounds and Features. Cambridge, Mass.: MIT.

Ganong, W.F. III (1980). Phonetic categorization in auditory word perception, Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.

Gleitman, L.R., & E. Wanner (1982). Language acquisition: The state of the state of the art. In E. Wanner and L. R. Gleitman (Eds.), Language acquisiton: The state of the art, Cambridge: Cambridge University Press.

Goguen, J.A. (1969). The logic of inexact concepts, Synthese, 19, 325-373.

Hary, J.M., & D.W. Massaro (1982). Categorical results do not imply categorical perception, Perception & Psychophysics, 32, 409-418.

Isenberg, D., E.C.T. Walker, & J.M. Ryder. (1980).A top-down effect on the identification of function words, Journal of the Acoustical Society of America, 68, AA6 (abstract).

Klatt, D.H. (1977). Review of the ARPA speech understanding project, Journal of the Acoustical Society of America, 62, 1345-1366.

Ladefoged, P. (1975). A Course in Phonetics. New York: Harcourt, Brace, and Jovanovich.

Liberman, A.M., F.S. Cooper, D.P. Shankweiler, & M. Studdert-Kennedy (1967). Perception of the speech code, Psychological Review, 74, 431-461.

Lopes, L. (1981). Decision making in the short run, Journal of Experimental Psychology:Human Learning and Memory, 7, 377-385.

Luce, R.D. (1959). Individual Choice Behavior. New York: Wiley.

Marslen-Wilson, W. & A. Welsh (1978). Processing interactions and lexical access during word recognition in continuous speech, Cognitive Pychology, 10, 29-63.

Massaro, D.W. (1975a). Experimental Psychology and Information Processing. Chicago: Rand-McNally.

Massaro, D.W. (Ed.) (1975b). Understanding Language: An Information Processing Analysis of Speech Perception, Reading

and Psycholinguistics. New York: Academic Press.
Massaro, D.W. (1979). Reading and listening (Tutorial paper). In P. A. Kolers, M. Wrolstad, & H. Bouma (Eds.) Processing of Visible Language, 1. New York: Plenum, 331-354.
Massaro, D.W., & M.M. Cohen (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction, Journal of the Acoustical Society of America, 60, 704-717.
Massaro, D.W., & M.M. Cohen (1983). Categorical or continuous speech perception: A new test, Speech Communication, 2, 15-35,
Massaro, D.W., & M.M. Cohen (in press). Phonological constraints in speech perception, Perception and Psychophysics.
Massaro, D.W., & G.C. Oden (1980a). Evaluation and integration of acoustic features in speech perception, Journal of the Acoustical Society of America, 67, 996-1013.
Massaro, D.W., & G.C. Oden (1980b). Speech perception: A framework for research and theory. In N.J. Lass (Ed.), Speech and Language: Advances in Basic Research and Practice. New York: Academic Press.
Morton, J. (1969). Interaction of information in word recognition, Psychological Review, 76, 165-178.
Oden, G.C. (1977). Integration of fuzzy logical information, Journal of Experimental Psychology: Human Perception and Performance, 3, 565-575.
Oden, G.C., & D.W. Massaro (1978). Integration of featural formation in speech perception, Psychological Review, 85, 172-191.
Palmer, S.E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B.B. Lloyd (Eds.), Cognition and categorization. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Pisoni, D.B., R.N. Aslin, A.J. Perey, & B.L. Hennessy (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants, Journal of Experimental Psychology: Human Perception and Performance, 8, 297-314.
Pollack, I., & J.M. Pickett (1964). The intelligibility of excerpts from conversation, Language and Speech, 6, 165-171.
Roberts, A.H. (1965). A Statistical Analysis of American English. Paris: The Hague.
Rosen, S.M., & P. Howell (1981). Plucks and bows are not categorically perceived, Perception & Psychophysics, 30, 156-168.
Rumelhart, D.E., & D.A. Norman (1983). Representation in memory, Center for Human Information Processing , 116, 1-117.
Stevens, K.N. (1981). Constraints imposed by the auditory system ontheproperties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics, In T.Myers, J.Laver, &J.Anderson (Eds.) The Cognitive Representation of Speech, Amsterdam: North-Holland.
Summerfield, Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops, Journal of the Acoustical Society of America, 72, 51-61.
Williams, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals, Perception & Psychophysics, 21, 289-297.
Zadeh, L.A. (1965). Fuzzy sets, Information and Control, 8, 338-353.

18.

VOWELS IN CONTEXT: DYNAMICS, STATISTICS, AND RECOGNITION

David J. Broad
1627 Bath Street
Santa Barbara, CA 93101

ABSTRACT

Vowels occur as dynamic gestures in the speech stream and their "steady states" are only fleeting and usually realized as reduced for perturbed versions of their idealized "targets". Fortunately, variability due to context seems to display some important regularities: (1) Effects from preceding and following sounds seem to combine linearly to a good approximation. These effects can be represented as initial- and final-context transition functions. (2) While such transition functions are not shaped the same for all consonants, many of them are and there is some hope that they may be shaped similarly for all vowels on a "per-consonant" basis. (3) There is evidence that the vowel target undershoot may be directly proportional to the distance between the vowel target and the adjacent consonant locus, with each consonant having its own constant of proportionality. (4) Perturbation of the endpoints of vowel formant trajectories away from their respective consonant loci may also be proportional to target-locus distance on a per-consonant basis. (5) Undershoot is a decaying exponential function of vowel duration. The time constant and scale of the exponential depend on the consonant.

The linearity of the combination of initial and final consonant gestures in CVC syllables has three major implications: (1) The description of formant trajectories is greatly simplified. (2) Data bases for modeling CVC syllables can be built "additively" from data on CV's and VC's, avoiding combinatorial explosion of studying all C1VC2's. (3) Such an "additive" data base provides a natural way to incorporate an unnormalized time scale, thus providing a way to model duration-dependent dynamics. An implication of the duration effects is that dynamic programming, or time warping without alternation of the formant configuration, will introduce systematic errors into vowel patterns.

These ideas have so far only been developed and tested on a limited set of speakers and languages with only a thin sampling of context and vowel combinations. Nevertheless they seem promising and amenable to reasonably scaled data bases, which should let us see how they work under a usefully general set of conditions and to determine how they can be used to handle context effects in automatic speech recognition.

1. INTRODUCTION

1.1. Coarticulation

It has long been known that vowel configurations are affected by context. Vowel characteristics change dynamically through the vowel gesture and change from instance to instance as the vowel occurs in different environments. This poses problems for simple methods for automatic vowel recognition. On the other hand, these variable characteristics carry information about the vowel and the sounds adjacent to it. Therefore it seems likely that good ways to handle contextual variations of vowels should provide better recognition of both the vowels and their adjacent sounds.

The characterization of context phenomena is a fundamental problem in acoustic phonetics. Its solution involves the systematic examination of sounds in various contexts and the development of formal models to characterize the results.

The generally accepted articulatory explanation for context effects on vowels is that each vowel has a certain "target" configuration toward which the articulators will head from some preceding sound. Before this target can be realized, the articulators will start to anticipate the following sound and the vowel target will be undershot.

This dynamic explanation for context effects on sounds is sometimes called "coarticulation" because at any given time the articulatory and acoustic configurations are functions not only of the "current" sound, but of the preceding and following ones as well. That is, the current sound is "coarticulated" with its predecessors and successors.

1.2. Acoustic Parameters

This paper deals with vowels and context effects in terms of formant frequencies. This is in part because the formants have been effective parameters for describing vowels, and in part because of their consequent use as the "universal language" in which past studies have been formulated. These past studies include nearly all the ones cited here.

There are many other acoustic parameters that can be, and are, used in speech analysis and which could be applied to vowels and coarticulation. Pols (1977) has done a nice study of Dutch vowels in CVC context using as parameters the principal components of the signal spectrum.

It would seem naive to expect that context effects might be banished by the use of a different parameter set. What seems more compelling for the use of non-formant parameters is that formant tracking is such a difficult and still unresolved problem. Given this, it becomes not only interesting but necessary to study the behavior of vowels in context using other parameters.

Regardless of the acoustic parameters used to study or recognize vowels in context, the same issues must be addressed: How do overlapping phonetic gestures combine in the acoustic domain? How can they be characterized in simple forms? How can such knowledge of context-sensitive characteristics be effectively used in speech recognizers? Since these questions are independent of the particular parameters used to represent vowels acoustically, what we learn below about the formant frequencies of vowels in

context should provide a useful framework for characterizing context effects with any reasonable acoustic representation.

2. SINGLE-CONTEXT FORMS

2.1. Vowel Reduction

Stevens and House (1963) found that the short vowels /I/ and /U/ were more subject to contextual shifts than long vowels, and suggested that this meant that undershoot was related to the time available for a vowel gesture.

Duration dependency was studied systematically in one Swedish speaker by Lindblom (1963), who measured vowel formant frequencies at the onsets, steady states, and ends of vowels in the symmetric contexts /b--b/, /d--d/, and /g--g/. He elicited each item at 4 different speaking rates. The second formant at the vowel midpoint, F20, was found to be well characterized by the relation:

$$F20 = F2t + k(F2i-F2t)\exp(-b\tau) \qquad (1)$$

where k and b are constants depending only on the consonantal context, τ is the vowel duration in milliseconds, F2t is the target frequency for the vowel's second formant, and F2i is the second formant measured at the vowel onset. For contexts /b--b/, /d--d/, and /g--g/ the values of k were found to be 5.0, 2.0, and 1.5, and those for b were 0.021, 0.012, and 0.010, respectively.

The second term, which represents the perturbation of F20 from the target F2t, decreases in magnitude with duration. For a given context, the size of the term is directly proportional to the distance between the target F2t and the F2 starting point F2i. Undershoot therefore depends on the scale of the phonetic gesture and on the time available for it.

The initial and center values of F2 show a remarkable relationship in Lindblom's data. Figure 1A shows these values plotted against each other for a vowel duration of 150 ms. Within each vowel category, the realized center frequency increases gently with initial frequency. It is easy to see how vowel recognition would be confused on the basis of the center frequency alone. For example, F20 of /a/ in context /g--g/ is nearly the same as that of /œ/ in context /b--b/.

The regular arrangement of vowels in the figure suggests the coordinate transform shown in Figure 1B. The D1 axis is the image of the straight line fit to the /b--b/ data in Figure 1A. D1 is distance along this line from the point (820,600) in Figure 1A. D2 is distance from the D1 axis along a line whose slope is the average of the slopes of the 8 vowel lines in Figure 1A.

Vowels are well separated by D1, except for /ɐ/ and /Y/, which are separated by their first formants. At the same time, most contexts seem fairly well separated by D2 on a per-vowel basis.

This example shows that even minimal attention to the dynamics of formant frequencies --looking at trajectory endpoints in addition to the vowel center point-- might be useful in handling context effects in automatic recognition.

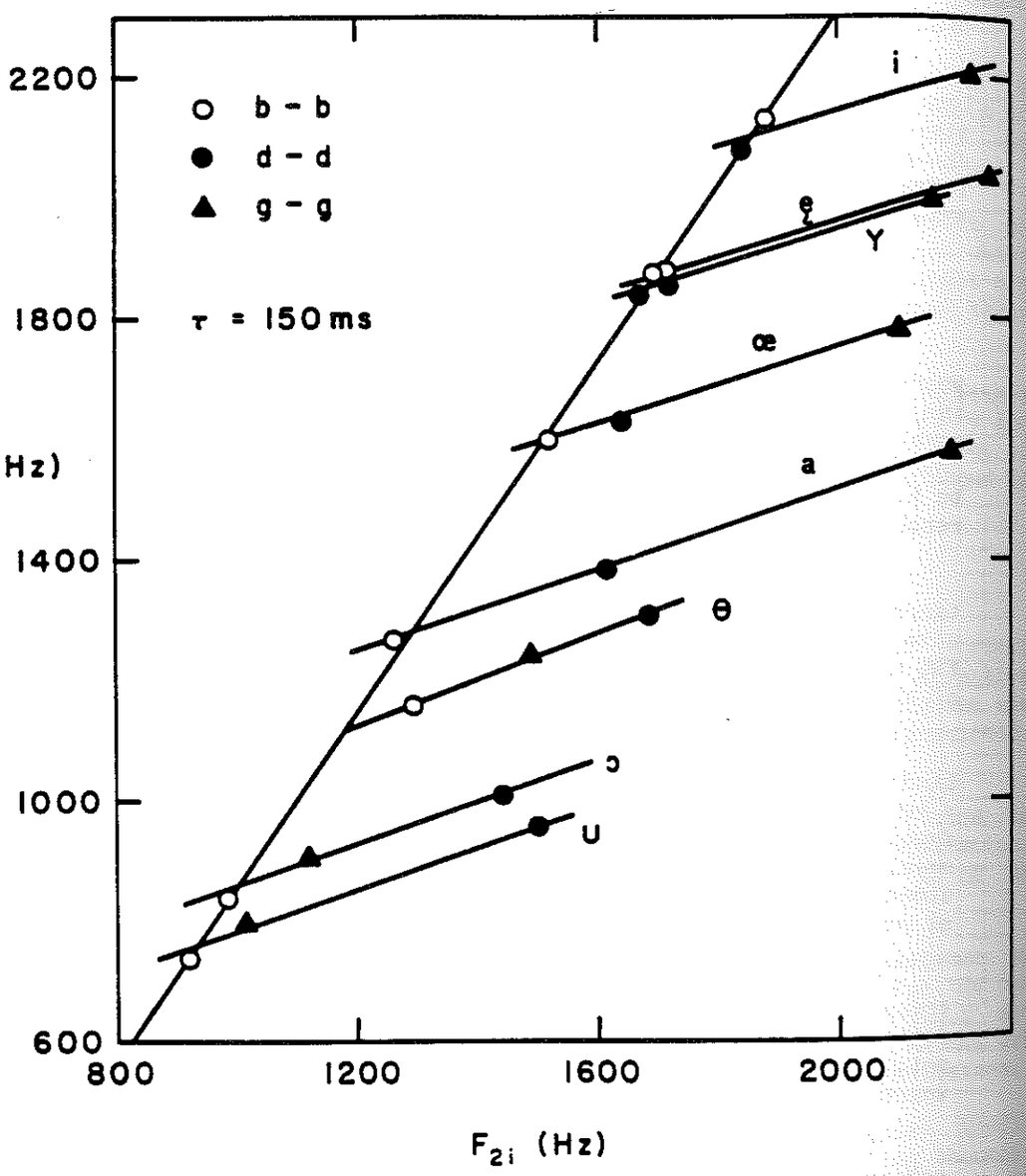

Figure 1A. Second formant at the vowel center F20 for a duration of 150 ms plotted against the second formant at the vowel onset, F2i, using the data and model of Lindblom (1963).

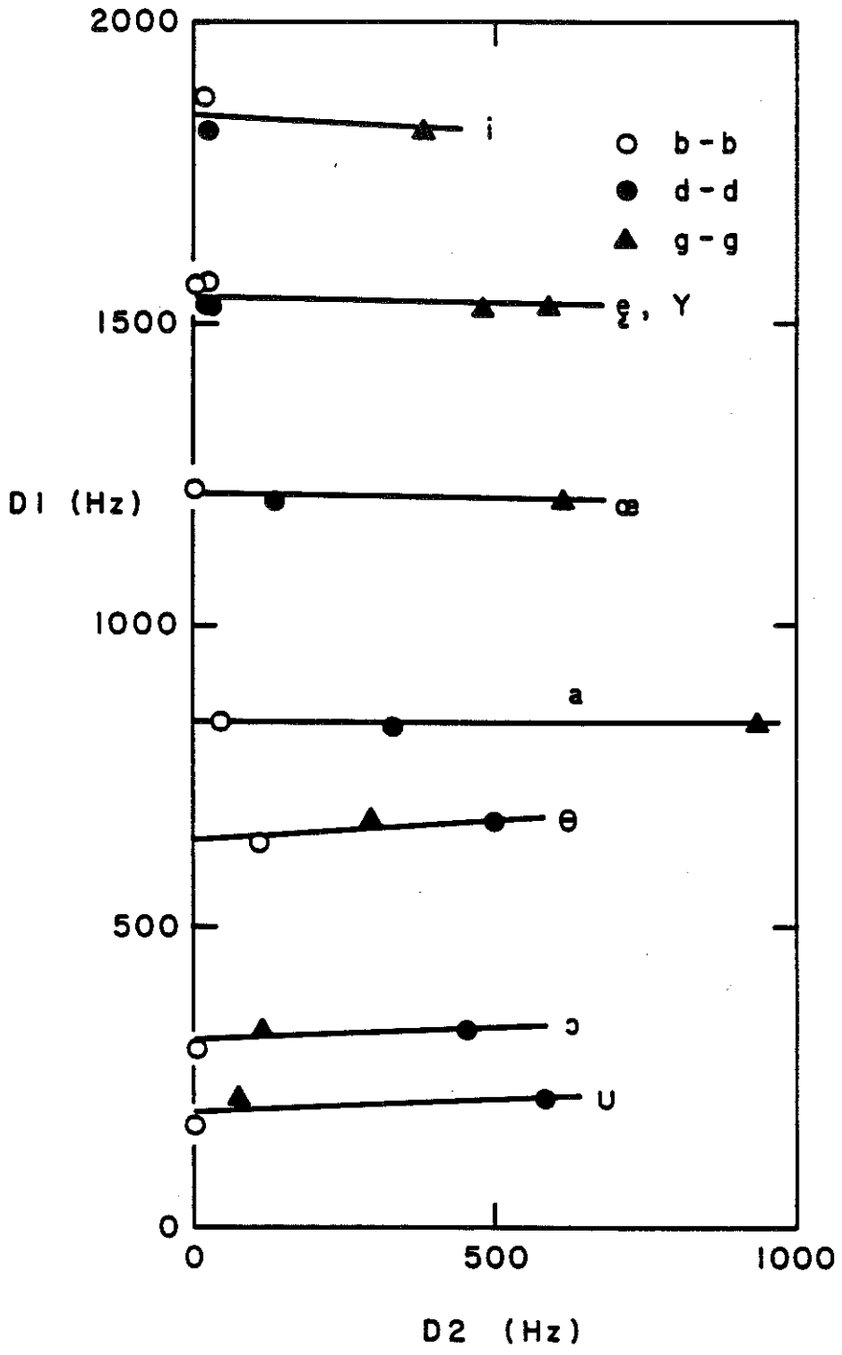

Figure 1B. Figure 1A with transformed coordinates (see text).

2.2 Articulatory Modeling

Öhman (1967) has formulated a coarticulation model at the level of vocal tract area functions measured from x-ray motion pictures of VCV utterances. The model is of the form:

$$s(x;t) = v(V,x) + k(t)w(C,x)\,[c(C,x) - v(V,x)] \qquad (2)$$

where s(x;t) is the vocal tract area as a function of the distance x along an axis from the glottis to the lips, and of time t; v(V,x) is the target area function for vowel V; c(C,x) is a target area function for consonant C; w(C,x) is a coarticulation function for consonant C (it is a weighting on the distance between the consonant and vowel targets), and k(t) is a function of time denoting the extent of completion of the gesture from V to C.

2.3. Conceptual Similarity of Single-Context Forms

Equation 2 resembles term for term Equation 1 for Lindblom's vowel reduction results: As illustrated in Table I, each involves a vowel target added to a perturbation term which is a product of an inter-target scale, a consonant-dependent scale factor, and a "dynamic" factor involving duration or time. Although the equations are formulated for different domains, formant frequencies and area functions, they share strong conceptual similarities involving hypotheses of per-consonant similarity of gestures and distance-proportional scaling. These notions will be taken up again in Sections 4 and 5.

Table I. Structural correspondences between Equations (1) and (2), which are, respectively, Lindblom's (1963) formula for F2 vowel-target undershoot and Öhman's (1967) coarticulation model for vocal tract area functions.

Term	Lindblom (1963) Eq. (1)	Öhman (1967) Eq. (2)
Realized Configuration	F20	s(x;t)
Vowel Target	F2t	v(V,x)
Locus-Target Distance	F2i - F2t	c(C,x) - v(V,x)
Dynamic Factor	exp(-bτ)	k(t)
Consonant-Dependent Scale Factor	k	w(C,x)

2.4. Effects of Consonant Features

2.4.1. Consonant Features As Additive Effects. Stevens and House (1963) found systematic effects on vowel formant frequencies attributable to the features of voicing and place and manner of articulation of the influencing consonant. Voicing usually moved F1 down, possibly because of larynx lowering. For F2, fricatives had greater perturbing effects than stops, possibly because they require more precision of movement and therefore more time than stops. The labial consonants tended to have greater effects on F2 than did consonants with other places of articulation. These results suggest that coarticulation might be resolved into separate effects from the consonant features. If so, it would be desirable to express these effects quantitatively.

Purcell (1979) has pursued this notion by representing additive consonant effects as linear combinations of effects of consonant features. He measured the formants in Russian VCV utterances and found that the consonant effects were characterized by relations of the typical form:

$$F2(TV1) = 322 - 277(\text{front/back V1}) + 1.041(\text{place of art}) + 290(\text{pal/unpal}) \qquad (3)$$

In this example, Purcell relates the second formant measured at the vowel-to-consonant transition for the initial vowel (TV1). The parameter "front/back V1" is set equal to 1, 2, 3, 4, or 5 for /i/, /e/, /a/, /o/, or /u/, respectively. "Place of art" is set to 1423, 1834, or 1637 for /b/, /d/, or /g/, respectively. "Pal/unpal" is set to 1 for an unpalatalized consonant and to 2 for a palatalized one. The values of the coefficients and the parameters are determined for a best fit to the data.

Models of this general form fit the F1 data very well: rms errors were near 20. Hz. The fits for F2 were not so good; rms errors were around 225 Hz. Nevertheless, most (66-80 percent) of the variance was accounted for. Significant improvements in this type of model might be expected if the constraints of the language mentioned by Purcell were incorporated. For example, some items of the script were for non-occuring sequences in Russian and were uttered as the nearest occuring sequences. This shows that coarticulation models cannot be expected to pick up much of the burden more properly carried by phonological rules.

2.4.2. Palatal and Velar Allophones. Models of coarticulation will indeed run into trouble unless phonological rules governing allophone selection are first taken into account. An example is the conditioning of /k/, /g/, and /ŋ/ in English by the horizontal place of articulation of the adjacent vowels: In the context of front vowels these sounds are realized by their palatal allophones, and in the context of back vowels by their velar allophones. One way to handle this would be to treat these allophones as the different consonants that they are at the phonetic level. One would then include phonological rules in the model to select the appropriate allophones depending on the frontness or backness of the contextual vowels.

These sounds may actually behave in a more complex manner. Houde (1968) has made x-ray motion pictures of VCV utterances in which he observed that the closure for /g/ was accomplished mainly by a vertical tongue-body

movement, with the horizontal place of closure representing a continuum conditioned by the horizontal place of articulation of the contextual vowel. Thus it seems likely that the discrete form of a simple phonological rule might more realistically have to be represented by some continuous function.

Öhman (1967) has suggested that this might be accomplished for /g/ in the formalism of Equation 2 by having the /g/ target area function c(/g/,x) depend also on a "velar/palatal" parameter which would be set equal to his parameter q2, which can be interpreted as a back-front parameter of the adjacent vowel sound. The target function would then have the form c(/g/,q2;x). This appears to agree with Houde's observations.

A continuous conditioning of palatal and velar consonants by the adjacent vowel is an example of coarticulation effects of vowels on consonants. Whether a given conditioning phenomenon should be modeled as a phonological rule of assimilation or as a coarticulation effect would depend on how it can actually be observed to behave in a body of data.

In general, the articulations of bilabial, alveolar, and palatal or velar consonants involving complete oral closure have mechanisms and degrees of freedom that differ considerably from each other: In the bilabial sounds, the entire tongue body is free to assume any shape or position short of closure; in the alveolar sounds the apico-alveolar contact places a "fixed-end" constraint on the tongue; in the palatal and velar sounds the tongue is, as just noted, apparently free to adopt any of a wide choice of horizontal places for closure. Such differences of mechanism may underlie the necessity for abandoning notions of universal shape similarities and proportionalities of scale in favor of "per-consonant" versions of these notions, as discussed later in Sections 4.3 and 5.

3. SUPERPOSITION OF PHONETIC GESTURES

3.1. Initial Formulation

3.1.1. Whole Gesture Versus Center Sample. Without data on more contexts, it is impossible to determine how much of the vowel target undershoot to attribute to the initial context and how much to the final. Also, it is desirable to measure more time samples of the formant trajectories to study how the balance among initial context, final context, and vowel target shift from beginning to end of the vowel. More samples allow the whole formant trajectory to be characterized as a phonetic gesture. From this point of view, it is the whole gesture and not just the vowel midpoint or steady state that is most useful for characterizing the phonetic category of the vowel. Because the center or steady state value generally undershoots the vowel target, this single sample by itself can be misleading and can be given its proper phonetic interpretation only when it is seen as a part of the trajectory, the trajectory that contains information on the degree of undershoot, and hence on the vowel target. If a vowel is to be characterized by a single set of formant frequencies, then, it would appear to be better to use the targets than any single sample of the frequencies actually realized during the course of the vowel's phonetic gesture. It is then the job of a recognition-directed coarticulation model to specify the mapping of vowel gestures onto vowel targets. It will also be seen in Section 4.2.2 that more time samples of the formant trajectories indeed capture more of the phonetic information carried by the formant contour.

Stevens, House, and Paul (1966) embodied the whole-gesture idea by sampling vowel formant trajectories every 8.3 ms through the vowel gesture in CVC utterances. These trajectories were then fit by segments of parabolic arcs.

3.1.2. Separating Initial and Final Contexts. Broad and Fertig (1970) pursued the separability of initial and final contexts by studying a single vowel in a variety of initial and final contexts. Fertig (1976) has described the details of measurement and segmentation. The utterances were the 576 possible C1/I/C2 syllables made up of combinations of 24 C1's with 24 C2's. Each context was recorded 3 times from a single speaker. Finally, the whole-gesture idea was incorporated by sampling the first three formant frequencies of each vowel at 11 equally spaced time points.

3.1.3. Analyses of Variance. Two-way analyses of variance of the vowel formant data provided a sensitive measure of the effects of C1 and C2 on the formant frequency trajectories. These analyses are summarized in Figure 2 in which the F-ratios (in the usual statistical sense of variance-estimate ratios, not formant frequencies) for F1, F2, and F3 (formants!) are plotted as functions of time for the C1 and C2 main effects and for the C1-C2 interaction. All three formants show the same pattern: The C1 main effect decreases with time as the C2 main effect increases. The two sets of main effects make an X-shaped pattern in the plot. Even where the main effects are the weakest, at their trans-vowel boundaries, they are much larger than the interactions, which are the more or less horizontal functions at the bottom of the graph. This shows that there is both "memory" of C1 and "anticipation" of C2 throughout the vowel duration.

The F's for C1-C2 interaction are highly significant statistically, even though they are numerically small. For simplicity, C1-C2 interaction terms are omitted from the model that follows. Though this is not strictly justified, the model turns out to be fairly robust under this departure from rigor, which, according to Broad and Fertig, increases the rms error of the model by an average of only 15 percent. The maximum increase in error is 32 percent, which occurs for time point 4 of F3. Figure 2 shows that this also corresponds to the highest value for the C1-C2 interaction F.

3.1.4. Linear Model. The analyses of variance lead to a very simple coarticulation model in which the effects of the initial and final contexts combine additively throughout the vowel duration. The model is given by the relation:

$$F(C1,C2,i;t) = T(i) + f(C1,i;t) + g(C2,i;t) + X(i;t) \qquad (4)$$

where F(C1,C2,i;t) is the ith formant-frequency trajectory for /I/ in the context of consonants C1 and C2, T(i) is the ith target frequency for the vowel /I/, f(C1,i;t) is an initial-consonant transition function for consonant C1, g(C2,i;t) is a final-consonant transition function for consonant C2, and time t is normalized to run from 0 to 1 from the beginning of the vowel to its end. The 11 equally spaced points therefore correspond to times t = 0, 0.1, ..., 0.9, 1.0. The final term X(i;t) is the error in the representation. It is a zero-mean nearly-gaussian random variable, composed mainly of random variations among repetitions of the script items by the speaker along with some measurement noise. It also contains a relatively small component attributable to the ignored interactions between C1 and C2.

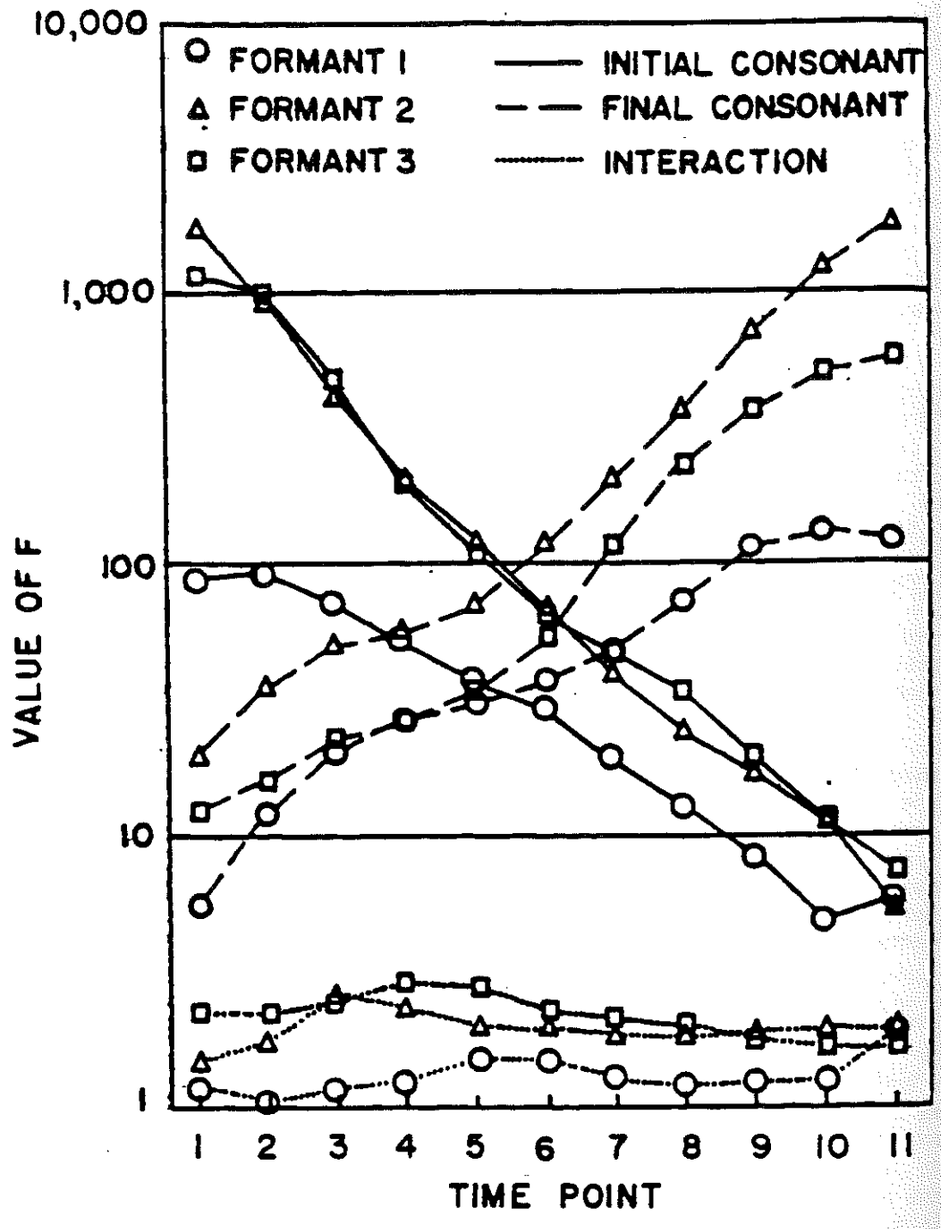

Figure 2. F-statistics for 33 two-way analyses of variance on the first three formant frequencies of C1/I/C2 syllables classified by C1 and C2. The time axis represents the 11 equally spaced samples of the formant trajectories. For each time point, each formant has its own analysis of variance, which results in three F-statistics: one for the C1 main effect, one for the C2 main effect, and one for the C1-C2 interaction effect. From Broad and Fertig (1970). Reproduced by permission of the American Institute of Physics.

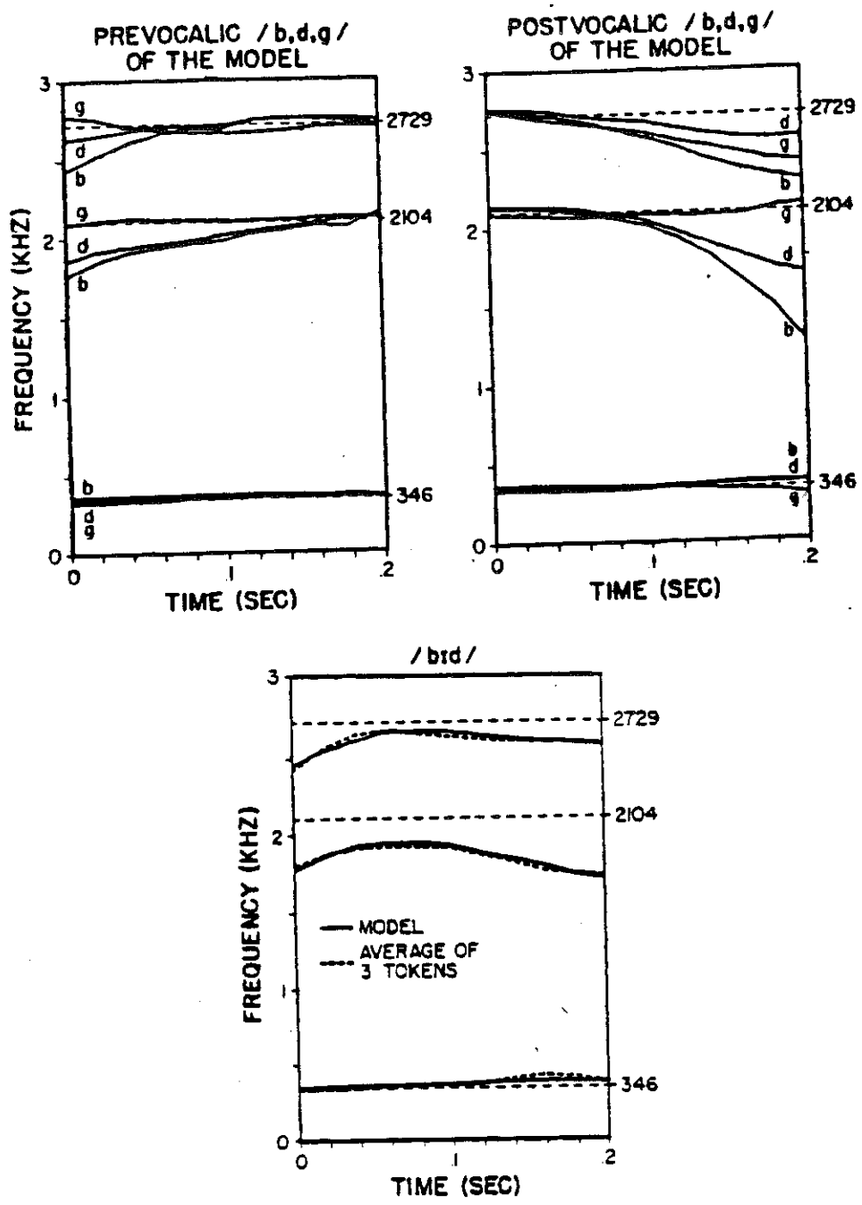

Figure 3. Example of the linear model of Broad and Fertig (1970). The top two panels show the initial- and final-consonant transition functions for the first three formant frequencies of the vowel /I/ with consonants /b/, /d/, and /g/ plotted using the respective vowel-target frequencies as baselines. The bottom panel shows how the formant trajectories for the sequence /bId/ are constructed by adding the transition functions for /bI/ and /Id/ from the top panel to the target frequencies. Also shown is the average trajectory for /bId/ measured from 3 tokens in the data. Figure courtesy of Dennis H. Klatt.

Figure 3 shows how the model in Equation 4 works. The figure, which is the best one I have seen for explaining the model, was drawn by Dennis Klatt using the data and model in the study by Broad and Fertig. The horizontal dashed lines represent the targets for the /I/ formant frequencies. The curves in the top two graphs show initial and final consonant transition functions for a small subset of the consonants studied: /b, d, g/. The transition functions for each Fi are drawn using their respective target frequencies as baselines. Thus they represent idealized C/I/ and /I/C trajectories. The bottom graph shows how the formant trajectories for the sequence /bId/ are constructed by superposing the F1, F2, and F3 f's for /bI/ and the respective g's for /Id/. Also shown is the actual average trajectory measured from three tokens of /bId/ in the data.

The fit between the model and the data is quite good. Indeed, the rms errors of the superposition model are only slightly larger than the rms variation among repetitions of the same item by the same speaker. (The increase is due to the omission of the C1-C2 interaction term in Equation 4.) That is, the model is nearly ideal in consideration of the limits on the human talker's ability to reproduce a given pattern from time to time.

3.1.5. Cluster Contraction. To see how this kind of model might aid automatic vowel recognition, consider the special case in which we know C1 and C2 beforehand. Then if Fi is a formant frequency measured at the vowel midpoint, t = 0.5, the corresponding target Ti can be estimated from Equation 4 as:

$$T(i) = F(C1,C2,i;0.5) - f(C1,i;0.5) - g(C2,i;0.5) \qquad (5)$$

Figure 4 shows the result of this operation on the /I/ data. Figure 4A shows the first two formant frequencies actually measured for the center point of the vowel /I/ for all the 1,728 tokens. As mentioned above in Section 3.3.1, there is no reason to consider the center or steady state formant values to characterize the vowel's phonetic category. In this case, the center value is distinguished from the others in that it has smaller variance than any of the other time samples. Figure 4A therefore shows the most compact distribution of formants for /I/ that can be obtained from single samples of the realized formant frequencies.

When Equation 5 is applied to each point in this plot, Figure 4B is obtained. Clearly it is a substantially tighter cluster than the one in Figure 4A. It represents the distributions of the error terms X(1;0.5) and X(2;0.5) from Equation 4. Subtracting the transition functions thus may be an effective method for estimating vowel targets, provided we know which ones to subtract! In a pure phonetic sense, it is also seen that the estimated targets provide a better acoustic characterization of the vowel than do the realized center frequencies.

This illustration is a long way from being a practical recognition technique. Besides needing to know C1 and C2, one really needs to know that the vowel is /I/, as f and g in Equation 4 have been constructed only for this vowel. This does not mean, however, that a more general coarticulation model will suffer the same limitation. Indeed, it is to be hoped that it will not and will achieve results such as those shown in Figure 4 without having to know the phonetic units in advance. Our ability to do this will depend on knowing regularities of contour shapes and scales, notions to be reviewed below. In particular, it will be shown that shape similarity among

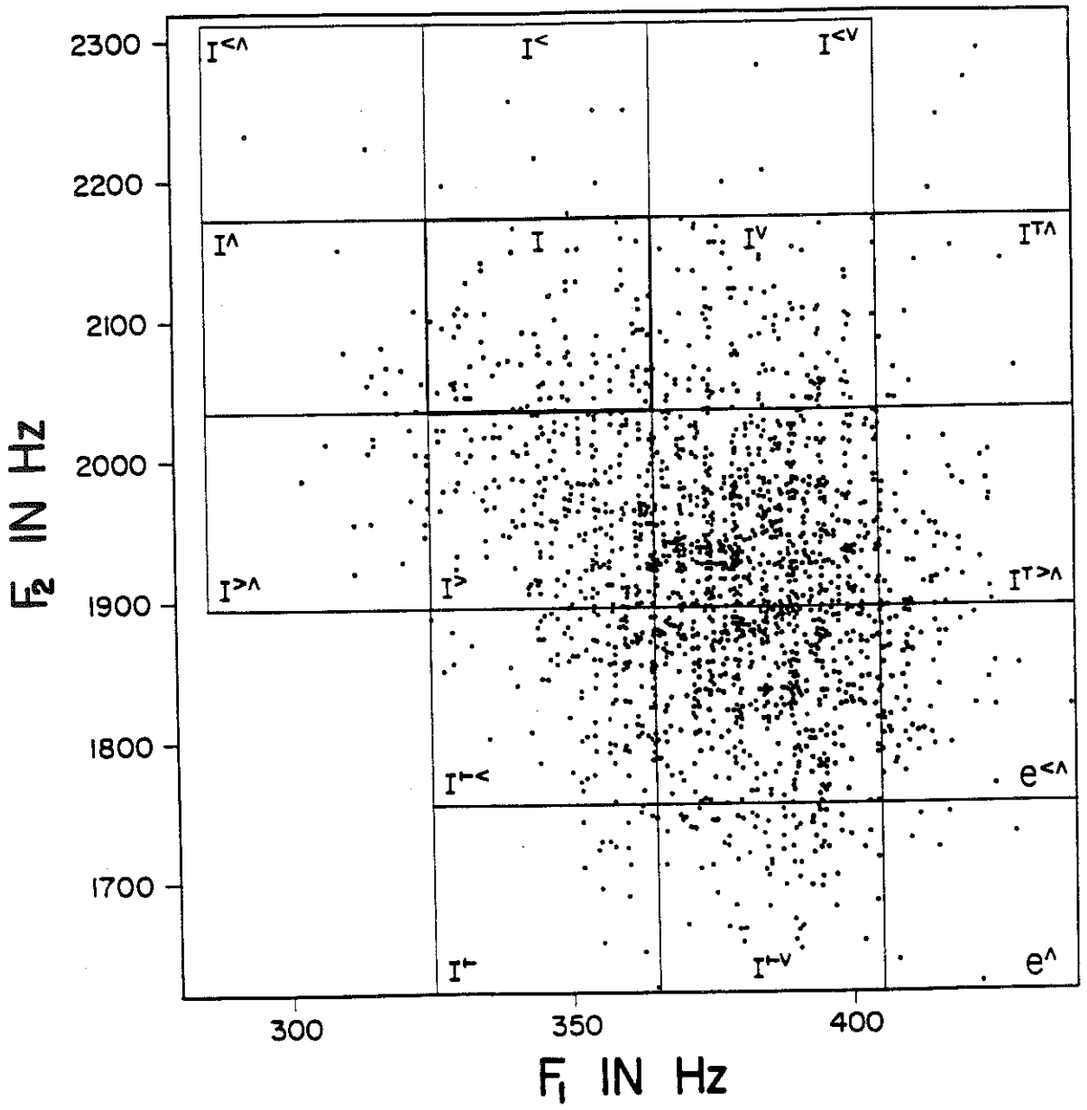

Figure 4A. Second formant frequency plotted against the first for the actually realized vowel centers of the 1,728 tokens of /I/ in the data of Broad and Fertig (1970). From Broad (1976). Reproduced by permission of S. Karger AG, Basel.

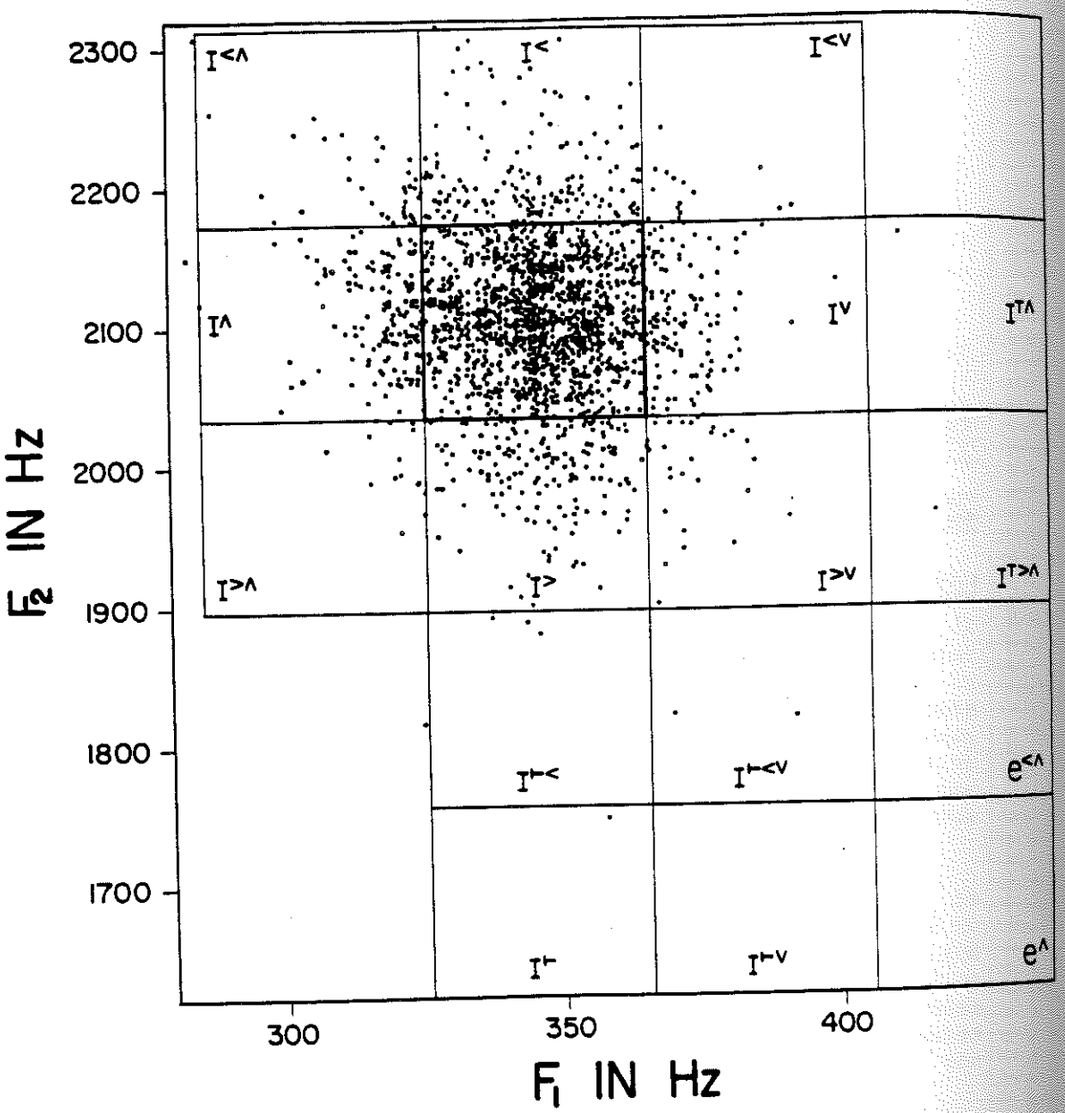

Figure 4B. The corresponding F1 and F2 target frequencies estimated by subtracting the appropriate C1 and C2 transition functions from each point in Figure 4A. From Broad (1976). Reproduced by permission of S. Karger AG, Basel.

transition functions would be one factor that would allow the vowel target to be estimated solely from samples of the realized formant contour.

The lesson of the present example is: There exists systematic structure in vowel parameters attributable to context. It is up to us to find a way to characterize and exploit this structure.

These data and the coarticulation model for them were derived for just this one vowel for this one speaker. Nevertheless, they are the starting point for a new study that is just getting under way.

3.1.6. Applicability to Other Vowels. There is some indication from the literature that additive consonant effects might characterize the steady states of other vowels. Öhman (1966) measured the vowel formant frequencies in 375 Swedish VCV utterances (5 repetitions x 5 initial vowels x 3 consonants x 5 final vowels). He found that the vowel formant frequencies at the consonant boundaries were affected by the trans-consonantal vowel. The effect seems to be significant even after an error in Öhman's statistical reasoning is allowed for: The effect was tested by computing the mean boundary values for each of the 5 trans-consonantal vowels and then applying Student's t-test to the two means that were most different. Selecting the extreme values after the fact biased the test toward a significant result. Nevertheless, the claimed coarticulation of the vowel with the transconsonantal vowel does seem to be present (Broad, 1972), and it would be interesting to construct a statistical model of Öhman's VCV data along the lines of Broad and Fertig's model.

For our present purposes, it is useful that Öhman's vowel data are given in enough detail (his Table II) to calculate approximate 2-way analyses of variance of the data classified by vowel and by associated consonant. The results of these analyses are condensed in Table II, where it is seen that the F-ratios (F again in its statistical sense) for interaction between vowel and consonant are generally either not significant (F1) or small in relation to the main effects of the vowels and consonants (F2 and final-position F3).

Table II. F-ratios for analyses of variance of the vowel formant data in Öhman's (1966) Table II. Separate analyses for vowels in initial and final positions in the VCV sequences are given.

Formant	1		2		3	
Position	init.	fin.	init.	fin.	init.	fin.
Effect						
Vowel Main*	1,813	1,302	12,628	7,993	238	112
Consonant Main**	234	22	51	120	26	54
Interaction***	2.77	1.52	10.9	4.76	31	11.1

*critical F, 99.5% level, = 3.8
**critical F, 99.5% level, = 5.4
***critical F, 99.5% level, = 2.8

This suggests that most of Öhman's vowel steady-state data can be represented by a simple additive model similar to Equation 4 in which only the transition function for one nearby consonant is needed:

$$F(V,C,i) = T(V,i) + f(C,i) \qquad (6)$$

where T(V,i) represents the ith target frequency for vowel V, and f is a consonant-related perturbation. T(V,i) is computed for each vowel V as F(V,C,i) averaged over all consonants C; f(C,i) is F(V,C,i) - T(V,i) averaged over all vowels V for each consonant C. Time is suppressed as a variable in Equation 6 because the published data are for only single time samples in the vowel steady-states.

The results of constructing the model in Equation 6 are summarized in the top row of Table III which shows the rms errors in estimating vowel targets for each CV or VC combination in Öhman's table by two different methods:

(A) T(V,i) is estimated by the realized frequency F(V,C,i).

(B) T(V,i) is estimated by subtracting the consonant perturbation f(C,i) from each realized F(V,C,i).

The success of the model is shown by the reduction in rms error from method (A) to method (B): by 48, 58, and 31 percent for F1, F2, and F3 respectively.

The bottom row of Table III gives the corresponding rms errors for the scatter plots for the /I/ data shown in Figure 4. The results seem to be comparable to those just described for Öhman's data, as in each case the error is substantially reduced by subtracting consonant effects from realized frequencies.

There is therfore some indication that additive context effects apply to situations more general than the single vowel /I/ studied by Broad and Fertig.

Table III. Rms errors in Hertz of vowel targets estimated via (A) realized steady-state formants, and (B) subtraction of consonant-dependent perturbations.

Formant	1		2		3	
Method	A	B	A	B	A	B
Study						
Öhman (1966)	13.1	6.8	38.6	16.3	49.6	34.4
Broad & Fertig (1970)	22	16	105	63	108	78

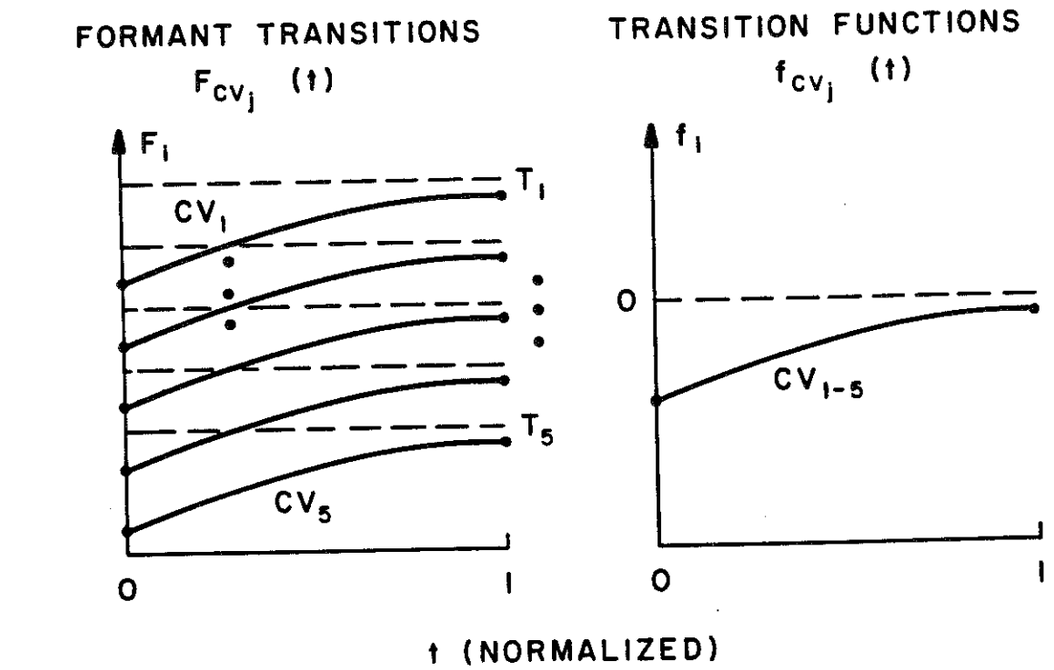

Figure 5A. (Right) Assumed global consonant transition function for consonant C. (Left) Resulting set of formant trajectories for consonant C going into five vowels, V1,...,V5.

3.2. Extended Superposition Model

3.2.1. Need for Vowel-Dependent Transition Functions. Equation 6 might raise the hope that a coarticulation model might be very simple indeed: that a given consonant might have the same additive effect on each vowel. Table III suggests that this will succeed for the vowel steady state. There is, however, a simple reason that consonant transition functions as used in Equation 4 will have to be different for each vowel. Figure 5A shows how formant trajectories would look if a given consonant had a fixed transition function for all vowels. The trajectories on the left correspond to the single transition function shown on the right. The trajectories would all be parallel to each other, and their endpoints at the consonant boundaries would be displaced a constant distance from their respective vowel targets. This is not at all the picture suggested by experience. Instead, we expect formant trajectories and their associated transition functions to look something like the schematic shown in Figure 5B.

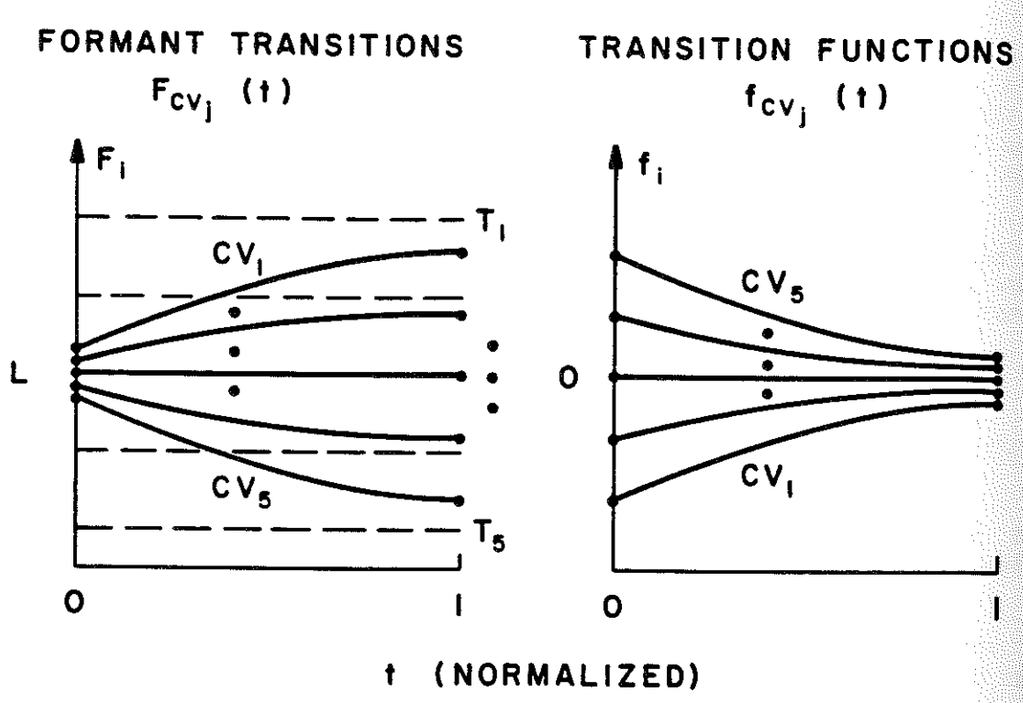

Figure B. (Left) Schematic of more realistic forms for formant trajectories for consonant C going into different vowels V1,...,V5: the onsets are drawn toward a locus L. (Right) The corresponding set of transition functions: a different function is needed for C going into each Vi.

Figure 5B embodies the notion that the formant trajectories for a given initial consonant will all be pulled toward some locus associated with that consonant. This is quite different from the situation shown in Figure 5A and shows why each combination of vowel and consonant must have its own transition function. Consonant C is associated with locus L and the jth vowel Vj is associated with target T(Vj), though L and T(Vj) are not necessarily exactly realized. The total gesture trajectory is denoted by F(C,Vj;t). The transition functions, in the sense of Equation 4, are obtained from F(C,Vj;t) via the vertical translations:

$$f(C,Vj;t) = F(C,Vj;t) - T(Vj) \qquad (7)$$

3.2.2. Model. The conceptualization shown in Figure 5B suggests a linear model that is a slight generalization of Equation 4:

$$F(C1,V,C2,i;t) = T(V,i) + f(C1,V,i;t) + g(V,C2,i;t) \qquad (8)$$

where F(C1,V,C2,i;t) is the ith formant-frequency trajectory for C1VC2, T(V,i) is a target frequency for vowel V, f(C1,V,i;t) is an initial-consonant transition function, but now indexed by the vowel V as well as by the consonant C1, and g(V,C2,i;t) is the corresponding final-consonant transition function. Time t can either be normalized to run from 0 to 1 or unnormalized to run from 0 to the vowel duration τ. The latter case involves a refinement discussed below in Section 5.4.

The model could become unwieldy if all the transition functions, one for each CV and VC combination, turned out to be unrelated to one another. We expect that this will not be the case. Indeed, the idealization in Figure 5B suggests a hopeful set of hypotheses: that the transition functions form simple families of curves, all of the same shape and hence characterized by their scales and positions. The scales and positions in turn would behave predictably from their associated consonants and vowels. This set of acoustic hypotheses is an analogy to the locus theory developed at Haskins Laboratories in the 1950's (Delattre, Liberman and Cooper, 1955), hence Figure 5 resembles those used to illustrate that theory. The hypotheses naturally break down into shape hypotheses and scaling hypotheses, which are taken up in more detail in Sections 4 and 5. First, we will examine some implications of the linearity of the model.

3.3. Additive Versus Multiplicative Data Bases

3.3.1. Avoiding Combinatorial Explosion. There is a fortunate spinoff from the linearity of the way that the initial and final contexts interact: this means that a data base for studying coarticulation can be built up in an "additive" rather than a "multiplicative" mode. Before we had any evidence that superposition could work so well, it was conceivable that it would be necessary to study all possible combinations of initial and final contexts. Thus for the 24 initial and final contexts measured by Fertig (1976), it was necessary to look at all 576 (=24x24) combinations. Now that those measurements have shown the relative adequacy of a linear model, i.e., that CVC's can be built up by adding CV's and VC's, it should be possible to derive the same sort of model from a suitable collection of VC and CV sequences. Then, for example, the 576 contexts could be built up from a script of 24 CV and 24 VC sequences: only 48 items. This represents a reduction in script size by an order of magnitude and indicates that even comprehensive coarticulation studies need not suffer combinatorial explosion.

3.3.2. Implications for Time Scale. A normalized time scale in Equation 4 was dictated by the necessity for a uniform way to analyze CVC utterances of varying durations. The normalization accomplished this by aligning the V's to all start and end at the same normalized times (t=0 and t=1). As noted by Broad and Fertig, there are reasons to prefer an absolute or unnormalized time scale.

Figure 6 shows the problem with a normalized time scale. It shows two different ways to "warp" the formant trajecory on the left to a different duration, with the results shown on the right. The one labeled "wrong way" is the original trajectory replotted on a linearly squeezed time scale.

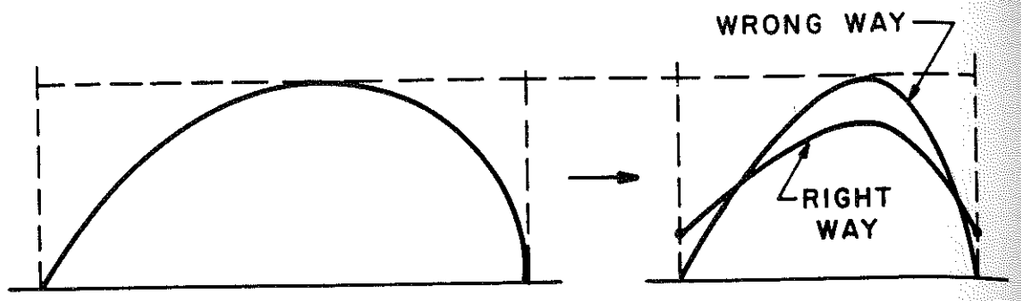

Figure 6. Two ways of normalizing a formant trajectory to a different duration. "Wrong Way": the trajectory is normalized by a linear scaling of the time axis without altering the frequencies. "Right Way": the trajectory is normalized by both a linear scaling of the time axis and by adjustments to the vowel-target undershoot and to the consonant-locus perturbations according to duration dependencies observed in speech production.

Nothing has been done to the frequency scale. Yet from Lindblom's study, we know that vowel target undershoot is a function of duration. Also, as will be seen later, the endpoints of the trajectory might be expected to be perturbed differently for different durations. The true situation is therefore better represented by the scheme labeled "right way". Here, in addition to the linear compression of the time scale, the trajectory has been warped in frequency to adjust the target undershoot at the center and the locus perturbations at the endpoints. This could be realized by using transition functions in Equation 8 that are truncated rather than time-normalized to a needed duration. Truncated transition functions are discussed in Section 5.4.

The difficulty just described shows that time normalization by itself does not realistically characterize the acoustic configurations of vowels uttered at different rates. This poses a theoretical and methodological problem for using data bases of CVC syllables for studying coarticulation. It also stands as an objection to the usual methods of time warping or dynamic programming used in speech recognition.

As it is usually applied in speech recognition, dynamic programming attempts to prepare an utterance for comparison with a template by a linear or non-linear warping of the time axis to synchronize events in the input with comparable events in a template. As has just been seen, however, warping only the time axis without taking into account the way that formants are changed by different timings leads to an incorrect picture of the formant trajectory. Dynamic programming will therefore introduce systematic error into vowel patterns. It is sometimes thought that template matching to whole utterances is a way to avoid problems connected with fine phonetic detail. The force of the present argument, however, is that dynamic programming will not avoid the problems posed by coarticulation. Some way of handling coarticulation will ultimately be needed if we are ever to do realistic time warping, that is, if we are ever to do it the "right way" shown conceptually in Figure 6.

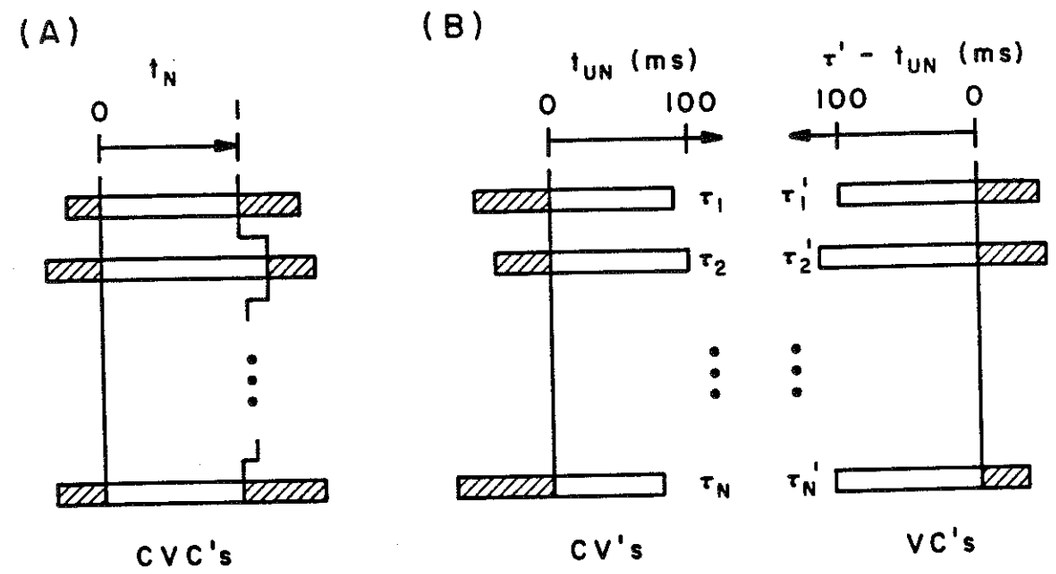

Figure 7. Schematic time alignments of vowel data. (A) CVC syllables aligned by both onsets and terminations. All items are thereby normalized to a common duration. (B) CV and VC syllables aligned by only one segmentation point. No time normalization of duration is involved.

This problem with time warping also poses a methodological difficulty for using CVC data bases to study coarticulation. To treat a body of data statistically, one must decide what items are to be grouped together. This is a decision about which measurements are to be considered comparable. In using a normalized time scale, one groups together measurements made a certain fraction of the vowel duration from the vowel onset. In an unnormalized time scale one groups together measurements made so many milliseconds from some event, such as a segmentation point.

In a body of CVC's consisting of all combinations of initial and final consonants, it makes conceptual sense to ask about the average behavior of, say, F2 at 80 percent of the way from the onset to termination of the vowel. It seems to make less sense to ask about the average behavior of F2 at 100 ms after the vowel onset: Owing to duration differences, this could be near the vowel midpoint, its termination, or even into the final consonant. Not only would such an F2 average be more variable than one taken 80 percent of the way through the vowel, but it would be harder to interpret conceptually. Therefore time warping to a scale of percent or fraction of vowel duration seems to be the only way to make statistical and conceptual sense of a C1VC2 data base. This is so in spite of the theoretical difficulties just outlined with normalized time scales.

A C1VC2 data base would evidently then have to be prepared for statistical treatment by simultaneously aligning both the C1-V and V-C2 boundaries as shown conceptually in the left column of Figure 7.

A way out of this difficulty would be to structure the data base in such a way that comparisons of events measured at the same unnormalized times would make conceptual sense. This might be done by taking advantage of the expected linearity of the effects of C1 and C2 in C1VC2 sequences. The linearity allows us to study the C1 and C2 effects separately through a data base consisting of CV's and VC's. This has already been discussed for its advantage in reducing the scale of a coarticulation data base. As shown in the middle column of Figure 7, the CV's can be aligned by their C-V boundaries. Their other ends, the V terminations, are then left free to fall at their respective vowel durations. Now repetitions of a CV or instances of different C's going into the same V can be compared or treated statistically without normalizing the time sacle, because now, without the complication of a variable final context, it does make conceptual sense to ask what happens, on the average, so many milliseconds after the C-V boundary.

In a similar fashion, the VC's can be aligned by their V-C boundaries, as shown in the right column of Figure 7. Now nothing is done to align the vowel onsets.

To generate a model of the form of Equation 8, the f's are built up from averages of the respective CV's and the g's from the VC's. To model a CVC of a given duration, the f's and g's are time-truncated to the desired duration before being added according to Equation 8.

The use of time-truncated transition functions involves an interesting hypothesis about the behavior of formant patterns. This is discussed later in Section 5.4. For the present, the point is that an "additive" data base of CV's and VC's is not only economical in scale but seems to provide a natural way to analyze context effects in items of variable duration without having to warp the time scale.

4. CONTOUR SHAPES

4.1. Shape Similarity

By hypothesizing the transition functions to be of the "same shape" we have to select one from the several equivalent formulations of similarity. One very simple one is to say two curves are similar if one curve can be transformed to another via a scalar mapping. A convenient form is to scale all the transition functions to extreme values of 0 and 1. Letting f(t) and g(t) be single curves of the families, the respective canonical forms f*(t) and g*(t) are given by:

(9)

$$f^*(t) = [f(t) - fmin]/[fmax - fmin]$$

$$g^*(t) = [g(t) - gmin]/[gmax - gmin]$$

where time t is normalized to run from 0 to 1, and fmax, fmin, gmax and gmin are the respective maxima and minima of f and g. In this form, fmin and gmin are mapped onto 0 and fmax and gmax onto 1.

4.2. Universal Similarity

4.2.1. Hypothesis. The simplest hypothesis would be that f* for all the initial-consonant transition functions f(C,V,i;t) will be the same to within some error that is close to the inter-repetition variability. Similarly, it would be hypothesized that the final-consonant transition functions g(V,C2,i;t) will map onto nearly the same g*. This type of normalization by extrema has been successfully used to map fundamental-frequency contours by different speakers onto each other (Earle, 1975).

4.2.2. Implications. If the universal same-shape hypothesis were confirmed, then the general coarticulation model for CVC syllables would be considerably simplified. To see this, let Equation 8 be rewritten using the extremum parameters and the hypothetically-known functions f* and g*:

(10)

$$F(t) = T + f^*(t)[fmax-fmin] + fmin + g^*(t)[gmax-gmin] + gmin$$

An automatic speech recognizer might have F(t) as a formant trajectory measured from incoming data. If f* and g* were known functions, then Equation 10 would be linear in the "unknown" targets and extrema: T, fmin, fmax, gmin, and gmax. These parameters could therefore be estimated from 5 time samples of the trajectory. The four extremum parameters could be used to reconstruct the initial- and final-consonant transition functions. These contain all the information that the formant trajectory has to offer about the consonantal contexts.

Equation 10 can also be used for a least-squares approximation if a larger number of time samples of F(t) is available. This would offset systematic and random errors due to measurement noise, inter-repetition variation, and residual deterministic error in the the model. The latter component might include departures from the same-shape hypothesis.

The least-squares analysis results in another system of 5 simultaneous linear equations in the unknown parameters. The deterministic and least-squares formulations lead to similar computational forms in which the 5x5 matrix needs to be inverted only once for all time, as in each case the matrix elements are all constants depending on f* and g*. Therefore the "per-vowel" computations would involve only matrix multipications and no matrix inversions.

The vowel target and consonant transition functions could therefore be estimated simultaneously, thus bypassing a familiar quandary: If we need the consonants to recognize the vowels, and vice versa, which do we do first?

The technique just outlined is not symmetric with respect to the vowel and the consonants. The estimation procedure gives an explicit result for the vowel target T, and each vowel is expected, for a given speaker, to have a unique value of T associated with it. We are not so lucky with the consonants, because for them we obtain only estimates of their transition functions. As discussed above, each consonant will have not a unique transition function, but a collection of them, one for each vowel. Therefore the estimation technique is still a step away from providing a unique consonant target or locus. It still provides information about the consonants, but not in quite as neat a form as it does for the vowel.

It will now be shown that this technique will not work in practice because the premise of universal shape similarity is not true. There is, however, a constructive lesson to be learned from the present example: It is well within the scope of imagination that a coarticulation model can be nicely enough structured to allow phonetically significant non-obervables, such as targets, to be quickly estimated from samples of observables, such as realized formant frequencies. It is the phonetically significant non-observables that would be most useful for automatic phonetic recognition.

4.2.3. Counter-Example. Unfortunately for the hopes just outlined, the data from Broad and Fertig show that transition functions for various initial consonants going into the vowel /I/ have different shapes. To see this, note that if transitions were of the same shape, then the re-write of Equation 9 with normalized time t=0.5:

$$f(C1;0.5) - f(C1;1) = K^*(0.5)[f(C1;0) - f(C1;1)] \quad (11)$$

where

$$K^*(t) = [f^*(0)-f^*(1)]/[f^*(t)-f^*(1)]$$

Equation 11 is a straight line through the origin with slope $K^*(0.5)$.

Figure 8 shows the second-formant f(0.5)-f(1) plotted versus f(0)-f(1) for all initial consonants C1. Clearly the data do not fall on a single straight line. The data do, however, fall into four interesting classes: (A) /w,r,ɱ/; (B) /h,p,t,k,j,ʃ,ʒ/; (C) /#,b,d,g,m,n,f,v,θ,ð,s,z,l/; and (D) /ŋ/. These will now be discussed in order.

(A) /w,r,ɱ/.

These consonants involve all the large total F2 transitions (|fmax-fmin| > 900 Hz). They also involve high initial slopes (|dF2/dt| > 250 Hz per 0.1 vowel duration). They do not, however, involve unusually large contributions to the vowel-target undershoot, essentially f(.5)-f(1). It may be that the undershoot at the vowel center "saturates" for large values of fmax-fmin.

(B) /h,p,t,k,j,ʃ,ʒ/

This grouping of consonants in Figure 8 is fairly well fit by the line:

$$f(0.5)-f(1) = 0.42[f(0)-f(1)] - 77 \quad (12a)$$

Except for the 77 Hz vertical displacement, this would be consistent with Equation 11, with $K^*(0.5) = 0.42$.

The production of /h/ and the aspiration of /p,t,k/ involve an interval of turbulence noise generated at the glottis, which leaves the supra-glottal articulators free to anticipate the vowel. The sonorant /j/ has a target near to that of /I/, again suggesting some sort of substantial anticipation of /I/ during the consonant. If this is a real grouping of consonants, it is not too clear why the homorganic sibilants /ʃ,ʒ/ should be included, unless their production can be accomplished with some extra freedom of the tongue shape that would again allow the /I/ to be more freely anticipated during the consonant.

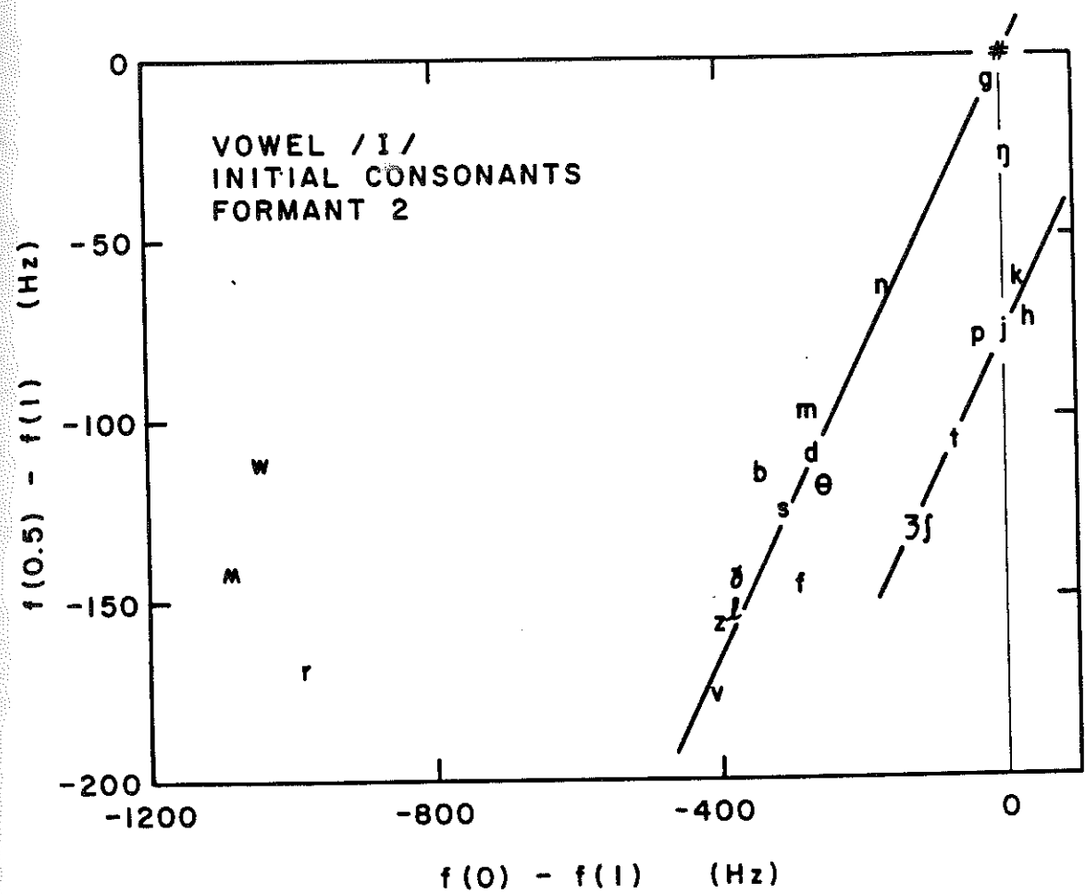

Figure 8. Second-formant difference f(0.5)-f(1) between the vowel center and its termination plotted against the second-formant difference f(0)-f(1) between the vowel onset and its termination for the 24 initial-consonant transition functions from the data of Broad and Fertig (1970).

(C) /#,b,d,g,m,n,f,v,θ,ð,s,z,l/

This set of 12 consonants plus silence (/#/) is fairly well represented by the line:

$$f(0.5)-f(1) = 0.42[f(0)-f(1)] \quad (12b)$$

This is the form of Equation 11 with $K^*(0.5) = 0.42$. The same-shape hypothesis might therefore be supported within this subset of initial consonants. The universal same-shape hypothesis, however, apparently must be abandoned in light of the totality of these results.

(D) /ŋ/

This consonant is by itself between groups (B) and (C). Perhaps it behaves differently because in English, the speaker's native language, /ŋ/ never occurs in syllable-initial position.

Groups (B) and (C) seem to be fit by lines of the same slope and these groups, which together contain 20 of the 24 contexts studied, seem to differ in some sensible phonetic dimension, a dimension perhaps of consonantal anticipatory freedom.

That there is some regularity to the relationship among f(0), f(0.5), and f(1) for the various consonants suggests that even with the universal shape-similarity hypothesis rejected, there is perhaps some hope for an economical representation of the f's and g's for the various consonant and vowel combinations, perhaps by a small collection of shapes rather than by one universal shape. Alternately, it might be found that a single shape will serve if it is allowed to be truncated or cut back in duration in a consonant-dependent way.

4.2.4. Another Counter-Example. The results from Lindblom's study, summarized above by Equation 1, also contradict universal shape similarity in that the inverse time-constant b varies from context to context. Within a context, however, the perturbation term of Equation 1 does describe a family of similarly shaped curves, differing only in scale.

4.3. Per-Consonant Similarity

These counter-examples do not yet contradict the hypothetical notions illustrated in Figure 5B, which suggest that the transition functions for a given consonant into or out of various vowels might follow trajectories of similar shape. Figure 8 shows that the converse is not true: that for a given vowel, the transition functions for the different consonants will differ in shape. It will therefore be interesting to test a more specialized hypothesis: that for a given consonant C the f and g functions will be of the same shape for all vowels V, i.e., that the f* and g* functions will be nearly the same in a "per-consonant" sense. As just noted, Lindblom's results are consistent with the notion of per-consonant similarity.

Similarly, Equation 2 for Öhman's articulatory model of context effects also suggests a per-consonant similarity of gestures in that it is formally similar to Lindblom's.

4.4. Closed-Form Representations

It would also simplify the coarticulation model if the f's and g's could be represented in closed forms. Parabolic (Stevens, House, and Paul, 1963) and exponential (implicit in Lindblom, 1963) forms have been used in the past. Any constants that characterize such forms then become parameters of interest. For example, Stevens, House, and Paul studied the systematic behavior of the "curvature parameter" of the parabolas they fit to the formant trajectories.

The first requirement for a closed-form representation of formant trajectories or transition functions is that it fit the data. Beyond this, one would hope for other desirable properties, such as the unique recoverability from measured formant contours of the separate terms

corresponding to the vowel target and to the initial and final transition functions.

4.5 Summary of Contour Shape Hypotheses

Starting from a definition and hypothesis of shape similarity among formant transition functions, it was easy to show that a similarity constraint would suffice for representing formant contours in a form from which the separate contributions of the vowel and the initial and final context could be uniquely estimated from formant data. Unfortunately, a re-working of some old data showed that while there exists a substantial set of transition functions that are shaped similarly to each other, there is also a substantial set of them that are not. Nevertheless, there still seems to be some hope of finding enough regularity among transition shapes for the separate elements of a coarticulation model to be recovered from formant data. If this can be done, the goal of handling coarticulation effects in automatic phonetic recognition will have been significantly advanced.

If such regularites among transition functions cannot be found, then it would still be possible to recover estimates of vowel and consonant contributions to formant contours in CVC syllables, but perhaps not elegantly. If the transition functions were known, then incoming formant contours could be fit by an expensive search for the best fit of the model to the data via an analysis-by-synthesis in which all possible CVC contours are computed, compared to the unknown contour, and ranked in order of increasing rms differnce. While this would show that coarticulation could be handled in principle even if the most convenient hypotheses should fail, it would be best if this cumbersome possibility could be avoided.

5. CONTOUR SCALING

We expect the endpoints of the transition functions, f(0), f(1), g(0), and g(1), to be predictable from case to case: T+f(0) and T+g(1) are expected to be pulled toward some initial and final consonant "loci" or targets; T+f(1) and T+g(0) are expected to be close to the vowel target. Furthermore, we hope that the perturbations of these values away from their respective targets will be predictable from context. In fact, from several previous studies, the differences between endpoint values and their respective targets will be expected to be proportional to the scale of the overall gesture. That is, vowel undershoot and perturbation of trajectory endpoints from respective consonant loci are hypothesized to be proportional to the differences between vowel targets and consonant loci.

5.1. Locus Hypotheses

Letting the respective initial and final consonant loci be L(C1) and L'(C2), we hypothesize that:

H1A: T + f(C1,V;0) will be pulled toward a locus L(C1)

H1B: f(C1,V;1) will be near zero

H1C: g(V,C2;0) will be near zero

H1D: T + g(V,C2;1) will be pulled toward a locus L'(C2)

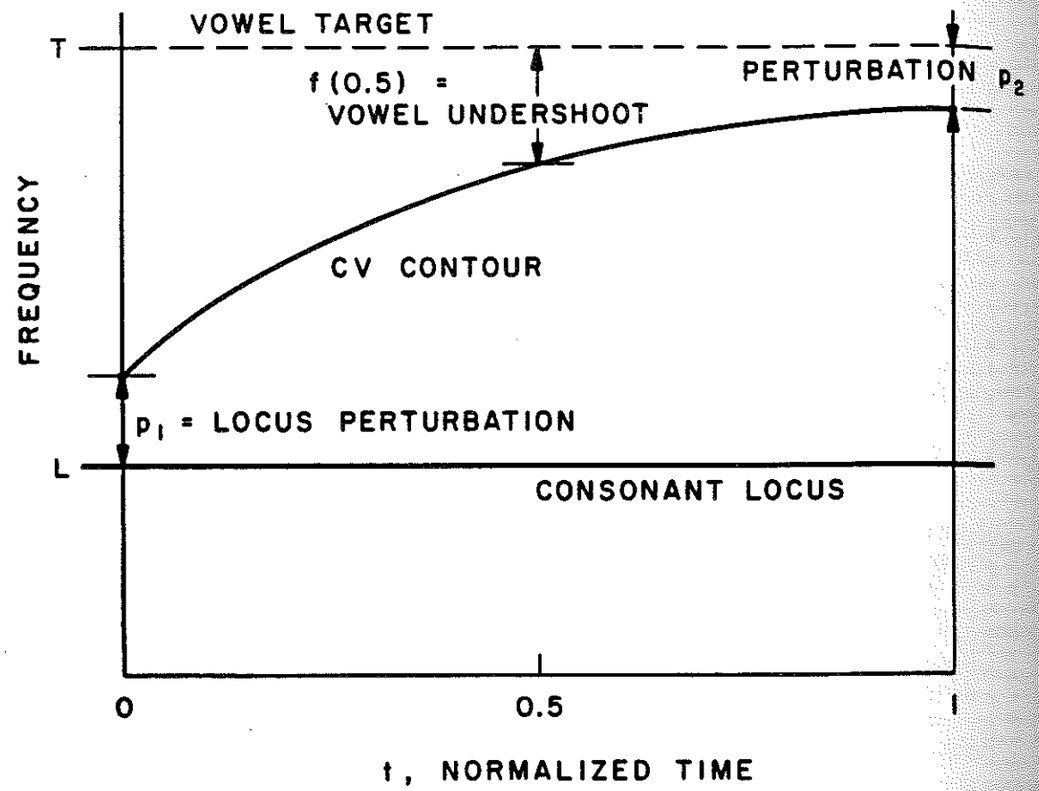

Figure 9. Perturbations p1 and p2 defined for an initial-consonant transition function. Initial consonant contribution to the vowel target undershoot is also shown.

The loci are only very loosely defined at this point. One operational approach to specifying them will be given in Section 5.2.

The perturbations of the endpoints of model CV and VC trajectories from their respective loci or targets are defined by:

(13)

$$p1(C1,V) = [T(V)+f(C1,V;0)] - L(C1)$$

$$p2(C1,V) = f(C1,V;1)$$

$$p3(V,C2) = g(V,C2;0)$$

$$p4(V,C2) = [T(V)+g(V,C2;1)] - L'(C2)$$

The first two p's, involving f, are illustrated in Figure 9. The p's involving g are exactly symmetric, with the roles of t = 0 and t = 1 reversed.

5.2. Effects Proportional to Locus-Target Distances

It is to be hoped that the p's can be expressed as simple functions of the vowels and consonants. Indeed, various studies suggest that vowel target undershoot is proportional to the consonant-vowel formant distance. Generalizing this idea to include the vowel endpoints results in a set of hypotheses about the linearity of the p's:

H2: There exist proportionality constants k1, k2, k3, and k4 such that

(14)

$$A: \quad p1(C1,V) = k1(T - L)$$

$$B: \quad p2(C1,V) = k2(T - L)$$

$$C: \quad p3(V,C2) = k3(T - L')$$

$$D: \quad p4(V,C2) = k4(T - L')$$

These proportionalities are similar to the factor k(F2i-F20) in Lindblom's equation (Equation 1), even though he presents no direct picture of how his data support this distance-proportional term. Equation 14 is a slight generalization in that target-locus distances are being hypothesized to apply at transition-function endpoints and not just at vowel centers. These proportionalities also differ slightly from Lindblom's in that they involve abstract and not directly observable targets and loci, while Lindblom's equation involves both a target F2t and a directly observable initial value F2i.

Lindblom's data and model can be re-worked to give direct support to coarticulation effects proportional to target-locus distances. As a preliminary step, Figure 10 shows a plot of Lindblom's tabulated F2t versus the difference between tabulated F2i and F2t. For the most part, the data fall on straight lines, one for each consonant. Context /g--g/ is separated into palatal and velar allophones [ɟ--ɟ] and [g--g]. The intercepts of these lines with F2i-F2t = 0 are plausible values of F2t to assign to the consonant loci L: when the onset and target frequencies are the same, it seems reasonable to say that the consonant has no perturbing effect on the vowel, and this would be expected when L = F2i = F2t. The loci thus derived are shown at their respective intercepts in the figure. The locus for [g] is undefined, as its 3 points are close to the vertical line F2i-F2t = -350 Hz. As an expedient in what follows, then, the value L(ɟ) will be operationally adopted for L(g).

These values of L, together with Lindblom's tables of F2t and F2i, give us the plot in Figure 11, which shows the locus perturbation F2i-L plotted against the locus-target distance F2t-L. For each consonant, a straight line fits the data for the 8 vowels fairly well. Considering the provisional definition of the [g] locus, not much significance should be attached to the straight line through its 3 data points. That each consonant has its own line supports the distance-proportional hypothesis in a per-consonant sense. Therefore each C might be expected to have a different ki in Equation 14. This plot supports the portrayal of vowel onsets schematized above in Figure 5B.

Similarly, the plot in Figure 12 is generated by using Equation 1 to calculate realized values of F2 at the vowel center for a fixed duration of

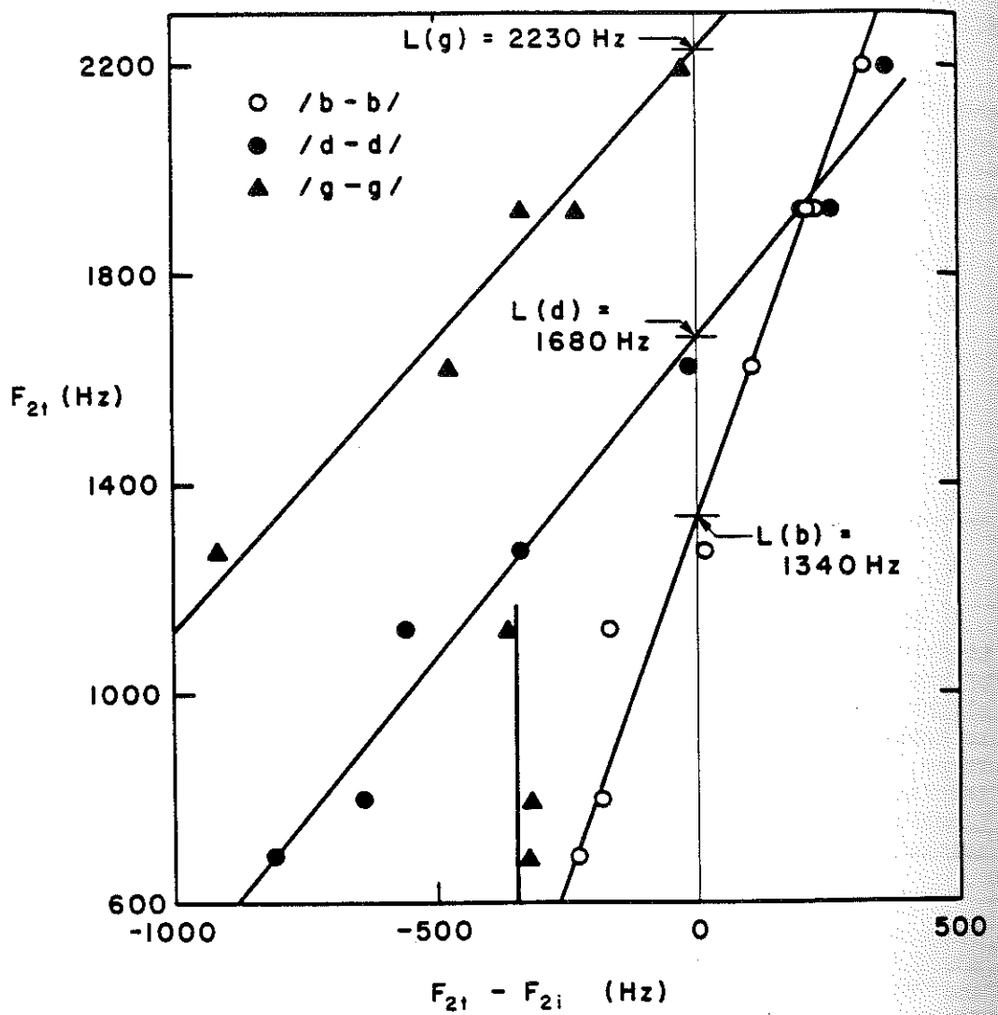

Figure 10. Second formant target frequency F2t plotted against the onset-to-target distance F2t-F2i for the 8 vowels and 3 contexts reported by Lindblom (1963).

150 ms. These values allow us to compute the vowel target undershoot. This is ploted against the locus-target distance in Figure 12. Again the data are mostly well represented by lines through the origin, one for each consonant. Therefore vowel undershoot seems to be proportional to locus-target distance, again in a per-consonant sense.

An exception to the proportionality is the insensitivity of the vowel target undershoot to T-L for /g/ in the context of back vowels. In this case, the perturbation seems to be "saturated" at -110 Hz, reminiscent of the behavior of /ʍ,w,r/ in Figure 8.

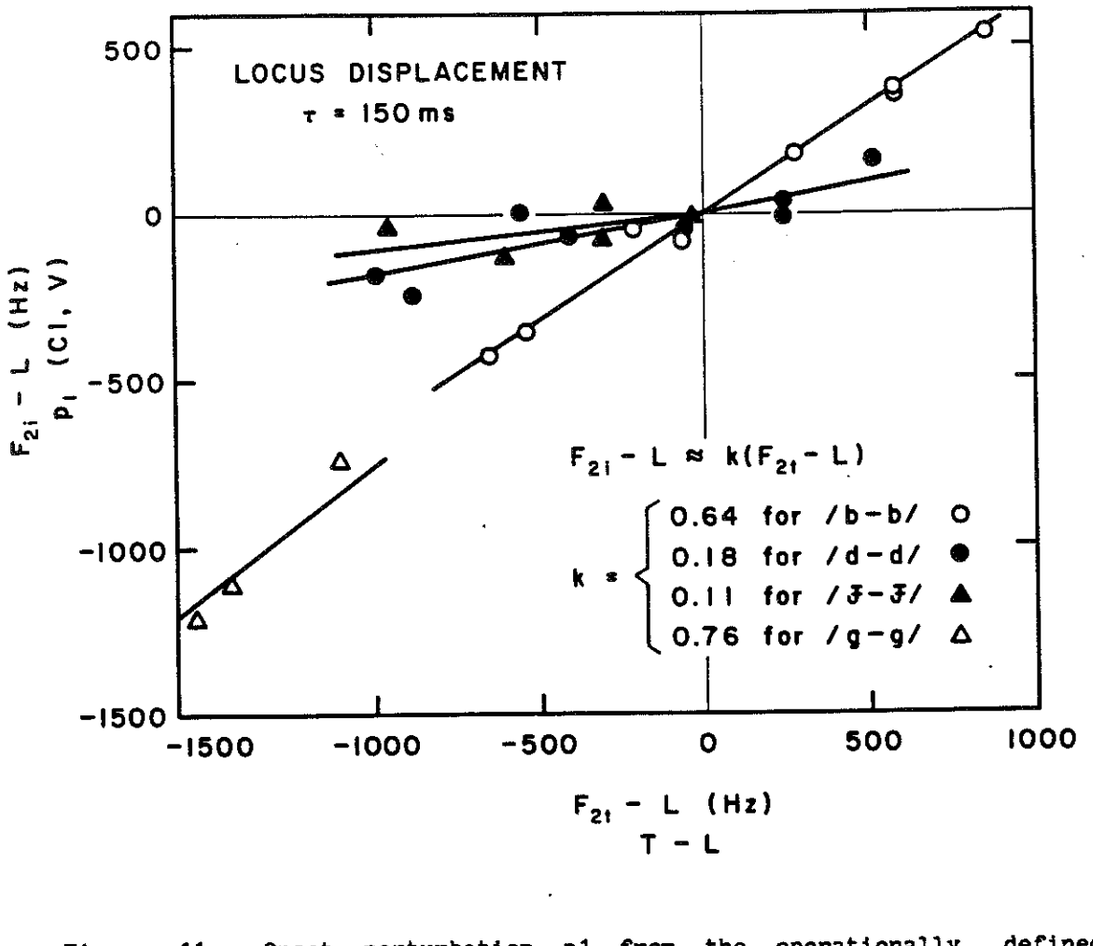

Figure 11. Onset perturbation p1 from the operationally defined consonant locus plotted against the target-to-locus distance for the 8 vowels and 3 contexts studied by Lindblom (1963).

These plots are the first I know of to directly demonstrate the proportionalities implied by Equations 1 and 2 for the models of Lindblom and Öhman.

A qualitative articulation of distance-proportional effects comes from Stevens and House (1963), who observed, "It is evident that the extent to which an ideal articulatory configuration is achieved for a vowel depends upon the effective 'distance' that the articulators must traverse during various phases of the syllable..." (p. 120). Later, Stevens, House, and Paul (1966) say: "...the displacement [from the loci] is in general greater when the distance between the consonant locus and vowel target is greater." (p. 128). Also, "The amount of displacement of F2m [the realized F2 steady state] is, in general, greater when [the vowel-initial and vowel-final values of F2] are farther from the target frequency." (p. 129).

As just seen, Lindblom's data lend quantitative support to these notions.

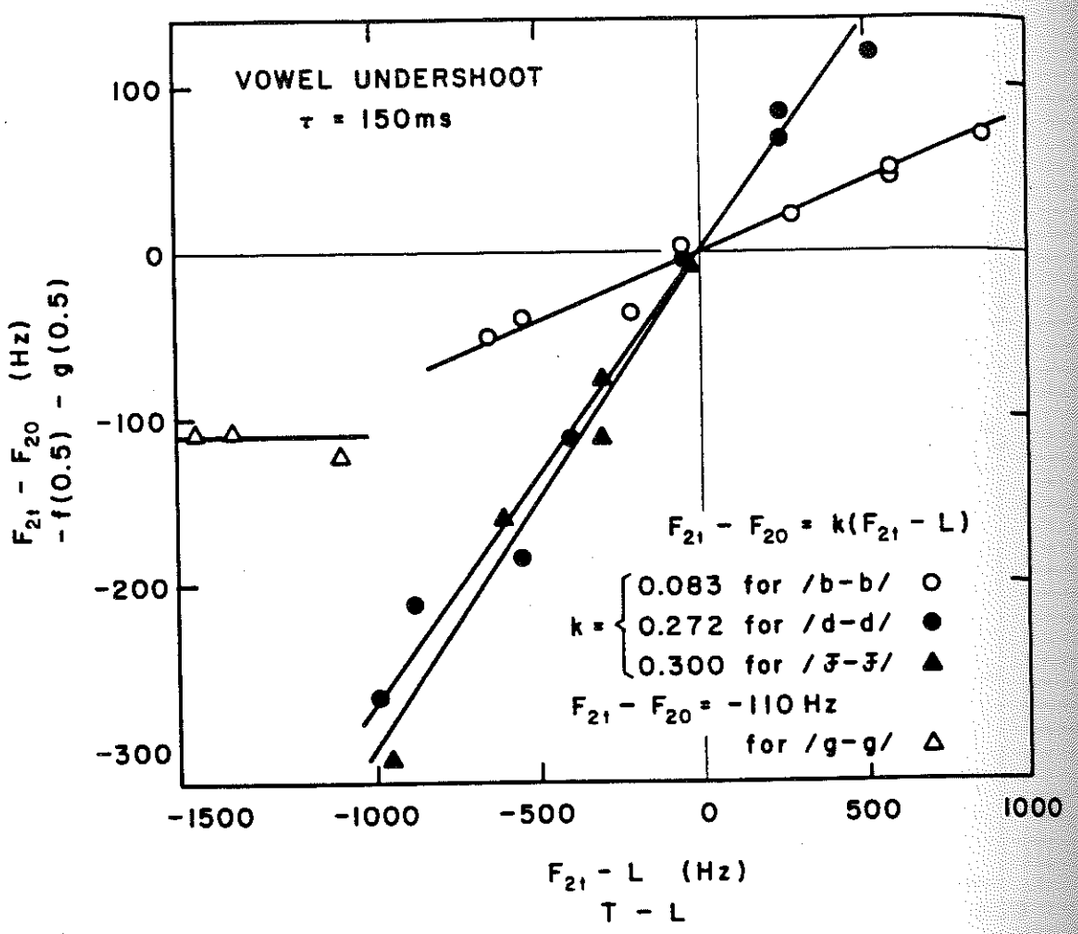

Figure 12. Vowel undershoot F2t-F20 plotted against the target-to-locus distance for the 8 vowels and 3 contexts studied by Lindblom (1963).

5.3. Implications for Modeling and Recognition

Clearly, distance-proportional effects simplify the coarticulation model. This was already seen in Equations 1 and 2 for the effects of a single context. For independent initial and final contexts, the transition functions can be cast in the similar forms:

(15)

$$f(C1,V,i;t) = k5(C1,i)\,[T(V,i) - L(C1,i)]\;\hat{f}(C1,i;t)$$

$$g(V,C2,i;t) = k6(C2,i)\,[T(V,i) - L'(C2,i)]\;\hat{g}(C2,i;t)$$

where $\hat{f}$ and $\hat{g}$ are consonant-dependent transition function shapes, k5 and k6 consonant-dependent scale factors, and T(V), L(C1) and L′(C2) the vowel and

consonant target and loci. The transition function shapes $\hat{f}$ and $\hat{g}$ in this instance are not given by f* and g* from Equation 9, but can be even simpler: by appropriate scaling of k5 and k6 they can be set equal to any representative f and g associated with C1 and C2. For example, f could be set equal to any one of the non-zero f′s shown in Figure 5B and, by appropriate scaling, could generate all the other f′s in the figure.

If f and g could be represented by Equation 15 it would still be true that each CV or VC combination would have to have its own associated transition function. But these functions would now be built up from simple elements, each dependent only on the individual vowels and consonants, and not on their combinations.

Equation 15 involves not only the factors proportional to target-locus distances, but also factors representing shapes of transition functions which are similar on a per-consonant basis. This equation substituted into the linear coarticulation model can be used to derive target undershoot or locus displacement as proportional to target-locus distances. Therefore the empirical proportionalities shown by the data in Figures 11 and 12 also lend some support to the notion of per-consonant shape similarity among transition functions.

Beyond their roles in simplifying the coarticulation model and in supporting per-consonant shape similarity of transition functions, it is not too clear what implications distance-proportional effects might have for recognition. The situation is complicated by the consonant-to-consonant variability of the constants of proportionality. One would expect that a conceptually simple form such as Equation 15 would aid in "solving" formant contours for the elements attributable to the vowel and consonants, perhaps along the lines tried in Section 4.2.2. At the moment, however, it is not clear how this could be done without going to a costly search for a best fit between a data contour and the full set of model contours.

5.4. Duration Dependence

Something must now be said about how duration effects might be handled in the linear model. How this is done will obviously depend on how formants are actually observed to act. The intuitive notion of explaining coarticulation in terms of the time available for completing a gesture suggests a natural hypothesis: Each CV or VC combination follows the same path, independent of duration, but completes a smaller or larger subset of that path depending on the available time. What this means for the linear model is that for a given duration τ, the transition functions f and g corresponding to C1, V, and C2 would be time-truncated to duration τ before being added together to form the formant contour. More precisely, the hypothesis might be stated:

H3A: For any given C1 and V, there exists a unique f(C1,V,i;t) such that:

(16)

$$F(C1,V,i;t) = T(V,i) + f|[0,\tau](C1,V,i;t)$$

where $f|[0,\tau]$ is the restriction of f to the time interval $[0,\tau]$, i.e., it is f truncated at $t=\tau$.

Thus the initial-consonant transition function would always start with the same value f(0), but end at the duration-dependent value $f(\tau)$. Will this be found to be true, or will f(0) itself be displaced by duration, just as the formant value at the vowel onset is displaced by target-locus distance? Observations made under duration changes will be needed to answer these questions and to characterize the duration dependence of initial-consonant transition functions.

Similar questions apply to the final-consonant transition functions, the g's. The hypothesis that is symmetric to the hypothesis of time-truncated f's would be that g's for a given VC all terminate on the same value, but begin at some duration-dependent position on the general g curve. This becomes easier to express if the time argument is taken to be τ-t instead of t. Then g(0) would represent the value of g at the vowel termination and $g(\tau)$ its value at the vowel onset. Using this convention for the time argument of g, the hypothesis is:

H3B: For any given V and C2, there exists a unique $g(V,C2,i;\tau-t)$ such that:

$$F(V,C2,i;t) = T(V,i) + g|[0,\tau](V,C2,i;\tau-t) \qquad (17)$$

where $g|[0,\tau]$ is g truncated at τ-t=τ.

If some form of these hypotheses is borne out, the incorporation of duration dependency into a coarticulation model will be fairly simple. This should also make it easier to apply such a model in automatic recognition. Equation 1 for Lindblom's duration-dependent model is consistent with the idea that the f's and g's might be represented by time-truncated exponential functions.

6. CONCLUSION

Review and re-interpretation of past work has shown a number of ideas that should be useful for modeling and recognizing vowels in context. The gestures into and out of a vowel appear to combine by superposition, promising to simplify a coarticulation model. Such linearity also opens the door to using smaller data bases and to using real time scales instead of scales normalized by durations.

The model will be further simplified if acoustic phonetic gestures can be characterized by a few simple shapes and if their scales turn out to be proportional to locus-target distances.

Evidence for these simplifications is promising, but extensive new data will be needed, first to test the validity of these notions under a wider variety of conditions, and second to obtain the actual numerical values of the model.

Present analysis tools coupled with the reductions in data base size promised by linearity should make this a reasonably scaled task. In terms of the proposed linear model, initial-consonant transition functions are essentially built up from the average behavior of CV utterances and final-consonant transition functions from that of VC utterances. ("Essentially" refers to other considerations that favor using CVK utterances instead of CV's, where K denotes some fixed final context whose corresponding g has been found from VK utterances.)

If the resulting f's and g's show sufficient structure among themselves, it may be possible to develop an analytic form of a coarticulation model from which vowel and consonant parameters can be estimated directly from incoming formant contours. Otherwise, vowels and their contexts might have to be analyzed by a cumbersome search-and-fit between the model and incoming data. A third possibility would be to use a purely empirical method to recognize vowel and context according to where they fall in a space defined by n time samples of the formant. Figure 1 represents such a method using 2 time samples. To do any of these requires a look at more data.

The ideas explored in this paper have so far been developed and tested on the basis of a limited set of observations involving only a thin sampling of languages, speakers, durations, and vowel-context combinations. It is therefore too early to tell how coarticulation models will finally look and how they will be used to enhance recognition. What does seem sufficiently proven is that context is a significant source of variability in vowel parameters. This means it is necessary at some stage to take on the problem of context. Context is also a largely predictable source of variability. This means that its pursuit should also be rewarding.

REFERENCES

Broad, D. J. (1972). "Analysis of Variance of Some Formant Transition Data of S. E. G. Öhman ["Coarticulation in VCV Utterances: Spectrographic Measurements," J. Acoust. Soc. Amer. 39, 151-168 (1966)], Research on Physiological Parameters in Phonetic Theory, Terminal Progress Report, USPHS Grant No. 5 R01 NS08036, J. E. Shoup, Principal Investigator, March 31, pp.61-75.

Broad, D. J. (1976). "Toward Defining Acoustic Phonetic Equivalence for Vowels," Phonetica, Volume 33, Number 6, pp. 401-424.

Broad, D. J., and R. H. Fertig (1970). "Formant-Frequency Trajectories in Selected CVC Utterances," The Journal of the Acoustical Society of America, Volume 47, Number 6, Part 2, June, pp. 1572-1582.

Delattre, P.C., A. M. Liberman, and F. S. Cooper (1955). "Acoustic Loci and Transitional Cues for Consonants," The Journal of the Acoustical Society of America, Volume 27, Number 4, July, pp. 769-773.

Earle, M. A. (1975). "An Acoustic Phonetic Study of Northern Vietnamese Tones," SCRL Monograph, No. 11 (Santa Barbara: Speech Communications Research Laboratory).

Fertig, R. H. (1976). "Temporal Interrelations in Selected English CVC Utterances," SCRL Monograph, No. 12 (Santa Barbara: Speech Communications Research Laboratory).

Houde, R. A. (1968). "A Study of Tongue Body Motion During Selected Speech Sounds," SCRL Monograph, No. 2 (Santa Barbara: Speech Communications Research Laboratory).

Lindblom, B. (1963). "Spectrographic Study of Vowel Reduction," The Journal of the Acoustical Society of America, Volume 35, Number 11, November, pp. 1773-1781.

Öhman, S. E. G. (1966). "Coarticulation in VCV Utterances: Spectrographic Measurements," The Journal of the Acoustical Society of America, Volume 39, Number 1, January, pp. 151-168.

Öhman, S. E. G. (1967). "Numerical Model of Coarticulation," The Journal of the Acoustical Society of America, Volume 41, Number 2, February, pp. 310-320.

Pols, L. C. W. (1977). Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words, Soesterberg: Institute for Perception TNO.

Purcell, E. T. (1979). "Formant Frequency Patterns in Russian VCV Utterances," The Journal of the Acoustical Society of America, Volume 66, Number 6, December, pp. 1691-1702.

Stevens, K. N., and A. S. House (1963). "Perturbation of Vowel Articulations By Consonantal Context: An Acoustical Study," Journal of Speech and Hearing Research, Volume 6, Number 2, June, pp. 111-128.

Stevens, K. N., A. S. House, and A. P. Paul (1966). "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," The Journal of the Acoustical Society of America, Volume 40, Number 1, July, pp. 123-132.

RULES AND STRATEGIES FOR SYLLABIC SEGMENTATION, PHONEME IDENTIFICATION AND TUNING IN CONTINUOUS SPEECH RECOGNITION

Guy Mercier
TSS/RCP
Centre National d'Etudes des Telecommunications
route de Tregastel, 22301 LANNION CEDEX (France)

ABSTRACT

This paper describes how to transform the continuous speech signal into a series of discrete units. The basic linquistic units frequently taken into account by the phonetic analyzers are the syllable and the phoneme. Unfortunately, various and irregular phenomena of speech such as coarticulation, speech rate, noise, stress, and dialectal variations affect the accustic signal, making a straight-forward link between these basic units and their acoustic or articulatory representation impossible. As a result, a larger set of intermediate units closer to the acoustic signal and easier to locate and to indentify has to be selected. The most frequently used units are the diphone, the cluster, the phone, the acoustic segment, etc. These units are themselves characterized by properties or attributes such as phonetic features which can be detected on the speech signal by means of acoustic cues.

The basic parameters and acoustic cues useful for interpreting the different sounds of a given language are presented. A way to get the allophones and their spectral characteristics from a set of phonetically balanced sentences is put forward. Special care will be paid to phonetic knowledge and to the rules which enable the segmentation of the signal into syllables and phonetic segments and the progressive recognition of the most important phonetic classes from the identification of the articulatory features : opening, nasalization, occlusion, voicing, frication, silence, burst. Context-dependent rules allowing the refinement of the segmentation and the classification achieved before are explained. The last two important points to be dealt with in this paper concern first the interactions between these different steps, namely the general organization of the acoustic-phonetic decoder and its integration into a continuous speech recognition system and secondly the problems of automatic speaker adaptation and the search for speaker independent cues.

Examples and results achieved either through the acoustic-phonetic decoder of the KEAL speech recognition system or through other systems illustrate the steps mentioned above. Thus, continuous speech segmentation into syllables with an error rate of about 5% can be achieved for a great number of speakers. As regards the segmentation into phonemes, an omission rate of about 5% and an insertion rate of 10% are usually obtained, in normal conditions. The percentage of correct identification for broad phonetic transcription ranges from 80 to 95% for a limited number of speakers. Excellent recognition percentages of the place of articulation of the plosive and nasal consonants have also been published. However, in normal conditions, the final percentage of correct classification for finer phonetic transcription can decrease below 60%. One of the reasons is that in a hierarchical process, errors of segmentation and identification are cumulative. Moreover, these results are too much affected by different factors such as recording conditions, noise, speakers, etc. Ways to overcome part of these difficulties and to make the cues and rules more reliable are suggested.