Reprint from

INFORMATION PROCESSING
IN THE NERVOUS SYSTEM

Edited by K. N. Leibovic

NOT TO BE SOLD

Springer-Verlag
New York • Heidelberg • Berlin
1969

WAYNE A. WICKELGREN
Massachusetts Institute of Technology*
Cambridge, Massachusetts

5 · *Context-Sensitive Coding in Speech Recognition, Articulation and Development*†

This paper describes a theory of the coding of speech at the phonetic level and applies that theory to the problems of speech recognition, articulation, and development. The theory specifies the elementary structural components of words and has nothing to say about the higher-level syntactic and semantic aspects of language. The theory is not contradicted by any of the facts that I know, and it provides a very simple explanation of many facts concerning speech recognition and articulation. However, the theory is extremely speculative, and no one should be misled concerning the amount of direct evidence for the theory. My enthusiasm for the theory is based largely on the clarity and simplicity with which it handles many of the basic problems of speech recognition and articulation, rather than on the definitiveness of its empirical support.

I. Context-Free and Context-Sensitive Coding of Ordered Sets

I define a context-free code for words to consist of an ordered set of symbols for every word, where some symbols in some words give insufficient information concerning the adjacent symbols to determine them uniquely out of the unordered set for the word. That is to say, the same

* Present address: The University of Oregon, Eugene, Oregon.
† This work was supported by grant, MH 08890–05, from the National Institute of Mental Health, U. S. Public Health Service.

symbol can be used in a variety of contexts of left and right adjacent symbols, and the ordering of the symbols in a word carries information not found in the conjunction of the unordered set of symbols with the sequential dependency rules.

I define a context-sensitive code for words to consist of an unordered set of symbols for every word, where each symbol restricts the choice of its left and right neighbors sufficiently to determine them uniquely out of the unordered set for any given word. In this case, the unordered set, in conjunction with the dependency rules, contains all the information necessary to reconstruct a unique ordering of the symbols for each word. Thus, context-sensitive coding provides a way to represent in a one-to-one mapping certain ordered sets by unordered sets in conjunction with some dependency rules.

This general formulation of the relationship between certain ordered and unordered sets seems to be of some value by itself. However, the general formulation was designed primarily to apply to a particularly simple example of a context-sensitive coding defined on a context-free coding for the vocabulary of a real language like English, and it is this particular example which is of primary interest here.

Let $x\ y\ z$ be adjacent context-free symbols in a word. In the context-free coding of the vocabulary, words always begin and end with the symbol #. The symbols $x\ y\ z$ could be adjacent letters in a written word or adjacent phonemes in a spoken word. For certain vocabularies, a context-sensitive coding for the word is obtained by mapping each context-free symbol y ($y \neq \#$) into a context-sensitive symbol ${}_{x}y_{z}$, where x and y are the left and right neighbors of y in that word and x and z may be #.

Note that a single-valued spelling of words with *context-sensitive symbols* does not imply that one has a *context-sensitive code*. The latter requires one to show that the mapping from the *context-sensitive spelling* back to the *context-free spelling* is also single-valued. If the vocabulary of the language was defined with ordered sets using n context-free symbols and if every possible triple of context-free symbols occurred in at least one word, then n^3 context-sensitive symbols would be required to define all the words in the vocabulary with unordered sets.

In general, reconstruction of the ordered set from the unordered set is easily accomplished by starting with the only ${}^{\#}u_{v}$ symbol in the word, then selecting the ${}_{u}v_{w}$ symbol in the word, assuming there is only one ${}_{u}v_{w}$ symbol in the word, and so on until all the symbols in the unordered set have been used and ${}_{y}z_{\#}$ has been written down as the last symbol. Whenever, by left-to-right generation, there is more than one choice for the next symbol in a word, the decision must be made by looking ahead to determine which choice leads to use of all of the symbols in the unordered set. It is possible to invent words with a context-free spelling for which the specified spelling with context-sensitive symbols is consistent with more

than one ordering of the symbols (i.e., spelling with context-sensitive symbols does not yield a context-sensitive code). However, this is an extremely rare event even if words are spelled randomly in the context-free code, and is easily avoided with non-random spelling.

These formulations of context-free and context-sensitive coding can be applied to both written and spoken English in the following manner. The spelling of written English words using letters (graphemes) is a context-free code. Similarly, the phonemic spelling of spoken English words is a context-free code. In both cases, the previously specified mapping from context-free symbols to context-sensitive symbols appears to produce a context-sensitive code. That is to say, the unordered sets of context-sensitive symbols for each word can be mapped back into one and only one ordering of the context-free symbols. Thus, the English word *stop* can be coded by the unordered set of context-sensitive symbols ${}_{\#}s_{t}, {}_{s}t_{o}, {}_{t}o_{p}, {}_{o}p_{\#}$, which is consistent with one and only one ordering of the associated context-free symbols */s, t, o, p/*.

I do not know of any words in either written or spoken English where the proposed context-sensitive spelling would be consistent with more than one context-free (graphemic or phonemic) spelling. In any event, the cases are so rare that, if any are found, they could probably be handled by special means, for example, slight modification of the defined set of phonemes or graphemes. In fact, in my original paper on context-sensitive coding (Wickelgren, 1969a), I proposed (for other reasons) that vowels with different stress be considered as different phonemes. As pointed out in that paper, there are actually very few cases in spoken English where one encounters any choice in the straight left-to-right generation of the order of the context-sensitive (and associated context-free) symbols. Out of the 3,800 words beginning with *b, d, f,* and *l* and occurring at least once in 10^6 words according to the Thorndike-Lorge (1944) count, there are only 12 words in spoken English where simple left-to-right generation would not be sufficient: *barnyard, brethren, fair-haired, farmyard, foreshorten, forlorn, fourscore, lampblack, Lapland, lifelike, limelight,* and *lullaby.* These are words which have two identical pairs of phonemes followed by a different phoneme. The choices in all of these cases can be resolved by "looking ahead" to see which choice uses all of the context-sensitive symbols. Alternatively, they can be resolved by a simple left-to-right associative process having no look-ahead capability of this type, provided one assumes that stress is a feature that distinguishes between vowel phonemes.

Granted that one can represent English words with unordered sets of context-sensitive symbols, does this accomplish anything? Is there any reason to think that human beings use a context-sensitive code for spoken or written words? The rest of this paper is devoted to making as strong a case as possible for the use of context-sensitive coding in the recog-

nition and production of spoken language. The case is nowhere near so strong for written language and will not be discussed in the present paper.

II. Speech Recognition

A. Acoustic Input

It hardly needs to be said that speech recognition in human beings has proven to be a very difficult problem to explain and to achieve artificially. Two of the principal problems discussed in the past have been: (*a*) the difficulty in subdividing the acoustic waveform for a word into segments corresponding to phonemes and (*b*) the lack of invariance in the acoustic cues for a phoneme across different left and right contexts of adjacent phonemes (Liberman, *et al.,* 1967). These are problems essentially because of the assumption that speech uses context-free (phonemic) coding. The second problem is completely eliminated by the assumption of context-sensitive (allophonic) coding, and the first problem is also eliminated for a device like the brain with parallel processing capability.

The fact that the cues for adjacent symbols in a word are often intermixed in time creates a problem if word recognition depends on recognizing the component symbols (phonemes) in a single correct temporal order. However, since context-sensitive coding permits recovery of the correct order from an unordered set, it is no longer very critical in what order the context-sensitive allophones are recognized. I will assume that the brain has an internal representative (one or more neurons) for each context-sensitive allophone. This internal representative is activated by some conjunction of acoustic features occurring over some maximum period of time (on the order of tens or hundreds of msec). The features are those characteristic of the context-sensitive allophone. For the brain, it is reasonable to assume that all the allophone representatives are "examining" the acoustic input in parallel, and when the word is finished some subset of the allophone representatives will have been activated above some variable threshold and the rest will not have been. Assuming no semantic context effects, the word representative which is maximally associated with this unordered set of context-sensitive allophone representatives will be selected. Nowhere in this process has it been necessary to subdivide the acoustic waveform for the word into segments, though, of course, it is still necessary in continuous speech to have marked the word boundaries.

For artificial speech recognition with serial devices, segmentation is highly desirable in order to reduce recognition time, even with context-sensitive coding. There are numerous ways to approach this problem from the standpoint of context-sensitive coding, but artificial recognition is beyond the scope of the present paper.

The success of context-sensitive coding obviously depends upon how invariant the acoustic cues are for a context-sensitive allophone. Frankly,

I do not know how invariant these cues are across different remote (non-adjacent) phonemic contexts, different syntactic and semantic contexts, different rates of talking, different speakers and different, recognizable dialects. However, the type of context-conditioned variation in the acoustic features of phonemes that has been most frequently discussed in the past (see Liberman, *et al.,* 1967), namely, dependence on adjacent phonemes, has been directly incorporated into the theory of context-sensitive coding and is no longer a problem. It remains to be seen how well the theory of context-sensitive coding will handle speech recognition, when extensive data become available on the effects of remote phonetic contexts, syntactic and semantic contexts (with their associated effects on suprasegmental stress), rates of talking, different speakers (particularly, men *vs.* women *vs.* children), and different, recognizable dialects. Nevertheless, we can say a little bit concerning how certain effects of these variables could be handled by a theory that assumed context-sensitive coding of speech.

Some remote phonetic context effects on acoustic features could be handled by expanding the number of phonemes (allophone classes) to include consonant clusters and distinguish between vowels with different segmental stress, for example.

If syntactic and semantic context greatly change the acoustic cues that are characteristic of particular context-sensitive allophones, then some basic modification would have to be made in the theory. To the extent that syntactic and semantic context affect cues that are not essential to recognition of each allophone, there is no problem. Thus, suprasegmental stress need not be a problem, if the cues for suprasegmental stress do not strongly interact with the cues for context-sensitive allophones.

The effects of rate of talking might be handled by variation in the time interval over which a context-sensitive allophone representative examines the acoustic input for its defining features. Faster rates of talking should be associated with shorter time-windows for each allophone representative. This hypothesis assumes that acoustic features for context-sensitive allophones can be defined which are invariant over rate of talking, except for the time interval over which they are found. If this hypothesis is false, then the theory of context-sensitive coding will require an important modification, at the very least.

Conceivably, the features of context-sensitive allophones can be defined so as to be invariant across different individual speakers and different dialects. However, such features would have to consist of relations (e.g., differences, ratios, rates of change) between formants at the same and different times, with a rather wide range of acceptable absolute formant values. Some of the capacity to recognize words in different dialects or spoken by different individuals might be handled by associations from different sets of context-sensitive allophone representatives to the same word. However, if neither of these two approaches is completely successful

in accounting for human speech recognition capacity in the face of individual and dialect differences, then it might be necessary to assume that the input to each context-sensitive allophone representative is a disjunction of conjunctions of acoustic features. In a recent paper (Wickelgren, 1969b), I discussed some reasons for thinking that many single cortical neurons might have the powerful logical capability of computing disjunctions of conjunctions. But it certainly would be simpler if it were not necessary to assume that context-sensitive allophone representatives had this capability.

B. Contextual Input

Another major problem that has frequently been discussed in connection with speech recognition is the rather large contribution of the context, in addition to the specific features of the acoustic input, in determining what word representative or component (allophone or phoneme) representative will be activated. Note that I am not now referring to the context effects on speech recognition which are mediated by their effects on the specific acoustic features of each word or word-component (allophones or phonemes). These were discussed in the previous section. I am now concerned with the context effects which raise or lower the probability of recognizing certain words or word-components, without having any effect on the specific acoustic features of the word or word-component being recognized.

It seems useful to distinguish between phonetic context on the one hand and syntactic and semantic context on the other hand. Phonetic context refers to the known or previously recognized allophones or allophone-classes (phonemes) in a word which could assist in the recognition of other allophones or phonemes in the word because of the previously learned non-random probabilities of different allophones or phonemes when preceded or followed by other allophones or phonemes. These effects would most likely be due to a subset of fully recognized word components being sufficient to activate the correct word representative, which, in turn, is associated to all of its allophone representatives, rather than being due to associations directly from one allophone or phoneme representative to another.

Certain syntactic and semantic context effects on word recognition can be handled by assuming that the context sets *thresholds* (biases) for different word classes, which, in turn, set thresholds for different words, which, in turn, might or might not set thresholds for different context-sensitive allophone representatives. It hardly needs to be said that this is a sketchily presented, unoriginal idea for the solution of a very difficult problem. The same idea has been described in more detail by Morton and Broadbent (1967), though not in relation to context-sensitive coding.

The principal contribution of context-sensitive coding to the problem of explaining context effects is the somewhat greater ease of incorporating backward phonetic, semantic, or syntactic context effects because of the

lack of any necessity to preserve the temporal order of the recognized and unrecognized allophones in a word.

III. Speech Articulation

A. Words

Context-sensitive coding in conjunction with an associative memory also provides a very simple theory of speech production, exclusive of the syntactic and semantic factors involved in the selection of word representatives. The application of context-sensitive coding to speech production has been discussed in more detail in a previous paper (Wickelgren, 1969a). This paper presents a briefer, slightly modified version of that application.

Once a word representative has been selected, for whatever reason, the ordered articulation of its components (allophones) is explained as follows: First, the word representative "primes" (partially activates) all of the context-sensitive allophone representatives either as an unordered set or with a slight temporal ordering favoring the earlier allophone representatives. The selection of the correct unordered set of around 7 allophone representatives from a total set in the tens of thousands is obviously an extremely important step, but one which is easily achieved by an associative memory.

The slight temporal ordering could come about because the long-term associations between the word representative and its allophone representatives are ordered in strength by degree of remoteness from the beginning of the word. There are reasons for thinking that this slight ordering of strength could not, by itself, account for the ordered generation of allophone representatives, though it could play a small role in helping to discriminate the order of non-adjacent allophone representatives. The basic mechanism by which a word's unordered set of context-sensitive allophone representatives is converted into an ordered set is by starting with the initial allophone representative ${}_{\#}u_{v}$, which activates ${}_{u}v_{w}$, and so on to the terminal allophone representative ${}_{y}z_{\#}$.

As mentioned earlier in the paper, there are rare instances in English where this simple left-to-right associative generation of an ordered set of allophones from the unordered set of allophones will be very slightly ambiguous, if one defines context-sensitive allophones on the usually accepted set of context-free phonemes. These rare cases can be handled in a variety of ways: (a) by expanding the number of phonemes to include vowel stress and perhaps also consonant clusters, (b) by assuming some "look-ahead" capability in the associative generation scheme, of the type previously described, or (c) by using any slight gradient of strength of association from the word representative to the allophone representatives as a function of remoteness from the beginning of the word.

In addition to explaining how a word representative could lead to the

ordered activation of its component vocal gestures, the theory of context-sensitive coding provides a mechanism for achieving roughly the same vocal gesture from different starting positions of the articulators. At the same time, the theory explains why there are coarticulation effects both in the vocal tract and in the patterns of firing of speech motor neurons (Harris, 1963; Fromkin, 1966; Harris, *et al.* 1966; Liberman, *et al.* 1967; MacNeilage and DeClerk, 1967). With a different central articulatory representative for each context-sensitive allophone, there obviously can be differences in the pattern of speech motor neuron activity for each allophone. However, it can be (and is) the case that the allophones of the same phoneme are generally quite similar in speech motor neuron activity (MacNeilage, 1963; Harris, *et al.* 1965; Fromkin, 1966; Liberman, *et al.*, 1967).

Certain aspects of articulation, namely, the control of timing and suprasegmental features, are not handled by the present theory, but it is important to note that this in no way contradicts the theory.

Although one can regard the order of vocal gestures as a part of the general question of the timing of vocal gestures, it is also possible to regard the control of speech rate as being quite separate from the control of the order of speech. It is the latter assumption which is made by the present theory. I assume that timing is accomplished by some kind of neural clock that regulates the rate of switching from one allophone representative to the next in the series.

The present theory assumes that suprasegmental representatives do not interact with segmental representatives at some level of the articulatory system. That is to say, the theory assumes that a suprasegmental stress representative is associated to the word representative in the input to the articulatory system. The word representative selects the segmental allophone representatives in the manner described and the suprasegmental stress representatives are simply activated along with the allophone representatives. Speech motor activity depends on both segmental and suprasegmental representatives which are activated at the same time. This hypothesis needs to be made more specific to be tested, but the general outline is clear: At some level of the articulatory system, segmental and suprasegmental representatives are additive, though this does not necessarily imply additivity in the acoustic waveform, vocal trace configuration, or pattern of motor neuron activity.

B. Phrases

Context-sensitive coding, in conjunction with some assumptions about priming and short-term and long-term associative memory, also allows a human being to select an ordered set of word representatives for a novel phrase and then articulate the entire phrase as an automatic process. During the articulation process, the higher cognitive level that selected the word representatives for the phrase can be selecting the words for the next phrase,

without the necessity of continued direction of the articulation process for the last phrase. This is achieved as follows:

The ordered activation of word representatives produces an ordered priming (partial activation) of the unordered sets of allophone representatives for each word in the phrase. The priming process selects the correct unordered sets of allophone representatives for the phrase (about 10^2) out of the vastly larger totality (about 10^4) of all allophone representatives. It also establishes short-term associations (due to relative contiguity of activation) among the allophone representatives. These short-term associations are strongest among the allophone representatives of each word and next strongest from the set of allophone representatives of one word to the set of allophone representatives of the next word in the phrase. Also, the unpronounced representative of the concept "begin" is primed before the priming of the first set of allophone representatives, so "begin" has its strongest short-term association to the allophones of the first word.

When the priming process is completed, "begin" is activated. This leads to the activation of the set of allophone representatives for the first word, with the ordering of full activation for the allophone representatives within a single word being determined by long-term associations in the manner described in the previous section. Then, the first word's allophone representatives will activate most strongly the set of allophone representatives for the second word, because they have stronger short-term associations to the set of allophone representatives for the second word than to the set of allophone representatives for any other word. The same process continues to the end of the phrase, when a new priming process can occur.

IV. Speech Development

No attempt will be made here to describe even an approach to a theory of the development of speech in children. However, it is worthwhile to point out that context-sensitive coding does somewhat simplify the problem facing the child in coming to understand and articulate words and word components. Throughout the following discussion, I will assume, for the sake of parsimony, that there is only one set of context-sensitive allophone representatives; not two sets, one for sensory functions and another for motor functions. I do not see how to distinguish these alternative hypotheses.

After development, there must be connections to the allophone representatives from lower-level auditory feature-representatives and connections from the allophone representatives to lower-level articulatory feature-representatives. We must be able to inhibit the motor output from these allophone representatives because we can perceive or think of words without repeating them aloud. These connections might be formed innately at some stage of maturation or be established by learning—both alternatives seem plausible.

Assume that the input and output connections of the allophone repre-

sentatives are specified independently of experience. In this case, there is little more to be said from a psychological point of view, except that the complexity of the mapping is simpler on both ends with allophonic coding than with phonemic coding.

If we assume that the input and output connections of allophone representatives are specified by learning, then there is a great deal more which must be said than I am prepared to say at the present time. However, it is again the case that the input and output connections for context-sensitive allophone representatives will be simpler than for context-free phonemic representatives.

The sensory specification of the allophone representatives might come about in the following way. First, we must assume that there are a large number of free neurons in the cortex whose inputs are not specified innately. Neurons standing for features of the acoustic signal send their axons into this region of free neurons and tend to grow toward a common point (their "center of gravity") when activated at the same time. If the same (or a highly similar) acoustic pattern is repeated sufficiently, then these axons will come very close together. When they are very close, there will be some nearest free neuron. The axons will all zoom down onto this free neuron, synapse with it and specify it to stand for that pattern of acoustic input. In such a manner, the sensory input of allophone representatives might be specified. An even grander version of this wild idea is discussed in an earlier paper (Wickelgren, 1969b).

Learned specification of the motor output of these allophone representatives would seem to be more complex, which fits with the fact that speech recognition precedes speech production by several months in child development. Sets of speech motor feature representatives must be activated in a variety of at least semi-random patterns. Those patterns of articulatory feature representatives which lead to sounds sufficiently similar to an allophone representative to activate it to some degree, acquire connections from the allophone representative. This is simply the old motor-sensory feedback loop for speech postulated by many to account for the development of sensory-motor connections in the development of speech. The idea is just somewhat more plausible with allophonic coding than with phonemic coding because the pattern of connections is simpler. However, it seems clear that this learning is not a one-step process. The series of successive approximations that the child makes to adult speech makes it clear that the motor output connections of allophone representatives must be assumed to be undergoing a series of changes.

V. Conclusion

It is quite obvious that context-sensitive coding does not solve all the problems of speech recognition, articulation, and development at even the phonetic level. However, it does solve some problems, and this makes it worthy of consideration as a theory of one basic phonetic unit.

REFERENCES

Fromkin, V. A. (1966). Neuro-muscular specification of linguistic units. Language and Speech, *9:* 170–199.

Harris, K. S. (1963). Behavior of the tongue in the production of some alveolar consonants. J. acoust. Soc. Am., *35:* 784 (abstract).

Harris, K. S., Lysaught, G. and Schvey, M. M. (1965). Some aspects of the production of oral and nasal labial stops. Language and Speech, *8:* 135–147.

Harris, K. S., Huntington, D. A. and Scholes, G. N. (1966). Coarticulation of some disyllabic utterances measured by electromyographic techniques. J. acoust. Soc. Am., *39:* 1219 (abstract).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychol. Rev., *74:* 431–461.

MacNeilage, P. F. (1963). Electromyographic and acoustic study of the production of certain final clusters. J. acoust. Soc. Am., *35:* 461–463.

MacNeilage, P. F. and DeClerk, J. L. (1967). On the motor control of coarticulation in CVC monosyllables. Unpublished paper presented at the 1967 Conference on Speech Communication and Processing, Massachusetts Institute of Technology.

Morton, J. and Broadbent, D. E. (1967). Passive versus active recognition models or is your homunculus really necessary? in W. Wathen-Dunn (ed.), Models for the perception of speech and visual form. Cambridge: MIT Press, 103–110.

Thorndike, E. L. and Lorge, I. (1944). The teacher's word book of 30,000 words. New York: Columbia University Bureau of Publications.

Wickelgren, W. A. (1969a). Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychol. Rev., *76:* 1–15.

Wickelgren, W. A. (1969b). Learned specification of concept neurons. Bull. math. Biophys., *31:* 123–142.

DISCUSSION

BOYNTON: What happens if you record a set of allophones and then put them together in a different order to correspond to new words?

WICKELGREN: This sort of thing has been tried and it was found that it could not be done. If you cut, say, one section from one phoneme and put it together with another section from another phoneme, you do not get the sound you might expect. This is in line with my model, in which I would have to cut sound segments three phonemes wide and match them appropriately. Maybe Al Liberman will have some more to say about this problem. If one takes a large set of allophones, instead of some 40 or 50