Reprinted from

DYNAMIC ASPECTS OF SPEECH PRODUCTION

Current Results, Emerging Problems, and New Instrumentation

edited by

MASAYUKI SAWASHIMA and FRANKLIN S.COOPER

UNIVERSITY OF TOKYO PRESS, 1977

SPEECH SYNTHESIS AS A TOOL FOR THE STUDY OF SPEECH PRODUCTION

Franklin S. Cooper, Paul Mermelstein and Patrick W. Nye

Haskins Laboratories, New Haven, Connecticut

Our intent in this paper is to describe some research studies that we are undertaking and to explain our reasons for choosing speech synthesis and the class of research questions that synthesis as a methodology implies. Briefly, we wish to learn what parts of the complex articulatory events of speech production are actually carrying the message, i.e., what articulatory cues the speaker must produce in order that the listener will understand what was said. We think of this as a search for the articulatory cues that parallels earlier work we have done on searching for the acoustic cues in speech.

There are close parallels between the two kinds of search, and we have found it useful in planning the work on articulatory cues to draw analogies with our experience in searching for acoustic cues. Hence, we will speak of that experience in presenting our plans. We will even digress into a brief description of a new pattern playback we have built; it will be useful in the planned studies even though it was designed primarily for research on acoustic cues.

We have usually spoken of speech synthesis as a tool for the study of speech perception. But the acoustic cues we found all seemed to point back to articulation, implying that we were, in fact, studying production by way of perception. Thus, the parallels between our earlier work and the planned work can be viewed in this way: both were concerned with speech production, though the earlier work was on cues at the acoustic level, whereas the planned work is on cues at the articulatory level.

In either case, the distinguishing characteristics of the methodology are that it seeks to find the principal carriers of information, that it tests for these cues by perceptual methods, and that it uses synthetic speech to do so. Obviously, speech is the required stimulus when the perception of a message is to be tested, and synthetic speech has the very great advantage that systematic manipulation of the stimuli is possible, either at the acoustic level or at the articulatory level that precedes it.

RESEARCH METHODS: FROM ACOUSTIC CUES TO ARTICULATORY CUES

The method we used in searching for the acoustic cues, often called "hypothesize-and-test," proved well suited to that task (Liberman and Cooper, 1972). We think it will be equally effective in the search for the articulatory cues. The earlier work was, in fact, modeled on the chemist's customary technique of testing his analytic conclusions by synthesizing the suspected compound and comparing properties. We started with the patterns we thought we could see in sound spectrograms and regenerated sound from such patterns with a device we built for that purpose, namely, the Pattern Playback. In using it, a speaker produces an utterance from which the experimenter prepares a spectrogram. Guided by this spectrogram, a schematic copy is painted and passed to the Pattern Playback for conversion into synthetic speech. Now the two speech samples, the natural and the synthetic, are compared to determine <u>by ear</u> whether the essential acoustic cues have survived in the painted copy. The procedure is highly interactive. The user is given the opportunity to rapidly insert or delete spectral features at will and to immediately assess their importance by listening to the synthetic output and comparing it with the natural speech sample.

The principal ways in which we propose to model our new procedures on the old are by providing the means to obtain results

quickly, to make modifications to the data interactively by hand, and to compare the outputs at a variety of different levels, but especially at the perceptual level. The organization of the research method is illustrated in Fig.1, which shows three ways to experiment on speech that is generated by a real speaker, an articulatory model, or a terminal analog speech synthesizer.

For articulatory synthesis, we may compare the articulatory control data for a particular articulation with EMG data from our physiological research, especially as to relative timing of events. Likewise, we may compare, almost directly, the vocal tract shape for articulatory synthesis with x-ray and fiberoptic data measured from an actual vocal tract. Differences in the moment-by-moment vocal tract configurations will indicate where improvements might be made in the synthesis. When rules have been used to compute the control signals and vocal tract configurations, means will also be available to override these controls and to make changes in the vocal tract shape directly by hand. This facility will be useful in a number of experimental situations where it is desirable to examine the acoustic effects of individually specified articulatory movements.

As a final step in the above procedures, the output signal is presented to listeners, who are asked to make relative judgments about the speech, or absolute judgments about its intelligibility or adequacy. Exploratory manipulations and informal listening will usually be followed by formal group tests.

MODELING THE SPEECH PROCESS

In representing the speech process by a model (or synthesizer) and in manipulating it with control parameters that specify the phonetic elements of the message, the choice of level of representation is crucial. Moreover, that choice hinges on a number of considerations: intended use, feasibili-

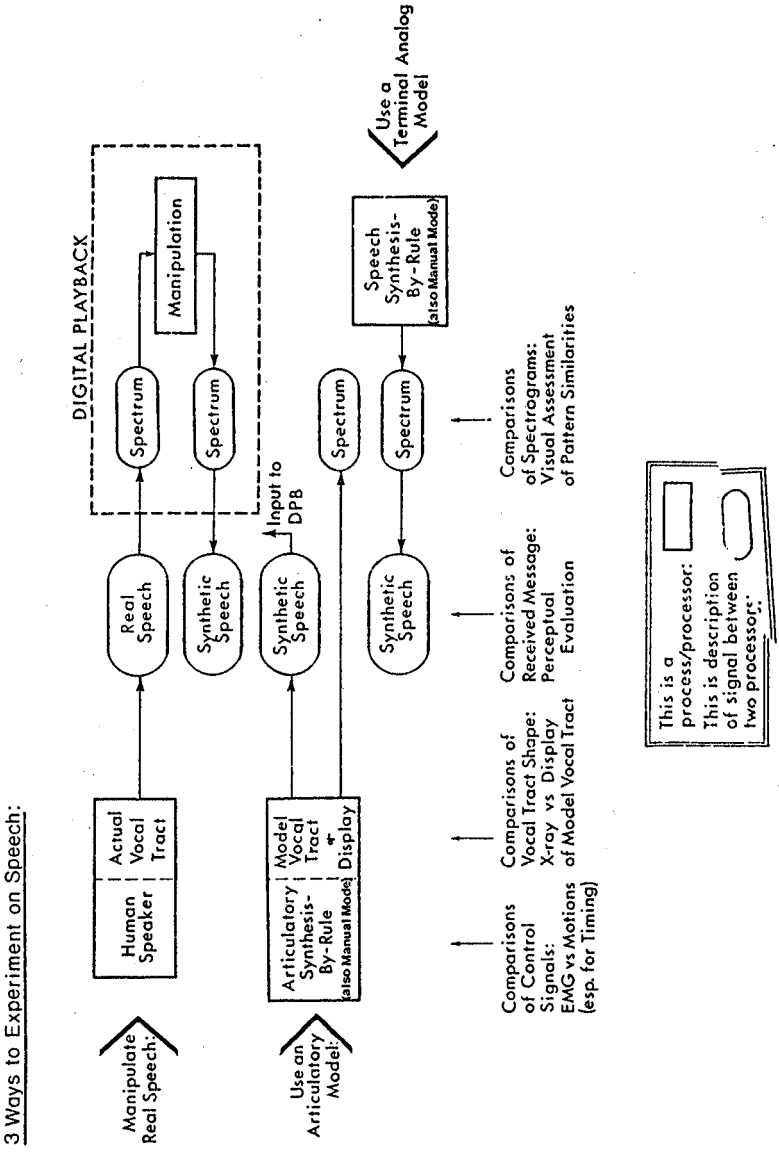
3 Ways to Experiment on Speech:
Manipulate Real Speech:
Human Speaker
Actual Vocal Tract
Real Speech
DIGITAL PLAYBACK
Spectrum
Manipulation
Spectrum
Synthetic Speech
Comparisons of Control Signals: EMG vs Motions (esp. for Timing)
Use an Articulatory Model:
Articulatory Synthesis-By-Rule (also Manual Mode)
Model Vocal Tract + Display
Comparisons of Vocal Tract Shape: X-ray vs Display of Model Vocal Tract
Input to DPB
Synthetic Speech
Comparisons of Received Message: Perceptual Evaluation
Spectrum
Use a Terminal Analog Model
Speech Synthesis-By-Rule (also Manual Mode)
Spectrum
Synthetic Speech
Comparisons of Spectrograms: Visual Assessment of Pattern Similarities
This is a process/processor:
This is description of signal between two processors:

Fig.1.

ty, conceptual convenience, and available knowledge are the primary desiderata.

If we consider human speech production, we find three distinct levels of the articulatory process that lie downstream from the presumed neural levels (to which we have little or no direct experimental access):

1. The activity of the individual muscles (in response to neuromotor commands).
2. The positions of the articulators and their movement in responses to muscle activity.
3. The corresponding vocal-tract shape in terms of the cross-sectional area function of the vocal tract.

For the research purposes we have in mind, namely an exploratory search for the articulatory cues, the third and lowest articulatory level is not very useful since, at the level of vocal-tract area functions, the conceptually important entities--the positions and movements of individual articulators--have already been merged into a single continuum. We will certainly wish to observe the performance of the model at this level, and even to exercise supervisory control over the area functions, but primary conceptualization and control can be done to better advantage at the next higher level, i.e., by manipulating the articulators themselves.

Would we gain by working at still higher levels of muscle activity or of the neuromotor commands that control the muscles? The philosophical question of where maximum simplicity is eventually to be found is yet to be answered convincingly. For the present, then, we rely on the practical considerations that our knowledge (from articulatory phonetics and cinefluorography) is better at level two than at level one, and that starting higher in the speech process would require more parameters and an additional stage of computation (to reach level two) without compensating advantages other than that electromyographic information could be applied more directly. For all these reasons, we intend to concentrate on the representa-

tion of phones and features at the second of the levels listed above and on transformations from that level to the speech signal.

DESIGN CONSIDERATIONS FOR AN ARTICULATORY SYNTHESIZER

We intend to start our research using an articulatory model developed by Mermelstein (1973) that allows parametric specification in the midsagittal plane for the position of the lips. tongue tip, tongue body, velum, jaw, and hyoid; we will extend the model by the addition of a variable that produces concave/convex arching of the tongue blade. These parameters permit the computation of vocal-tract transfer functions for laryngeal excitation or for fricative excitation at points internal to the tract (Mermelstein, 1972). The model has already demonstrated a capability for matching vocal tract configurations seen in x-ray movies and for generating highly intelligible VCV syllables.

The model does not simulate the entire speech-production system in man. In particular, it separates control of the sources of excitation of the vocal-tract resonances from control of the changes in those resonances with time. Since many aspects of coarticulation depend only on the interaction of the supraglottal articulators, only the positions of these articulators are computed, starting from a phonetic specification. Laryngeal excitation parameters (amplitude, fundamental frequency, onset, and duration) are specified explicitly, and effects of the supraglottal system back on the excitation source are neglected. Similarly, for frication, the amplitude and spectrum of the noise is explicitly specified; the output spectrum will, of course, reflect not only the source spectrum but also the filtering action of the vocal-tract cavities posterior and anterior to an assumed frication source at the point of maximum constriction along the tract.

The prime reason for not modeling directly the effects of articulator movement on the characteristics of the sound gen-

eration process is to limit the complexity of the simulation. Thus, we do not for the present intend to model laryngeal action because we do not think that it plays a central role in the coarticulation processes that we plan to study first. An exception may be the relative timing of laryngeal and supralaryngeal events, but this does not require detailed simulation of laryngeal mechanisms. Similarly, although the generation of frication is directly dependent on appropriate articulatory conditions, its accurate modeling requires very precise timing and positioning of the articulators. For these reasons, we rely on explicit control over the excitation signal itself rather than over the generative processes. The perceptual effects of simultaneous excitatory and articulatory variations can still be evaluated quite adequately, despite these substitutions for aerodynamic effects that link excitation to articulation in real speech.

The considerations that led to modeling articulatory movements exclusively in terms of the resulting vocal-tract shape in the midsagittal plane were primarily based on observational limitations. That is, the model was originally developed on the basis of a systematic examination of a series of midsagittal x-ray tracings of the vocal tract, in conjunction with the time-synchronized speech signal. By working interactively with the model during its development, it could be shown that displacements of the midsagittal vocal-tract outline can be derived from the movements of the independently controlled articulators.

Primary control of the model in terms of positions and movements of the principal articulators is, of course, an essential design consideration: this mode of control is conceptually convenient for the experimenter and it is the natural framework for the application of structural and dynamic constraints. The articulatory model we will be using builds on the "ball-in-mouth" model of the articulators that was introduced by Coker and Fujimura (1966), but uses a more nearly complete set of

articulators. The parameters assigned to these articulators are position variables which indicate the position of the structure in fixed space or relative to some other articulatory structure to which the articulator is primarily attached. For example, lip and tongue-body positions are specified with respect to the moving jaw. This representation allows an active mode of movement when an articulator's own parameters are changing; alternatively, a passive mode of movement may be executed relative to the fixed articulators as a result of movement of the structure to which that articulator is attached, but relative to which its position remains unchanged.

The model first computes the midsagittal outlines that result from the momentary positions of the articulators and then computes the midsagittal separations relative to an essentially fixed outer structure. Published information is used to convert these distances, measured at a large number of points along the vocal tract, to a continuous cross-sectional area function along a center-line distance function between the glottis and the lips. Up to 25 area samples spaced 0.875 cm apart are now computed and used in a nonuniform acoustic transmission line representation. Appropriate lumped terminations model the larynx at one end of the tract and the lips at the other. Articulations accompanied by a velar opening are modeled with the aid of an acoustic sidebranch which parallels the nonuniform transmission line for the oral tract. The cross-section of this nasal branch is assumed to be fixed except for a region near the velum.

THE CONTROL AND DISPLAY OF ARTICULATORY SYNTHESIS

The articulatory process will, of course, be simulated on a computer, since digital simulation provides flexibility and convenience that is not attainable through the use of physical models. For the model to be a truly useful tool, it must be equipped with displays that allow observation of the conse-

quences of input instructions at all levels of the synthesis process. Further, to facilitate the hypothesize-and-test mode of experimentation, the model is controllable by interactive graphical editing at either the level of the individual articulators or of vocal-tract shape; also, when comparison with spoken utterances is desired, changes can be made directly in the spectral representation. Finally, the model must provide an acoustic output promptly on demand. Only on this basis can the user readily assess the perceptual consequences of the synthesis process, and only when the synthesizer responds promptly to changes in the control parameters is it easy to maintain a conceptual link between the hypothesis being tested and the result of the test.

The control and display facilities we will use are best considered in terms of the functional modes in which the model is to be operated. At the articulatory level, there is need for convenient control in terms of articulatory parameter values (for the individual articulators) and their allowed variation. As an aid in visualizing these numerical specifications, the corresponding vocal-tract outline (midsagittal) will be displayed. To change or improve the synthetic sound, interactive graphical editing will allow the user to redraw part or all of a midsagittal vocal-tract outline. X-ray information can be conveniently introduced at this point. Generally, also, the articulatory parameters may be quickly determined from such an outline.

When the model is used in its dynamic mode, the specified articulations for allophones will be supplemented by a set of rules that govern the time-variation of the articulatory system. With the specified articulations stored in a table of parameter values, one for each specified allophone, the rules will operate on the selected sets of these parameters to yield continuous functions of time.

An alternative procedure to the automatic generation of articulatory sequences by rule--one that will be especially use-

ful in trying out new ideas or making detailed improvements to rule-generated sequences--is to redraw individual parameter "tracks" on the interactive graphic display. The total effect (on the other parameters as well) can then be seen in the mid-sagittal display and heard as the synthetic speech output.

At the spectrum level, it will be useful to view the spectral consequences of the articulatory movements. For voiced articulations, it is advantageous to view the spectral envelope without regard to fundamental-frequency variations. Such a spectral envelope and the corresponding formant frequencies can be derived from the model without the need for generating the actual signal waveform. Since the formant frequencies are the terms in which the acoustic cues are best known, this also makes for easy comparisons. When the results, as viewed in the above representations, are acceptable, we will generally want to generate the acoustic signal itself. This is so because, by listening to the signal, we may quickly judge its quality or naturalness and assess its identifiability.

At this point we can make good use of another research tool we are just completing: the Digital Pattern Playback (Nye et al., 1975). This device stores the speech spectrum in computer core memory and so can immediately display a conventional gray-scale spectrogram for interactive graphical editing. Visual comparisons can then be made between the original spectrogram (generated, in this case, by articulatory synthesis) and either (1) the same spectrogram after it has been edited to improve intelligibility, or (2) a spectrogram of human speech of the same sentence. The Digital Pattern Playback also provides for comparison <u>by ear</u> of the sounds that correspond to the spectrograms. In other ways, too, the DPP's capabilities for display and manipulation of speech data make it a useful companion device to the articulatory synthesizer.

IDEAL SYNTHESIZERS AND RESEARCH SYNTHESIZERS: WHY THEY MAY DIFFER

Speech synthesis based on articulatory models has, of course, a considerable history. Some of the major contributions to the design and use of articulatory synthesizers, and to the underlying knowledge about relations between articulation and sound, are listed under "References." There have been varied reasons for building such synthesizers; in some cases, the reason was to demonstrate that synthesis could be done in a particular way; in some, to mimic human production as seen by x-ray; and in some to attempt the production of more natural speech than is easily obtainable from terminal analog synthesizers or to control synthesis at a lower bit rate. Usually, some part of the effort has been directed to getting natural-sounding speech, i.e., to approximating the performance of an ideal synthesizer.

It seems obvious that an articulatory synthesizer deserving the label "ideal" would have a capability for mimicking human speakers quite exactly. We would rate its performance initially on the naturalness of its "spoken" output; later, we would inquire about how accurately its chain of transformations (from phonetic sequence to sound) match those of the human speaker. Comparisons would be made at the levels of vocal-tract shape and acoustic spectrum--perhaps even at the level of the speech waveform. It hardly needs saying that no existing synthesizer comes near to meeting such criteria.

However, ideal performance is not necessarily what we most want from a research synthesizer; that is, the question of ideal performance needs reexamination when we ask, not about the naturalness of the speech, but about the usefulness of the synthesizer for research--in particular for searching out the articulatory cues. The objective of this latter task is to find the simplest possible description of articulatory events that will, despite crudities in the speech, let a listener recover

the phonetic message.

If we draw on our experience in searching for the acoustic cues, we will wish to manipulate these articulatory "events" in a variety of ways to study the perceptual consequences for the corresponding speech events. Sometimes this will involve efforts at simplification, for example, by allowing only the tongue or the lips to move in synthesizing a syllable that is normally spoken with some degree of movement by most of the articulators. Again, experimentation will involve stepwise variation in the relative timing of two component gestures, for example, of tongue and lip movements in synthesizing a syllable like (ibu), or an initial consonantal cluster such as (bl) in (bled). Here good discrimination of the time delay between lip and tongue release would imply a basically unified organization for the cluster, whereas poor discriminability would indicate an independence of the constituent gestures. Too much delay of the tongue-lip release would result in the insertion of a vowel in the perceived phonetic string (i.e., (bəled) instead of (bled)); of course, this must be avoided, since it would cue a phonetic distinction.

Obviously, experimental manipulations of this kind do not mimic natural speech. Often they call for an independence of control or a grading of spatial and temporal relationships that a human speaker could hardly achieve. To be sure, they ought not violate physiological constraints but, short of that, we will want to put the articulators through their paces in order to assess the perceptual consequences. Our expectations about the resulting sounds is that many, perhaps most, will sound "strange" but that some, perhaps many, will be clearly identifiable.

Thus, simplicity, both conceptual and operational, is a primary requirement in a research synthesizer. We expect to employ the fewest independent articulators, and the fewest control parameters to position and move them, that will still generate acceptable tokens for all the syllables of the language,

i.e., that will generate all the phones in the full range of phonologically allowed contexts.

Indeed, it is the essence of modeling to try for the maximum simplicity that will still give the required performance. In a search for the cues, performance is properly judged at the level of intelligibility, which is different from, and less demanding than, naturalness. Hence, naturalness in synthesis is for us not a primary short-term goal, nor was it needed in our earlier search for the acoustic cues. We found, in our work with the Pattern Playback, that the pursuit of the acoustic cues could proceed in the presence of a somewhat unnatural speech quality that even lacked pitch inflection. Intelligibility was the important requirement and proved to be nearly orthogonal to the dimension of naturalness.

Departures from naturalness are not, of course, a virtue, nor do we wish, when manipulating the articulators, to depart unnecessarily from the general configurations we see in x-ray movies. The guidance that level-by-level comparisons (of synthesis vs nature) can give us is too valuable to be ignored. Indeed, we will sometimes want to manage the articulation so as to make it come quite close to the human model. The problem in designing a research synthesizer was to retain this capability, or as much as could be had, without paying too high a price in complexity of representation and control.

SUMMARY

Our reasons for undertaking a search for the articulatory cues are, firstly, that this will provide an insight into the nature of speech production comparable to the view we gained of speech perception when we succeeded in finding many of the major acoustic cues, and secondly, that the relationships between cues and phonetic elements should be simpler and more direct in the articulatory domain than they proved to be in the acoustic domain.

The origins of our interest in this undertaking lie in what we think we have learned about the nature of speech from two parallel lines of investigation. From studies of how speech is perceived, we learned that, although the accoustic signal contains a wealth of detail, only some of the things one sees in a spectrogram are important to the ear in identifying the phonetic content of the spoken message. These we have called the acoustic cues. Numerous other things that can be seen in the spectrogram are largely irrelevant, at least for intelligibility. By ignoring these things and synthesizing speech from patterns that contained only the acoustic cues, we were able to greatly simplify the acoustic signal and still retain most of the intelligibility. When we examined the nature of these acoustic cues, however, we found few one-to-one correspondences between them and the phones they represented; rather, the relationships were complex in ways that pointed to a reorganization and overlapping of articulatory gestures during speech production.

From physiological studies of how speech is produced, we have learned that articulatory events also seem complicated; thus, articulation, as seen in x-ray motion pictures or electromyographic recordings, involves most of the articulators most of the time. We can assume that here, too, some limited set of the component gestures provides the critical information (by way of sound as intermediary) on the basis of which a listener identifies the phonetic content of the message. These we refer to as the articulatory cues. If our assumption is correct, then much of the total articulatory description is also largely irrelevant, at least for intelligibility. But interest in the articulatory cues goes beyond stripping away irrelevancies. A more important point follows from our interpretation of the nature of their counterparts in the acoustic domain: if the acoustic cues do indeed reflect their articulatory origins, then the articulatory cues should show the simpler relationship with phonetic elements.

We think the time is right to undertake a search for the articulatory cues. There exists an extensive body of knowledge about both perception and production. There is a proven research method and experience with a computer-based articulatory synthesizer on which to implement it. Thus, both a significant problem and the means to probe it are at hand.

REFERENCES

Chiba, T. and Kajiyama, N. (1941) The vowel, its nature and structure (Tokyo: Tokyo-Kaiser Kan Pub. Co.).

Coker, C. H. and Fujimura, O. (1966) Model for specification of the vocal tract area function. J. Acoust. Soc. Am., 40; 1271.

Coker, C. H., Umeda, N. and Browman, C. P. (1973) Automatic synthesis from ordinary English text. IEEE Trans. Audio Electroacoust. AU-21, 293-298.

Dunn, H. K. (1950) The calculation of vowel resonances, and an electrical vocal tract. J. Acoust. Soc. Am. 22; 740-753.

Fant, G. (1960) Acoustic Theory of Speech Production ('s-Gravenhage: Mouton).

Flanagan, J. L., Ishizaka, K. and Shipley, K. L. (1975) Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell System Tech. J., 54; 485-506.

Fujimura, O. (1962) Analysis of nasal consonants. J. Acoust. Soc. Am., 34; 1867-1875.

Fujimura, O. (1975) Syllable as a unit of speech recognition. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 79-82.

Liberman, A. M. and Cooper F. S. (1972) In search of the acoustic cues. In Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, ed. by A. Valdman (The Hague: Mouton), pp. 329-338.

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

Lindblom, B. E. F. and Sundberg, J. (1971) Acoustical consequences of lip, tongue, jaw and larynx movements. J.

Acoust. Soc. Am., 50; 1166-1179.

Mermelstein P. (1971) Calculation of the vocal-tract transfer function for speech synthesis applications. In Proceedings of the Seventh International Congress on Acoustics, Vol.3 (Budapest: Akademiai Kiado), pp. 173-176.

Mermelstein, P. (1972) Speech synthesis with the aid of a recursive filter approximating the transfer function of the vocal tract. In Conference Record, 1972 Conference on Speech Communication and Processing, April 24-26, Newton Mass., pp. 152-155.

Mermelstein, P. (1973) Articulatory model for the study of speech production. J. Acoust. Soc. Am., 53; 1070-1082.

Nye, P. W., Reiss, L. J., Cooper, F. S., McGuire, R. M., Mermelstein, P. and Montlick, T. (1975) A digital pattern playback for the analysis and manipulation of speech signals. Haskins Laboratories: Status Report on Speech Research, SR-44, 95-107.

Öhman, S. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am., 39; 151-168.

Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Am., 24; 175-184.

Rice, L. (1971) A new line analog speech synthesizer for the PDP-12. Working Papers in Phonetics (Linguistics Department, University of California at Los Angeles) 17; 58-75.

Rosen, G. (1958) Dynamic analog speech synthesizer. J. Acoust. Soc. Am., 30; 201-209.

Stevens, K. N. (1968) On the relations between speech movements and speech perception. Zeitschrift f. Phon. u. Kommunikationsforschung, 21, 102-106.

Stevens, K. N. and House, A. S. (1955) Development of a quantitative description of vowel articulation. J. Acoust. Soc. Am., 27; 484-493.

Stevens, K. N. and House, A. S. (1963) Perturbation of vowel articulations by consonantal context. J. Speech Hearing Res., 6; 111-128.