REMARKS ON STOP CONSONANTS -- SYNTHESIS EXPERIMENTS AND ACOUSTIC CUES*

Osamu Fujimura

I. The Work at Haskins Laboratories

The systematic study at Haskins Laboratories of acoustic cues for the perceptual identification of stop and nasal consonants demonstrated, for the first time, the usefulness of speech synthesis experiments for exploring physical correlates of linquistic units.[5] By use of a unique technique of speech synthesis, important conclusions were derived, which, combined with the development of the acoustic theory of speech production,[7] led to an epoch-making advance in speech science. One of the essential contributions to our understanding of the nature of the sounds of speech was the finding that the perception of consonants, in particular stop consonants, crucially depends on the formant transition characteristics.[20] The notion of formant frequency loci was thus deduced from psychological experiments where naive subjects listened to artificially synthesized stimuli and identified them as syllables containing one of several consonants specified as candidates.[6]

As for the distinction between [p] and [b], [t] and [d] or [k] and [g], using native speakers of American English as subjects, their experimental results showed, in terms of the spectrographic patterns as specified for the Play-back synthesizer, that the transitional characteristics of the first formant bar constituted the common cue. In particular, they concluded that the "voiced" stops were characterized by a very low F_1 locus, i. e. , marked and invariably negative F_1 transitions, whereas the "voiceless" stops were simulated successfully by "cutting-back" an uninflected first formant bar for the

*To appear in "Festschrift for Eli Fischer-Jørgensen".

initial period of 30-50 msec.[21] For this rather unexpected conclusion about the characteristics of "voicelessness" some partial but quite plausible explanations have been forwarded by speech scientists.* Before discussing this problem, and adding some new points based on later experimental findings about sound perception, I shall first review my own experinece relating to the perception of synthetically produced stop consonants, which has thus far been only informally reported,[13),14)] but in my opinion is quite relevant to the issue.

II. Voice Onset Time and Voiced-Voiceless Distinction

The experiment to be described briefly here was conducted at the Speech Communication Group, Research Laboratory of Electronics, MIT, mostly in 1959. Some of the relevent technical details are reported in the Quarterly Progress Report of the Laboratory.[14)]**

An electrical analog of the vocal tract[25)] was used to synthesize a set of test stimuli which had controlled values of voice onset time associated with a fixed (40-msec linear) transition of the vocal tract shape from a palatal stop [k/ɟ] configuration to a vowel [ɛ] configuration. For the optimal values within the range of control of the variables, we obtained quite natural sounds for [kɛ] and [ɟɛ]. Actual test stimuli were prepared with 10 different buzz onset times (in steps of 10 msec), and the presence or absence of a fundamental frequency inflection was also controlled. The fundamental frequency, when given an inflection, was shifted linearly from 70 Hz to 100 Hz during the initial 150-msec period of voicing. A white-noise source was applied at the glottis input of the electrical vocal tract during the initial portion of the syllable. The noise started abruptly (with an onset of 1 msec) and simultaneously as the start of the articulatory change,

---

*See, e.g., f.n. (5), p. 154 in Liberman et al.,[21)] and discussions in Fischer-Jørgensen.[10)]

**I would like to take this opportunity to express my thanks to K. N. Stevens, Morris Halle and A. S. House for their leadership and valuable discussions during my work with the research group.

and then immediately decayed in an amplitude linearly to zero in 50 msec, regardless of the voice onset time.

Eighty test items containing 4 each of 20 different stimuli were recorded in a randomized order. The subjects were asked to decide whether the stimulus sounded like [ʤɛ] or [kɛ]. Ten American subjects and three Japanese subjects participated in the listening test. The result for two subjects who gave particularly interesting responses are shown in Figure 1. Both subjects

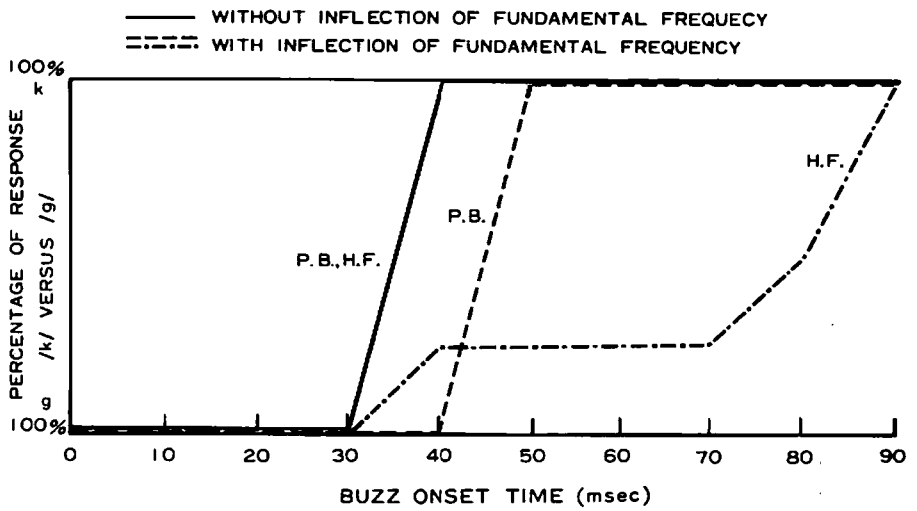

Figure 1. Two samples of different responses in respect to the effect of the fundamental frequency inflection in judgments [kɛ] vs [ʤɛ].

P. B. and H. F. showed the same complete (4 votes) jump from [ʤɛ] to [kɛ] in one step, that is, with a 10-msec shift of the buzz onset time, when there was no fundamental frequency inflection. When the inflection was present, this boundary shifted invariably to a later onset time for all subjects, but the curve of the stimuli with inflection present differed considerably from subject to subject. Subject P. B. did not give any [ʤ]-responses if the onset time was

50 msec or more, after the start of noise, whereas H. F. was reluctant to give a [k]-vote, even with an unvoiced period as long as 80 msec, even though he had no problem in giving [k]-judgments when there was no pitch inflection. Subject H. F. was Japanese, but the other Japanese subjects did not show this marked behavior. Also this tendency was often seen among the American subjects, although less markedly. Thus the two response patterns cited above represented the two extremes, and general response patterns revealed somewhat intermediate characteristics; the average curve for 10 American subjects as shown in Figure 2 seemed to represent a typical response curve for both American and Japanese subjects.

A similar experiment was conducted for the alveolar stops combined with the vowel [i]. In this experiment the noise was inserted at the articulatory constriction. The result was very similar to that of the experiment

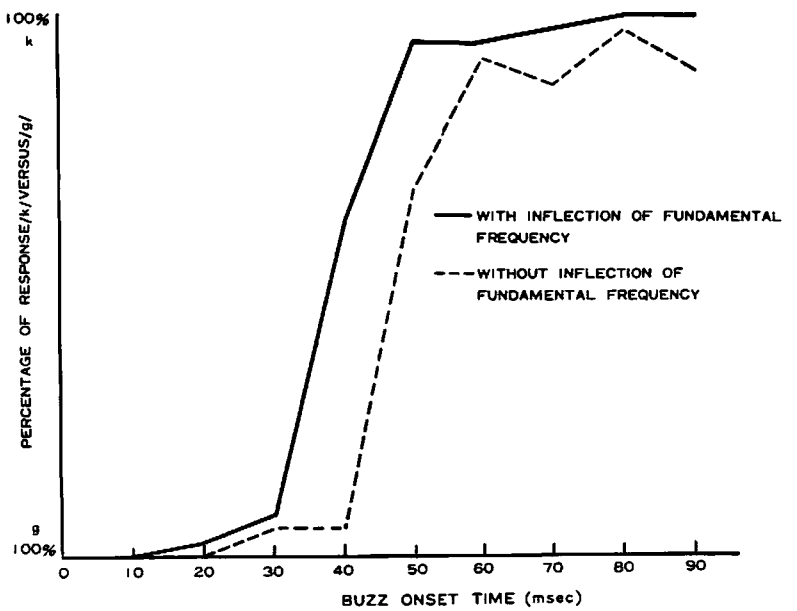

Figure 2. Average responses of 10 subjects for the same experiment as in Figure 1.

just described, except that the shift from [di] to [ti] occurred at an earlier value of the buzz onset time in relation to the moment of expolsion. *

It would appear that the fundamental frequency inflection, as used in these experiments, is not favorable for the naturalness of syllables with an initial [k] or [t], while it is favorable for [ɡ] or [d]. Also, it may be mentioned, that a judgment can be switched from [k] to [ɡ] or [t] to [d] just by changing the fundamental frequency inflection, if the buzz onset is given a value in the "boundary region". **

This experimental result suggests in this situation a rather clear-cut distinction between the primary and the secondary cues. When the primary cue, viz., an appropriate specification of the buzz onset time, exists either for the so-called voiced or voiceless stop, adding the secondary cue, viz., the pitch inflection for the voiced judgment, is either immaterial, or perturbs naturalness and causes some partially inconsistent responses for the judgment. The effect of the secondary cue is often easily overridden, even when it is provided with the optimal value of the pertinent variable, if the primary cue has a counteracting effect. Sometimes by some reason. the primary cue may not be clear enough for the identification of either of the candidates, and it is in such cases that the secondary cue casts a deciding vote.

III. Cues in Synthesis of [ɡɑ]

There are cases, however, where there is difficulty in finding acoustic characteristics that can be interpreted as the primary cues in the above

---

* This result concerning dependence of the voice onset delay on the place of articulation conforms with data obtained by the analysis of natural utterances. 22), 8) It also compares well with the results of more recent and extensive synthesis work. 1)

** T. Sato, in studying the waveform characteristics of natural utterance, had concluded that the pertinent class distinction of initial stops of Japanese (which is traditionally labeled "pure vs. impure") was determined by the pattern of the fundamental frequency inflection. 26)

sense. The notion of cue, in general, has to be clearly defined for further discussion, because it is necessary to define a level of description within which the characterizations of stimuli are given, in order to compare effects of different factors. To obtain some insight into this problem, let us consider cues for identification of [b], [d] and [ɡ] in an isolated CV syllable.

The synthesis-listening experiments at Haskins Laboratories have demonstrated that this type of syllable generally can be synthesized successfully by means of linear formant transitions of a fixed duration for all formants, and that the patterns of the transitions, furthermore, may be prescribed by the notion of formant loci that are inherent to the consonants. If we could assume this framework of acoustic specification to be complete for a certain set of consonants in certain contexts, then we would be in a position to evaluate contributions of, e. g. , different formants to perceptual identification of the phonemes or of particular distinctive features, by controlling at one time one locus frequency with all other loci fixed as parameters. For this purpose, theory advocates the use of a series connected terminal analog synthesizer which automatically determines the formant levels under the constraints of the natural production mechanism of vocalic sounds.[11] The use of this particular type of synthesizer defines the framework of describing the stimuli. If we wish, we may also add to this framework of description some other independent acoustic features such as an explosive burst or the "spike fill" in the visible speech terminology,[24),18)] the selection of its energy-concentrated frequency portion being controlled, and look for the relative importance of such features in comparison with the transition features.

It may be remarked that in order to study intricate perceptual characteristics quantitatively, it is in general necessary to prepare an experimental situation where we can precisely control the physical characteristics of stimuli that sound natural and convincing as representing intended phonetic forms. Since the superposition principle in predicting the perceptual results of combining different factors is not to be counted on, it follows that it is not generally the case that we can compare contributions

of different factors in a realistic perceptual mode by estimating effects of the controlled variables under influences of some important but unknown factors for which the parameters are fixed at values far from the optimum ones. Thus the desirable experiment will be one employing a synthesizer by which natural production processes can be actually simulated when the appropriate values of the controlled variables are selected, and at the same time, the important factors to be studied are represented by a small number of independently controllable variables. The series type synthesis scheme mentioned above seems to satisfy these conditions for studying the characteristics of voiced stops in CV-type syllables.

In this somewhat better defined working framework, we may follow the Haskins experiments and try to reevaluate the contributions of different acoustic features to stop identification. In fact, the simple model of the linear and simultaneous formant transition (without a burst) is effective for many stop-vowel combinations with satisfactory intelligibility and naturalness in the resulting synthesized syllables.[14] But some particular CV combinations are difficult to simulate by this method. A typical example is the syllable [ga] with a back open vowel.

If we expand the scheme of acoustic specification by adding a noise burst feature to the synthesis scheme, we can quite successfully simulate this syllable. This was demonstrated by an experiment at MIT, using their terminal analog synthesizer (POVO) controlled by trapezoidal signal generators.[25] In order to obtain optimal results within the modified linear transition model, we employed the following technique:[14] (a) a short burst of noise was mixed at the input of the synthesizer with the buzz excitation at the initial portion of the syllable; (b) a higher starting point for the first formant, e. g. , 450 Hz instead of 200 Hz (value for [ba] and [da]) was used; and (c) the start of the formant transition was delayed, thereby leaving a short buzz-excited portion with stationary formants.

The resulting sound was evaluated by a formal listening test using native speakers of American English. In the test tape, the [ga] stimulus with the special features, which may be labeled [g'], was mixed with three other

syllables, labeled [b], [d], and [ɡ]. All of the latter three were synthesized by optimal selections of variables within the Haskins type synthesis framework for representing the syllables [bɑ], [dɑ], and [ɡɑ], respectively.

The scores of the test clearly showed that the stimuli [b], [d], and [ɡ'] were practically perfect for identification, giving scores of 99-100% correct responses. The [ɡ]-syllable, however, was ambiguous, and the score for this stimuli varied from subject to subject, the number of [ɡ]-responses varying from 7 to 0 votes out of 10 for each subject, the rest being mostly [d]-responses. For most subjects, it was clear that for distinction between [dɑ] and [ɡɑ], the three special features, viz., the burst of noise, the transition delay, and the higher F_1, were quite helpful.

Among these, it was clear that the presence of the noise burst was necessary for producing a clear [ɡɑ]. The relative importance of the two others was evaluated by formal pair-comparison tests. The results strongly indicated that the higher location of the starting frequency of F_1 improved the phonetic quality of [ɡɑ] regardless of whether there was formant delay combined or not. The delay in the start of the formant transition was clearly preferred only when the other two features were present.

The weaker two cues, if we should call these two acoustic features so, are ad hoc, viewed from the natural production mechanism. The first vocal tract resonance for [ɡ]-closure is certainly significantly higher than for [b] or [d], as revealed by direct acoustic measurements,[15] but it is not higher than 250 Hz.* The delay of the formant transition simulates the natural acoustic signal, but this reflects the non-linear characteristics of the articulatory-acoustic conversion,[7),27)] and is strongly dependent on the particular CV combination. A more invariant characterization of the acoustic structures of syllables in reference to the inherent phonetic properties of the constituent phonemes or distinctive features will be given in terms of the articulatory variables as functions of time.[2),3)] The introduction of the

*For discussions concerning interpretations of the formant loci in terms of actual vocal tract resonances during closure, see Stevens and House,28)

noise burst, which seems particularly important for the consonant in this context, might also be automatically accountable if we adopted an appropriate physical model for the production mechanism, considering the interaction between the vocal cord vibration and the vocal tract.[12]

Acoustic cues, if they should be somehow defined, must ultimately pertain to perceptual effects, however. In order to speak of a perceptual cue and discuss some hierarchy among cues in any consistent way, we will need to establish an effective model of auditory processing, so that we can always know exactly how the acoustic (or articulatory) conditions are transformed in the auditory domain. There is no a priori reason to assume that the spectral components as such are particularly relevant in describing either acoustic or auditory correlates of phonetic features. The pertinent level of the auditory domain might not be expected to be single or clear-cut, even if we eventually come to understand the physiological mechanism exactly and in detail.

IV. The Cutback Technique and Sensation of Periodicity

As discussed in the first section, the apparently marked effect of the F_1 cutback as a common cue of "voiceless" stops requires some explanations. It may be recalled first, that the synthesizer that was employed in the famous study by Cooper and others at Haskins Laboratories about twenty years ago was by obvious reasons highly limited in technical performance when viewed from the presently available synthesis techniques. In fact, essentially the synthesizer was appropriate only for quasi-periodic sound with a fixed fundamental frequency, which was set at 120 Hz, i. e. , for voiced sounds by a typical male voice.[4] Thus "voiceless" sounds, according to the most straightforward phonetic interpretation, cannot be appropriately produced by this equipment under quantitative and systematic control. Liberman et al. later used another synthesizer called Voback in their experiment, in order to estimate the effect of using a random noise source in place of buzz for simulating voiceless stop releases.[21] By this study they

concluded that the use of the noise source for the F_2 and F_3 components improved the quality of voiceless stops, but F_1 cutback was still necessary for the perception of voicelessness.

In a recent preliminary study of aperiodicity of voice source signals, conducted at the Bell Telephone Laboratories by use of computer-simulated vocoder techniques, we have found some interesting facts about the periodicity of natural speech signals and the consequent perceptual effects.[16] The portion of the speech signal, even a portion of an open vowel articulation, that evokes perfect phonetic perception of voicing, is sometimes essentially aperiodic in the higher frequencies. This is so, not only in the sense that it lacks regular periodicity when observed by accurate analysis, but also in that the time portion of the signal in the frequency band can be artificially replaced by completely random noise (with appropriate spectral shaping according to the channel-vocoder principle) without any significant perceptual difference or degradation. Generally, the higher the range of frequency, the less periodic the "voiced" signal is. In the frequency range between 1 kHz and 2 kHz, only limited time portions in sentences are significantly periodic in the sense that replacing the periodic source by a random noise source in the frequency-time domain gives rise to a degradation of quality. Above 2 kHz, there are only occasional occurrences of periodic segments in this perceptual sense. It may be mentioned, incidentally, that the vertical striation in the wide-band spectrograms can be deceptive in judgments of periodicity, because there can be buzz-modulated random noise that shows perfectly lined up striation, but has no periodicity at all. The horizontal striation in the narrow-band spectrogram is more reliable in this respect. The analysis mentioned above was performed by the cepstrum technique[23] applied separately for different frequency ranges by use of appropriate weighting windows in the frequency domain.

Thus one may assume that the sensation of "voicedness"* primarily

*"Voicedness" (Stimmhaftigkeit) in this context is a phonetic (or physical) measure, 9) and it does not pertain directly to the so-called "voiced-voiceless" opposition. The latter notion is open to deeper discussions, which will be far beyond the scope of this paper. 9)

depends on the so-called baseband, which includes frequencies up to the F_1 region. It is, of course, true that one may hear something like pitch, a sensation related to regular waveform repetition, when a buzz-excited signal is passed through a high pass filter with a cut-off frequency of say 1,000 Hz. This is not surprising because the resulting waveform has apparent periodicity whether there is a fundamental component or not, and the auditory nerves conduct a set of partially synchronized periodic pulses from nonlocalized portions of the cochlea. The phonetic perception of "voicedness" as such can be different from such periodicity sensations. It will also depend on the environment of listening, or the mode of perception.

Under experimental conditions where the subjects always hear sound that has a fixed periodicity in the physical sense, like in the Playback synthesis experiments, the contrast between the presence and absence of the periodic F_1 component (in the baseband) may, by itself, well be responsible for a clear and consistent voiced-unvoiced discrimination for the time portion immediately after the explosion. If we adopt this assumption, together with what is known about the voice onset time as a dominating cue for the so-called voiced-voiceless opposition, then the effect of the F_1 cutback is satisfactorily explicable. The high-stressed energy balance in the cutback portion, of course, simulates the natural energy distribution of hiss in aspiration. In the case of a natural voiceless stop followed by a vowel, the first formant transition is perceptually obscure, though physically it does exist, partly because of a weaker distribution of excitation energy compared with the voice source signal, and partly because of some absorption of the first formant energy due to the open glottis. The latter point has been corroborated by direct acoustic measurements of the vocal tract transfer characteristics.[15]* It may also be mentioned that not only dissipation loss, but also the resonant characteristics of the subglottal system, contribute to energy absorption in the low frequencies. When the frication noise is produced at the articulatory constriction, the effects of the resonant characteristics of the cavity behind the constriction, viz., the effects of an anti-

*Some further discussions on this point are given in a paper to be published by the same authors.

formant,[19] also have to be considered for the time portion near the explosion.

The conclusion by Liberman et al. that the use of noise as the source did not evoke the "voiceless" judgment when it was not accompanied by the F_1 cutback, suggests that the energy balance gave a stronger cue than aperiodicity for "voicelessness", under the conditions of energy balance used in the experiments. The channel type synthesizer Voback, by its nature, does not guarantee a natural level balance for itself. Thus we will need an accurate control of levels with quantitative reference to a theoretical or experimental simulation of the natural production system, in order to obtain comprehensive conclusions concerning human acuity to any deviation from the natural conditions.

References

1) Arthur S. Abramson and Leigh Lisker, "Voice Onset Time in Stop Consonants: Acoustic Analysis and Synthesis," 5e Congrès International d'Acoustique: A51 Liège 7-14 (1965).

2) C. H. Coker and O. Fujimura, "A Model for Specification of the Vocal-Tract Area Function," J. Acoust. Soc. Amer., Vol. 40, 1271 (1966).

3) C. H. Coker, "Speech Synthesis with a Parametric Articulatory Model," Preprints: Speech Symposium, Kyoto, A-4-1 -- A-4-6 (1968).

4) Franklin S. Cooper, Alvin M. Liberman, and John M. Borst, "The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech, "Proceed. Nat. Academy of Sciences, Vol. 37, 318-325 (1951).

5) F. S. Cooper, P. C. Dellattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am., Vol. 24, 597-606 (1952).

6) P. C. Delattre, Alvin M. Liberman, and Franklin S. Cooper, "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am., Vol. 27, 769-773 (1955).

7) G. Fant, Acoustic Theory of Speech Production, The Hague: Mouton and Co., 1960.

8) Eli Fischer-Jørgensen, "Acoustic Analysis of Stop Consonants," Miscellanea Phonetica II, 42-59 (1954).

9) Eli Fischer-Jørgensen, "Beobachtungen über den Zusammenhang zwischen Stimmhaftigkeit und intraoralem Luftdruck," Z. Phonetik, Sprachwissenschaft Kommuniationsforschung, Vol. 16, 19-36 (1963).

10) Eli Fischer-Jørgensen, "Voicing, Tenseness and Aspiration in Stop Consonants, With Special Reference to French and Danish," Annual Report of the Institute of Phonetics, University of Copenhagen, Vol. 3, 63-114 (1968).

11) James L. Flanagan, "Note on the Design of 'Terminal-Analog' Speech Synthesizers," J. Acoust. Soc. Am., Vol. 29, 306-310 (1957)

12) J. L. Flanagan and L. Cherry, "Excitation of Vocal-Tract Synthesizers," J. Acoust. Soc. Am., Vol. 45, 764-769 (1969).

13) O. Fujimura, "Some Characteristics of Stop Consonants," J. Acoust. Soc. Am., Vol. 31, 1568 (1959).

14) O. Fujimura, "Some Synthesis Experiments on Stop Consonants in the Initial Position," Quarterly Progress Report of Research, Laboratory of Electronics, M. I. T., No. 61, 153-162 (1961).

15) O. Fujimura and J. Lindqvist, "Experiments on Vocal Tract Transfer," Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, Sweden, 1-7, (October, 1964).

16) O. Fujimura, "An Approximation to Voice Aperiodicity," IEEE Transactions on Audio and Electroacoustics AU-16, 68-72 (1968).

17) O. Fujimura and J. Lindqvist, Sweep-tone Measurements of the Vocal Tract Characteristics (to be published).

18) M. Halle, G. W. Hughes, and J. P. A. Radley, "Acoustic Properties of Stop Consonants," J. Acoust. Soc. Am., Vol. 29, 107-116 (1957).

19) S. Hattori, K. Yamamoto, and O. Fujimura, "Nasalization of Vowels in Relation to Nasals," J. Acoust. Soc. Am., Vol. 30, 267-274 (1958).

20) A. M. Liberman, P. C. Delattre, F.S. Cooper, and L. J. Gerstman, "The Role of Consonant-Vowel Transitions in the Perception of Stop and Nasal Consonants," Psychological Monographs, Vol. 68, 1-13 (1954)

21) A. M. Liberman, P. C. Delattre, and F. S. Cooper, "Some Cues for the Distinction between Voiced and Voiceless Stops in Initial Position," Language and Speech, Vol. 1, 153-167 (1958).

22) Leigh Lisker and Arthur Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements, Word, Vol. 20, 384-422 (1964).

23) A. Michael Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am., Vol. 41, 293-309 (1967).

24) R. K. Potter, G.A. Kopp, and H. G. Green, Visible Speech, D. Van Nostrand Co., Inc., 1947.

25) G. Rosen, "Dynamic Analog Speech Synthesizer," J. Acoust. Soc. Am., Vol. 30, 201-209 (1958).

26) T. Sato, "On the Differences in Time Structures of Voiced and Unvoiced Stop Consonants," J. Acoust. Soc. Japan, Vol. 14, 159-164 (1958).

27) K. N. Stevens and A. S. House, "Development of a Quantitative Description of Vowel Articulation," J. Acoust. Soc. Am., Vol. 27, 484-493 (1955).

28) Kenneth N. Stevens and Arthur S. House, "Studies of Formant Transitions Using a Vocal Tract Analog," J. Acoust. Soc. Am., Vol. 28, 578-585 (1956).