XII. SPEECH COMMUNICATION

Prof. M. Halle	R. Capraro	K. Nakata (visitor)
Prof. K. N. Stevens	J. Emerson	G. Rosen
Dr. T. T. Sandel	J. M. Heinz, Jr.	C. E. Persons
G. W. Hughes	D. T. Hess	M. G. Schachtman
	P. Lieberman	

A. INTRODUCTION

For the past few years, part of the speech research at M.I.T. has been administered through the Acoustics Laboratory, although close collaboration with the Speech Analysis group of the Research Laboratory of Electronics has always existed. The Acoustics Laboratory group has recently transferred to RLE, and their research is now reported in the RLE Quarterly Progress Report. The general objectives of the speech communication research of the combined group are the same as those outlined by Professor Halle in the Quarterly Progress Report of January 15, 1958. Research topics transferred from the Acoustics Laboratory include studies of the acoustics of speech production, the electrical synthesis of speech, and contributions to the development of systems for bandwidth compression of speech.

K. N. Stevens

B. DYNAMIC ANALOG SPEECH SYNTHESIZER*

The main part of this synthesizer is an electric transmission-line analog of an acoustic tube that represents the human vocal tract. The line has eleven L-C sections consisting of Miller-effect capacitors and saturable inductors. These elements are variable over a 100:1 range and respond to a set of 13 articulatory control signals while satisfying certain constraints between element values. The line has one output, one input for periodic impulse excitation, and several alternative inputs for noise excitation.

The control system consists of a multiple-output trigger pulse source and a group of circuits that respond to the trigger pulses to generate continuous voltages for excitation and control. The pulse source has a scale-of-100 counter that generates a 100-point time base. Four trapezoidal function generators are used to: (a) generate the temporal envelope of glottal pulse excitation, (b) generate the temporal envelope of turbulent noise excitation, (c) program inflection patterns (changes in the pulse rate essential for naturalness), and (d) program movements of the vocal tract from one articulatory configuration to another. A glottal pulse generator produces a pulse train in response to the frequency and amplitude control signals from the trapezoidal function generators. A noise generator is controlled by an amplitude control signal and generates noise with the proper time and frequency envelopes. The remainder of the control system consists of a resistance matrix for storing several configurations of the vocal tract, and circuits for insuring proper transitions between those configurations.

A new method of obtaining linear modulation of noise amplitude was recently incorporated in the control system; precise control of noise excitation of the dynamic analog

*This work was supported in part by Air Force Contract AF19(604)-2061.

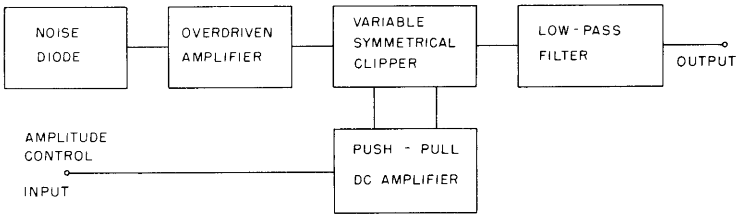

Fig. XII-1. Block diagram of symmetrical-clipping noise modulator. The lowpass filter acts as a summing device for noise samples.

was not possible with classical modulation schemes. The new modulator is a physical implementation of the central limit theorem (1). The theorem states that if $y = \sum_{1}^{n} x_i$ and if the x_i are statistically independent and are distributed with finite mean and variance, then the distribution of y approaches the Gaussian as n approaches infinity, regardless of the exact form of the x_i distributions. Figure XII-1 illustrates the principle of operation. The noise diode feeds an overdriven amplifier whose output is ideally a rectangular wave with random zero-crossing times and whose two values are equiprobable and symmetrical about zero. Such a wave has a zero mean. The actual output distribution clusters about two values in a manner that approximates the ideal distribution, and the output waveform has significant components up to 200 kc. Symmetrical clipping of the ideal rectangular wave does not distort it but merely changes its amplitude scale. The push-pull dc amplifier acts as a threshold voltage source for the clipper so that the clipping limits can change in response to the control voltage. Hence this voltage controls the amplitude scale of the clipper output.

The effect of the lowpass filter can be seen with the help of Shannon's sampling theorem (2). The input and output of the filter are bandwidth-limited to 200 kc and to 10 kc, respectively. Each sample of the filter output can be regarded as a weighted sum of effectively twenty samples of the filter input, although no lumped parameter filter exhibits an ideal square-cornered impulse response. The conditions of the central limit theorem hold sufficiently well for the intended purposes, and the distribution of the filter output is approximately Gaussian, with the variance determined by the amplitude control voltage. Symmetrical circuitry operating on symmetrical waveforms yields an output with no dc component. The circuit is free from the "thump" that arises when a portion of the control voltage appears as an unwanted component of the output.

This circuit can be compared with the classical balanced modulator whose signal-to-"thump" ratio is often less than one. The classical modulator has a highly nonlinear control voltage-versus-output characteristic with tails in the region of cut-off. Also, it has distortion whose character varies with gain. Thus the shape of the noise amplitude distribution at the output of a classical modulator depends on its gain setting. In our circuit these difficulties have been circumvented by taking advantage of the properties of the signal.

G. Rosen

References

1. W. B. Davenport, Jr. and W. L. Root, An Introduction to the Theory of Random Signals and Noise (McGraw-Hill Book Company, New York, 1958).
2. C. E. Shannon, Communication in the presence of noise, Proc. IRE 37, 10-21 (1949).

C. SYNTHESIS OF NASAL CONSONANTS BY TERMINAL ANALOG SYNTHESIZER

Terminal analog speech synthesizers are usually designed for the generation of vowel sounds. In this series of experiments we are examining the extent to which certain classes of consonants can be generated by a terminal analog synthesizer. We hope that the experiments will also provide some indication of the cues that are important for the perception of certain speech sounds. The present study is concerned with a particular class of consonants – the nasals /m/, /n/, and /ŋ/.

The terminal analog synthesizer consists of two parts: a variable electronic circuit excited by an electrical "buzz" source and a control device. The variable circuit is a cascade connection of four electronically controlled tuned circuits, each of which simulates one of the first four vocal tract resonances or formants. The control portion of the synthesizer controls the variable resonances and the excitation to generate synthetic syllables of the type schematized by the intensity-frequency-time pattern shown in Fig. XII-2.

The synthetic syllable has three major parts: an initial segment in which the resonances are stationary; a transition segment during which the resonances move

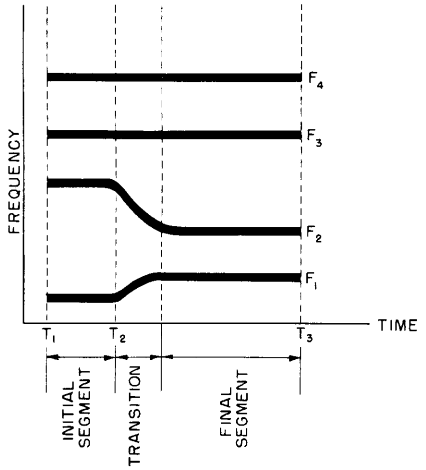

Fig. XII-2. Schematized intensity-frequency-time pattern of a typical syllable, generated by terminal analog synthesizer. The formants are designated by $F_1 \ldots F_4$. No attempt is made in this pattern to reproduce relative formant intensities.

approximately exponentially to steady positions; a final segment during which the resonances are stationary at new frequency locations. The frequency of the buzz source that excites the resonances is held constant at about 125 cps in all stimuli. The following parameters can be controlled: the frequencies of each of the four formants during both initial and final segments; the initiation and termination times of the buzz excitation, T_1 and T_3; the time T_2 at which the transition begins; and the transition rate, or the duration of the transition segment. In many respects the synthesizer performs the same functions as other synthesizers that have been described previously (1). Other formant synthesizers have not, however, utilized the cascade connection for the resonances, and hence the amplitudes of the resonances have not been constrained to vary in a natural manner.

The acoustical properties that appear to distinguish the nasal consonants from other classes of speech sounds are (a) a low resonance in the frequency range 200-300 cps, (b) a broadening of formant bandwidths caused by excessive damping in the nasal cavity, and (c) the existence of a zero in the transfer function of the configuration used to generate nasal consonants. In order to verify these statements, we carried out a preliminary listening test in which stimuli of the type shown in Fig. XII-2 were generated with different starting frequencies and bandwidths for F_1. The results showed that with a starting frequency of 200 cps for F_1, nasal consonants can be distinguished from voiced stop consonants if the bandwidth of F_1 is sufficiently large, even if there is no zero in the transfer function. It should be noted that when F_1 is at a low frequency, the over-all amplitude is relatively low compared with the amplitude of the adjacent vowel, and the amplitudes of higher formants are also very low. These amplitude relations are a consequence of the cascade connection of the formant resonant circuits.

Following the preliminary listening tests, we performed a further test in which 99 stimuli were synthesized with the following characteristics:

1. Frequency of first formant in initial segment: 200 cps;
2. Bandwidth of first formant during entire stimulus: 300 cps;
3. Bandwidth of all other formants: 50-100 cps;
4. Frequency of second formant in initial segment: 900, 1100, 1300, 1500, 1700, 1900, 2100, 2300 cps;
5. Duration of initial segment: 20, 50, 100 msec;
6. Duration of transition: 20, 50, 100 msec;
7. Formant frequencies during final segment: 730, 1090, 2500, 3500 cps (vowel /ɑ/).

The stimuli were presented in random order to a group of subjects, who were asked to identify each stimulus as one of /mɑ/, /nɑ/, or / ŋ ɑ/.

Results are summarized in Fig. XII-3. Each point represents an average of responses for all subjects over the three initial time intervals and the three transition

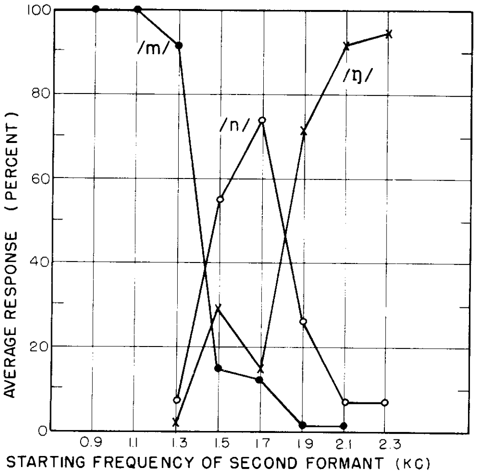

Fig. XII-3. Percentage of /m/, /n/, and / ŋ / responses as a function of starting frequency of second formant. Data are averaged over several consonant durations and transition times.

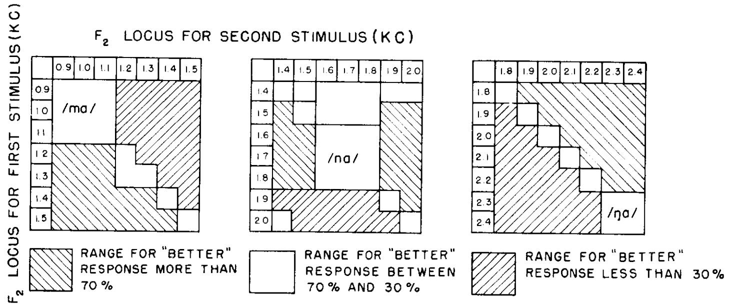
F_2 LOCUS FOR SECOND STIMULUS (KC)
F_2 LOCUS FOR FIRST STIMULUS (KC)
0.9 1.0 1.1 1.2 1.3 1.4 1.5
/ma/
1.4 1.5 1.6 1.7 1.8 1.9 2.0
/na/
1.8 1.9 2.0 2.1 2.2 2.3 2.4
/ŋa/
RANGE FOR "BETTER" RESPONSE MORE THAN 70 %
RANGE FOR "BETTER" RESPONSE BETWEEN 70 % AND 30 %
RANGE FOR "BETTER" RESPONSE LESS THAN 30 %

Fig. XII-4. Results of paired comparison tests. The F_2 locus is plotted vertically for the first member of a pair and horizontally for the second member. The percentage of better or worse responses is coded as shown. Areas for best /ma/, /na/, and / ŋa/ stimuli are indicated. See text.

durations. The figure shows that the starting frequency of the second formant (which we shall call the F_2 locus) has a marked effect upon the identification of the syllable, /m/, /n/, and / ŋ / responses being associated with low, medium, and high F_2 loci, respectively. These results follow a general trend noted in other studies (2, 3), and yield response peaks that are more sharply defined than the data of the previous investigators. The initial time interval and transition time do not have a marked effect upon the responses, although intermediate values of these variables appear to yield the highest responses. The preferred duration for the initial segment was 50 to 100 msec; the preferred transition time was 20 to 50 msec for /m/ and 50 to 100 msec for /n/ and / ŋ /.

In an attempt to specify the best F_2 locus more closely for each nasal consonant, we carried out further listening tests by a method of paired comparisons. In each test, the

F_2 loci for all stimuli were within the ranges appropriate for either /m/, /n/, or / ŋ / responses as determined from Fig. XII-3. Stimuli with different F_2 loci were presented in pairs, and subjects were asked to judge whether the second member of the pair was a better or worse version of the nasal consonant-vowel syllable.

Results of the paired-comparison tests are given in Fig. XII-4. These data again indicate ranges of F_2 loci that are appropriate for each of the nasal consonants. A 50 per cent response in this test means that both members of the pair are either good versions or poor versions of the nasal consonant under test. The regions associated with "good" versions can easily be identified in Fig. XII-4. The following ranges of F_2 loci appear to be appropriate for the various nasal consonants:

/mɑ/	: 900-1100 cps
/nɑ/	: 1600-1800 cps
/ ŋɑ/	: 2300-(2500) cps

These results are entirely consistent with data on voiced stop consonants obtained by other investigators (3), who used a different type of synthesis device.

In summary, the following conclusions emerge from the present data:

1. Very low (about 200 cps) starting frequency and broad bandwidth for the first formant are adequate cues for distinguishing nasal consonants from other consonants.

2. Nasal consonants can be distinguished from each other on the basis of the starting frequency for the second formant. Well-defined frequency ranges for the F_2 locus are associated with each of the nasal consonants, at least in the vowel context studied here.

K. Nakata

References

1. F. S. Cooper, Spectrum analysis, J. Acoust. Soc. Am. 22, 761-762 (1950).

2. A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, The role of consonant-vowel transitions in the perception of the stop and nasal consonants, Psychol. Monographs 68.8, 1-13 (1954).

3. A. Malécot, Acoustic cues for nasal consonants, Language 32, 274-284 (1956).