Speech Synthesis — A Critical Review of the State of the Art

Mark Tatham

Reproduced from International Journal of Man-Machine Studies (1970) Vol. 2, pp. 303-308.

INTRODUCTION

The present paper is divided into three parts: a. the synthesiser, b. control of the synthesiser and c. use of synthesizers. There is no attempt to give a detailed account of the history of speech synthesis (for this, see: Flanagan, 1965 and Mattingly, 1968), nor any account of the details of computer programming for the control of synthesizers: what this paper is mainly concerned with are the strategies involved in what has become known as "rule synthesis" and the effect of these on the use to which speech synthesis might be put.

THE SYNTHESIZER

Currently there are two basic types of speech synthesizer in use as laboratory research tools: i. vocal tract analogue (VTA) synthesizers and ii. terminal analogue (TA) or resonance synthesizers. The VTA is designed to simulate the human vocal tract by means of a cascade of resonant circuits arranged to perform analogously to a human vocal tract quantized into a number (usually 18-20 — corresponding to the average length in centimetres of the tract itself) of (approximately) cylindrical sections. The resonant circuits are driven by a signal derived from a circuit whose output corresponds to the waveform produced at the vocal cords — that is, a quasi-periodic source having a harmonic structure, whose harmonics decrease in amplitude at a rate of approximately 6 dB/octave above 100 Hz. (For details of VTA synthesis see Rosen, 1958 and Kelly and Lochbaum, 1963.) A TA synthesizer, on the other hand, is designed to simulate the acoustic output of the human speech process. A typical configuration for a TA would consist of three of four resonant circuits each having a frequency range corresponding to the first three of four formants of average human speech and a bandwidth of some 70-100 Hz; and possibly a further resonant circuit corresponding to the (fixed frequency) nasal formant. These resonators are driven by a larynx pulse generator similar to that required for a VTA. A filtered white noise source is usually provided for the simulation of fricative sounds (see Lawrence, 1953; Fant *et al*., 1963; McKinney *et al*., 1966). Both VTAs and TAs exist either as hardware or in the form of computer simulations — the latter having the advantages of stability and ease of change of configuration.

CONTROL OF THE SYNTHESIZER

Historically (that is, over the last 25 years or so) several methods of synthesizer control have been employed, but since about 1960 computer control has been gaining ground and provides the most versatile and comprehensive method. Whichever method of control is used it is generally the case that temporally governed varying voltages provide inputs to the various synthesizer circuits for holding or changing the circuit variables. In the early days of speech synthesis values for these variables were obtained as continuous functions from X-ray motion pictures (in the case of a VTA) of spectrographic analysis of human speech (in the case of a TA) and often supplied to the synthesizer by manual programming. This was obviously a long and tedious process involving complete prior analysis of a human version of the utterance required from the synthesizer.

Kelly and Gerstman (1961) and Holmes *et al*. (1964) established that it was possible to generate time governed control signals from a simple set of typical control values for individual sounds and a set of rules establishing the temporal variations in these values for sound segments in juxtaposition: in other words and as an example, from typical values for the sound [a] and for the sound [i], given a rule about what happens to these values when [a] and [i] occur in a word in succession as [-ai-] then intermediate, transitional values could be calculated. It became possible to input to the control computer only simple strings of symbols (sequentially, but not temporally specified) corresponding roughly to normal orthography, to obtain running speech as the output from the synthesizer. At the present time there are several such systems in operation and research is continuing in an attempt to improve the rules to produce more human-like speech (see, for example, Werner and Haggard 1969). It is significant that the demands of economy of space in the computer and time on the part of the researcher had forced some of those working in speech synthesis to adopt a generative strategy which is common enough in the physical sciences but which now constitutes a theoretical mainstay of the most productive area of modern linguistics — Transformational Generative Grammar (Chomsky, 1967; Chomsky and Halle, 1968).

Chomsky, as linguist, has noted that a human being is capable of uttering sentences which he himself has never heard of spoken before; his obvious conclusion is that this capability is best accounted for by a grammar involving the rearrangement by rule of a stored set of items. A very large or infinite set of sentences can be generated using a small finite set of rules. This notion is precisely that underlying speech synthesis strategy: utterances can be generated from a set of units and a set of rules which interact to produce speech not contained within the memory of the computer. The utterance exists potentially only by reason of the computer's ability to interact the two sets of items. Thus, whereas in the previous systems of strategy utterance memory was involved (in the form of the X-ray motion picture of the spectrogram), now no memory is involved where the speech generated might be the simple reproduction of stored information derived externally as a complete succession of units. Too little has been made of the enormous theoretical implications of this change of strategy — whatever might have been the reasons for its happening (see Kim, 1966, where this fact is implied, possibly for the first time).

USE OF SYNTHESIZERS

The convergence of a strategy in speech synthesis and a theoretical principle in linguistics now makes sense of attempts to use speech synthesis as a working model of speech production. The TA synthesizer reproduces the waveform of speech more or less accurately; the VTA synthesizer is a model of the acoustic properties of the vocal tract itself — but neither on its own hitherto could be described as a model of speech production. The area of speech synthesis of principal interest is of course the development of control programs which in some way are analogous to current linguistic theory concerning the behaviour of human beings.* [**footnote:* It should point out here a personal view that the use of speech synthesis to overcome transmission problems in telecommunications is largely an anachronism — there are cheaper and better ways of bandwidth compression and communication under high static conditions.

Two principal uses emerge for speech synthesis as a research tool:

- the use of synthesis (by which is now meant the synthesizer and its control system) as a developmental model for a theory of speech production and a means of testing theory worked out in abstract; and
- as a means of providing stimulus items in perceptual experiments.

The use of synthetic speech as an aid to the development of perception theory is not new and many researchers have adopted this technique (see, for example: Broadbent and Ladefoged, 1960, Haggard 1970; Liberman *et al*, 1954). What any researcher must be extremely careful of is an experiment which provides information about the human perception of *synthetic* speech: that is, the stimulus items used must be identical to human speech — which may involve producing them in a (near) identical way. It is probably too early yet to expect reliable information about speech perception derived from experiments using synthetic speech; though I am not implying that nothing reliable has yet been found out about perception using this technique.

A practical use of synthetic speech (as opposed to the above two major research tool uses) is in man-machine communication. As with automatic speech recognition as a practical exercise the criteria upon which the system is based may be different from the criteria demanded by the linguist (and apt to be considered trivial by him). Perhaps the most important criterion is whether of not the system works — *how* it works may be of secondary importance. But it may well be the case with speech synthesis [as it is with automatic speech recognition (see Sparkes, 1969)] that the system will not work unless account is taken of linguistic theory: this is particularly likely to show up in synthesis when it comes to generating the prosodic features such as stress and intonation where semantic and syntactic categories may need to be incorporated in the control strategy.

Returning to the use of synthetic speech as an aid to the development of a theory of speech production, it will be necessary to take into account the various competing theories which exist in linguistics at the present time concerning speech production. The acoustic output of speech is, crudely, a fairly well understood (Fant, 1960; Flanagan, 1965) result of the interaction of an airstream mechanism and co-ordinated movement of the vocal tract. What is interesting is just how this vocal tract movement is achieved. Notice that if our synthesizer is of the TA type then only the acoustic theory is built into the synthesizer (the building of theoretical constraints into the synthesizer is a major achievement of synthesis strategy with wide implications for linguistic theory), and that if it is a VTA type then control is in terms of the variation of the cross-sectional area of the various cascaded cylinders of cylinder analogues. No synthesizer to date that I am aware of is controlled directly in terms of articulators of muscles controlling the articulators (but see Flanagan and Landgraf, 1968).

Haggard's work (1969, 1970) is ultimately in terms of articulator movement to achieve a particular acoustic output, but does not solve the current problem in the theory of the control organisation of articulator movement (Tatham, 1969). Muscle contraction for articulator movement is not a continuously updated system but a "GOTO" ballistic system (Ladefoged, 1967; Tatham, 1968a, b; Wickelgren, 1969) — this much of the theory is satisfied by the "target" and rule synthesis described above. But, as Wickelgren points out, there are several competing ways of explaining the overlap or interdependence between adjacent articulations.

This survey of speech synthesis has attempted to show that the major interest lies not in the design of the synthesizer hardware, but in its control. This control strategy may rather trivially aim at nothing but the production of lifelike synthetic speech, but to reach a level of importance and to be of interest and use to the linguist, it must take account of the facts of linguistic theory and the facts of neuro-physiological theory. It has so far really only taken account of the facts of acoustic theory.

References

Broadbent, D. E. and Ladefoged, P. (1960). Vowel judgements and adaptation level. *Proc. R. Soc. B* vol. 151.

Chomsky, N. (1967). The formal nature of language. In *Biological Foundations of Language*, Ed. E. H. Lenneberg. New York: John Wiley.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Fant, C. G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, C. G. Martony, J., Rengman, U. and Risberg, A. (1963). OVE II synthesis strategy. *Proc. Speech Commun. Seminar* 1962. Stockholm: Royal Institute of Technology.

Flanagan, J. L. (1965). *Speech Analysis, Synthesis and Perception*. Berlin: Springer Verlag.

Flanagan, J. L. and Landgraf, L. (1968). Self-oscillating source for vocal tract synthesizers. *IEEE Trans. Audio and Electroacoustics* AU-16, pp. 57-64.

Haggard, M. (1969, 1970). Various papers in Speech Synthesis and Perception. *Progress Reports* No. 1 and No. 2. University of Cambridge: Psychological Laboratory.

Haggard, M. (1970). The use of voicing information. *Speech Synthesis and Perception Progress Report* No. 2. University of Cambridge: Psychological Laboratory.

Holmes, J. N. Mattingly, I. G. and Shearme, J. N. (1964). Speech synthesis by rule. *Language and Speech* 7, 127.

Kelly, J. L. and Gerstman, L. J. (1961). An artificial talker driven from a phonetic input. *JASA* 33, 835.

Kelly, J. L. and Lochbaum, C. (1963). Speech synthesis. *Proc. Speech Commun. Seminar* 1962. Stockholm: Royal Institute of Technology.

Kim, C.-W. (1966). The linguistic specification of speech. In *Working Papers in Phonetics* No. 5. Los Angeles: University of California.

Ladefoged, P. (1967). Linguistic Phonetics. *Working Papers in Phonetics* No. 6. Los Angeles: University of California.

Lawrence, W. (1953). The synthesis of speech from signals which have a low information rate. In *Communication Theory*, Ed W. Jackson: New York and London.

Liberman, A. M., Delattre P. and Cooper, F. S. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr*. 68.

Mattingly, I. G. (1968). Synthesis by rule of General American English. *Supplement to Status Report on Speech Research*. New York: Haskins Laboratories.

McKinney, N., Tatham, M. and Ladefoged, P. (1966). Terminal analog speech synthesizer. *Working Papers in Phonetics* 4. Los Angeles: University of California.

Rosen. G. (l958). Dynamic analog speech synthesizer, *JASA* 30, 201

Sparkes, J. J. (1969). Pattern recognition and a model of the brain. *Int. J. Man-Machine Studies* 1, 263.

Tatham, M. A. A. (1968a). Classifying Allophones. *University of Essex Language Centre Occasional Papers* No. 3.

Tatham, M. A. A. (1968b). The control of muscles in speech. *University of Essex Language Centre Occasional Papers* No. 3.

Tatham, M. A. A. (1969). Experimental phonetics and phonology. *University of Essex Language Centre Occasional Papers* No. 5.

Werner, E. and Haggard, M. (1969). Articulatory synthesis by rule. *Speech Synthesis and Perception Progress Report* No. 1.

Wickelgren, W. (1969). Context-sensitive coding, associate memory, and serial order in (speech) behavior. *Psychol. Rev*. 76 No. 1, 1-15.