Review

Holmes, J. N. (1972) *Speech Synthesis*
London: Mills and Boon

Mark Tatham

Reproduced from *Journal of Phonetics* (1973) Vol. 1, pp. 369-373

The eleven Chapters of *Speech Synthesis* can be divided roughly into six main parts:

1. the first three Chapters cover phoneme theory, the acoustic theory of speech production and the acoustic analysis of speech.
2. Chapters 4 and 5 deal with a history of speech synthesis and the methods and uses of electrical synthesis (as opposed to earlier non-electrical methods).
3. Chapters 6, 7 and 8 deal with articulatory synthesis, resonance synthesis and vocoders respectively.
4. Chapter 9 dwells on a difficult aspect of any kind of synthesis — generating an appropriate excitation signal.
5. Chapter 10 sketches the notion of synthesis-by-rule.
6. and the half-page eleventh Chapter draws some conclusions.

1. In his introduction to the phoneme, Holmes finds it appropriate to distinguish two classes of phoneme: those derived as the result of acoustic or articulatory similarity between sounds, making it useful to group them into a single phoneme, and those which form a class of sounds whose distribution or function in a language is a shared property — although the sounds themselves may be dissimilar from the point of view of articulation or acoustic features. Such linguistically motivated groupings accord well with naive users' concepts of speech sounds even though the actual sounds within the class (i.e. under the same phoneme heading) may differ considerably. This heightens the problem of the complex relationship between phoneme and acoustic signal: a problem yet further enhanced by the wide variability of 'same' signals across different speakers or even within one speaker.

Holmes identifies a genuine speech synthesis system by defining that the transmission of synthesizer control signals should require less information carrying capacity than would be needed to 'specify an arbitrary audio waveform of about the same bandwidth and signal-to-noise ratio as the resultant speech signal'. This definition excludes normal radio or telephone systems, recording systems and any artificial moment-by-moment building up of a complete speech waveform, say, digitally by computer. The properties of speech must be taken into account in the design of synthesis systems.

A brief traditional description of the vocal organs follows, stressing the contributions of various articulators to the speech wave and explaining the resonance effect of the vocal tract on the source sounds produced either at the vocal cords (voicing) or elsewhere in the system (turbulence). The analysis Chapter concluding the section limits itself to acoustic analysis and to the sound spectrograph as a hardware analyzer and to analysis by calculation of the Fourier transform, the autocorrelation function and the cepstrum.

2. A very brief (two pages) history of speech synthesis covers some early non-electrical attempts to simulate speech, and the author moves now (Chapter 5) to modern electrical systems. After stating that "all electrical synthesizers incorporate methods of simulating the essential human speech functions electrically", Holmes continues with outlines of three main

areas of application for synthesis: (a) phonetic research, (b) telephone systems and (c) output devices for machine-man communication. (a) deals with a useful property of synthesizers that individual modification of each parameter is usually possible enabling various kinds of experiment to be performed; (b) deals with various special uses in telephone systems; and (c) deals with the transformation of otherwise encoded information into speech signals for transmission from machine to man. There follows a brief introduction to the three types of synthesizer in use today: articulatory and resonance synthesizers and vocoders; a short rundown of their essential differences is given. Before moving to a more detailed description of synthesizers Holmes explains the reason why computer simulation of the hardware is more convenient and reliable for research purposes than the time consuming construction of relatively inflexible special purpose synthesis systems.

3. Chapters 6 and 7 introduce in more detail articulatory (difficult, but theoretically better) and resonance (easier, but less theoretically well motivated) synthesis respectively. Resonance synthesis is given a comparatively full account with detailed remarks on higher pole correction and the relative merits of parallel and series configurations. Chapter 8 discusses various vocoder systems.

4. A relatively full treatment is given in Chapter 9 of adequate excitation signals for vocoders and articulatory and resonance synthesizers.

5. Chapter 10 deals very briefly with synthesis-by-rule, outlining first of all the history of the idea. After a short description of rules based on acoustic and articulatory parameters Holmes spends one paragraph on linguistic rules. "The *most important feature* of the recent work (on synthesis-by-rule), which is not directly concerned with the actual speech synthesis process, is the use of rules to convert from conventional spelling and punctuation to a phoneme sequence with suitable timing and pitch information" (italics mine). Synthesis-by-rule is a system for deriving control signals for the synthesizer automatically "from a phonetic specification" or from conventional spelling.

> "The fair measure of success achieved (in synthesis-by-rule) is a tribute partly to the ingenious programs that have been developed, but even more to the fact that the achievement of highly intelligible speech without any requirement to copy a particular utterance or talker's voice allows an enormous tolerance on the result. This is further helped by the redundancy due to linguistic context. Some unnatural features of such synthetic speech, such as the separate release of each stop consonant in English stop clusters, actually aid intelligibility even though they would never occur in the speech of a native English talker."

6. Finally the half page concluding Chapter ends:

> "Advances in linguistics, coupled with greater understanding of articulatory dynamics, could well lead to the production of good quality synthetic speech, with typical natural intonation and rhythm, from written text by a practical (if rather complicated) system."

For me this is one of those books about which there is either nothing to say or a great deal to say. It seems to be aimed at engineers rather than linguists, with the object of acquainting them with some of the problems associated with the design and use of synthesizers and with some of the successes achieved. To this end perhaps it was found necessary to include chapters concerning the phoneme and the acoustic theory of speech production. The latter is scant, but probably adequate if the readership is more concerned with the hardware (or its computer simulation) than with the speech end of the theory of linguistics. But those parts of the book which talk about linguistics or phonetics are inadequate enough to be misleading to the naive reader and perhaps distressing to the linguistically sophisticated.

The phoneme theory outlined in the first Chapter is a little old-fashioned and rather confused to say the least: the lack of clarity is not necessarily Holmes' fault — had he chosen a more recent model and perhaps one more appropriate for speech synthesis his task of explaining what it was all about might have been that much the simpler. The model he selects, of course, is barely descriptively adequate and stems from an analysis-only orientation that cries out for leaving it alone when it comes to synthesis. The arguments against the chosen model (however appropriate it may be for some purposes) are well rehearsed in the linguistics literature, if not in the engineering oriented synthesis literature. It is a great pity that some of the advances in linguistics (mentioned in the Conclusion) which *have already happened* could not have been included if any linguistics was to be included at all in the book.

One of the most interesting aspects of synthesis in general is just how and why such systems as articulatory and resonance synthesizers are possible *at all* and where the control signals they require should be derived. Holmes implies that the synthesizer represents a generalization of an aspect of speech production (articulatory or acoustic — depending on the type). This is a significant point and would have been worth making more of.

For the purposes of telephone engineering it is interesting that it can be assumed (and has always been assumed) that the system is required to transmit only the human voice — it is exceptional for, say, music to be transmitted on ordinary phone links. (Data transmission in some digital form, however, is a recent additional use.) For reasons of economy it has been found necessary to limit the bandwidth of telephone channels and because of the speech-only requirement the method adopted was to make a generalization about speech — namely that the majority of the *perceptually relevant* information occurs within a bandwidth of, typically, 300-4000 Hz. Any relevant (that is, linguistically significant or distinctive) information occurring outside this bandwidth turns up comparatively rarely and can usually be supplied by the perceiver (i.e. is describable by general rules of a linguistic kind).

Speech synthesizers, on the other hand, rely on generalizations for the most part not about the perception of speech but about its production. Within certain reasonably well defined limits all human beings have similar articulatory mechanisms which they use for making speech. The rules of acoustics also operate similarly for all human beings. There is a very close correlation between articulations possible by every human being and the linguistic (specifically, phonological) requirements of language — the absurdity of a language "requiring" a sound which most of its speakers cannot make is obvious. There is, however, an important aspect of human speech not subject to such broad generalization (and which is largely, therefore, phonologically irrelevant): individual speaker characteristics — those peculiarities, usually of a phonetic nature, which enable us to identify a particular speaker rather than what he or she is saying.

Now, in the transmission, as we know it, of speech over telephone circuits a large amount of the bandwidth — probably the majority — is taken up with information which simply says that this is a human being talking and that it is a particular human being (the latter itself taking less bandwidth than the former). In other words the generalizable characteristics of speech (the properties of vocal tracts and their acoustics), the generalizable characteristics of a particular individual (that he has a small/large head, a quiet/soft voice, a lisp, etc.) are continuously or repeatedly transmitted — when such repetition, or in the case of the former, transmission at all, is quite unnecessary or redundant.

The idea of synthesis is therefore to remove this "unnecessary" information and reduce the bandwidth to the essential linguistic information which is not subject to such generalization because there is no way of predicting exactly all of what a person is going to say on any one occasion. The concept is therefore simple: if *one person* is going to use the channel you simply build a synthesizer replicating the general characteristics of speech and the individual speech producing characteristics of the single person. In practice, however, the channel will probably be used by many people (as in the public network) and the individual characteristics cannot be built in — i.e. they must form a part of the transmitted information along with the linguistic information (though perhaps they need be transmitted only once and not repeatedly).

The conceptually simplest method of this type of bandwidth reduction therefore would be to create the vocal apparatus at the receiving end and transmit control signals of linguistic derivation and make it talk. Speech synthesis sets out to simulate just this situation. And in doing so it is making the theoretical claim that the generalizations I have been talking about are both accurate and possible — i.e. it is testing a theory. What is being implied is the, by now, generally accepted model of a peripheral output device with its own idiosyncrasies (the vocal apparatus) which is under independent linguistic control.

The object of fundamental research for this particular use of speech synthesis is to determine exactly what phonetic generalizations are possible for incorporation in the synthesizer and just what in normal speech is phonological and therefore requires transmission. In the future we may see sophisticated receiver systems which *can* incorporate phonological as well as phonetic rules — these will undoubtedly include those rules able to specify phonological redundancy and are likely to be those which are psychologically (i.e. perceptually) motivated. This is one area in which speech synthesis is useful as a research tool in linguistics .

Linguistic research using speech synthesis presupposes that the synthesizer itself (of whatever type) correctly and adequately captures the phonetic generalizations about the vocal apparatus and its functioning — or at least sufficiently correctly and adequately so as not to interfere with the phonology. The field is not quite so narrow, however, as implied by Holmes and there is much more to it than simply that "in the investigation of any feature of human speech, with the object of assessing its possible role in determining the phonemic pattern, dialect, stress, etc., of any utterance, it is necessary to perform experiments in which the chosen feature can be varied in a controlled way". A complex relationship exists between production and perception of speech and linguists are particularly interested in what knowledge of perception is required by the speaker in order to speak, and what knowledge of production by the perceiver in order to perceive — and, of course, how this knowledge is used.

Of central importance is surely the possibility with speech synthesis of simulation of the production process (both phonological and phonetic) in experiments designed to test the validity of production and perception models: by using simulation we can learn more about the nature of the peripheral device (the vocal apparatus), more about which of its characteristics must be, can be, and are taken into account in the formulation of the control signals necessary for its operation. We can throw light on important questions such as: in what way is the sound pattern of language (the phonology) constrained by the nature of the output device and what, if any, is the hierarchy of such constraints? What are the limits of perception of speech and to what extent do these constrain the production of speech — among others? In short, what is it that a speaker needs to know about the perception of speech and the nature of the vocal apparatus to produce speech which is acceptable and what is it that the perceiver needs to know about the nature of perception of speech, its production, etc., to be able to use the incoming signal for linguistic purposes? And finally, how do we use such knowledge, under what conditions and by what set of strategies can we ignore constraints and make judgements about what is satisfactory communication and what is not? Speech synthesis as a simulation of production and as a generator for perception research offers enormous potential as a bond between studies of both ends of the speech communication process.

Machines that talk for the sake of it might prove a gimmick detracting from the serious study of machine-man communication. The field of research called Artificial Intelligence — also open to gimmickry and self-devaluation — will clearly benefit from integration with speech synthesis research. But there is little point in not recognizing the two (equally — who knows?) worthwhile directions developments are taking. The general object of having a machine speak is to establish either a more effective communication with man for practical reasons (or, perhaps, more philosophically, to have the machine communicate with man on *his* terms rather than *its*), or as the final stage of an artificial intelligence system to study how man does the job by simulation of the process. The quick-results middle-of-the-road method, involving satisfaction in making a machine talk by whatever method, is what brings discredit. This is not to say that for practical and economic reasons developing the cheapest speaking machines — whether or not they simulate human beings — is not a good idea.

However, whether you want to find out how human beings produce speech, under what constraints they do so and why human speech is as it is, or whether you want to make a machine talk as cheaply as possible, the common ground lies in establishing the generalizations about speech which can be made — leading on the one hand to explaining human rule governed speech producing ability and behaviour and on the other to cheaper methods of generating speech by whatever method works best for practical reasons. It is probably the case that the differing research motivations will result in different answers: it will be important to remember why.

Using speech synthesis is a unifying link, therefore, for the three areas of research Holmes outlines (linguistics, telephones, machine-man communication). Linguistics here is concerned with establishing a theory of how the human being gets language beyond his brain into the outside world and why phonology takes the form it does; machine-man communication with the simulation of the human process with a view to either testing models of that process or to making machines talk because life is better if they do; and the development of telephone systems in which establishing just what need and what need not be transmitted may be crucial economically.

My conclusion is that Holmes' monograph is superficial. There is not enough detail to inform the linguist, psychologist or engineer in any meaningful way, not a competent enough overview to stimulate thought, and in general not enough discussion of the quality of speech synthesis. Clearly, though, Holmes *could* have written about these matters — one wonders why he did not. There is nothing here that is not better found (albeit inconveniently, because the papers are scattered) elsewhere.

1997 NOTE

After this review was published I learned that John Holmes had been heavily constrained by his publishers and was not himself responsible for the overall balance of the book. Twenty-five years on I am much clearer myself that the balancing act he attempted between linguistics and engineering is much more difficult than I realised. As a linguist I wanted a different kind of book, and my final paragraph criticising John for not writing it was grossly unfair.