COMPUTER SPEECH SYNTHESIS: AN OVERVIEW

L. Robert Morris and Jean-Pierre Paillet

Carleton University, Ottawa

ABSTRACT

Techniques for computer synthesis of high quality speech are reviewed. A bibliography of relevant papers is included.

ABRÉGÉ

Les techniques de la syntèse du parler de bonne qualité sont examinées. Une bibliographie des communications pertinantes est donnée.

COMPUTER SPEECH SYNTHESIS: AN OVERVIEW

L. Robert Morris and Jean-Pierre Paillet

Carleton University, Ottawa

"I CAN TELL FROM YOUR VOICE HARMONICS, DAVE, THAT YOU'RE BADLY UPSET. WHY DON'T YOU TAKE A STRESS PILL AND GET SOME REST?"

- HAL 9000 Computer, 2001: A Space Odyssey

Although efficient computer recognition of human speech may be as far off as 2001, computer speech synthesis is a reality.

Background: Mechanical production of speech-like sounds was first demonstrated in the latter part of the 18th century by Kratzenstein and von Kempelen.[1] Each constructed a machine which produced a vowel by using an air stream passing through a vibrating reed to excite an acoustic resonator whose shape was analogous to that assumed by the human vocal tract when that particular vowel is enunciated. Such anthropomorphic devices were an attempt to approximate the actual properties of the human vocal system. For example, the quasi-periodic stream of air pulses which excites the human vocal tract in voiced sound production is produced when air expelled from the lungs causes vocal cord vibration. The acoustic properties of the human vocal system are now well understood[1,2] and its behaviour can be modelled in terms of differential equations with pressure, volume velocity and acoustic impedance the relevant variables. Similar equations describe the behaviour of electrical networks in terms of voltage, current and electrical impedances. For this reason, later efforts at "mechanical" production of speech employed electrical/ electronic circuits.

Speech Analysis/Synthesis: The Speech Spectrograph, invented in 1946, produces a graphic record of the short-term amplitude spectrum of about 2 seconds of speech signal. Wide band spectrograms (see fig. 1) reveal that vowels (or voiced consonants) are characterized by 3 or 4 resonances (formants) whose centre frequencies are time varying, and that these resonances are excited quasi-periodically; narrow band spectrograms (fig. 2) clearly demonstrate that energy is concentrated at harmonics whose frequencies are integral multiples of the reciprocal of the excitation or pitch period. Thus, in cascade type electrical synthesizers, vowels and voiced consonants are produced by exciting a series of R-L-C two-pole resonators with an analog of the quasi-periodic glottal (vocal cord) waveform and loading the network with an impedance which is analogous to the acoustic radiation impedance of the mouth. Nasals require the insertion of a pole-zero pair in the transmission path while

other speech sounds may require a change in excitation as well as in network configuration. Voiceless sounds, for example, require a noise-like aperiodic excitation function.

Thus speech sounds can be synthesized by a linear electrical network whose excitation, resonances and anti-resonances are periodically updated. Fourier and cepstral analysis techniques may be utilized (with the aid of the fast Fourier transform) to analyze a particular spoken word and so calculate the network parameters necessary to synthesize a good approximation of that word.[3,4] The differential equations describing such a system can be transformed into difference equations and a sampled data simulation of the system then "calculated" on a general purpose digital computer. For example, the current output sample $y(nT)$ of a sampled-data pole-pair resonator with resonant frequency F_1 Hz and bandwidth σ_1 Hz can be computed as follows:

$$y(nT) = x(nT)C+y([n-1]T)\cdot 2\cdot e^{-2\pi\sigma_1 T} \cdot \cos 2\pi F_1 T - y([n-2]T)\, e^{-4\pi\sigma_1 T}, \qquad (1$$

where T is the sampling period of the system, $x(nT)$ is the current system input sample, and $y([n-1]T)$, $y([n-2]T)$ are the last, and second last, system output samples. C is set equal to

$$1 - 2\cdot e^{-2\pi\sigma_1 T}\cdot \cos 2\pi F_1 T + e^{-4\pi\sigma_1 T} \qquad (2$$

to give unity gain at 0 Hz and T must be less than $(2W)^{-1}$, where W is the system bandwidth. For 4 KHz speech synthesis the system throughput is 8000 samples/ sec (T=125 usec) and each sample output requires three multiplications and two additions per pole-pair. The calculated waveform is usually stored and later output via a D/A converter followed by a lowpass filter set at W Hz. Alternatively, the system parameters may be transmitted periodically to an independent hardware synthesizer (interfaced to the computer) capable of real-time synthesis.[5]

Recently, the speech analysis - synthesis problem has been formulated from a different viewpoint. Atal[6] has shown that the coefficients of a difference equation whose solution is a minimum mean-square error estimate of the speech waveform over a short time interval (e.g., a pitch period for vowels) can be obtained by solving the set of equations.

$$\sum_{k=1}^{P} \phi_{jk}\cdot a_k = \phi_{j0} \quad , j= 1, \ldots p \qquad (3$$

Where

$$\phi_{jk} = \langle s_{n-j}\cdot s_{n-k}\rangle_{av}, \text{ and } s_1, \ldots s_N$$

are the speech samples over the relevant interval. p=10 is sufficient for 4 KHz speech.

The analyzed speech can then be resynthesized (to a m.m.s.e. approximation) using

$$s_n = \sum_{k=1}^{10} a_k \cdot s_{n-k} + e_n, \quad N \geq n \geq 1, \qquad (4$$

where $\{e_n\}$ are the appropriate excitation function samples. For vowels and voiced consonants, $e_1 = G$ (a gain factor) and $e_n = 0$, $N \geq n > 1$; for unvoiced speech sounds, the $\{e_n\}$ are Gaussian noise samples. This formulation assumes that the vocal system can be adequately modelled as an all pole system or, equivalently, that any existing zeros (e.g. as in nasals) may be approximated in terms of poles. Note that this representation automatically incorporates source and radiation characteristics and that only the appropriate excitation need be chosen according as the sound is voiced (an impulse), unvoiced (Gaussian noise), or both. Morris and Paillet have demonstrated that real-time software 4 KHz speech synthesis (from stored $\{a_k\}$ coefficients) is possible using in-line assembly language programming and fixed point arithmetic on certain minicomputers.[7,8] The inordinate CPU overhead required for this purpose (e.g. 100% for the PDP-15) may be unacceptable for some applications. In this case, an economical hardware synthesizer (economical because of the simplicity of the synthesis algorithm) may be required.[9]

The main value of speech synthesis as opposed to sampled speech storage and playback is in data rate. Ten 16-bit predictor coefficients (plus a gain and segment length parameter) serve for synthesis of about 80 (or more) 4 KHz speech samples and coding methods can further increase this compression factor. Pitch can be altered easily and, conversely, the speech rate can be slowed or increased without altering pitch by repeating or skipping pitch periods. The advantage of real-time software synthesis is that messages can be synthesized without the storage requirement for non-real-time synthesis and without the expense of any extra hardware peripheral (save a D/A converter).

Synthesis by rule.

The storage of parameters required for creation of enunciated messages demands a large amount of memory if each pitch period (for voiced sounds) or equivalent time segment of every word to be spoken has to be described by a separate set of parameters. Phonetic analysis shows that very few (of the order of 50) types of sounds are used in any one language. Accordingly, it would be desirable to use as input data to the synthesis scheme a string of symbols representing the succession of sound-types which constitutes the linguistic description of the word to be realized.

What is needed then is a set of routines to generate the parameters required as input for the sound synthesis scheme from the string of sound symbols. This area of research has received the name of synthesis by rule.

Formant targets. The early attempts[10] in this direction consisted of assigning to each sound symbol a typical description in the form of a a formant pattern (for steady state sounds) or a assemblage of several pieces for more complex sounds such as stops (e.g., p,t,k).

It was then necessary to compute a continuous (or stepwise, in the case of digital systems) transition between any two successive sounds. Since all sounds do not have the same importance in actual speech, nor take up the same time, the computation of transition involves modifying the duration of each formant pattern in a context sensitive fashion. It may happen that the formant target for a particular sound is not reached before the transition to the next sound has to begin.

Recently this principle has been incorporated into an otherwise, classical terminal analog synthesizer. The result, VOTRAX[11], provides relatively inexpensive intelligible voice output for a computer. However the compromises necessary to achieve this result in a definite machine-like accent.

Articulatory model. The formant target method lacks realism, because formants are only intermediate variables. The dynamics of speech is ultimately controlled by the nervous impulses sent to the muscles of the vocal tract and by the inertia of the vocal organs. Thus, realistic synthesis by rule can only be obtained through the modeling of the human vocal tract. Orders for various articulatory movements are modelled by step functions in appropriate sequences, to which the various organs (more exactly, the geometric variables representing the movement of organs) respond each with its characteristic inertia (modelled by overdamped filters). We thus can obtain a time varying representation of the geometry of the vocal tract, from which a set of resonances is computed at short intervals, for input to the sound synthesis.

The difficulties in this approach lie mostly in obtaining appropriate values for the inertia of the vocal organs, and, especially, for the timing of the input orders. Although very crude in comparison to the actual human vocal apparatus, such a model gives evidence of the overriding importance of elaborate time patterns in the commands sent by the brain to the numerous muscles involved in speech.

Syntactic considerations. These time patterns are not dependent only on the phonetic description of individual words. The rate of production of individual words, as well as the changes in pitch (intonation) and in loudness (stress), are strongly conditioned by the syntactic, and ultimately the semantic, organisation of the message. Thus an adequate program for speech synthesis by rule, in addition to the sound synthesis routine itself and to the vocal tract model, must also include a (at least partial) syntactic analyser, which will provide information on the location of the main syntactic boundaries (clauses, nounphrases, etc.). This, together with the dictionary, will constitute the input to the purely phonetic part of the synthesis.

Dictionary. Suppose we want our computer to be able to take English text as input. It would be redundant to store all the word forms which might be found in a text. Indeed most plurals of nouns, tenses and particples of verbs, etc. are regular; that is, they are formed in a systematic fashion from the root forms. Some interesting work[12] has been done on algorithms designed to undo these forms, and retrieve roots, prefixes and suffixes for looking up in a simplified dictionary. Naturallly, such

algorithms must embody some of the principles of English spelling such as doubling of consonants, or change of final -y to -i- before a suffix.

While the last step described (from English spelling to phonetic transcription) does not result in storage space savings (on the contrary) it offers the possibility of the most natural input for the human user, namely ordinary printed text (for instance through the use of optical character readers).

References

1. J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd Edition, Springer-Verlag, New York, 1972.

2. G. Fant, Acoustic Theory of Speech Production's-Gravenhage: Mouton and Co., 1960.

3. R. Schafer and L. Rabiner, "System for automatic formant analysis of voiced speech", J.Acoust, Soc. Am., Vol. 47, pp. 634-648, February, 1970.

4. J. Markel, "Digital inverse filtering - a new tool for formant trajectory estimation", IEEE Transactions on Audio and Electroacoustics, vol. AU-20, pp. 129-137, June 1972.

5. L.R. Rabiner, L.B. Jackson, R.W. Schafer and C.H. Coker, "A hardware realization of a digital formant s ynthesizer", IEEE Transaction on Communication Technology, vol. COM-19, pp. 1016-1020, Dec. 1971.

6. B.S. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust, Soc. Am., vol. 50, pp. 637-655, August 1971.

7. L.R. Morris and J.P. Paillet, "Real-time software speech synthesis", Conference Record, 1972 International Conference on Speech Communication and Processing, pp. 166-169, Boston, April 1972.

8. L.R. Morris and J.P. Paillet, "Real-time software speech synthesis on minicomputers", Speech Research Group, Carleton University, Monograph 1. Presented at Sixth Biennial Symposium on Communication Theory and Signal Processing, Queen's University, Kingston, August 1972.

9. L.R. Morris and M.J. Gough, "An economical hardware realization of a digital linear predictive speech synthesizer". Proceedings, International Communications Conference, Seattle, June 1973.

10. J. Holmes, I. Mattingly, J. Shearme, "Speech synthesis by rule", Language and Speech Vol. 7, pt. 3, pp. 127-143, 1964.

11. Federal Screw Works, Detroit Michigan.

12. F. Lee and J. Allan, "Machine-to-man communication by speech", SJJJ, 1968.

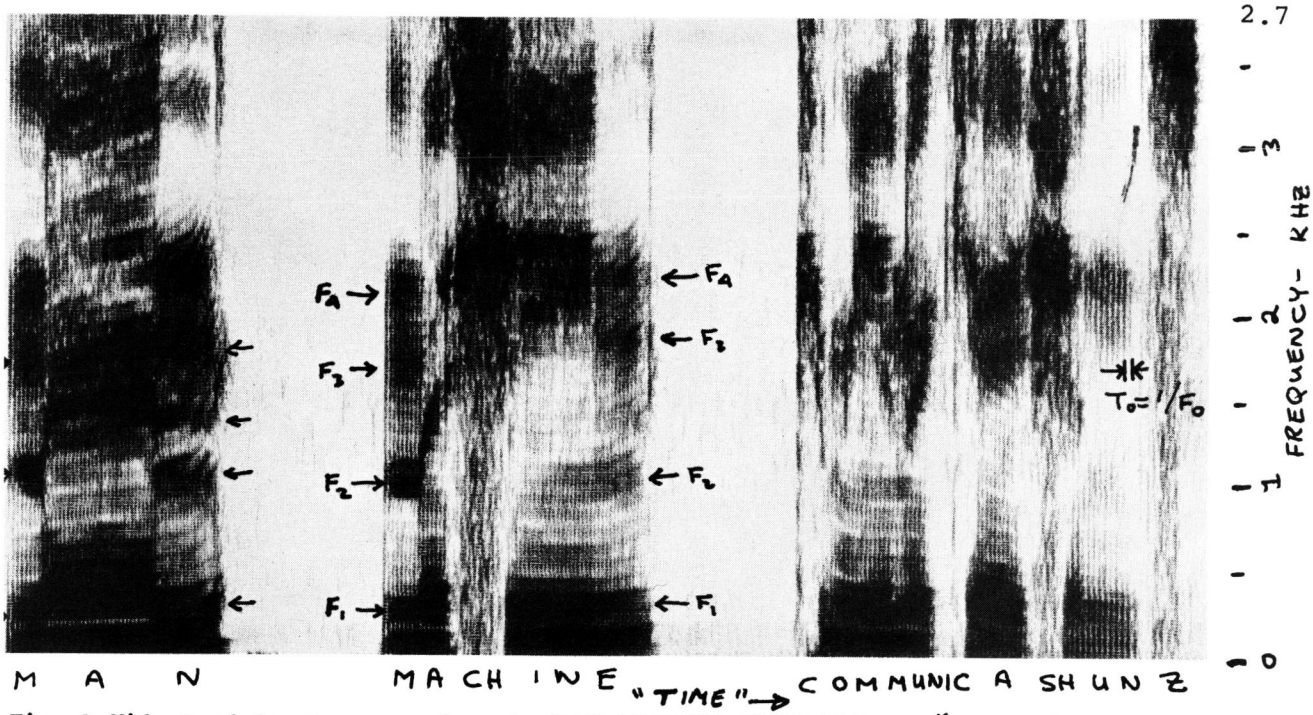

Fig. 1 Wide Band Spectrogram of words "MAN MACHINE COMMUNICATIONS" The wide horizontal bands are the trajectories of F_1, F_2, F_3, F_4, the formants or resonances of the vocal tract and are prominent in vowels, nasals and other voiced consonants. The thin vertical striations occur every T_o seconds where $T_o = 1/F_o$, F_o being the fundamental or "pitch" frequency.

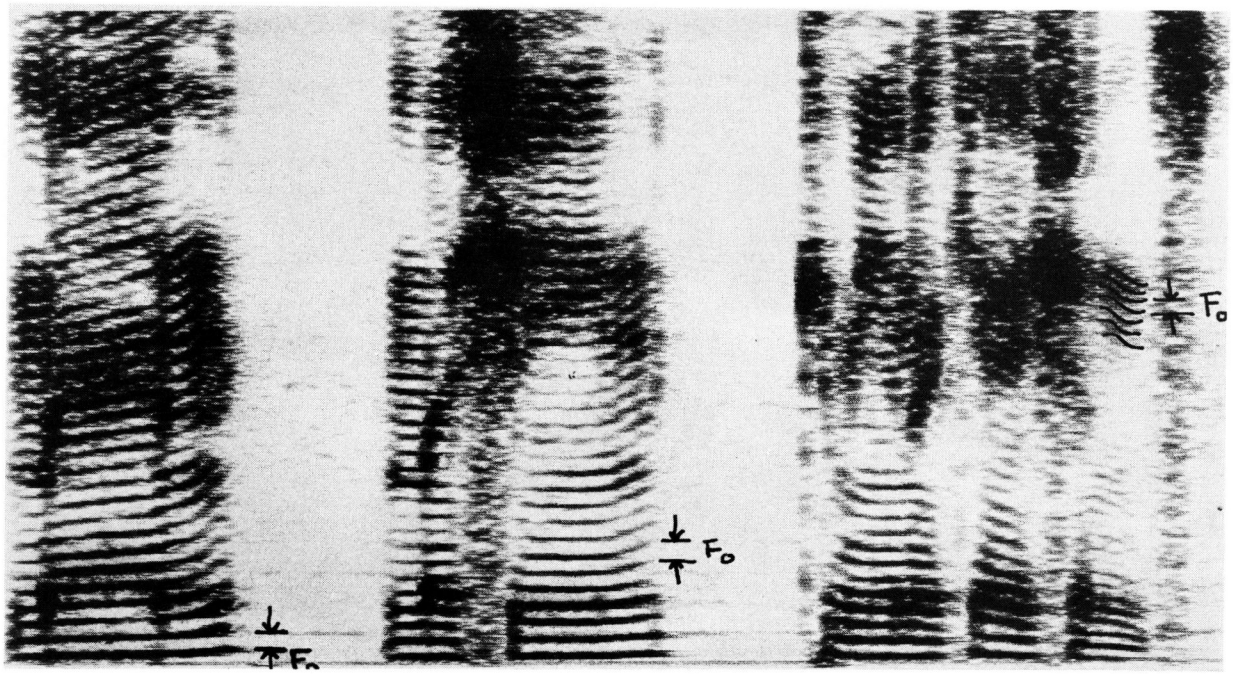

Fig. 2 Narrow Band Spectrogram of same words. The harmonics of the fundamental frequency F_o are resolved as thin horizontal lines F_o Hz apart. Every $T_o = 1/F_o$ sec the digital filter producing synthesized speech must be updated with new parameters.