COMPUTER SIMULATION OF SPONTANEOUS SPEECH PRODUCTION

Bengt Sigurd
Dept of Linguistics and Phonetics
Helgonabacken 12, S-223 62 Lund, SWEDEN

ABSTRACT

This paper pinpoints some of the problems faced when a computer text production model (COMMENTATOR) is to produce spontaneous speech, in particular the problem of chunking the utterances in order to get natural prosodic units. The paper proposes a buffer model which allows the accumulation and delay of phonetic material until a chunk of the desired size has been built up. Several phonetic studies have suggested a similar temporary storage in order to explain intonation slopes, rythmical patterns, speech errors and speech disorders. Small-scale simulations of the whole verbalization process from perception and thought to sounds, hesitation behaviour, pausing, speech errors, sound changes and speech disorders are presented.

1. Introduction

Several text production models implemented on computers are able to print grammatical sentences and coherent text (see e.g. contributions in Allén, 1983, Mann & Matthiessen, 1982). There is, however, to my knowledge no such verbal production system with spoken output, simulating spontaneous speech, except the experimental version of Commentator to be described.

The task to design a speech production system cannot be solved just by attaching a speech synthesis device to the output instead of a printer. The whole production model has to be reconsidered if the system is to produce natural sound and prosody, in particular if the system is to have some psychological reality by simulating the hesitation pauses, and speech errors so common in spontaneous speech.

This paper discusses some of the problems in the light of the computer model of verbal production presented in Sigurd (1982), Fornell (1983). For experimental purposes a simple speech synthesis device (VOTRAX) has been used.

The Problem of producing naturally sounding utterances is also met in text-to-speech systems (see e.g. Carlson & Granström, 1978). Such systems, however, take printed text as input and turn it into a phonetic representation, eventually sound. Because of the differences between spelling and sound such systems have to face special problems, e.g. to derive single sounds from the letter combinations th, ng, sh, ch in such words as the, thing, shy, change.

2. Commentator as a speech production system

The general outline of Commentator is presented in fig. 1. The input to this model is perceptual data or equivalent values, e.g. information about persons and objects on a screen. These primary perceptual facts constitute the basis for various calculations in order to derive secondary facts and draw conclusions about movements and relations such as distances, directions, right/left, over/under, front/back, closeness, goals and intentions of the persons involved etc. The Commentator produces comments consisting of grammatical sentences making up coherent and well-formed text (although often soon boring). Some typical comments on a marine scene are: THE SUB-

MARINE IS TO THE SOUTH OF THE PORT. IT IS APPROACHING THE PORT, BUT IT IS NOT CLOSE TO IT. THE DESTROYER IS APPROACHING THE PORT TOO. The original version commented on the movements of the two persons ADAM and EVE in front of a gate.

A question menu, different for different situations, suggests topics leading to propositions which are considered appropriate under the circumstances and their truth values are tested against the primary and secondary facts of the world known to the system (the simulated scene). If a proposition is found to be true, it is accepted as a protosentence and verbalized by various lexical, syntactic, referential and texual subroutines. If, e.g., the proposition CLOSE (SUBMARINE, PORT) is verified after measuring the distance between the submarine and the port, the lexical subroutines try to find out how closeness, the submarine and the port should be expressed in the language (Swedish and English printing and speaking versions have been implemented).

The referential subroutines determine whether pronouns could be used instead of proper or other nouns and textual procedures investigate whether connectives such as but, however, too, either and perhaps contrastive stress should be inserted.

Dialogue (interactive) versions of the Commentator have also been developed, but it is difficult to simulate dialogue behaviour. A person taking part in a dialogue must also master turntaking, questioning, answering, and backchannelling (indicating, listening, evaluation). Expert systems, and even operative systems, simulate dialogue behaviour, but as everyone knows, who has worked with computers, the computer dialogue often breaks down and it is poor and certainly not as smooth as human dialogue.

The Commentator can deliver words one at a time whose meaning, syntactic and textual functions are well-defined through the verbalization processes. For the printing version of Commentator these words are characterized by whatever markers are needed.

Lines	Component	Task	Result (sample)
10-35	Primary information	Get values of primary dimensions	Localization coordinates
100-140	Secondary information	Derive values of complex dimensions	Distances, right-left, under-over
152-183	Focus and topic planning expert	Determine objects in focus (referents) and topics according to menu	Choice of subject, object and instructions to test abstract predicates with these
210-232	Verification expert	Test whether the conditions for the use of the abstract predicates are met in the situation -(on the screen)	Positive or negative protosentences and instructions for how to proceed
500	Sentence structure (syntax) expert	Order the abstract sentence constituents (subject, predicate, object); basic prosody	Sentence structure with further instructions
600-800	Reference expert (subroutine)	Determine whether pronouns, proper nouns, or other expressions could be used	Pronouns, proper nouns, indefinite or definite NPs
700-	Lexical expert (dictionary)	Translate (substitute) abstract predicates, etc.	Surface phrases, words
900	Sentence connection (textual) expert	Insert conjunctions, connective adverbs; prosodic features	Sentences with words such as också (too), dock (however)
1000	Phonological (pronunciation, printing) expert	Pronounce or print the assembled structure	Uttered or printed sentence (text)

Figure 1. Components of the text production model underlying Commentator

3. A Simple speech synthesis device

The experimental system presented in this paper uses a Votrax speech synthesis unit (for a presentation see Giarcia, 1982). Although it is a very simple system designed to enable computers to deliver spoken output such as numbers, short instructions etc, it has some experimental potentials. It forces the researcher to take a stand on a number of interesting issues and make theories about speech production more concrete. The Votrax is an inexpensive and unsophisticated synthesis device and it is not our hope to achieve perfect pronunciation using this circuit, of course. The circuit, rather, provides a simple way of doing research in the field of speech production.

Votrax (which is in fact based on a circuit named SC-01 sold under several trade names) offers a choice of some 60 (American) English sounds (allophones) and 4 pitch levels. A sound must be transcribed by its numerical code and a pitch level, represented by one of the figures 0,1,2,3. The pitch figures correspond roughly to the male levels 65,90,110,130 Hz. Votrax offers no way of changing the amplitude or the duration.

Votrax is designed for (American) English and if used for other languages it will, of course, add an English flavour. It can, however, be used at least to produce intelligible words for several other languages. Of course, some sounds may be lacking, e.g. Swedish *u* and *y* and some sounds may be slightly different, as e.g. Swedish *sh-*, *ch-*, *r-*, and *l*-sounds.

Most Swedish words can be pronounced intelligibly by the Votrax. The pitch levels have been found to be sufficient for the production of the Swedish word tones: accent 1 (acute) as in *and-en* (the duck) and accent 2 (grave) as in *ande-n* (the spirit). Accent 1 can be rendered by the pitch sequence 20 and accent 2 by the sequence 22 on the stressed syllable (the beginning) of the words. Stressed syllables have to include at least one 2.

Words are transcribed in the Votrax alphabet by series of numbers for the sounds and their pitch levels. The Swedish word *höger* (right) may be given by the series 27,2,58,0,28,0,35,0, 43,0, where 27,58,28,35,43 are the sounds corresponding to h,ö:,g,e,r, respectively and the figures 2,0 etc after each sound are the pitch levels of each sound. The word *höger* sounds American because of the *ö*, which sounds like the (retroflex) vowels in *bird*.

The pronunciation (execution) of the words is handled by instructions in a computer program, which transmits the information to the sound generators and the filters simulating the human vocal apparatus.

4. Some problems to handle

4.1. Pauses and prosodic units in speech

The spoken text produced by human beings is normally divided by pauses into units of several words (prosodic units). There is no generally accepted theory explaining the location and duration of the pauses and the intonation and stress patterns in the prosodic units. Many observations have, however, been made, see e.g. Dechert & Raupach (1980).

The printing version of Commentator collects all letters and spaces into a string before they are printed. A speaking version trying to simulate at least some of the production processes cannot, of course, produce words one at a time with pauses corresponding to the word spaces, nor produce all the words of a sentence as one prosodic unit. A speaking version must be able to produce prosodic units including 3-5 words (cf Svartvik (1982)) and lasting 1-2 seconds (see Jönsson, Mandersson & Sigurd (1983)). How this should be achieved may be called the *chunking problem*. It has been noted that the chunks of spontaneous speech are generally shorter than in text read aloud.

The text chunks have internal intonation and stress patterns often described as superimposed on the words. Deriving these internal prosodic patterns may be called the *intra-chunk problem*. We may also talk about the *inter-chunk* problem having to do with the relations e.g. in pitch, between succesive chunks.

As human beings need to breathe they have to pause in order to inhale at certain intervals. The need for air is generally satisfied without conscious actions. We estimate that chunks of 1-2 seconds and inhalation pauses of about 0.5 seconds allow convenient breathing. Clearly, breathing allows great variation. Everybody has met persons who try to extend the speech chunks and minimize the pauses in order to say as much as possible, or to hold the floor.

It has also been observed that pauses often occur where there is a major syntactic break (corresponding to a deep cut in the syntactic tree), and that, except for so-called hesitation pauses, pauses rarely occur between two words which belong closely together (corresponding to a shallow cut in the syntactic tree). There is, however, no support for a simple theory that pauses are introduced between the main constituents of the sentence and that their duration is a function of the depthof the cuts in the syntactic tree. The conclusion to draw seems rather to be that chunk cuts are avoided between words which belong closely together. Syntactic structure does not govern chunking, but puts constraints on it. Click experiments which show that the click is erroneously located at major syntactic cuts rather than between words which are syntactically coherent seem to point in the same direction. As an illustration of syntactic closeness we mention the combination of a verb and a following reflexive pronoun as in Adam närmar+sig Eva. ("Adam approaches Eva"). Cutting between närmar and sig would be most unnatural.

Lexical search, syntactic and textual planning are often mentioned as the reasons for pauses, so-called hesitation pauses, filled or unfilled. In the speech production model envisaged in this paper sounds are generally stored in a buffer where they are given the proper intonational contours and stress patterns. The pronunciation is therefore generally delayed. Hesitation pauses seem, however, to be direct (on-line) reflexes of searching or planning processes and at such moments there is no delay. Whatever has been accumulated in the articulation or execution buffer is pronounced and the system is waiting for the next word. While waiting (idling), some human beings are silent, others prolong the last sounds of the previous word or produce sounds, such as ah, eh, or repeat part of the previous utterence. (This can also be simulated by Commentator.) Hesitation pauses may occur anywhere, but they seem to be more frequent before lexical words than function words.

By using buffers chunking may be made according to various principles. If a sentence termination (full stop) is entered in the execution buffer, whatever has been accumulated in the buffer may be pronounced setting the pitch of the final part at low. If the number of segments in the chunk being accumulated in the buffer does not exceed a certain limit a new word is only stored after the others in the execution buffer. The duration of a sound in Votrax is 0.1 second on the average. If the limit is set at 15 the system will deliver chunks about 1.5 seconds, which is a common length of speech chunks. The system may also accumulate words in such a way that each chunk normally includes at least one stressed word, or one syntactic constituent (if these features are marked in the representation). The system may be made to avoid cutting where there is a tight syntactic link, as e.g. between a head word and enclitic morphemes. The length of the chunk can be varied in order to simulate different speech styles, individuals or speech disorders.

4.2. Prosodic patterns within utterance chunks

A system producing spontaneous speech must give the proper prosodic patterns to all the chunks the text has been divided into. Except for a few studies, e.g. Svartvik (1982) most prosodic studies concern well-formed grammatical sentences pronounced in isolation. While waiting for further information and more sophisticated synthesis devices it is interesting to do experiments to find out how natural the result is.

Only pitch, not intensity, is available in Votrax, but pitch may be used to signal stress too. Unstressed words may be assigned pitch level 1 or 0, stressed words 2 or higher on at least one segment. Words may be assumed to be inherently stressed or unstressed. In the restricted Swedish vocabulary of Commentator the following illustrate lexically stressed words: Adam, vänster (left), nära (close), också (too). The following words are lexically unstressed in the experiments: han (he), den (it), i (in), och (and), men (but), är (is). Inherently unstressed words may become stressed, e.g. by contrast assigned during the verbalization process.

The final sounds of prosodic units are often prolonged, a fact which can be simulated by doubling some chunk-final sounds, but the

Votrax is not sophisticated enough to handle these phonetic subtleties. Nor can it take into account the fact that the duration of sounds seem to vary with the length of the speech chunk.

The rising pitch observed in chunks which are not sentence final (signalling incompleteness) can be implemented by raising the pitch of the final sounds of such chunks. It has also been observed that words (syllables) within a prosodic unit seem to be placed on a slope of intonation (grid). The decrement to the pitch of each sound caused by such a slope can be calculated knowing the place of the sound and the length of the chunk. But so far, the resulting prosody, as is the case of text-to-speech systems, cannot be said to be natural.

4.3. Speech errors and sound change

Speech errors may be classed as lexical, grammatical or phonetic. Some lexical errors can be explained (and simulated) as mistakes in picking up a lexical item. Instead of picking up höger (right) the word vänster (left), a semi-antonym, stored on an adjacent address, is sent to the buffer. Grammatical mistakes may be simulated by mixing up the contents of memories storing the constituents during the process of verbalization.

Phonetic errors can be explaned (and simulated) if we assume buffers where the phonetic material is stored and mistakes in handling these buffers. The representation in Votrax is not, however, sophisticated enough for this purpose as sound features and syllable constituents often must be specified. If a person says pöger om porten instead of höger om porten (to the right of the gate) he has picked up the initial consonantal element of the following stressed syllable too early.

Most explanations of speech errors assume an unconscious or a conscious monitoring of the contents of the buffers used during the speech production process. This monitoring (which in some ways can be simulated by computer) may result in changes in order to adjust the contents of the buffers, e.g. to a certain norm or a fashion. Similar monitoring is seen in word processing systems which apply automatic spelling correction. But there are several places in Commentator where sound changes may be simulated.

REFERENCES

Allén, S. (ed) 1983. Text processing. Nobel symposium. Stockholm: Almqvist & Wiksell

Carlson, R. & B. Granström. 1978. Experimental text-to-speech system for the handicapped. JASA 64, p 163

Ciarcia, S. 1982. Build the Microvox Text-to-speech synthesizer. Byte 1982:Oct

Dechert, H.W. & M. Raupach (eds) 1980. Temporal variables in speech. The Hague: Mouton

Fornell, J. 1983. Commentator, ett mikrodator-baserat forskningsredskap för lingvister. Praktisk Lingvistik 8

Jönsson, K-G, B. Mandersson & B. Sigurd. 1983. A microcomputer pausemeter for linguists. In: Working Papers 24. Lund. Department of linguistics

Mann, W.C. & C. Matthiessen. 1982. Nigel: a systemic grammar for text generation. Information sciences institute. USC. Marina del Ray. ISI/RR-83-105

Sigurd, B. 1982. Text representation in a text production model. In: Allén (1982)

Sigurd, B. 1983. Commentator: A computer model of verbal production. Linguistics 20-9/10 (to appear)

Svartvik, J. 1982. The segmentation of impromptu speech. In Enkvist, N-E (ed). Impromptu speech: Symposium. Åbo: Åbo akademi