A DEMISYLLABLE-BASED TEXT-TO-SPEECH SYNTHESIS SYSTEM FOR ENGLISH

S.J. Eady, P. Ollek and J.R. Woolsey

Speech Technology Research Ltd., Suite D, 1623 McKenzie Avenue, Victoria, B.C. V8N 1A6, Canada

Introduction

Synthesis of English speech by computer can be accomplished in several different ways, depending on the size of the speech units that are used to produce voice output. The most widely used units for speech synthesis are phonemes (i.e., small speech units corresponding to individual phonetic items) [1]. An alternate method of producing computer-generated speech is to concatenate entire words of English in a method called "word-concatenation" synthesis [2]. A third strategy, the one described in this paper, is to use intermediate-sized units corresponding to half syllables, called "demisyllables" [3].

Demisyllables as Units of Synthesis

In demisyllable synthesis, each syllable of a word is composed of an initial demisyllable, which comprises the initial consonant and the first part of the following vowel, plus a final demisyllable, which includes the remaining portion of the vowel and any following consonants. The examples below illustrate this point:

SYLLABLE	INITIAL DEMISYLLABLE	FINAL DEMISYLLABLE
"bet"	BE	ET
"set"	SE	ET
"quench"	KWE	ENCH

Since all words of English are composed of syllables, and all syllables can be created from demisyllables, then it follows that this method can be used to produce any English word. This paper describes the various components that have been developed for microcomputer-based speech synthesis using demisyllables.

Hardware Requirements

The Demisyllable Synthesis System is designed for use on an IBM AT or compatible with a minimum of 640K of RAM. In addition, it also requires a TMS-320C25 DSP chip and a digital-to-analog converter. In its current configuration, the system will run on the Loughborough TMS320C25 Development Board or on the Kay Elemetrics CSL hardware.

Demisyllable Inventory

The inventory of demisyllable speech units consists of approximately 1400 prerecorded items that were produced in monosyllabic words by a male speaker of English. The recorded demisyllables were then digitized and encoded using pitch-synchronous LPC (10-pole, covariance method) [4].

Each encoded demisyllable unit consists of a number of 10-msec speech frames, and each frame contains quantized values for energy, pitch and 10 LPC reflection coefficients. Quantization of these values results in a storage requirement of 14 bytes per frame, and a corresponding transmission rate of 1400 bytes per second. The entire demisyllable inventory requires about 450 kilobytes of storage.

Text-to-Demisyllable Conversion

Voice output from the demisyllable synthesis system is initiated through a text-input module that accepts standard English orthographic text, as well as punctuation marks and symbols that commonly occur in written text (e.g., $, %, @, etc.). When text is input to the system, each word or symbol is first translated into its constituent demisyllable units. In addition, for each word, the system determines the stress pattern and the part of speech. All of this processing is done using a set of rules that have been developed for this purpose [5].

Demisyllable-to-Speech Rules

When an English sentence is entered into the text-input module described above, its constituent words are automatically translated into demisyllable units, and the designated units are then retrieved from the demisyllable inventory files.

Demisyllables are then transformed into complete sentences of English by means of a set of rules that are summarized below and described in greater detail elsewhere [6]. The rules are applied in the order given. The general strategy is to work from the smallest units (i.e., demisyllables) to progressively larger, more complex units (i.e., syllables, words and sentences).

Syllable Creation

The first step in the conversion from demisyllables to sentences is the creation of syllables. Each syllable is created by concatenating an initial and a final demisyllable from the demisyllable inventory. Since all initial demisyllables end in a vowel and all final demisyllables begin with a vowel, this concatenation is achieved quite simply by joining the two vocalic segments together and performing a spectral smoothing across the boundary between them.

Word Creation

Words are produced from the newly-created syllables by means of three different steps, which are described as follows:

1. Syllable Linking: A set of syllable-linking rules is used to modify phonetic segments at syllable boundaries within a word. These rules are formulated in terms of ten phonetic classes (i.e., voiced and voiceless stops, affricates and fricatives, as well as nasals, liquids, semivowels and vowels). Depending on the phonetic classes involved, the syllable-linking rules may act to delete certain speech frames, to smooth the energy contour at the boundary or to perform a spectral smoothing (i.e., smoothing of LPC reflection coefficients) at the syllable boundary.

2. Adjustment of Syllable Durations: The second stage in word creation is the adjustment in the length of each syllable in a word. This duration adjustment is required so that the syllables will have lengths that are appropriate for the stress pattern of the word in question.

3. Word-Level Pitch Assignment: The final step in the creation of words from demisyllables, is the assignment of appropriate pitch contours. As with duration adjustments,, the pitch contour of an English word is determined primarily by the stress pattern of its constituent syllables (see [6] for further details).

Sentence Creation

After words have been created from demisyllable units, the next task is to produce complete sentences from these words. This process involves three different steps.

1. Word Concatenation: The first step is to join together the word units that have been created by the components described above. When the words are concatenated, a set of word-linking rules is applied. These rules are very similar to the syllable-linking rules described above, in that they act to modify phonetic segments at syllable boundaries. In this case, however, the syllables in question are at word boundaries.

2. Sentence-Level Pitch Contour: This component is designed to provide an appropriate intonation pattern for each synthesized sentence. The method used here is very similar to that previously developed for a word-concatenation synthesis system (see [2] for details). In short, it works by overlaying a sentence-level pitch contour on top of the word-level pitch contours that are produced during the word-creation stage. The pitch level of each word is adjusted, depending on its function in the sentence. In addition, certain "tonic" pitch contours are applied at the end of each sentence to differentiate statements (which end in a falling pitch) from questions (which have rising terminal pitch contours). A third tonic contour, called a continuation rise, is also available, and may be used in the middle of a sentence at major clause boundaries.

3. Sentence-Level Timing Adjustment: The final step in sentence creation is the adjustment of word durations at different locations in a sentence. This primarily involves an increase in duration on the final word of a sentence or on any word within a sentence that occurs before a pause.

This "pre-pausal" lengthening is accomplished by adjusting the frame size of the demisyllable items that constitute the word in question. As noted above, the default frame size is 10 msec. By increasing this value to 15 msec, we can effect a 50% increase in the duration of a word or syllable. Frame-size adjustment of this magnitude is used to produce a duration increase for words that occur before a pause.

Summary

The microcomputer-based system described here has been developed to generate English speech from unlimited text input. The system uses prerecorded demisyllables as units of synthesis. With an inventory of approximately 1400 demisyllables, it can generate all possible syllables and words of the English language. By combining these units to form continuous speech, the system can produce any English sentence.

Acknowledgement

This work was supported by the Science Council of British Columbia, by NRC Canada and by NSERC of Canada.

References

[1] Klatt, D.H. (1987). "Review of text-to-speech conversion for English," Journal of the Acoustical Society of America, vol. 82, pp. 737-793.

[2] Eady, S.J., Dickson, B.C., Urbanczyk, S.C., Clayards, J.A.W. and Wynrib, A.G. (1987). "Pitch assignment rules for speech synthesis by word concatenation," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp. 1473-1476.

[3] Lovins, J.B., Macchi, M.J. and Fujimura, O. (1979). "A demisyllable inventory for speech synthesis," Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), pp. 519-522.

[4] Hunt, M.J. and Harvenberg, C.E. (1986). "Generation of controlled speech stimuli by pitch-synchronous LPC analysis of natural utterances," Proceedings of the 12th International Congress on Acoustics, paper A4-2.

[5] Hemphill, T. and Ollek, P. (1990). "Text-to-demisyllable conversion in the STR text-to-speech synthesis system,," STR Technical Report No. SS9001.

[6] Eady, S.J., Hemphill, T., Woolsey, J. and Clayards, J. (1989). "Development of a demisyllable-based speech synthesis system," Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 463-466.