DEVELOPMENT OF A DEMISYLLABLE-BASED SPEECH SYNTHESIS SYSTEM

S.J. Eady, T.M.S Hemphill, J.R. Woolsey and J.A.W. Clayards

Centre for Speech Technology Research
University of Victoria

This paper describes the development of a microcomputer-based voice output system for English that uses prerecorded demisyllables as units of synthesis. With an inventory of approximately 950 demisyllables, the system is capable of producing all possible syllables and words of English. By combining these units to form continuous speech, the system can produce any English sentence.

1. 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). An alternate method of producing computer-generated speech is to concatenate entire words of English in a method called "word-concatenation" synthesis. A third strategy, the one described in this paper, is to use intermediate-sized units corresponding to half syllables, called "demisyllables".

1.1 Demisyllables as Units of Synthesis

The rationale for the use of demisyllables as a unit for speech synthesis is that they strike a balance between the relatively high quality, but inflexible and memory-intensive nature of word-level synthesis, and the low memory load, but rule-intensive nature of phoneme-level synthesis. In fact, demisyllables maintain most of the positive attributes of both systems, with few of their weaknesses. Demisyllables are flexible in that the present selection is sufficient to produce any English word or sentence. The memory load is relatively low, with only 950 demisyllables (250 Kbytes of storage) necessary to produce that flexibility. Furthermore, the fact that demisyllables include all consonant vowel transitions and most consonant clusters in the data, ensures accurate coarticulatory effects and hence more natural sounding speech synthesis, all with a minimum of rules.

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:

Table 1.

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.

2. HARDWARE REQUIREMENTS

The Demisyllable Synthesis System is designed for use on an IBM XT/AT or compatible with a minimum of 512K of RAM. In addition, a TMS-32020 development board (with digital-to-analog converter) must be mounted in the host computer. Speech output from this board may be filtered and amplified before being passed to an audio speaker.

3. DEMISYLLABLE INVENTORY

The inventory of demisyllable speech units consists of approximately 950 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).

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 250 Kbytes of storage.

4. TEXT INPUT

Voice output from the demisyllable synthesis system is initiated through a text-input module that accepts English orthographic text, as well as special diacritics that are used to indicate sentence-level intonation patterns and the location of any pauses in a sentence. An example of the text-input module is displayed in Figure 1.

```
THIS + IS THE MAIN +  MENU.//////

TOUCH THE SQUARE, NEXT TO THEE ITEM YOU WANT.//////

THE TOP ITEM,-\ IS AIRPORT - INFORMATION.//////

THE SECOND ITEM,-\ IS GROUND TRANSPORTATION -.//////

THIRD ITEM,-// ASSOCIATIONS - FOR THE DISABLED.//////

LAST ITEM,-// HOTEL INFORMATION -.
```

FIGURE 1: An example of input text for the demisyllable synthesis system. Standard English orthography is augmented with diacritics to indicate pauses (/ \), emphasized and deemphasized words (+ -), continuation rise (,) and statement intonation patterns (.).

5. TEXT-TO-DEMISYLLABLE CONVERSION

At the present time, the conversion of English text into a demisyllable representation is accomplished by means of a "Lexicon" file, which is created and modified by the user. This file contains a separate entry for every English word that is to be synthesized in a particular application. An example of a Lexicon file is displayed in Figure 2.

```
AIRPORT         E ER 40 @ PA ORT 30  #

INFORMATION     I IN FEO EORM ME EI 35 @ SHOE OEN 30 #

ITEM            A AIT 10 @ TE OEM #

MAIN            ME EIN 0 @#

MENU            ME EN @ NYU UU #
```

FIGURE 2: An example of a Lexicon file, containing word items for the demisyllable synthesis system. Each word item contains its English orthography (on the left) and a demisyllable translation (on the right). The demisyllable translation consists of an initial and final demisyllable for each syllable, an optional duration reduction value for each syllable (30,40,etc.), an indication of the syllable that bears primary stress (@), and a symbol to indicate the end of a lexicon item (#).

As this figure illustrates, each entry of the Lexicon consists of the English orthography for a word (on the left), plus a transcription of the word in demisyllable notation (on the right). The demisyllable notation includes a listing of the initial and final demisyllable items for each syllable, an indication of which syllable of the word carries primary stress (designated by an "@" symbol following the primary-stressed syllable), and an optional number listed after each syllable which determines syllable duration (see below).

The Lexicon is created by an interactive program that allows the user to create lexicon items, listen to their pronunciation and modify them as necessary. This provides a flexible method for specifying the pronunciation of English words.

The Lexicon approach, while providing great flexibility, does impose limitations on voice output. That is, before a word can be synthesized, it must occur in the Lexicon. In order to overcome this limitation, an algorithm is being developed that will automatically convert English text into demisyllable notation. This component will allow for unlimited text-to-speech capability, and it will be implemented in the near future.

6. 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 by means of the Lexicon. The designated demisyllable 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. 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).

6.1 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. Spectral smoothing is done by an algorithm that calculates a weighted average of LPC reflection coefficients for 5 frames (i.e., 50 msec) on both sides of the boundary between the initial and final demisyllable items.

6.2 Word Creation

Words are produced from the newly-created syllables by means of three different steps (i.e., Syllable Linking, Syllable Duration Adjustment, and Word-Level Pitch Assignment). The first step is designed to ensure that coarticulation effects at syllable boundaries are adequately modelled. The last two steps are intended to produce appropriate prosodic features to account for the different syllable-stress patterns of English words.

6.3 Syllable Linking

The syllable-linking rules are 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). Each item in the demisyllable inventory is coded with respect to one of these ten classes.

The particular rule that will apply at a given syllable boundary depends on the phonetic items that are present at that boundary. 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.

6.4 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.

Syllable stress is an important component of English words. It can be illustrated by the difference in the noun "SUBject" versus the verb "subJECT". The noun has primary stress on the first syllable, while the verb has it on the second. The difference in stress is realized acoustically by differences in syllable duration and also pitch contour (see below for details about pitch contour). In particular, a syllable that has primary stress will be longer than when it is unstressed. Thus, modification of syllable durations is an important component in word creation.

The strategy for modifying syllable durations has been to record the original demisyllable speech items with relatively long durations (i.e., longer than required for most primary-stressed syllables), and then to shorten them when required. Shortening of syllables is achieved by an algorithm that selectively deletes up to 66 percent of the voiced frames from each syllable (see Urbanczyck, S.C. et al for a more detailed description).

The amount of duration reduction that is applied to each syllable of a word is thus expressed as a percentage of the number of voiced frames in the syllable. The percent reduction in syllable duration is determined automatically, depending on the stress pattern and the number of syllables in a word. Primary-stressed syllables typically have reduction values of 20-25%, while unstressed syllables are reduced by 40-50% in duration.

6.5 Word-Level Pitch Assignment

The final step in the creation of words from demisyllables, is the assignment of appropriate pitch contours. As indicated above, the pitch contour of an English word is determined primarily by the stress pattern of its constituent syllables. That is, in general, a syllable with primary stress will have a higher pitch value than unstressed syllables. An algorithm has been developed that assigns pitch contours so that the highest pitch of a word is always on the primary-stressed syllable (i.e., the first syllable of the noun "SUBject", but the second syllable of the verb "subJECT"). Unstressed syllables are assigned lower pitch values.

7. 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.

7.1 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.

The particular rules that apply at a given word boundary depend on the phonetic items that are present. The rules are formulated in terms of the ten phonetic classes described above, and they operate to delete certain speech frames, to smooth the energy contour or to perform spectral smoothing at the word boundary.

7.2 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. 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.

7.3 Sentence-Level Timing Adjustments

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.

8. SUMMARY

The speech synthesis system described here can be used to produce computer-generated voice output for English. At present, this output is produced with the aid of a separate user-specified Lexicon file that determines the pronunciation of each word. We are currently proceeding with development of a text-to-demisyllable component that will eliminate the need for the Lexicon and will allow unlimited text-to-speech capability. Work is also underway to create a new demisyllable inventory produced using a female speaker. This will enable us to produce synthesized voice output with either a male or a female voice.

NOTES

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

REFERENCES

Eady, S. J. and Dickson, B. C. (1987). Speech synthesis of sentence focus in English declaratives. Proceedings of the 11th International Congress of Phonetic Sciences, 3, 262-265.

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, 3, 1473-1476.

Hemphill, T. (1988). The DEMI320 Program: An interim report on development of a demisyllable-based speech synthesis system. CSTR Technical Report No. SS8802.

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.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, **82**, 737-793.

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.), 519-522.

Urbanczyk, S. C., Eady, S. J. and Dickson, B. C. (1987). Durational adjustment in a demisyllable-based speech synthesis system. Proceedings of the Canadian Acoustical Association, 79-84.