SPEECH SYNTHESIS OF SENTENCE FOCUS IN ENGLISH DECLARATIVES

S.J. EADY and B.C. DICKSON

Centre for Speech Technology Research, University of Victoria,
P.O. Box 1700, Victoria, B. C., V8W 2Y2, Canada

ABSTRACT

This paper describes an algorithm that has been developed to synthesize sentences containing focused words in varying locations. Synthetic speech is generated by concatenating LPC-encoded words and phrases of English. The method involves the assignment of one of three accent levels to each syllable of all encoded vocabulary items. The highest accent marking for a given word depends on its grammatical category. The accent assignments are then used to determine the relative pitch for each word in a synthesized sentence. A method is described for using this system of pitch assignment to synthesize sentence focus in English declaratives. The pitch assignment algorithm is currently being used in applications requiring automated speech output.

INTRODUCTION

The use of synthesized speech for applications such as computer-assisted language instruction [1] and automated information delivery [2] requires a synthesis system that is capable of generating utterances in which the location of sentence focus may vary. For example, the English sentence "I LIKE TO PLAY FOOTBALL" would be produced in a different way for each of the following contexts:

1. What do you like to play?

2. Do you like to play football or watch football?

3. Do you like or dislike playing football?

For each of these questions, a different word would be focused in the declarative response. The ability to convey this difference in the location of focus is an important aspect of a speech synthesis system.

Previous work on the production of sentence focus in English [3-5] shows that the focusing of a word results in changes to the durations and pitch patterns of the sentence components. In this paper, we describe how the acoustical patterns of sentence focus are generated using a word-concatenation synthesis method [6,7].

SPEECH SYNTHESIS METHOD

Synthesized speech is generated on a microcomputer using a Texas Instruments TMS-5220C speech synthesis chip. A control program is used to provide the synthesis chip with a series of quantized values for pitch, energy and ten LPC reflection coefficients. These parameter values are stored as preprocessed vocabulary items corresponding to individual words or short phrases. English sentences are synthesized by concatenating these preprocessed vocabulary items in a specified order and then applying rules to produce appropriate pitch patterns and to eliminate spectral discontinuities at word boundaries.

Vocabulary Production

Vocabulary items required for a particular application are first embedded in carrier sentences and read by a male speaker whose voice is recorded on a digital PCM recorder. Each item is then digitized (at a 10-kHz sampling rate with 12-bit resolution) and excised from its sentence environment. The digitized vocabulary items are then analyzed using the autocorrelation method of LPC [8] to derive values of energy, pitch and 10 LPC reflection coefficients at 20-msec intervals. These parameters are quantized for output on the synthesis chip.

Each encoded vocabulary item is then edited to eliminate any spectral discontinuities, and to provide a uniform energy maximum and a neutral pitch contour. During the editing process, every syllable is assigned one of three levels of "accent". An accent marker is assigned to the "nuclear" frame of each syllable (usually the frame containing the highest energy in the syllable). In general, the primary accent level is assigned to the highest stressed syllable of nouns, adverbs and adjectives, whereas verbs are marked with secondary accent, and the third level of accent is assigned to function words, such as conjunctions, prepositions and modal verbs.

Word Concatenation

At the time of synthesis, encoded vocabulary items are concatenated to form complete sentences of English. The input to the system is standard English spelling augmented by diacritics to specify sentence type (i.e., statement or question), the identity of any focused words, and the location and duration of any pauses within a sentence. The system verifies the existence of each word in the list of encoded vocabulary items, and the requested items are joined together in the order specified. At this point, the encoded vocabulary items are modified by a number of phonetic liaison rules (described elsewhere [7]) and by a set of rules for pitch assignment.

Pitch Assignment

The pitch assignment rules utilize the accent levels assigned to each vocabulary item to adjust the pitch level of each word in a sentence. For the purposes of the pitch assignment rules, the sentence is divided into two major components, called the "head" and the "tonic" (following Halliday [9] and Young and Fallside [10]). The head comprises all syllables from the start of the sentence up to and including the penultimate primary-accented syllable. The tonic is made up of the final primary-accented syllable and any other following syllables in the sentence.

Head. Within the head, the pitch rules utilize the accent level assigned to each vocabulary item to determine the highest pitcl with respect to a predetermined fundamental frequency (FO) topline, midline and baseline. Each of these lines has a gradual FO declination over the duration of the sentence head. These three lines are used to determine the relative pitch level of each word in a synthesized sentence. Vocabulary items containing a primary accent have their maximum pitch on the topline; those with secondary accent are set to the midline; and words with only tertiary-accented syllables fall to the baseline. This method maintains the overall shape of the pitch contour that was given to each vocabulary item during the original recording procedure, while at the same time modifying the relative pitch level for each word with respect to other vocabulary items in the sentence.

Tonic. A special series of rules applies to the last primary-accented syllable in a sentence. This syllable, known as the tonic, is set apart from the sentence head and is not bound by the limits of the topline and baseline. Instead, it is given a distinctive pitch contour depending on the sentence type. If the tonic occurs at the end of a declarative sentence, it is given a falling pitch contour. For sentences requiring an interrogative intonation pattern (i.e., yes-no questions), the tonic is given a rising pitch contour.

The difference between the tonic pitch contours for declarative and interrogative sentences is shown in Figure 1. The pitch display at the bottom of the figure illustrates the difference between the declarative tonic (on the words PREFER and SOCCER) and the interrogative tonic (on the word FOOTBALL).

Synthesis of Sentence Focus

The strategy for synthesizing sentence focus is based on recent studies [3-5] showing that focused words in declarative English sentences are characterized primarily by two factors. These are an increase in the duration of the focused item and a relatively low, flattened pitch contour for that part of the sentence following the focused word. In attempting to synthesize sentence focus, we have found that the second factor (i.e., the lowering of the post-focus pitch contour) is more important than the first for focus perception.

The method that we have used to synthesize sentence focus utilizes the accent levels that are assigned to the words of an utterance, as described above. The strategy is to assign the focused word an accent level of 1, and to reduce any primary-accented syllables that follow the focused word to a level of 2. The result is to shift the tonic syllable of the sentence so that it occurs within the focused word. Consequently, the focused item coincides with a local peak in the pitch contour and is followed by a relatively low flattened pitch for any words that follow the focus. The effect of this algorithm on a sentential pitch contour is illustrated in Figure 2. This figure shows how the pitch contour for a sentence is modified when the focus is shifted from the end to the middle or to the beginning of the utterance.

Informal listening tests indicate that this strategy of modifying only the pitch contour of a sentence is successful in eliciting the required focus perception. In some cases, it is also useful to provide an increase in the duration of the focused word, and this option has been incorporated into the speech synthesis system. In general, however, the modification of the sentential pitch contour is usually sufficient to produce the desired effect.

CONCLUSION

The method described here for synthesizing sentence focus in English declaratives has been developed for use in a word-concatenation synthesis system. In this system, the pitch contour for a synthesized sentence is determined by preassigned accent levels associated with each word. The location of focus within a sentence is overtly specified at the time of input. The sentence focus algorithm exploits the presence of the preassigned accent levels and modifies the sentential pitch contour to produce the desired focus effect.

The word-concatenation synthesis system is presently being used in applications requiring automated speech output. The pitch assignment rules described here are also being incorporated into a demisyllable-based text-to-speech synthesis system that is currently under development.

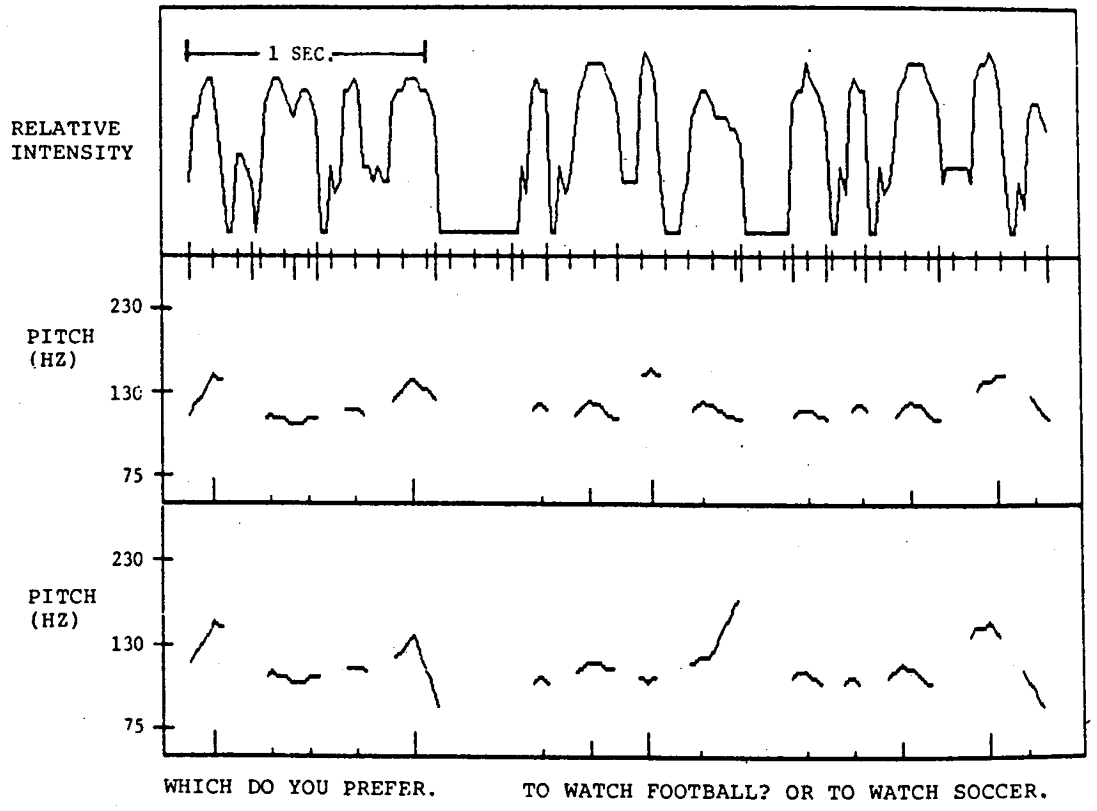

FIGURE 1: Display of relative intensity and pitch contours for a sentence synthesized by the word-concatenation method. The uppermost pitch display shows the pitch contours for the concatenated vocabulary items prior to the application of the pitch assignment rules. The bottom pitch display shows the modifications produced by the pitch assignment rules. Note the vertical markers located along the bottom of each pitch display, which indicate the accent level that has been assigned to each syllable. The tallest markers denote syllables with primary accent; markers of intermediate length are for secondary accent; and the shortest markers designate syllables with tertiary accent.

ACKNOWLEDGEMENT

This work was funded in part by the Science Council of British Columbia and by the Natural Sciences and Engineering Research Council of Canada. We thank Allan Wynrib, Jocelyn Clayards and Sue Urbanczyk for technical assistance.

REFERENCES

[1] Eady, S.J., Dickson, B.C. and Walraven, J. (1987). "Use of speech synthesis for language instruction on a microcomputer," Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, in press.

[2] Dickson, B.C., Eady, S.J., Clayards, J.A.W., Urbanczyk, S.C. and Wynrib, A.G. (1987). "Use of speech synthesis in an information system for handicapped travellers," Proceedings of the 11th International Congress of Phonetic Sciences, in press.

[3] Cooper, W.E., Eady, S.J. and Mueller, P.R. (1985). "Acoustical aspects of contrastive stress in question-answer contexts," J. Acoust. Soc. America, vol. 77, pp. 2142-2156.

[4] Eady, S.J. and Cooper, W.E. (1986). "Speech intonation and focus location in matched statements and questions," J. Acoust. Soc. America, vol. 80, pp. 402-415.

[5] Eady, S.J., Cooper, W.E., Klouda, G.V., Mueller, P.R. and Lotts, D.W. (1986). "Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments," Language and Speech, in press.

[6] Olive, J.P. and Nakatani, L.H. (1974). "Rule-synthesis of speech by word concatenation: A first step," J. Acoust. Soc. America, vol. 55, pp. 660-666.

[7] 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, in press.

[8] Markel, J.D. and Gray, A.H. (1976). Linear Prediction of Speech (New York).

[9] Halliday, M.A.K. (1967). Intonation and Grammar in British English (Paris).

[10] Young, S.J. and Fallside, F. (1980). "Synthesis by rule of prosodic features in word concatenation synthesis," International Journal of Man-Machine Studies, vol. 12, pp. 241-258.

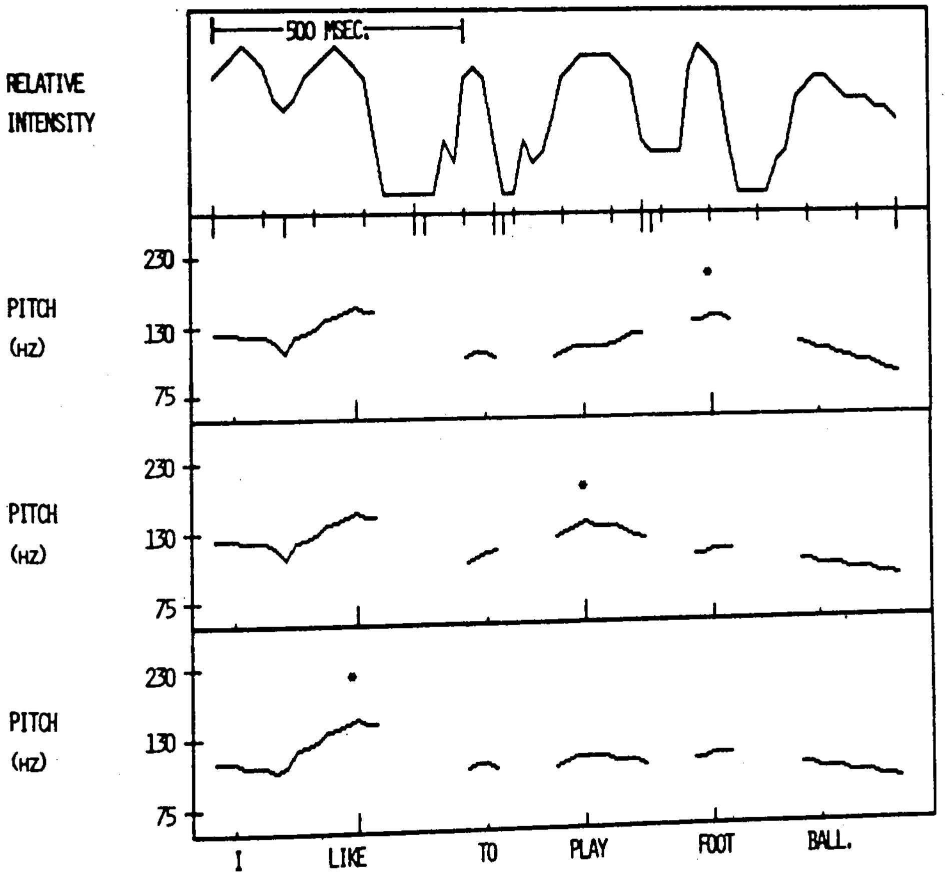

FIGURE 2: Display of relative intensity and pitch contours for three synthesized versions of the same sentence. The asterisk in each pitch display indicates the location of focus in each version, as described in the text. As in Figure 1, the vertical markers located along the bottom of each pitch display indicate the accent level assigned to each syllable. Note how the focused word in each version always corresponds with a primary-accented syllable (as denoted by the tallest accent markers), whereas all post-focus words have lower accent levels.