USE OF SPEECH SYNTHESIS IN AN INFORMATION SYSTEM FOR HANDICAPPED TRAVELLERS

B.C. DICKSON, S.J. EADY, J.A.W. CLAYARDS, S.C. URBANCZYK AND A.G. WYNRIB

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

ABSTRACT

Communicaid is an interactive microcomputer-based information retrieval system that employs synthesized speech and visual displays to assist handicapped travellers at an international airport. Audio output is accomplished using an LPC-based synthesis system in which the units of synthesis are individual words and phrases of English or French. Concatenated synthesis units are modified by a set of phonetic liaison rules and by a pitch smoothing algorithm.

INTRODUCTION

This paper describes a microcomputer-based system, called Communicaid, that has been designed to provide information to handicapped travellers at the Vancouver International Airport. Communicaid uses a combination of visual displays and synthetic speech output to provide information in English and French on topics such as transportation, accommodation and airport facilities. The video presentation is designed for travellers with hearing impairments, whereas the audio presentation is intended for those with sight impairments. Both modes of presentation are automated on a microcomputer. Since video mode of presentation does not involve audio speech output, it will not be discussed further here. The remainder of this paper deals exclusively with the audio portion of the Communicaid system.

DESIGN CONSIDERATIONS

Information Format

Communicaid presents audio information to the traveller in the form of information menus [1], similar to the one shown in Figure 1. Each menu provides a list of topics from which to choose. The user listens to the menu and selects an item by pressing a button after hearing a topic of interest. This process is repeated several times with increasingly specific menus, until the user obtains the desired information. At that point, the user can return to the main information menu and choose another topic of interest.

The information menus have been designed to take into account the temporary nature of the speech signal (compared to written text) and the way this interacts with the limitations of human information processing [1]. Consequently, each menu contains a maximum of seven information items, so as not to exceed the limits of human short-term memory [2]. In addition, the user has the option of listening to each menu several times. During the repetition, menu items are presented in a more cryptic form to introduce some variability into the audio presentation. Finally, for each menu item, the audio presentation is augmented by the visual display of a key word in large letters on a video monitor. All of these features are designed to facilitate easy information access in the audio mode.

Voice Output

In the initial stages of this project, we considered several different types of automated voice output for the presentation of audio information. Due to the interactive nature of the audio presentation, we needed a method that would provide fast access to the audio messages in an order that would be determined by the user. Thus, analog recordings of the information menus on audio tape would not be appropriate. On the other hand, commercially available text-to-speech conversion systems were judged unsuitable for this project, because of relatively low speech quality [3].

We then considered the use of digitally recorded speech materials, a method that has been demonstrated in applications such as telephone directory assistance [4], weather forecasts [5] and time-of-day announcements [6]. The advantage of using digitally recorded speech is that the quality is quite high, and the audio material is also easy to access in an interactive application of this kind. The major drawback with digitized speech, however, is that it requires enormous amounts of storage space. It is possible to reduce storage requirements by recording the speech materials in the form of isolated words and phrases, each of which can then be used in a number of different contexts. However, the problem with this strategy is that the prosodic aspects of each digitized word (i.e., pitch, duration and intensity) are fixed and cannot be easily modified to produce an intonation pattern that is appropriate for different sentence contexts. Speech output systems that make use of digitally recorded words typically require several versions of each word with a different intonation contour for each version [4].

MAIN INFORMATION MENU

YOU ARE NOW / AT THE START. (600) / TO CHOOSE INFORMATION ABOUT / ONE /
OF THE FOLLOWING / SEVEN / TOPICS (100)/ PRESS THE SELECT BUTTON /
AFTER YOU HEAR / THE TOPIC / YOU WANT. (600)/ TO END THIS SESSION /
AT ANY TIME (100)/ JUST RETURN / THE HEADSET / TO THE HOOK. (600)/

ONE (100)/ FOR FRENCH (100)/ POUR CONTINUER / EN FRANCAIS (100) / APPUYEZ /
SUR LE BOUTON DE SELECTION. (600) /
TWO (100) / LOCATIONS OF / AIRPORT FACILITIES. (600)/
THREE (100) / GROUND TRANSPORTATION / TO AND FROM / THE AIRPORT. (600)/
FOUR (100) / LOCATIONS OF / AIRLINE / TICKET COUNTERS. (600) /
FIVE (100) / HOTELS / IN THE VANCOUVER AREA. (600) /
SIX (100) / ASSOCIATIONS OF / THE BLIND AND PARAPLEGIC. (600) /
SEVEN (100) / FOR INFORMATION ON / FLIGHT DEPARTURES AND ARRIVALS (100)/
PLEASE CONTACT / THE TRANSPORT CANADA / INFORMATION
BOOTH (100) / DIRECTLY BEHIND YOU. (600) /

THIS IS THE END OF / THE LIST. (600) /
TO REPEAT / THE LIST / YOU HAVE JUST HEARD (100) / PLEASE WAIT. (600)/

FIGURE 1: Contents of the main audio information menu for the Communicaid Centre. The diagonal slashes demarcate the preprocessed vocabulary items that are concatenated to produce the speech output for this menu. Numbers in parentheses indicate the location and duration (in msec) of pauses in the audio output.

Due to space limitations imposed by the present application, this method of voice output was not suitable. Communicaid's information menus required a vocabulary of some 1,100 different words and phrases, each of which would be used in an average of seven different contexts. The digitization of several different versions of each vocabulary item would require an inordinate amount of storage space.

As an alternative to the use of unmodifiable digitally recorded speech items, we chose to produce the voice output for this system by encoding prerecorded words and phrases using linear predictive coding (LPC). This method maintains a relatively high quality of speech output. At the same time, it allows modifications to the prosodic aspects of encoded words and phrases, so that they can be joined together to form complete sentences. In addition, this technique greatly reduces the storage requirements for the encoded speech materials. Thus, our strategy for voice output was to use LPC-encoded words and phrases to generate sentences in a variation of the word-concatenation method proposed by Olive [7,8].

As pointed out by Olive, the production of good-quality synthetic speech using the word-concatenation technique requires that certain modifications be made to the word units when they are concatenated. The parameters that require modification include the pitch contour of each word, the spectral shape and amplitude at word boundaries, and the duration of each syllable.

Due to time constraints for this project, we were not able to develop complex prosody rules for this application. Instead, we chose to address this issue by making the basic units of synthesis (called "vocabulary items") as large as possible.

Thus, each vocabulary item could contain as many as four or five words (see Figure 1 for examples). By using large synthesis units, we were able to reduce the need for modifying prosodic features, because each vocabulary item would already contain pitch and duration patterns that would be appropriate for one or more sentence contexts. The rules that we did develop for modifying prosody acted mainly to alter pitch and energy parameters at the boundaries between vocabulary items.

SPEECH SYNTHESIS METHOD

Synthesized speech for this system 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 and French sentences are synthesized by concatenating these preprocessed vocabulary items in a specified order and then applying rules to modify pitch patterns and to eliminate spectral discontinuities at word boundaries.

Vocabulary Production

As indicated above, the strategy for speech synthesis was to employ preprocessed vocabulary items that contained as many words as possible. Considerable care was taken to parse the audio menu scripts so as to maximize the length of vocabulary items, while also ensuring that each item would be used in several different contexts, within the various menus. The chosen vocabulary items were then embedded in carrier sentences, which replicated as closely as possible the sentential environments in which they would be found in the information menus. This strategy ensured an appropriate intonation pattern for each item.

The carrier sentences were then read by a male speaker whose voice was recorded on a reel-to-reel tape recorder. Each item was digitized (at a 10-kHz sampling rate with 10-bit resolution) and excised from its sentence environment. The digitized vocabulary items were then analyzed using the autocorrelation method of LPC [9] to derive values of energy, pitch and 10 LPC reflection coefficients at 20-msec intervals. These parameters were quantized for output on the synthesis chip. Each encoded vocabulary item was then edited to eliminate any spectral discontinuities, and to provide a uniform energy maximum. This method of digital encoding produces a compression ratio of approximately 80 to 1, compared to the original sampled speech data.

Concatenation of Vocabulary Items

At the time of synthesis, encoded vocabulary items are concatenated to form complete sentences of English or French. The input to the system is a set of command files corresponding to the audio information menus. Each command file contains the vocabulary items that are to be concatenated, along with diacritics to indicate the ends of sentences, and the location and duration of pauses. 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 and by a pitch smoothing algorithm.

Phonetic Liaison

The phonetic liaison rules act to change energy and segment durations at the boundaries between vocabulary items. The application of each rule is determined by the particular phonetic segments that are present at each boundary. The phonetic segments at the beginning and the end of each item are identified by a rule and classified into one of several categories. There are five phonetic categories for item-initial segments and four such categories for item-final segments. These are listed in Table 1.

A total of 13 phonetic liaison rules handle all possible combinations of initial and final phonetic segments. Eight of these rules involve a smoothing of the energy contour at the boundary between vocabulary items (e.g., for a vowel-vowel or fricative-fricative combination). The other five rules involve the replacement of aspiration with silence (e.g., for a combination of two voiceless plosives), the insertion of a short silent interval (e.g., for a fricative-voiced plosive combination), or the repetition of a voiced frame immediately preceding the boundary (e.g., for a vowel-voiced plosive combination).

TABLE 1

Classification of Segments
for Phonetic Liaison Rules

Item-Initial Segments	Item-Final Segments
1. Vowels and [j].	1. Vowels, liquids, nasals and glides.
2. Nasals and liquids.	2. Fricatives and affricates.
3. Fricatives.	3. Voiced plosives.
4. Voiced plosives and [w].	4. Voiceless plosives.
5. Voiceless plosives, affricates and [ð].	

Pitch Smoothing

Following application of the phonetic liaison rules, a pitch smoothing algorithm modifies the pitch contours at the boundaries between vocabulary items. The aim of this algorithm is to eliminate abrupt pitch changes that may occur at such boundaries. Modifications of this type are required when a vocabulary item recorded for use in

sentence-final position occurs in a sentence-medial location. Sentence-final items are characterized by a falling terminal pitch contour. This falling contour must be flattened somewhat if the vocabulary item is to be used in the middle of a sentence.

In general, this relatively simple pitch-smoothing strategy was found to be adequate for concatenating the large synthesis units used in the present application. This was due to the fact that the information delivery system did not require the use of an interrogative intonation pattern. The intonational requirements for the synthesized speech included a declarative (falling) pitch contour for major syntactic boundaries in non-final position [10]. The former was handled by the falling pitch contour that accompanied most vocabulary items; the latter was generated as a result of the pitch smoothing algorithm and by the insertion of a pause at the major syntactic boundary.

The success of the simple pitch-smoothing rule used in this application is due to the restricted intonational requirements and the use of relatively large synthesis units. A more elaborate pitch assignment algorithm has since been developed for English to handle the concatenation of smaller synthesis units (i.e., individual words), as well as interrogative pitch patterns and the synthesis of focused words in English sentences [11,12]. It is anticipated that this more complex algorithm will be incorporated into the Communicaid system at a later date.

SUMMARY

The Communicaid system uses speech synthesis to provide audio information to handicapped travellers at an international airport. The traveller specifies the desired information by choosing from a list of items presented in an audio menu. Voice output is generated by concatenating LPC-encoded words and phrases of English or French. The concatenated vocabulary items are modified by a set of phonetic liaison rules and by a pitch smoothing algorithm. This application illustrates the use of speech technology for automated information delivery.

ACKNOWLEDGEMENT

This project was done under contract to Rutenberg Design Inc. of Montreal, with funding from the Transportation Development Centre of Transport Canada. We thank Mr. Uwe Rutenberg and Dr. Ruth Heron for technical advice and assistance.

REFERENCES

[1] Waterworth, J.A. (1982). "Man-machine speech dialogue acts." Applied Ergonomics, vol. 13, pp. 203-207.

[2] Miller, G.A. (1956). "The magical number seven, plus or minus two: Some limits to our capacity for processing information," Psychological Review, vol. 63, pp. 81-97.

[3] Carlson, R. and Granstrom. B. (1984). "Text-to-speech conversion in telecommunications," Behaviour and Information Technology, vol. 3, pp. 73-78.

[4] Waterworth, J.A. (1983). "Effect of intonation form and pause durations of automatic telephone number announcements on subjective preference and memory performance," Applied Ergonomics, vol. 14, pp. 39-42.

[5] Andersen, D.P. (1984). "A talking computer gives weather forecasts by telephone," First International Conference on Speech Technology (J.N. Holmes, ed., Brighton, U.K.).

[6] Waterworth, J.A. (1984). "Interaction with machines by voice: A telecommunications perspective," Behaviour and Information Technology, vol. 3, pp. 163-177.

[7] Olive, J.P. (1974). "Speech synthesis by rule," Proceedings of the Speech Communication Seminar, Stockholm, vol. 2, pp. 255-260.

[8] 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.

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

[10] Eady, S.J. (1986). "The influence of syntactic structure on fundamental frequency patterns of Canadian French sentences," Proceedings of the 12th International Congress on Acoustics, paper A6-7.

[11] 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.

[12] Eady, S.J. and Dickson, B.C. (1987) "Synthesis of sentence focus in English declaratives," Proceedings of the 11th International Congress of Phonetic Sciences, in press.