Personalized Synthetic Speech for AAC devices: The ModelTalker Project

H. Timothy Bunnell, Jane McNicholas, James Polikoff, Jason Lilley, George Oikonomou

NIDRR grant # H133G020167

Nemours Children's Clinic-Wilmington, Alfred I. duPont Hospital for Children

INTRODUCTION

The ModelTalker project involves the application of modern concatenative speech synthesis technology to assistive and augmentative communication (AAC). The ModelTalker text-to-speech synthesizer and its associated voice capture software, called InvTool, provide a method for individuals to use their own voice or the voice of another person (with their cooperation) in creating a personalized synthetic voice that will ultimately be usable in an AAC device.

Because the synthesis technology is based on recordings of natural speech, it is capable of producing highly natural-sounding and intelligible speech, at least under ideal conditions. We exploit this capability in ModelTalker by recording desirable AAC words and phrases in their entirety, thus ensuring that those utterances will sound fully natural when reproduced. Utterances produced by ModelTalker that were not originally recorded sound less natural than recorded speech but are generally intelligible and recognizable as the voice of the speaker who recorded the speech inventory.

BACKGROUND

Concatenative speech synthesis using diphones was first proposed by Peterson, Wang, and Silverman (1958). In diphone synthesis, speech is constructed by concatenating preselected segments of recorded speech that represent the region from the middle of one phoneme to the middle of an adjacent phoneme. An inventory of about 2500 such segments, extracted from natural speech and stored in a database, is adequate to allow synthesis of most English words in phrasal context.

Despite its early development, diphone synthesis did not become popular for several reasons, two of which are perhaps most crucial. First, it is virtually impossible to extract diphones that fit smoothly together in all possible contexts. As a result, diphone synthesis has sometimes quite jarring acoustic boundaries, and there is a boundary in every phoneme. Second, because the diphone segments must be stored in a database, diphone synthesis systems tended to require significantly more memory than competing technology based on parametric rule synthesis such as DecTalk. This was an especially significant consideration in earlier assistive and augmentative (AAC) devices.

Several factors have changed since diphone synthesis was first proposed that make concatenative synthesis more attractive today. First, computer memory has become very inexpensive compared to memory costs in the 1980s and early 1990s. Second, significant improvements in the design of concatenative synthesis systems were made that have moved away from diphone synthesis toward variable length unit concatenation systems (Takeda, Abe, & Sagisaka, 1992). Modern unit concatenation systems require larger databases of recorded speech but are capable of avoiding much of the acoustic boundary discontinuities associated with strict diphone synthesis.

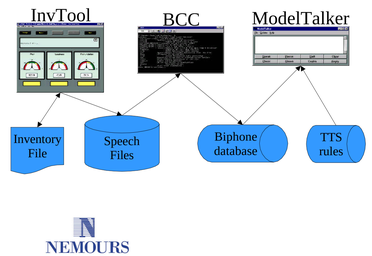
Despite the significant advances in concatenative synthesis systems over the last decade, we have only just begun to see them used in AAC devices. Further, although concatenative systems have the ability to capture an individual's voice characteristics, they are not being used for this purpose in AAC devices. At present, it costs about $200,000 for a commercial firm to produce a single, high quality synthetic voice based on unit concatenation. The primary goal of the ModelTalker project is to bring concatenative synthesis technology to the AAC user in a way that will allow the development of acceptable personalized voices without this great expense.

THE MODELTALKER SYSTEM

Figure 1 illustrates the components of the ModelTalker system and how they are related to one another. Broadly, the three components are the voice capture tool (InvTool), the text-to-speech synthesis tool (ModelTalker), and a program called biphone constrained concantenation (BCC) that is responsible for converting speech recorded by InvTool to a database that can be used by ModelTalker. Each of these three components, in turn, can be described in terms of several functional subunits.

Fig. 1

Overall System Design



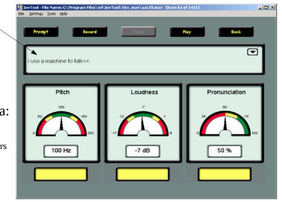
Nemours is the largest established group of pediatric specialists in the United States, serving patients in Delaware, Maryland, New Jersey, Pennsylvania, Florida, and Georgia. Visit us online at http://www.nemours.org, http://www.KidsHealth.org, and http://www.PedsRef.org.

Figure 2 illustrates the components that comprise InvTool. These are (a) a prompting system; (b) an audio recorder that creates waveform files; (c) speech analysis systems for detecting intonation, loudness, and phonetic content; and (d) feedback meters to indicate possible problems with recorded utterances. Additionally, InvTool itself has several special features to make it more useful for the evaluation study we describe below.

Visual and Aural prompt

Fig. 2

Feedback via:
- dials,
- digital readouts
- warning indicators

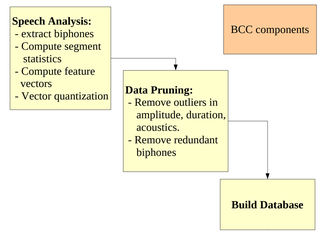


Builtin Speech analysis measures F0, RMS amplitude

Recognition module does phonetic alignment

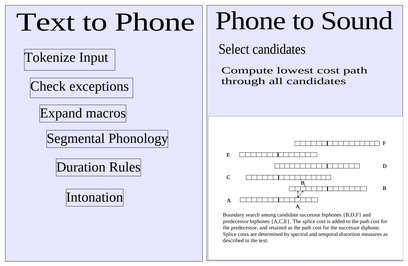
Figure 3 shows a diagram indicating how the BCC program converts recorded speech to a database for use by ModelTalker.

Fig. 3



The text-to-speech synthesis process itself is illustrated in Figure 4, which shows that ModelTalker includes a user interface, text-to-phoneme module, and phoneme-to-sound system. The text-to-phoneme module is responsible for parsing standard English text and generating a phonetic description of speech associated with the text. This is accomplished via the application of a large number of linguistic rules (and rule exceptions!) that are stored in data files and can be altered as needed to improve or adjust ModelTalker's pronunciation. The phoneme-to-sound module in ModelTalker is the component that directly operates on speech data by locating candidate portions of recorded utterances in the database and selecting the specific combination of those candidates to concatenate to produce fluent synthetic speech.

Fig. 4



While the ModelTalker system has been under continuous development and testing in our laboratory for some time, we have only just reached the stage where we feel it is useful to distribute this technology to potential users and evaluate how well the technology actually is able to perform "in the field." To that end, we have begun an evaluation study in which people diagnosed with amyotrophic lateral sclerosis can use InvTool and ModelTalker to create and listen to their own personalized synthetic voice. Below is a brief description of this study.

EVALUATION STUDY

Objectives

Broadly, there are three objectives: **1)** to assess the user's satisfaction with the process of generating a personalized voice, as well as the satisfaction, both of the user and of his or her friends and family, with the resulting synthetic speech; **2)** to determine speaker/voice characteristics that are predictive of satisfactory synthetic voices; and **3)** to obtain information to help improve the speech analysis and synthesis process.

Approach

After a client registers for the experiment, we obtain basic data about the client's health and speech status either from the client's clinician or via telephone interview with the client. Clients then download the software, work through a tutorial on the recording process, and record their voice for the speech synthesizer. Speech Laboratory staff are available by email and phone to assist clients in this process. The actual recordings may take place in a clinic or in the client's home and must be completed within two weeks. During the two-week recording period, clients periodically upload their completed utterances to the laboratory so a copy of the speech inventory is maintained exactly as recorded. InvTool has a built-in feature that automates the upload process. Once the speech is recorded, clients use the software to create a personalized voice that can be used with ModelTalker. Clients and their associates listen to a standard passage (the Grandfather passage) as synthesized by ModelTalker and then each fill out a web-based survey regarding the synthetic voice and the recording process. Participants are then encouraged to use the synthesizer as part of their daily activities. Once each week for the next eight weeks, clients and their associate receive email from our laboratory with a web link to a series of sentences spoken in the client's "synthetic voice." Clients (and, separately, their associate) listen to each sentence, rate its naturalness on a five-point scale, and type what they believe the sentence said. At the end of the experiment, the client and associate each fill out a final survey over the web. The survey questions address original dialect, intelligibility of the synthesizer, naturalness and sound quality of the voice, and usability of the overall system.

Status

To date, about 400 people have downloaded the InvTool and ModelTalker software, and about a quarter of these have expressed interest in participating in the study. Nine clients and their associates are presently registered and actively participating in the research protocol. Five clients have completed recording their speech inventory with InvTool, uploaded it to the Speech Lab, and entered into the user evaluation phase of the study.

SUMMARY

The images and text in this presentation guide viewers in considering the following: 1) using the InvTool software to capture the user's premorbid voice and prepare a synthetic speech database that incorporates the user's speech; 2) using the ModelTalker speech synthesizer to accept typed text from the user or the user's AAC device and produce unlimited synthetic speech; and 3) the user evaluation phase of our research project and how clinic-based speech language pathologists can collaborate in the project.

REFERENCES

Beukelman, Yorkston, Poblete, & Naranjo (1984). Frequency of word occurrence in communication samples produced by adult communication aid users. *Journal of Speech and Hearing Disorders*, 49, 360-367.

Bunnell, T., Hoskins, S., & Yarrington, D. (1996). A biphone constrained concatenation method for diphone synthesis. Proceedings of the 3rd International Workshop on Speech Synthesis, Jenolan Caves, Australia, November 1998, pp. 171-75.

Fletcher, S.G. (1972). "Time-By-Count Measurement of Diadochokinetic Syllable Rate". *Journal of Speech and Hearing Research*, 15, pp. 763-70.

Hillel, A.D., Miller, R.M., Yorkston, K., et al. (1990). Amyotrophic Lateral Sclerosis Severity Scale. In: Rose, C.F. (Ed). Chapter 11. *Amyotrophic Lateral Sclerosis.* Demos Publications. New York, N.Y

Hirano, M. (1981). Clinical Examination of Voice. In G.E. Arnold, F. Winckel, and B.D. Wyke (Eds.) *Disorders of Human Communication 5.* Springer-Verlag/Wein. New York.

Peterson, G., Wang, W., and Silversten, E. (1958). Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America*, 30, 739-742.

Strand, Edythe A. et al. (1996). Management of Oral-Pharyngeal Dysphagia Symptoms in Amyotrophic Lateral Sclerosis. *Dysphagia*. 11:129-39.

Takeda, K., Abe, K., and Sagisaka, Y. (1992). On the basic scheme and algorithms in nonuniform unit speech synthesis. In G. Bailly, C. Benoit and T.R. Sawallis (eds.), Talking Machines: theories, Models, and Designs. Amsterdam: Elsevier, 93-105.

Yamaguchi, H. (1995). GRBAS training method. Indiana University, Bloomington.

Additional references and publications available at http://www.asel.udel.edu/speech