Automatic Construction of Concatenative Speech Synthesis Databases for AAC

H. Timothy Bunnell, John Gray, Christopher Pennington, Debra Yarrington

*Speech Research Lab, Alfred I. duPont Hospital for Children, Nemours Children’s Clinic, 1600 Rockland Rd., Wilmington, DE 19803; AgoraNet Inc., 102 East Main St., Suite 303, Newark, DE 19711

ABSTRACT

Following realignment of the boundaries, statistics are computed for several values related to each minor SLP professional oversight, this software will permit virtually automatic generation of synthetic voices for use in AAC devices. In CSS, recorded natural speech is stored in a database structure that allows phonetic units to be selected and concatenated to form novel utterances (i.e., utterances that were not originally recorded for the speech database). Because it is based on natural recorded speech, CSS can result in highly intelligible and natural sounding personalized voices, provided that the concatenated speech units blend smoothly together. However, both intelligibility and naturalness suffer greatly when the concatenated units do not blend smoothly together.

How well units blend together is a function of several factors, but primarily hinges on two: (a) the consistency of the speaker in recording the speech database; and (b) the accuracy of the phonetic tags in the database that identify the locations of phonetic segments within the recorded speech. Because of this, constructing high-quality CSS voices is typically a time-consuming and costly process. Commercially available CSS systems are typically constructed with speech recorded by professional speakers whose recordings are carefully monitored to ensure consistent voice quality, amplitude, and pronunciation. To acquire a uniform and consistent voice, a technically skilled laborer is needed to verify and correct acoustic phonetic markers in the recorded corpus of utterances. These factors have put the construction of personalized synthetic voices out of reach both technically and financially for most AAC users.

Our goal for the ModelTalker project is to develop software that will permit individuals at risk of losing their voice to create a professional voice based on their speech with minimal supervision by a Speech-Language professional or appropriately trained para-professional. The voices so created will then be usable with Microsoft Windows-based AAC devices.

INTRODUCTION

Concatenative speech synthesis (CSS) is ideal for Augmentative and Alternative Communication (AAC) devices. In CSS, recorded natural speech is stored in a database structure that allows phonetic units to be selected and concatenated to form novel utterances (i.e., utterances that were not originally recorded for the speech database). Because it is based on natural recorded speech, CSS can result in highly intelligible and natural sounding personalized voices, provided that the concatenated speech units blend smoothly together. However, both intelligibility and naturalness suffer greatly when the concatenated units do not blend smoothly together.

CONCLUSIONS

We have developed a Concatenative Speech Synthesis system along with software that is intended to make it possible for individuals to develop their own personalized voice for use in AAC. To date, we have shown that the synthesis technology itself is comparable in intelligibility to commercially available systems, and significantly more intelligible than a DECTalk female voice that is in use in many AAC devices. Some voices generated with this process are already in use by augmented communicators, and more are being developed. In the next phase of our project, we will be working to further improve the voice capture process by increasing the efficiency of InvTool, and hopefully reducing the number of utterances needed to create a high-quality voice. We also expect to work more closely with AAC device manufacturers to ensure that ModelTalker synthesis technology is available to a wide range of AAC device users.

ACKNOWLEDGEMENTS:

This work has received support from the U.S. Department of Education (grant # H133E00010), from the National Institutes of Health (R41-DOCO56183), and from Nemours Biomedical Research. We are especially indebted to the individuals with ALS who have supported this research by investing the time necessary to record their voices and evaluate the voice generation process.