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

We describe techniques used to optimize the accurate alignment of phonetic markers to speech recorded for concatenative synthesis. With minor SLP professional oversight, this software will permit virtually automatic generation of synthetic voices for use in AAC devices by clients who are at risk of losing their voices.

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.

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. Then technically skilled labor 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 CSS voice based on their speech with at most minor 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.

APPROACH

In the *ModelTalker* project, we address the issue of speaker consistency in the design of a program we developed called *InvTool* that guides speakers in the process of recording speech for a CSS database. *InvTool* (see Figure 1) prompts the user to record each utterance that is needed for a CSS database and monitors the user's productions for pitch range, amplitude, and pronunciation accuracy. If the user's utterance falls outside acceptable bounds in any of these dimensions, they are asked to re-record it. A calibration procedure is used to set the acceptable bounds for pitch and amplitude to values that are appropriate for each user. Assessment of pronunciation is based on measurements from a speaker independent speech recognition (SR) system that is part of *InvTool*. The SR system aligns each recorded utterance to a phonetic transcription of the text that the user was asked to produce. If the segmental acoustic structure of a recorded utterance is a good match to the transcription of the requested utterance, it receives a high pronunciation score. Otherwise, if the pronunciation score falls below a threshold, the speaker is asked to repeat the utterance.

Figure 1. Illustration of *InvTool* interface, which assists speakers in recording a speech corpus with consistent voice quality and pronunciation.

Nemours is one of the largest established groups 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 and http://www.PedsEducation.org.

The output of *InvTool* is a corpus comprising a speech waveform file, pitch period location file, and phonetic boundary locations for each utterance. This is illustrated in Figure 2, which shows part of a labeled waveform for the utterance *Hi* as produced by *InvTool*.

Figure 2. Speech waveform and spectrogram indicating information assigned by *InvTool*. In this figure, pitch marks are shown in the top panel, the phonetically labeled waveform in the middle panel, and spectrogram in the lower panel. Each phonetic segment has a two-character phonetic symbol code (0B = beginning silence; HX = /h/; AI = /aɪ/).

Once an entire corpus has been recorded, the second stage of processing—handled by a program called *bcc*—takes the output of *InvTool* and constructs a CSS database. In the process, *bcc* attempts to adjust phonetic boundary information to improve its accuracy and removes segments (or even whole utterances) from the database if they appear to be unusual in any one of several ways.

For a majority of phoneme boundaries, the phonetic alignment assigned by *InvTool* is acceptable; however, errors of several types do occur. First, errors can be the result of the talker not producing exactly the expected phoneme sequence. Even careful, professional talkers such as radio news announcers will make a few of these errors in the course of a 1650 phrase corpus. While the pronunciation assessment in *InvTool* usually catches mispronunciations and word errors, it is not infallible: some will incorrectly be accepted.

Second, errors can result from misalignment of the phonetic boundaries to a correctly produced utterance. Although the recognition engine used in *InvTool* has been trained on a large amount of speech from many talkers, there is nonetheless enough variability in normal speech to lead to ambiguity as to exactly the "best" phoneme boundary locations. The most common boundary errors are relatively small in size (on the order of 50 msec or less); however, much larger errors are sometimes observed.

To minimize the impact of these possible errors in the data collected by *InvTool*, the *bcc* program begins by completely reanalyzing the speech corpus, optimizing the acoustic analysis to the characteristics of the speaker and recording environment (e.g., the specific microphone and sound card the speaker used). Using optimal acoustic features, *bcc* then retrains the SR system to the characteristics of the corpus. As a result of this, we obtain a substantial reduction in the number of phoneme alignment errors and end up with boundaries that are more consistent throughout the corpus.

To assess how well this process works, we selected about 650 utterances from a speech corpus that had been recorded with *InvTool*. All of the segment boundaries in these utterances were adjusted manually by one of the authors (DMY) who has extensive experience in labeling speech for synthesis databases. Most of these utterances were chosen because they were flagged as having possible segment alignment errors. That is, the majority of these utterances contained one or more segments that apparently were not labeled correctly. Thus, these utterances represent the worst cases from the speech corpus in terms of possible boundary or other errors.

The dotted line in Figure 3 shows the cumulative proportion of segment boundaries that fell within successively wider regions surrounding the hand assigned boundaries. For example, almost no automatically assigned boundaries fell within 5 msec of the hand adjusted boundaries, and only about 10% of the automatic boundaries were within 40 msec of the corresponding hand alignment. Indeed, fewer than 60% of the original phoneme boundaries in these utterances were within 100 msec of the boundaries assigned by an expert.

The solid line in Figure 3 shows the same information following reanalysis and adjustment of segment boundaries. Clearly, even in these worst-case instances, substantial improvements are made in the segment alignment.

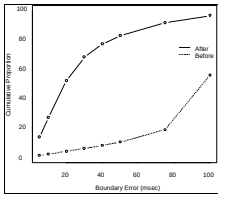

Figure 3. Boundary errors before (dotted line) and after (solid line) reanalysis for selected utterances from one speaker's corpus.

Following realignment of the boundaries, statistics are computed for several values related to each phoneme. In particular, we calculate mean and standard deviations for segment duration, amplitude, proportion of voiced pitch periods, and log likelihood (a measure of goodness of fit to the recognition models). Based on these, units are removed from the database if they fall outside 3 standard deviations from the mean of any measure. When recorded phrases contain too many segments identified as "outliers" in this manner, the entire sentence is removed from the database; otherwise, only the outlier segment is removed.

To date, we have tested these automated procedures on speech recorded by a variety of talkers, including one radio announcer (hired to create a "professional" voice for our system), two elderly individuals (a male aged 89, and female aged 72), and several individuals with ALS. Recordings from several additional ALS talkers have been obtained but have not yet been processed using the new procedures. Live synthesis examples are available for voices generated by our female professional speaker and for one ALS patient (used with her permission). Formal intelligibility testing is under way comparing our professional female voice with several commercially available female synthetic voices. A preliminary comparison of the ModelTalker female voice with four other commercially available female voices (AT&T Crystal, Cepstral Linda, DECTalk Betty, and Microsoft Mary) is shown in Figure 4. Of the five voices tested, ModelTalker is second highest in intelligibility, but not significantly different in intelligibility from the Microsoft Mary voice.

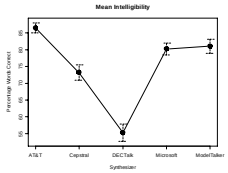

Figure 4. Mean Intelligibility (percentage of content words correctly identified in SUS sentences) of five synthetic voices. Error bars indicate the 95% confidence intervals associated with each mean. Each mean is the average of data from 30 listeners transcribing 20 sentences.

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 # H133E30010), from the National Institutes of Health (R41-DC006193), 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.