Automatic personal synthetic voice construction

H. Timothy Bunnell[1], Chris Pennington[2], Debra Yarrington[2], and John Gray[2]

Speech Research Laboratory
A. I. duPont Hospital for Children, Wilmington DE, USA[1]
AgoraNet, Newark DE, USA[2]
bunnell@asel.udel.edu; (penningt,yarringt,gray)@agora-net.com

Abstract

We describe techniques used for automatic personal synthetic voice creation in our laboratory. These techniques are implemented in two pieces of software. One, called InvTool, guides novice users in the process of recording a corpus of speech that is appropriate for creation of a concatenative synthetic voice. The other program, called BCC, compiles a speech corpus recorded with InvTool into a database appropriate for use with the ModelTalker TTS system. Our primary goal in this project is to develop software to support "voice banking" wherein individuals at risk to lose the ability to speak will be able to record their own personal synthetic voice for later use in voice output communication devices.

1. Introduction

Augmented communicators—individuals who cannot produce understandable speech and instead use synthetic speech generated by an Augmentative and Alternative Communication (AAC) device—have for years relied on a small number of commercially available synthetic "voices" for use in their AAC devices. Mostly, these devices have used rule-based formant synthesis systems such as DECTalk to generate synthetic speech. Thus, many AAC devices have relied upon synthesis technology that is decades old and demonstrably less intelligible and less natural sounding than more recently developed systems that use unit concatenation [1-4].

Some recent AAC systems are now providing users with more options for synthetic speech including concatenative voices (e.g., voices from Cepstral now ship with some DynaVox AAC systems). Moreover, as AAC device technology evolves to piggyback on standard operating systems such as Windows CE, it opens up the possibility of using virtually any Microsoft SAPI compliant voice as the synthetic voice for the AAC device.

The current systems still fall short of the ideal goal of providing every AAC device user with a personal voice, that is, one that no other augmented communicator is also using. The ModelTalker project is designed to address this goal by providing the capability of recording a corpus of speech from an individual talker and automatically converting it into a concatenative synthetic voice. The potential for rapid automatic concatenative voice creation resonates most strongly with individuals who have neurodegenerative diseases such has Amyotrophic Lateral Sclerosis (ALS) or Lou Gerhig's disease. These individuals are typically diagnosed while their ability to speak is intact and they thus have an opportunity to record their own voice for later use in an AAC device.

In the following sections, we describe the overall design of the software we have developed for voice creation, and present some recent results on intelligibility and naturalness of the resulting synthetic speech.

2. Voice Generation

Broadly, there are two components to developing a concatenative synthetic voice: 1) acquisition of a speech corpus; and 2) acoustic phonetic indexing of the corpus. We address these two components of the process using two distinct applications. The InvTool program guides users in the process of recording a corpus of utterances. The system is highly configurable in terms of the speech corpus content and is intended to allow users who are unfamiliar with speech and language to successfully record a speech corpus.

The second application in this process, called BCC, converts a corpus of speech to a concatenative synthesis database usable by the ModelTalker TTS system. BCC has the task of adjusting and verifying all the acoustic phonetic information of the speech corpus to arrive at a database that is phonetically accurate and internally consistent. We describe each of these programs in greater detail below.

2.1. InvTool

Figure 1 displays the InvTool user interface. It presents both a written prompt and an aural model of the utterance to be recorded. The user then records the utterance and InvTool analyses it. The analysis consists of (a) pitch analysis and tracking, (b) tests for amplitude levels, and (c) forced recognition using a set of Hidden Markov Models (HMMs) to align a phonetic transcription of the requested utterance to the received acoustic token. Results of these analyses are presented to the user via three controls in the form of graphical meters or gauges. The gauges provide visual feedback on the measured average pitch of the utterance relative to the user's calibrated pitch range, amplitude on an absolute decibel scale, and pronunciation on a percentage scale. Each of the three gauges will give either a green "Good" feedback or the gauge will be red and indicate what the problem with the utterance was. If an utterance passes all of the screening tests, InvTool automatically moves to the next prompt in the inventory list, otherwise, InvTool does not automatically advance and the user is expected to rerecord the utterance.

2.1.1. Standard InvTool Corpus

The recording process is controlled by a stored list of utterances along with their phonetic transcriptions. InvTool reads this list and prompts for utterances in the order they are

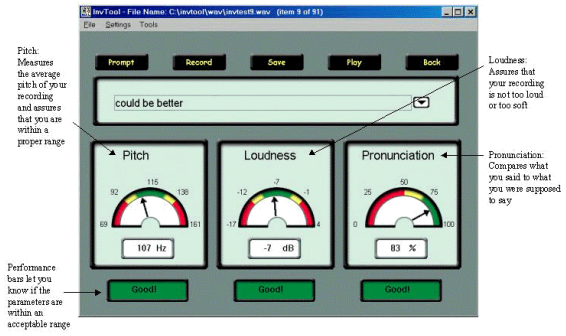

Figure 1. InvTool user interface illustrating controls used to assist in capturing valid recordings for each utterance in the corpus.

read. Users may also add their own utterances to the "inventory" list using a selection in the InvTool tools menu.

The default inventory list contains 1650 discrete utterances of varying length. The first 80 utterances in the list are phrases and sentences that are likely to be of need to users of AAC devices [5]. The remainder of the corpus was chosen to afford broad diphone coverage for American English in a variety of prosodic contexts. Specifically, four types of materials comprise the non-AAC-specific portions of the corpus: 1) 50 isolated high frequency words; 2) about 150 utterances of the form <syllable>-<Art|Prep>-<syllable> where an utterance medial article or preposition is embedded between syllables that are either real words or nonsense forms; 3) about 600 high-frequency word pairs spoken in isolation; and 4) the remaining nearly 800 utterances are semantically anomalous and meaningful phrases or short sentences. Bi-gram frequencies for the word pairs and diphone frequencies were estimated from the Brown Corpus [6]. Depending upon speaking rate, a complete recorded corpus typically contains between 40 and 50 minutes of actual speech.

2.1.2. *Data Screening in InvTool*

As each utterance is recorded, it is analyzed and results of the analysis are used to screen utterances for possible errors. The three aspects that are closely monitored by InvTool:

F0 Monitoring. An average F0 and F0 range is computed dynamically from the first few utterances recorded and thereafter used to screen for utterances in which the average F0 falls outside the expected F0 range.

Amplitude Monitoring. Peak amplitude is monitored to ensure that it is high enough to provide a good signal to noise ratio while not allowing digital clipping.

Pronunciation Monitoring. The expected phonetic transcription of each utterance is aligned to the acoustic speech signal using forced recognition with a set of discrete HMMs that were trained on the TIMIT training set [7]. Based on the obtained alignment, a second pass algorithm estimates the probability that each aligned segment is the expected segment. From these per-segment estimates, a global probability measure is obtained. This global probability, reported as a percentage in the displayed gauge, is a non-linear combination of the per-segment probabilities that heavily weights segments that are identified as having very low probability of being correct. This avoids a tendency to over-rate long utterances that contain one or two highly questionable segments.

2.2. BCC

The corpus received from InvTool, while screened, is nonetheless assumed to contain a variety of errors of various types ranging from pitch tracking errors to misaligned phonetic boundaries to incorrect phonetic content. Our process for converting this corpus of speech to an acceptable synthetic speech database employs the following steps.

2.2.1. *Acoustic reanalysis*

All utterances are reanalyzed using a pitch synchronous 32-channel Bark-weighted filter bank. A second dataset is computed from this 32-band spectrum by taking the time derivatives of the log spectral amplitudes in each filter channel. The dimensionality of the spectral measures and the delta-spectral measures is then reduced from 32 coefficients per analysis frame to 8 coefficients by principal components analysis and decomposition (separate PCA solutions are computed for each of the two datasets). The PCA feature vectors are then vector quantized to 256 element codebooks. Thus, we end up with a description of the original speech data in terms of a pair of code words reflecting the PCA feature vectors for the Bark Spectrum and Delta Bark Spectrum measures of each pitch-synchronous analysis frame.

2.2.2. *Speaker-specific HMMs*

Beginning with the phonetic label alignment assigned by InvTool, new speaker-specific discrete HMMs are trained using the newly obtained speaker-tuned acoustic features.

2.2.3. *Outlier removal*

Based on the final phonetic boundary alignment derived from HMM training, means and standard deviations are computed for segment duration, average RMS amplitude, proportion of voiced frames, and log likelihood. Segments that are outliers on any of these measures are then removed from further consideration as long as there are at least five other examples of the same phoneme in the dataset.

2.2.4. *Redundancy reduction*

While the outlier removal is based on individual phoneme statistics, this step considers pairs of adjacent phonemes (biphones). In this stage, all instances of a particular biphone sequence are compared with all other instances of the same sequence to form a similarity matrix. The cells of this matrix represent the acoustic similarity of each biphone with every other biphone of the same type. A hierarchical clustering algorithm is then applied to the similarity matrix to identify clusters of highly similar biphones. We typically prune the cluster tree at 20 clusters and retain the most prototypical biphone of each cluster. All biphones that are not prototypical of one of the 20 clusters are then removed from the dataset.

2.2.5. *Final processing.*

The result of all the above steps is to identify biphone sequences to retain for the final speech database. In the process, questionable phonetic segments and unnecessary biphones are removed from the dataset, and a report file is

written which lists all of the rejected segments along with the reason(s) for rejection. This report can then be used to locate and hand correct problems if the voice is to be hand corrected.

In our present system, the raw speech data for concatenation are coded and stored as windowed waveform packets for PSOLA processing. However, alternative speech coding strategies are being implemented for any commercial release of this system.

3. System Evaluation

For a realistic overall evaluation of the ModelTalker system and its associated voice creation process, at least two distinct issues must be addressed. First, one must establish how the synthesis system itself compares to other synthesis systems, particularly those using unit concatenation, and/or those being used in AAC devices. This comparison must establish the best-case expectations for the ModelTalker TTS system, given the specifics of the synthesis technology being used and the limitations of the standard inventory. To examine this question, one must use a speech corpus that has been carefully corrected to eliminate errors due to mispronunciation, segment misalignment, poor pitch tracking, and so forth.

The second issue to examine is how automatically generated voices compare to carefully constructed voices, given the standard corpus. To date, we have completed one formal evaluation of the system addressing the first of these issues[3, 4], and have less formally evaluated the second issue for a single automatically generated voice. We are now collecting additional automatically generated voices and expect to have a sufficient number of these to do an initial formal evaluation in the near future. Examples of passages synthesized with several of these automatically generated voices are available at http://www.asel.udel.edu/speech/mt/).

An alternative evaluation of the automatic voice creation process, but not the inventory selection or recording process, was afforded by the Blizzard challenge. In the following, we describe the procedures used in preparing the ModelTalker voices for Blizzard, focusing on aspects of the process that differ from those described above.

3.1. Methods

Talkers. Voices were generated from the four talkers: SLT, BDL, RMS, and CLB. One of the authors listened to every sentence recorded by each talker and compared the utterance to a transcript of the standard Arctic corpus. Instances where the talker deviated from the transcript were noted.

Corpora. One of the authors listened to ModelTalker as it synthesized each of the 1132 sentences of the standard Arctic corpus to identify words that needed to be added to our pronunciation dictionary. About 65 words were identified in this process. Additionally, several instances were identified where ModelTalker chose the wrong form of a word (e.g., bow, read) and consequently the correct form needed to be coerced. After fixing these problems, the text of the corpus was transcribed by ModelTalker to generate a control file containing text and transcription information for use by BCC.

Four versions of the original control file were then generated, one for each talker. These separate versions contained talker specific adjustments to the English transcript and phonetic transcription in the cases where it was noted that the talker deviated from the standard transcript. No attempt was made to make fine phonetic adjustments; corrections were restricted to cases where talkers used different words and/or word order from the expected text.

Acoustic analysis and labeling. Standalone versions of the same pitch tracking and phonetic labeling routines that InvTool uses were run on the waveform files for each talker. These programs provided the pitch period location data used by BCC for pitch synchronous analysis, and aligned phonetic transcriptions to the waveforms respectively. Following these analyses, BCC processing followed the steps described above.

Data correction. For each talker processed by BCC, a report file was generated indicating all segments found to be statistical outliers (and consequently removed from consideration for use in synthesis). For one talker (BDL), this report file was used to guide hand editing of the speech data. Approximately 50 utterances that were flagged as containing multiple errors were examined and fixed by hand. In this case, the fixes consisted entirely of adjusting phonetic segment boundary alignment and/or correcting disagreements between the canonical transcription and the received pronunciation. Although pitch-tracking errors were also observed, they were not corrected.

The hand-corrected sentences were flagged for BCC to indicate that the segment alignment should not be automatically adjusted and BCC was then rerun to produce a new version of the BDL voice.

3.2. Results.

Since complete results of the Blizzard listening tests will be presented elsewhere, we concentrate on just the overall results for the ModelTalker system in this report.

Of the six synthesizers and corresponding natural voices used in the listening tests, ModelTalker consistently received the lowest mean opinion scores (MOS). That is, stimuli produced by ModelTalker were consistently rated the least natural sounding of those presented.

On the other hand, intelligibility, as indicated by WER scores was consistently above average for stimuli produced by ModelTalker. This is illustrated in Figure 2, which shows WER averaged over all listener types, voices, and tasks. Only

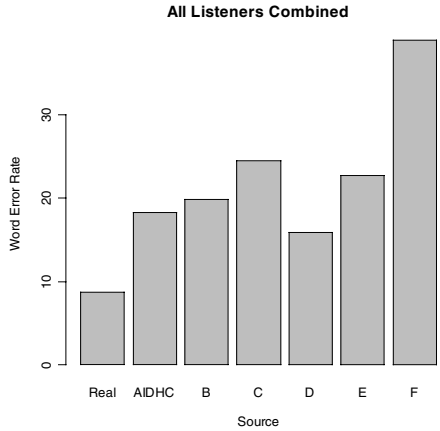

Figure 2. Composite WER results from Blizzard listening tests. Real designates results for the natural speech of real talkers, AIDHC designates results for the ModelTalker stimuli.

the system labeled 'D' in Figure 2 achieved a lower overall WER score (ModelTalker is the system labeled 'AIDHC' in this figure).

Interestingly, the system labeled 'F' in Figure 2 was the system that scored most similarly to ModelTalker on the MOS tasks.

We also examined overall word error rates for each of the four ModelTalker voices (Figure 3). The BDL voice (which was partly hand corrected) was the worst overall, having a WER almost twice that of the best voice (RMS).

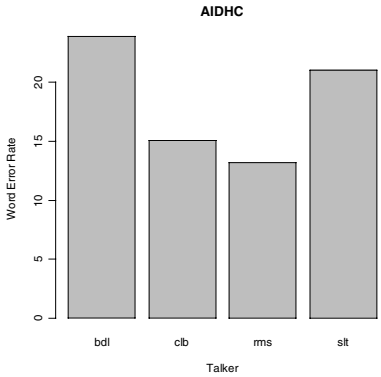

Figure 3. Word error rate by talker for the ModelTalker voices.

4. Discussion

While the ModelTalker system made a competitive showing in terms of WER, it's ranking in MOS tasks was disappointing. Several factors probably contribute to the poor MOS scores. First, unlike most current unit concatenation systems, the ModelTalker system normally runs in a full synthetic mode in which both timing and intonation are imposed on the synthetic speech. It is our impression, especially with smaller corpora, that this approach trades naturalness (in terms of voice quality) for smoother prosody and possibly higher intelligibility, but this has not been tested.

The pitch-tracking program we used did not always perform well, particularly with talker SLT. We noticed in particular that voiced/unvoiced decisions were often in error in the direction of labeling voiceless regions as voiced. This error, coupled with the control of segment duration and F0, causes significant buzziness, when using PSOLA processing.

The WER scores obtained for ModelTalker stimuli in the Blizzard challenge are very consistent with percentage words correct scores we have obtained using similar speech materials (SUS stimuli [8]) and a carefully hand corrected corpus [4]. In that study, we compared the ModelTalker 'Kate' voice to female voices produced by several commercially available TTS systems (Figure 4).

5. Conclusions

Despite weaknesses in naturalness and voice quality that may be due to our present speech coding methods, the results of the Blizzard challenge suggest that the BCC program is performing well in the task of identifying correct acoustic phonetic boundaries, and similarly doing an effective job of correctly rejecting mislabeled or misaligned segments.

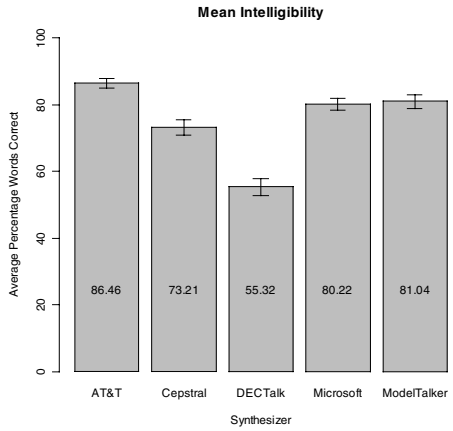

Figure 4 Overall mean intelligibility scores (percentage of words correct) for the five synthetic voices. Error bars are the 95% confidence intervals around each mean.

Coupled with the InvTool program to guide the recording process and reduce errors on the input side, this process has been used successfully by AAC users for voice banking.

6. Acknowledgements

This work was supported by grant R41-DC006193 from NIH/NIDCD and by Nemours Biomedical Research.

7. References

[1] H. S. Venkatagiri, "Segmental intelligibility of four currently used text-to-speech synthesis methods," *Journal of the Acoustical Society of America*, vol. 113, pp. 2095-2104, 2003.

[2] H. S. Venkatagiri, "Speech recognition technology applications in communication disorders," *American Journal of Speech-Language Pathology*, vol. 11, pp. 323-332, 2002.

[3] H. T. Bunnell, J. Gray, C. Pennington, and D. Yarrington, "Automatic construction of concatenative speech synthesis databases for AAC.," presented at American Speech Language and Hearing Association Conference, Philadelphia, PA, 2004.

[4] H. T. Bunnell, C. Pennington, and D. Yarrington, "An evaluation of emerging speech synthesis technology for AAC," in preparation.

[5] K. M. Yorkston, D. R. Beukelman, K. Smith, and R. Tice, "Extended communication samples of augmented communicators. II: Analysis of multiword sequences," *J Speech Hear Disord*, vol. 55, pp. 225-30, 1990.

[6] W. Francis and H. Kucera, *Frequency Analysis of English Usage*. Boston: Houghton Mifflin Company, 1982.

[7] X. Menéndez-Pidal, J. B. Polikoff, and H. T. Bunnell, "An HMM-based Phoneme Recognizer Applied to Assessment of Dysarthric Speech," presented at Proceedings of EuroSpeech '97, Rhodes, Greece, 1997.

[8] C. Benoit, M. Grice, and V. Hazan, "SUS test: A method of the assessment of text-to-speech synthesis using semantically unpredictable sentences," *Speech Communication*, vol. 18, pp. 381-392, 1994.