Personalized Synthetic Voices for AAC

Debra Yarrington, Steven R. Hoskins, James B. Polikoff, H. Timothy Bunnell
University of Delaware & duPont Hospital for Children
Wilmington, Delaware USA

INTRODUCTION:

The ModelTalker speech synthesis system creates individualized synthetic voices for augmentative communication devices. To emulate the voice of an individual, that person records an inventory of words and phrases. The recorded inventory is then converted to a speech database that is used by ModelTalker. The system presently can work with augmentative communication tools such as the Prentke-Romich Wivik program. Ultimately, it will be extended to work with emerging computer based AAC devices as a "plug compatible" synthesizer to provide augmented communicators with the alternative of using a personalized voice in their AAC device. The system is useful for anyone who wishes to use an individualized synthetic voice, but may prove especially desirable to people who are diagnosed with a speech disabling condition before the loss of their voice, for instance, persons diagnosed with ALS.

BACKGROUND:

Present AAC devices can provide unrestricted text to speech synthesis with unnatural speech quality, or restricted speech using prerecorded utterances with natural quality. The predominant speech synthesizer in these devices is DecTalk, a system based on technology developed in the 1970's and refined only a little since then. While DecTalk speech is highly intelligible, each voice is based on a set of complex rules. In order to generate a new voice, a new set of rules must be generated as well. Thus the DecTalk system only permits users of AAC devices a limited choice of possible synthesized voices, and those voices do not take advantage of the advances made in speech synthesis over the last 2 decades.

A number of the most natural-sounding and intelligible synthesizers presently available commercially and in research laboratories use a technology called 'unit concatenation' (Takeda, 1992). Unit concatenation is a technique in which recorded speech is partitioned into small units that can be concatenated to produce novel utterances. One type of unit concatenation, diphone concatenation, was first proposed in the 1950's (Peterson, Wang, and Siversten, 1958) and has been used in a number of commercially available synthesizers (e.g., SmoothTalker, Lernout & Houspie, and AT&T's TruVoice). For diphone concatenation, all the units being concatenated are specifically diphones (i.e., brief samples of speech spanning the region from the middle of one phoneme to the middle of the subsequent phoneme). Unfortunately, diphone concatenation often sounds very "choppy" because diphones don't always blend together smoothly.

As an alternative to predefining specific units such as diphones for concatenation, some recent unit concatenation systems use variable-length units for synthesis. As an utterance is being synthesized, efficient computer search algorithms identify optimal arbitrary-sized concatenation units in the speech database (Campbell and Black, 1996, Beutnagel, Conkie, and Syrdal, 1998) as a complete word or phrase if a unit of that length in the database matches the desired synthetic output. When words or even phrases are found to match the desired synthetic output the synthesizer is really just playing back a recording and sounds very natural. Thus, by selecting concatenation units "on the fly" during the synthesis process, the synthesizer is able to make optimal use of the natural speech recordings within the speech database, resulting in synthetic speech which can be very smooth and natural sounding when the concatenation units turn out to be large.

With variable length unit concatenation then there is a trade off between the quality of the speech output and the size of the speech database. The larger the speech database, the more likely it is to contain specific words or phrases and the less choppy the resulting synthesized speech will sound. Unfortunately, very large databases (e.g., hours of speech) take a long time to prepare, require substantial amounts of memory to store, and very significant computational resources to perform the optimal unit search in real time along with overhead involved in the actual concatenation synthesis of the speech.

To obtain the advantages of variable length unit concatenation systems with a reasonably sized corpus of recorded speech, the ModelTalker system uses elements of both strategies. ModelTalker includes high frequency words and user-desired words and phrases in its corpus. It also includes a complete set of smaller (diphone-like) concatenation units needed to produce unrestricted speech, but with less fluent quality. Thus, the ModelTalker system allows high frequency words and selected utterances to be synthesized with recorded speech quality, and blends smoothly into unrestricted text-to-speech as needed. To optimize the output quality of ModelTalker for unrestricted speech while minimizing the size of the speech database required, we are conducting research to map the relationship between an objective measure of acoustic distortion in the synthetic output of ModelTalker and perceptual evaluations of the speech. Results of this research will guide decisions regarding how much time and effort must be put into preparing a speech database for acceptable quality synthetic speech output.

SPEECH DATABASE: STRUCTURE AND OPTIMIZATION

The present speech database is built from a corpus of utterances comprising:

1) A list of user selected phrases and words.
2) 250 high frequency words (not counting articles and prepositions).
3) A collection of short phrases involving real words and nonsense syllables bounding articles and prepositions to provide diverse transitions into and out of these very common words (approximately 300 in number).
4) An additional collection of real and nonsense word forms and phrases needed to provide completeness in the inventory of biphones contained in the database (approximately 750 in number).

Items 3) and 4) in this corpus are relatively open ended; by increasing the size of these lists, the quality of the synthetic speech should be improved, but it is not certain how much improvement in speech quality is gained from a given set of list extensions. To provide insight into this trade-off, research is presently underway in which a group of 9 talkers (3 adults between the ages of 18 and 40, 3 children between the ages of 8 and 12, and 3 elderly adults, ages 65-85) will record very extended speech corpora of no more than 5000 utterances. The research will explore, from a computational perspective, how to minimize acoustic distortion in the speech data for a variety of list lengths, and, from a perceptual perspective, how our objective measures of acoustic distortion relate to the perceived quality of synthetic speech produced by ModelTalker. Results for these studies will be presented at the conference.

CONCLUSION:

The ModelTalker Speech Synthesis System uses the latest speech synthesis technology while addressing the specific needs and desires of users of augmented communication devices. The current trend in speech synthesis technology is a database-driven approach, in which large inventories of speech are recorded and stored, and then speech is synthesized using the longest stretches of recorded speech possible given the content of the database. ModelTalker is based on this database-driven approach. However, because recording huge databases could prove prohibitive to those with easily-fatigued voices, and the size of these databases could be too large for augmentative communication devices to handle, ModelTalker limits the size of the database and the amount of recording required by using biphones. Biphones allow for high quality, natural sounding synthesized speech from unrestricted text for words and phrases not in the database. Because the biphone speech segments originated in words recorded by the same person who recorded the words and phrases for the database, the "voice" of the synthesized speech will remain uniform throughout. ModelTalker will be able to plug into existing augmentative communication devices, allowing augmented communicators to communicate with a personalized voice.

ACKNOWLEDGEMENTS:

Work supported by NIDRR and the Nemours Research Programs.

REFERENCES:

Beutnagel, M., Conkie, A. and Syrdal, A.K. (1998). Diphone Synthesis using Unit Selection. Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, Australia, 185-190.

Campbell, W.N. and Black, A.W. (1996). Prosody and the selection of source units for concatenative synthesis. In Progress in Speech Synthesis. Berlin: Springer-Verlag.

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

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.