

ModelTalker Voice Recorder (MTVR) -
A System for Capturing Individual Voices for Synthetic Speech

Debra Yarrington, Chris Pennington, John Gray (AgoraNet, Inc)

H. Timothy Bunnell, James Polikoff, Kyoko Nagao, Jason Lilley
(Speech Research Laboratory, Al duPont Hospital for Children)

Abstract:

We describe the ModelTalker Voice Recorder (MTVR) software, a voice banking system for easily creating a personalized synthetic voice from recordings of an individual's speech. The resulting synthetic voice is SAPI 5 compliant and will work with Windows Operating Systems. The system has been updated both in terms of the interface that guides users through the process of creating a voice and the algorithms that are used to create the unique synthetic voice. A beta version of the system and over 60 synthetic voices are currently in use, largely by people with ALS but also by blind and low vision individuals. We will report the development progress on MTVR in particular while presenting synthetic voices created with our system, and discussing the successes and possible difficulties that potential users should be aware of to obtain successful outcomes.

Extended Abstract

Currently there are over 2 million people in the United States with severe communication disorders (AAC Demographic Information, 2007). Many of these individuals rely on Speech Generation Devices (SGD) to help them communicate. Some of these SGD are fairly low tech, in which a user has a board of images or alphanumeric characters and communication involves the user pointing to the board and the communication partner observing what the user is pointing at. Most high tech SGD try to facilitate communication with the addition of a number of different features, including synthetic speech output. Speech output allows users to communicate more naturally and lets the user communicate with those who cannot observe the user's gestures. Unfortunately, many current SGD are still using synthetic speech technology that is decades old. The synthetic speech in these SGD is both less intelligible and less natural sounding than state-of-the-art speech synthesizers (Bunnell et al., 2005).

Currently SGD users do not have the option of choosing their own unique voice. It is rare for someone using one of these voices to attain a unique sense of individual identity usually associated with one's voice, especially if the loss of speech occurs later in life. For example, the loss of identity can be more devastating for those who are losing their voice due to a degenerative disease such as Amyotrophic Lateral Sclerosis (ALS). For those with ALS, 75% use a SGD (AAC Demographic Information, 2007). To go from one's own voice to a synthetic sounding generic voice can be quite a loss to the individual and to family and friends who associate the voice with that person.

The ModelTalker Voice Recorder (MTVR) addresses several shortcomings inherent in many synthetic voices currently used with Speech Generation Devices. The MTVR system incorporates state of the art speech technology into creating unique, personal synthetic voices. Individuals record their own speech, which is labeled and stored in a database for concatenation. This voice banking process helps the resulting synthetic speech to retain many of the voice characteristics of the individual who made the recordings. Individuals who foresee a loss of voice function can record their voice ahead of time and use the resulting synthetic voice when needed. Those who either have always been nonspeaking or who have already lost their voice can either have a friend or family member use MTVR to record a voice for them or, as with other speech synthesis systems, they can choose from a number of previously created voices (selecting the one they feel best represents them).

This paper will report on the current state of the MTRV system, including updates and improvements to the user interface as well as advances in the voice creation algorithms. We will also present lessons learned from the growing number of users who have already used the ModelTalker System to create a personal synthetic voice.

Current State of MTRV

The current MTRV System guides users through the process of recording approximately 1650 words and phrases that will be used to create their synthetic voices. To ensure an optimal recording level, the system leads a user through a semiautomated audio calibration process. Once calibration is completed, the user is instructed to record and upload an initial set of phrases. Those phrases are quality checked and, if considered acceptable, the user receives the entire speech inventory to record. When the full inventory is recorded and uploaded, a synthetic voice customized for the user is generated. The synthetic voice is SAPI 5 compliant and will work with any current Microsoft Windows based hardware.

The consistency of the database of recorded speech is crucial to the quality of the resulting synthetic voice. Thus, much of the current work on the system has been focused on improving the consistency and accuracy of the recordings. Major improvements have been made to the calibration process and feedback functions. For example, a semiautomatic calibration procedure has been added to the system to check for volume level, silence threshold (background noise) and pitch range at the start of each session. MTRV also provides the user with feedback on loudness, pitch, and pronunciation accuracy for each recording.

In many database-based synthesis systems, a single user (generally a professional speaker) records hours of speech in a carefully controlled audio environment. The speech is both automatically labeled and (time intensively) hand corrected before being stored in a database for synthesis. Although the resulting voice is often of very high quality and can sound very natural with a versatile range of pitch and amplitude, this process can be both labor-intensive and expensive. Unfortunately, due to the realities of the population MTRV has been designed for, our system rarely has the luxury of a perfect audio environment or hours of actual speech recordings to work with. The database has a limited amount of recorded speech as a tradeoff to reduced recording time. Individuals with ALS and similar disorders who will be recording using the MTRV system do not have the time, nor do they often have the vocal strength to record hours' worth of speech. Our system requires about 45 minutes of recorded speech, but even that can take up to 6 hours of recording time. Since we are using a limited database of speech to create the synthetic voice, the recordings in the database must be relatively uniform in pitch and amplitude. Speech segment concatenation can become difficult if this is not the case. Accurate pronunciation is also critical. If a user mispronounces a word, with a limited database the system may use the mispronounced phoneme string during the synthesis process. For these reasons, accurate feedback for the user during the recording process is critical. Over the past year, we have improved the accuracy of feedback for these three factors so that the user can successfully record high quality speech at home without professional supervision.

Lessons Learned from Beta Testing

A beta version of the ModelTalker Voice Recording System has been distributed to users. Through this process we have identified the following critical factors that significantly affect the quality of the resulting synthetic voice.

1. Background Noise: While we as humans have the ability to adaptively filter out extraneous noises such as traffic, the television, or even children crying and dogs barking, when those noises are captured in recorded speech, they result in strange and unusual jarring noises that disrupt the flow of the synthetic speech. There is no way to filter out these noises once they are in the recordings. Furthermore, these noises make it difficult for the system to accurately label

segments of speech and may lead to segments being mislabeled. Thus it is very important that users record in a quiet environment.

2. Voice Quality: We have found that certain qualities lead to better sounding synthetic voices. Relatively strong and consistent voices usually produce high quality synthetic voices, although different voice qualities result in voices that sound more like the original speaker than others. However, voices that are weak or dysarthric do not always yield understandable synthetic voices. This is especially a problem for people with ALS. Often people with ALS do not begin to record until their voices have already degraded substantially or have become weak. Weak or dysarthric voices are difficult to label correctly, and often result in a poor quality synthetic voice. Even if the synthetic voice is of reasonable quality, it may not be representative of the speaker's original voice. Hence it is crucial that users with ALS or other degenerative disorders begin banking their voice as soon as possible before progression of the disease adversely affects their voice and stamina.

3. Hardware Issues: The MTRV System is designed to be downloaded and installed on any PC with the appropriate resources. Once installed, users can use their own microphone to record speech. A poor quality microphone or soundcard will result in low quality recordings. The recordings may include a low frequency humming noise. Like background noise, this noise is impossible to filter out completely, which will produce poor quality synthetic speech. Thus it is advisable to use the equipment recommended for use with MTRV. In general, a moderately priced (\$50 - \$100 retail) USB microphone can help avoid many of the above problems.

Future Directions

Based on feedback from users of the current beta version (downloadable from <http://www.modeltalker.com>), the user interface will be refined. The underlying speech synthesis engine will be ported to Windows Mobile based devices for ModelTalker voice playback. The length and complexity of the current set of required recorded utterances will be reduced. The ultimate goal is to develop a small set of simple phrases required for recording. The phrases will be ordered so that users can record them progressively, with the first set being crucial to the development of a synthetic voice and each subsequent set incrementally improving the quality of the resulting voice. Finally we will continue to investigate what voice qualities lend themselves to superior synthetic voices so we may better evaluate the chances of success for users in advance.

References

- 1 .AAC Demographic Information, Retrieved Oct. 8, 2007, from <http://aac.unl.edu/AACdemog.html>
2. Bunnell, H.T., Pennington, C., Yarrington, D. and Gray, J (2005). Automatic Personal Synthetic Voice Construction, InterSpeech 2005, 89-92.
3. ModelTalker Speech Synthesis System. Retrieved Oct. 8, 2007, from <http://www.modeltalker.com>