ModelTalker Voice Recorder – An Interface System for Recording a Corpus of Speech for Synthesis

Debra Yarrington, John Gray, Chris Pennington

AgoraNet, Inc.
Newark, DE 19711
USA
{yarringt, gray, penningt}
@agora-net.com

H. Timothy Bunnell, Allegra Cornaglia, Jason Lilley, Kyoko Nagao, James Polikoff,

Speech Research Laboratory
A.I. DuPont Hospital for Children
Wilmington, DE 19803, USA
{bunnell, cornagli, lilley,
nagao, polikoff}@asel.udel.edu

Abstract

We will demonstrate the ModelTalker Voice Recorder (MT Voice Recorder) – an interface system that lets individuals record and bank a speech database for the creation of a synthetic voice. The system guides users through an automatic calibration process that sets pitch, amplitude, and silence. The system then prompts users with both visual (text-based) and auditory prompts. Each recording is screened for pitch, amplitude and pronunciation and users are given immediate feedback on the acceptability of each recording. Users can then rerecord an unacceptable utterance. Recordings are automatically labeled and saved and a speech database is created from these recordings. The system's intention is to make the process of recording a corpus of utterances relatively easy for those inexperienced in linguistic analysis. Ultimately, the recorded corpus and the resulting speech database is used for concatenative synthetic speech, thus allowing individuals at home or in clinics to create a synthetic voice in their own voice. The interface may prove useful for other purposes as well. The system facilitates the recording and labeling of large corpora of speech, making it useful for speech and linguistic research, and it provides immediate feedback on pronunciation, thus making it useful as a clinical learning tool.

1 Demonstration

1.1 MT Voice Recorder Background

While most of us are familiar with the highly intelligible but somewhat robotic sound of synthetic speech, for the approximately 2 million people in the United States with a limited ability to communicate vocally (Matas et al., 1985), these synthetic voices are inadequate. The restricted number of available voices lack the personalization they desire. While intelligibility is a priority for these individuals, almost equally important is the naturalness and individuality one associates with one's own voice. Individuals with difficulty speaking can be any age, gender, and from any part of the country, with regional dialects and idiosyncratic variations. Each individual deserves to speak with a voice that is not only intelligible, but uniquely his or her own. For those with degenerative diseases such as Amyotrophic Lateral Sclerosis (ALS), knowing they will be losing the voice that has become intricately associated with their identity is not only traumatic to the individual but to family and friends as well.

A form of synthesis that incorporates the qualities of individual voices is concatenative synthesis. In this type of synthesis, units of recorded speech are appended. By using recorded speech, many of the voice qualities of the person recording the speech remain in the resulting synthetic voice. Different synthesis systems append different sized

Proceedings of the ACL-08: HLT Demo Session (Companion Volume), pages 28–31,
Columbus, June 2008.

segments of speech. Appending larger the units of speech results in smoother, more natural sounding synthesis, but requires many hours of recording, often by a trained professional. The recording process is usually supervised, and the recordings are often hand-polished. Because appending smaller units requires less recording on the part of the speaker, this is the approach the ModelTalker Synthesizer has taken. However using smaller units may result in noticeable auditory glitches at concatenative junctures that are a result of variations (in pitch, amplitude, pronunciation, etc.) between the speech units being appended. Thus the speech recorded must be more uniform in pitch and amplitude. In addition, the units cannot be mispronounced because each unit is crucial to the resulting synthetic speech. In a smaller database there may not be a second example of a specific phoneme sequence.

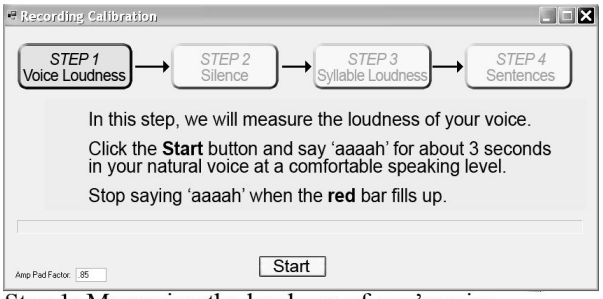
MT Voice Recorder expects that the individuals recording will be untrained and unsupervised, and may lack strength and endurance because of the presence of a degenerative disease. Thus the system is user-friendly enough for untrained, unsupervised individuals to record a corpus of speech. The system provides the user with feedback on the quality of each utterance they record in terms of pronunciation accuracy, relative uniformity of pitch, and relative uniformity of amplitude. Conference attendees will be able to experience this interface system and test all its different features.

1.2 Feature Demonstration

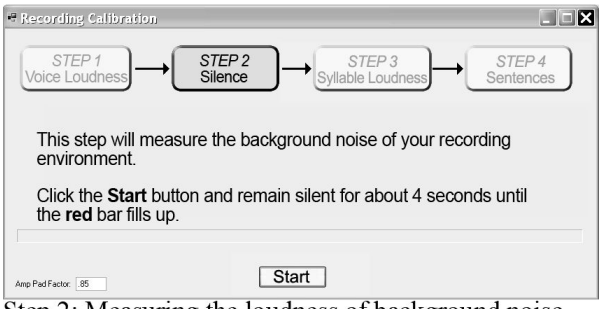
At the conference, attendees will be able to try out the different features of ModelTalker Voice Recorder. These features include automatic microphone calibration, pitch, amplitude, and pronunciation detection and feedback, and automatic phoneme labeling of speech recordings.

1.2.1 Microphone calibration

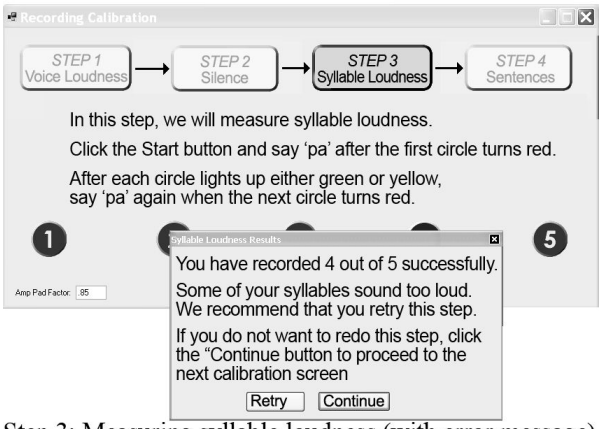
One important new feature of the MT Voice Recorder is the automatic microphone calibration procedure. In InvTool, a predecessor software of MT Voice Recorder, users had to set the microphone's amplitude. The system now calibrates the signal to noise ratio automatically through a step-by-step process (see Figure 1, below).



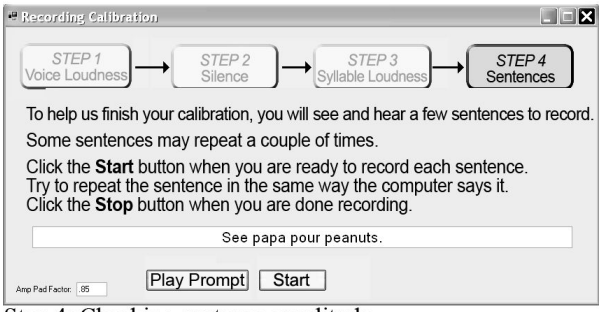
Step 1: Measuring the loudness of one's voice



Step 2: Measuring the loudness of background noise



Step 3: Measuring syllable loudness (with error message)



Step 4: Checking sentence amplitude

Figure 1: Automatic microphone calibration procedure

Using the automatic calibration procedure, the optimal signal to noise ratio is set for the recording session. These measurements are retained for future recording sessions in cases in which an indi-

vidual is unable to record the entire corpus in one sitting.

Once the user has completed the automatic calibration procedure, he will be able to start recording a corpus of speech. The interface has been designed with the assumption that individuals will be recording without supervision. Thus the interface incorporates a number of feedback mechanisms to aid individuals in making a high quality corpus for synthesis (see Figure 2, below).

1.2.2 Recording Utterances

The corpus was carefully chosen so that all frequently used phoneme combinations are included at least once. Thus it is critical that users pronounce prompted sentences in the manner in which the system expects. Alterations in pronunciation as small as saying /i/ versus /ə/ for "the," for example, can negatively affect the resulting synthetic voice. To reduce the incidence of alternate pronunciation, the user is prompted with both a text and an auditory version of the utterance.

1.2.3 Recording Feedback

Once an utterance has been recorded, the user receives feedback on the overall quality of the utterance. Specifically, the user receives feedback on the pitch, the overall amplitude, and the pronunciation of the recording.

Pitch: The user receives feedback on whether the utterance's average pitch is within range of the user's base pitch determined during the calibration process. Collecting all recordings within a relatively small pitch range minimizes concatenation costs during the synthesis process. MT Voice Recorder determines the average pitch of each utterance and gives the user feedback on whether the pitch is within an acceptable range. This feedback mechanism also helps to eliminate cases in which the system is unable to accurately track the pitch of an utterance. In these cases, the utterance will be marked unacceptable and the user should rerecord, hopefully yielding an utterance with more accurate pitch tracking.

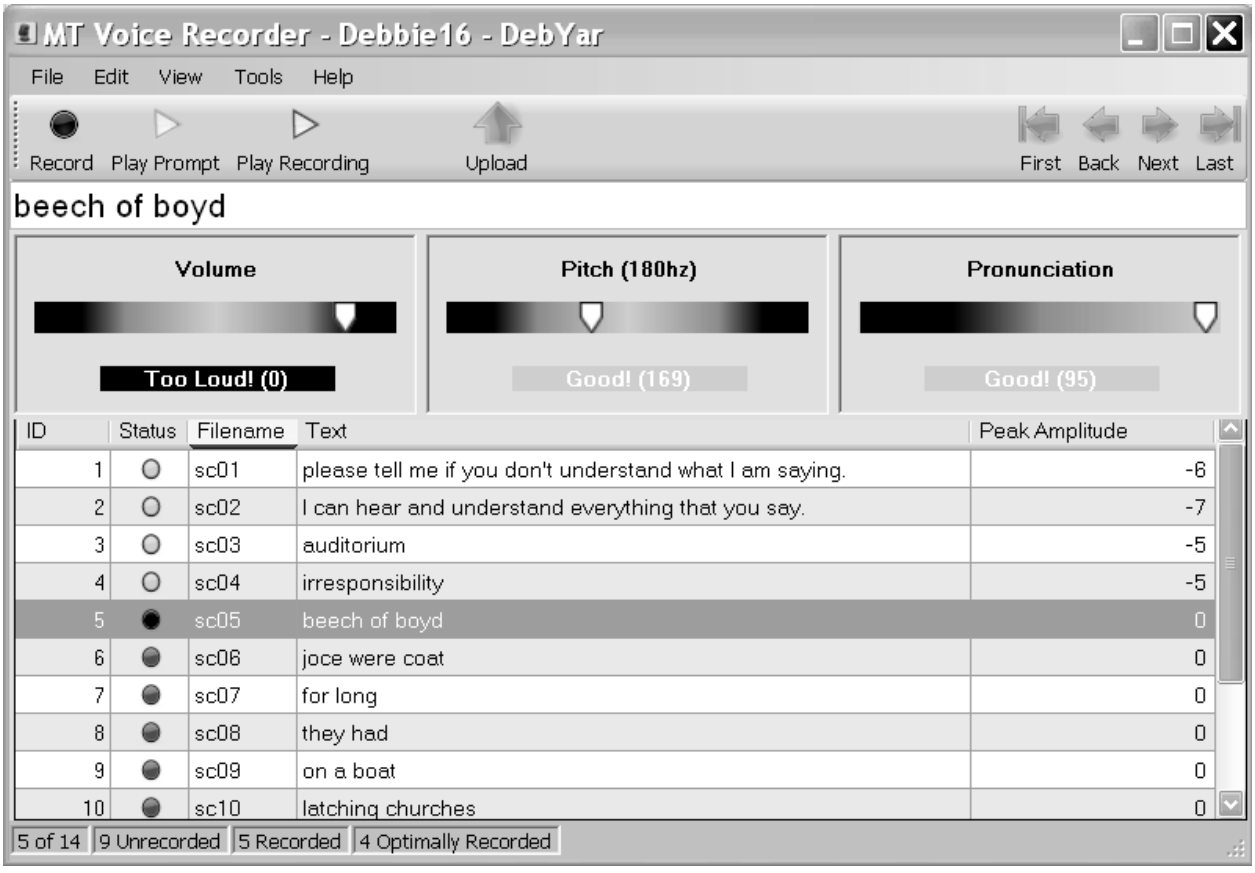


Figure 2: MT Voice Recorder User Interface

Amplitude: The user is also given feedback on the overall amplitude of an utterance. If the amplitude is either too low or too high, the user must rerecord the utterance.

Pronunciation: Each recorded utterance is evaluated for pronunciation. Each utterance within the corpus is associated with a string of phonemes representing its transcription. When an utterance is recorded, the phoneme string associated with the utterance is force-aligned with the recorded speech. If the alignment does not fall within an acceptable range, the user is given feedback that the recording's pronunciation may not be acceptable and the user is given the option of rerecording the utterance.

1.2.4 Automatic Phoneme Labeling

During the process of pronunciation evaluation, an associated phoneme transcription is aligned with the utterance. This alignment is retained so that each utterance is automatically labeled. Once the entire corpus has been recorded, alignments are automatically refined based on specific individual voice characteristics.

1.2.5 Other Features

The MT Voice Recorder also allows users to add utterances of their choice to the corpus of speech for the synthetic voice. These utterances are those the user wants to be synthesized clearly and will automatically be included in their entirety in the speech database. These utterances are also automatically labeled before being stored.

In addition, for those with more speech and linguistic experience, the system has a number of other features that can be explored. For example, the MT Voice Recorder also allows one to change settings so that the phoneme string, peak amplitude, RMS range, average F0, F0 range, and pronunciation score can be viewed. Users may use this information to more precisely adjust their utterances.

1.3 Synthetic Voice Demonstration

Those attending the demonstration will also be able to listen to a sampling of synthetic voices created using the ModelTalker system. While one of the synthetic voices was created by a professional speaker and manually polished, all other voices were created by untrained individuals, most of whom have ALS, in an untrained setting, with the recordings having no manual polishing.

2 Other Applications

Although the MTVR was designed specifically to record speech for the creation of a database that will be used in speech synthesis, it can also be used as a digital audio recording tool for speech research. For example, the MT Voice Recorder offers useful features for language documentation. An immediate warning about a poor quality recording will alert a researcher to rerecord the utterance. MT Voice Recorder employs file formats that are recommended for digital language documentation (e.g., XML, WAV, and TXT) (Bird & Simons, 2003). The recorded files are automatically stored with broad phonetic labels. The automatic saving function will reduce the time of recordings and the potential risk for miscataloging the files. Currently, the automatic phonetic labeling feature is only available for English, but it could be applicable to different languages in the future.

For more information about the ModelTalker System and to experience an interactive demo as well as listen to sample synthetic voices, visit http://www.modeltalker.com.

Acknowledgments

This work was supported by STTR grants R41/R42-DC006193 from NIH/NIDCD and from Nemours Biomedical Research. We are especially indebted to the many people with ALS, the AAC specialists in clinics, and other interested individuals who have invested a great deal of time and effort into this project and have provided valuable feedback.

References

Bird, S. and Simons, G.F. (2003). Seven dimensions of portability for language documentation and description. *Language*, *79*(3): 557-582.

Matas, J., Mathy-Laikko, P., Beaukelman, D. and Legresley. K. (1985). Identifying the nonspeaking population: a demographic study, *Augmentative & Alternative Communication*, *1*: 17-31.