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

Debra Yarrington*, John Gray*, Chris Pennington*, H. Timothy Bunnell, Allegra Cornaglia, Jason Lilley, Kyoko Nagao, James Polikoff
Speech Research Laboratory, Alfred I. duPont Hospital for Children, Nemours Children's Clinic, Wilmington, DE 19803
*AgoraNet, Inc. Newark, DE 19711

ABSTRACT

The ModelTalker Voice Recorder (MT Voice Recorder) is an interface system that lets individuals record and bank a speech database for the creation of a synthetic voice.

The system:

1. Guides users through an automatic calibration process that sets pitch, amplitude, and silence.
2. Prompts users with both visual (text-based) and auditory prompts.
3. Screens each recording for pitch, amplitude and pronunciation and gives users immediate feedback on the acceptability of each recording. Users can then rerecord an unacceptable utterance.
4. Automatically labels and saves recordings.
5. Creates a speech database 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.

BACKGROUND

- Approximately 2 million people in the United States have a limited ability to communicate vocally (Matas et al., 1985)
- 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 in their own unique voice, yet the restricted number of available voices lack the personalization they desire

Concatenative 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. However, with concatenative synthesis, noticeable auditory glitches may occur 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 uniform in pitch and amplitude. In addition, the units cannot be mispronounced or the resulting synthetic speech will contain the mispronunciation

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 (such as amyotrophic lateral sclerosis (ALS), or Lou Gehrig's disease) .

ACKNOWLEDGEMENTS

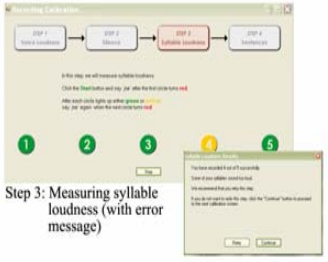
This work was supported by STTR grants R41/R42-DC006193 from NIH/NIDCD and from Nemours Biomedical Research. We are especially indebted to the many people with ALS, the AAC specialists in clinics, and other interested individuals who have invested a great deal of time and effort into this project and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

FEATURE OF MT VOICE RECORDER

AUTOMATIC MICROPHONE CALIBRATION PROCEDURE

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

Microphone Calibration
MT Voice Recorder now calibrates the signal-to-noise ratio automatically through a step-by-step process. With this, 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 individual is unable to record the entire corpus in one sitting

Recording Utterances
The recording corpus was carefully chosen so that all frequently used phoneme combinations are included at least once. Thus 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 pronunciations, the user is prompted with both a text and an auditory version of the utterance.

Recording Feedback
Once an utterance has been recorded, 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). 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.

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 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.

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.

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.

Recording Interface
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.

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.

CONTACT INFORMATION

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 or contact {yarringt, gray, penningt}@agora-net.com, {bunnell, polikoff, nagao, lilley, cornaglia}@asel.udel.edu

REFERENCES

[1] Bird, S. and Simons, G.F. (2003). Seven dimensions of portability for language documentation and description. *Language*, *79*(3): 557-582.
[2] Matas, J., Mathy-Laikko, P., Beaukelman, D. and Legresley. K. (1985). Identifying the nonspeaking population: a demographic study, *Augmentative & Alternative Communication*, *1*: 17-31.
and environment for statistical computing," 1.8.1 ed. Vienna, Austria: R Foundation for Statistical Computing, 2003.