INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

The ModelTalker Project: A web-based voice banking pipeline for ALS/MND patients

H. Timothy Bunnell, Jason Lilley, Kathleen McGrath

Nemours Biomedical Research, USA
{Bunnell,lilley,mcgrath}@nemoursresearch.org

Abstract

The Nemours ModelTalker supports *voice banking* for users diagnosed with ALS/MND and related neurodegenerative diseases. Users record up to 1600 sentences from which a synthetic voice is constructed. For the past two years we have focused on extending and refining a web-based recording tool to support this process. In this demonstration, we illustrate the features of the web-based pipeline that guides patients through the process of setting up to record at home, recording a standard speech inventory, adding custom recordings, and screening alternative versions of their voice and alternative synthesis parameter settings. Finally, we summarize results from 352 individuals with a wide range of speaking ability, who have recently used this voice banking pipeline.

Index Terms: Voice banking; speech synthesis; dysarthria; acoustic phonetics; speech disorders.

1. Introduction

Motor Neurone Disease (MND) or Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease of uncertain origin [1]. The disease is related to loss of motor neurons, resulting in muscle weakness and wasting. ALS/MND progresses rapidly, with 50% of patients dying within three years of disease onset. Depending on the onset pattern of the disease, patients may present with initial symptoms related to limb weakness, or with slurring of speech and difficulty swallowing. Most ALS/MND patients, but particularly the latter group of patients, with "bulbar" onset, rapidly develop significant dysarthria and typically require Augmentative and Alternative Communication (AAC) support for social interaction and communication with others.

Modern AAC devices or Speech Generating Devices (SGDs) allow users to communicate by inputting textual messages or selecting programmable symbols. Once a message has been entered, the device renders the message audibly either by playing back previously recorded speech, or by using text to speech synthesis. For many years, ALS/MND patients, if diagnosed before losing their speech, had the option of recording specific messages that could be played from an SGD on demand ("message banking" [2]) or using a device-supplied TTS "voice." However, as part of the ModelTalker project, our laboratory has provided an alternative to message banking called "voice banking" in which patients record enough speech to create a synthetic voice from the recordings.

The TTS technology used by the ModelTalker system has changed considerably since the system was first introduced as an extended diphone synthesis system [3] and is currently based on unit selection [4]. Additional changes such as parametric and hybrid synthesis are in development. One new design feature of the web-based pipeline is an auditioning process that will permit users to directly compare and select among different versions of their voice, including versions based on alternative synthesis technologies.

In this demonstration, we will describe and illustrate the current web-based pipeline for voice banking, provide examples of banked voices from users with speech ranging from normal to moderately dysarthric, and present results of analyses of the natural speech and synthetic voices from 352 recent users of the voice banking pipeline.

2. Voice Banking Pipeline

Figure 1 shows a flow chart of the major components of the current voice banking pipeline. On the left of Figure 1 are activities performed by registered users of the service from the point where they set up to record to the point where they are able to download a voice installer for Windows, Android, macOS, or iOS. Operations shown on the right in Figure 1 are those performed by staff or automated scripts in our lab.

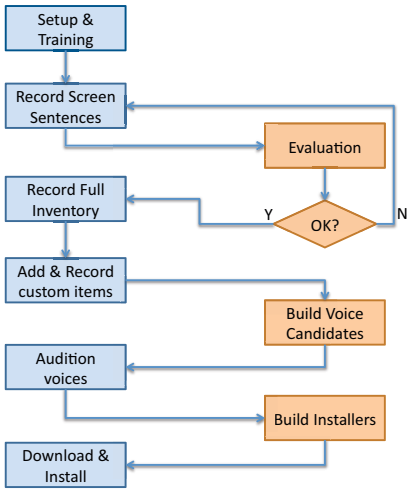

Figure 1: *Flowchart of voice banking pipeline*

Completion of the initial setup and screening steps notifies lab staff to review the user's recordings and determine if a) the audio is of acceptable quality, and b) if the user's speech and speaking style will allow creation of a unit selection voice. If either of these criteria are not met, feedback is provided to the user via a standardized email form. The screening process is iterated until both criteria are met or the user decides not to do voice banking.

Once approved for recording, users proceed at their own pace to complete the 1600 sentences we recommend (roughly one hour of running speech at a normal speaking rate). We do not actively monitor user progress during this stage and users

are quite variable in the amount of time they take to complete the recording, ranging from a few days to several months.

On completing the standard inventory, users are given an opportunity to provide custom recording material. A web form allows users to enter lists of person, place, or object names and a list of custom sentences or phrases to record. Recordings of these phrases are stored in the synthesis database in a manner that guarantees they will be rendered verbatim should the voice be asked to "synthesize" the exact phrase. Each of the person/place/object names is embedded in multiple brief sentence frames for recording.

When custom recording is completed (or declined by the user), a notice is sent to lab staff who then run the recordings through an automated voice construction process that results in the creation of six versions of the voice. Currently, the versions differ in terms of the criteria used for pruning questionable units from the synthesis database and whether or not an attempt is made to adapt our standard pronunciation dictionary to the speech of the user.

As a newly installed feature of the pipeline, all six voices from the build process are installed on the web server for use in an audition process that allows users to select options for synthesis (speaking rate, intonation control enabled, timing control enabled), and to compare all six voices with each other via a 2-alternative forced choice (2AFC) listening task in which the user hears 10 sentences from each of the six voices over a series of 30 paired comparison trials. The two highest scoring voices from the 2AFC task are then presented in an open listening task where the user can enter any text (up to about 2000 characters long) and compare its rendering by the two voices. The result of this is selection of a final voice version and synthesis parameter settings that are used to create voice installers.

3. Results

To summarize the results of operating the voice banking pipeline, we selected 352 recent users of the process. They varied in their speech from normal to moderately dysarthric. We used magnitude estimation based on the same two sentences as spoken by each individual to derive a dysarthria rating between 0 (no dysarthria) and 10 (moderate-severe).

For each user, we measured segment durations, jitter, shimmer, and amplitude modulation features in 200 sentences to examine the relationship between these features and dysarthria rating. All measures showed a significant linear trend associated with dysarthria rating. Figure 2 illustrates this for amplitude modulation rate in a mid-frequency band [5].

We also collected intelligibility measures for these users' synthetic voices using an SUS task. As expected (Figure 3), the synthetic speech becomes more difficult to understand as dysarthria rating increases. Notably this trend is well fit by a quadratic function and has no significant linear trend.

4. Conclusions

The web-based pipeline has allowed us to efficiently support voice banking for a substantial number of patients while requiring little staff time per user. Those who complete the process before they have become too dysarthric fare best in terms of the quality of their synthetic voice. For patients, recording as soon as possible after diagnosis is crucial because, while acoustic features associated with dysarthria progress linearly with increasing dysarthria, synthetic voices built from those recordings show quadratic growth in the difficulty with which they can be understood.

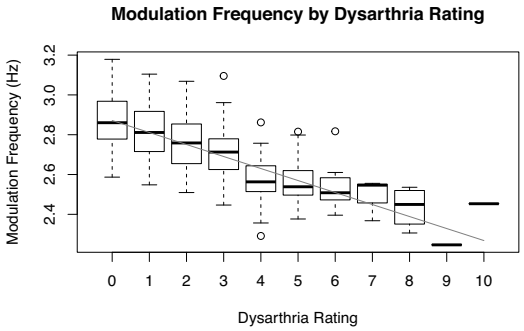

Figure 2: *Average mid-frequency modulation rate for speech recorded from voice bankers in each of 10 dysarthria rating categories.*

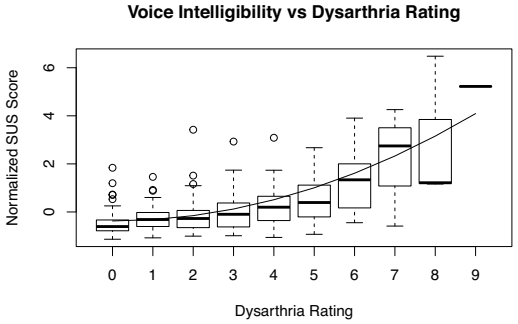

Figure 3: *Intelligibility as measured in terms of word-level edit distance between intended and transcribed synthetic speech from semantically unpredictable sentences.*

5. Acknowledgements

Work supported by Nemours Biomedical Research, and the Nemours Fund for Children's Health.

6. References

[1] J. D. Mitchell and G. D. Borasio, "Amyotrophic lateral sclerosis," *Lancet,* vol. 369, pp. 2031-41, Jun 16 2007.

[2] J. M. Costello. (2014, 4/1/2017). "*Message Banking, Voice Banking and Legacy Messages*." Available: http://www.childrenshospital.org

[3] H. T. Bunnell, S. R. Hoskins, and D. M. Yarrington, "A biphone constrained concatenation method for diphone synthesis," in *Proceedings of the 3rd International Workshop on Speech Synthesis*, ed Jenolan Caves, Australia, 1998.

[4] H. T. Bunnell, "Crafting Small Databases for Unit Selection TTS: Effects on Intelligibility," *Proceedings 7th International Speech Synthesis Workshop - SSW7,* p. CD only, 2010.

[5] J. M. Liss, S. LeGendre, and A. J. Lotto, "Discriminating dysarthria type from envelope modulation spectra," *J Speech Lang Hear Res,* vol. 53, pp. 1246-55, Oct 2010.