VocaliD: Personalizing Text-to-Speech Synthesis for Individuals with Severe Speech Impairment

Camil Jreige
Dept. of Electrical & Computer Engineering
Northeastern University
Boston, MA 02115
jreige.c@neu.edu

Rupal Patel
Dept. of Speech Language Pathology
Northeastern University
Boston, MA 02115
617-373-5842
r.patel@neu.edu

H. Timothy Bunnell
Speech Research Laboratory
A.I. DuPont Hospital for children
Wilmington, DE 19803
302-651-6835
bunnell@asel.udel.edu

ABSTRACT

Speech synthesis options on assistive communication devices are very limited and do not reflect the user's vocal quality or personality. Previous work suggests that speakers with severe speech impairment can control prosodic aspects of their voice, and often retain the ability to produce sustained vowel-like utterances. This project leverages these residual phonatory abilities in order to build an adaptive text-to-speech synthesizer that is intelligible, yet conveys the user's vocal identity. Our VocaliD system combines the source characteristics of the disordered speaker with the filter characteristics of an age-matched healthy speaker using voice transformation techniques, in order to produce a personalized voice. Usability testing indicated that listeners were 94% accurate in transcribing morphed samples and 79.5% accurate in matching morphed samples from the same speaker.

Categories and Subject Descriptors

[**Computer Applications**]: Life and Medical Sciences – Health

General Terms

Design, Human Factors, Algorithm

Keywords: Text-to-Speech Synthesis, Assistive Communication, Speech Generation Devices, Dysarthria

1. INTRODUCTION

Voice quality is unique to each individual and inextricable from personality, self-image, and the perceptions of others. While voice quality may not matter for many text-to-speech (TTS) applications, it is essential for over two million Americans who have severe speech and motor impairments that require them to use assistive communication aids with TTS synthesis to speak on their behalf [5]. Current commercially available devices use a limited number of synthetic voices that sound unnatural and do not represent the user in terms of age, gender or personality. Thus it is not uncommon to see a 9 year old girl using an adult male voice or for several children in a classroom to be using the same voice. Thus, many users lack a personal connection with the voice on their device which may lead to low technology adoption rates.

This project aims to build an adaptive text-to-speech synthesizer for individuals with severe speech impairment. Given the severity of their impairment, conventional methods of voice morphing and voice banking will not work. However, despite highly distorted speech sound production, many individuals retain control of prosodic cues such as pitch, loudness and duration [7, 8]. These prosodic cues are among the acoustic cues that signal speaker identity [6]. We leverage this preserved ability toward building a synthesizer that is intelligible yet conveys the user's identity.

2. APPROACH

The Source-filter theory of speech production asserts that the source (the vocal folds for voiced sounds) and the filter (the vocal tract) are independent in their contribution to the acoustic speech signal. Previous studies suggest that the phonatory source may be relatively spared in speech impairment [7, 8] and that source-related cues make an important contribution in conveying speaker identity [6]. It may be possible to harness this preserved ability toward building a personalized voice despite the speaker's inability to modulate filter characteristics which result in highly distorted speech sound articulation [1].

We have developed a system called VocaliD (for 'vocal identity') in which the source characteristics of a speech-impaired target user are modulated by the time-varying filter characteristics of a surrogate age and gender matched healthy speaker in order to produce articulate speech that retains the phonatory quality of the target user. Figure 1 provides a flow diagram of the system architecture in which sustained vocalizations produced by the target user (source) are used to transform a biphone inventory produced by an age and gender matched healthy speaker (filter) which are then concatenated to output the personalized voice.

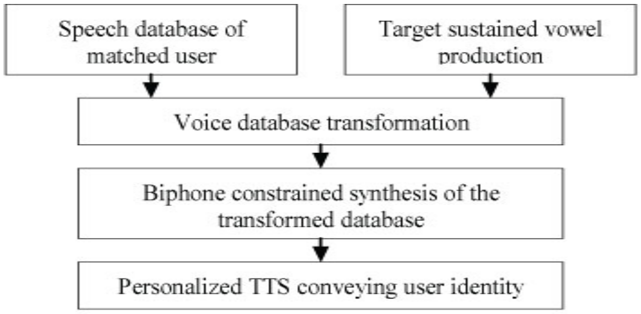
Speech database of matched user
Target sustained vowel production
Voice database transformation
Biphone constrained synthesis of the transformed database
Personalized TTS conveying user identity

Figure 1. Flow diagram of the VocaliD architecture.

3. IMPLEMENTATION

Traditional voice transformation techniques such as articulatory inversion and spectral mapping [4, 10] require relatively large databases of clearly articulated speech which is not feasible for

ASSETS'09, October 25–28, 2009, Pittsburgh, Pennsylvania, USA.
ACM 978-1-60558-558-1/09/10.

this population. To address this limitation, the speech database is produced by an age and gender matched healthy speaker. These productions are then used to create a blended database that imprints the spectral and prosodic properties of the fluent speech onto source information derived from the target speaker.

The process begins by recording a sustained vowel (approximately 2-3 seconds long) from a target speaker with speech impairment. The vowel is inverse filtered to remove formant-like features and approximate the target speaker's glottal source function (hereafter we will refer to this as the target glottal source). To capture a fluent database, we use the InvTool software from the ModelTalker project [2] to collect a corpus of 1050 words and sentences from an age and gender matched healthy speaker. Utterances in this corpus are phonetically labeled and pitch-tracked to derive segmental and prosodic information. For each utterance, we construct a new target source function by frequency modulating the target glottal source to match the fundamental frequency contour of the fluent utterance produced by the healthy speaker. This target source function is passed through a channel vocoder in which the channel amplitudes are modulated by the dynamic spectral properties of the fluent utterance. The result is an utterance that retains the segmental and suprasegmental properties of the fluent utterance while adopting source characteristics associated with the target speaker. When all utterances in the corpus have been processed in this way, the ModelTalker BCC (biphone constrained concatenation) program is used to create a database for unit selection synthesis to enable generation of unrestricted English text.

4. USABILITY TESTING

Usability testing consisted of assessing the intelligibility of the VocaliD system and determining whether listeners could accurately match pairs of morphed samples from the same speaker. Twenty-four monolingual English listeners were recruited to transcribe 220 synthetic sentences created by transforming the vocalizations of four target users with age and gender matched healthy speakers. The four target users consisted of two healthy children (1M, 1F: mean age 10.8 years) and two children with speech impairment due to cerebral palsy (1M, 1F: mean age 11.1 years). Thus, eight transformed voices were generated using combinations of gender matched and unmatched target and control speakers. Stimuli consisted of unpredictable sentences selected from the Harvard Sentence [3] and Semantically Unpredictable Sentences (SUS) generated using the susgen program [9]. Results indicated that listener accuracy ranged from 94% to 97.6% for the eight transformed voices. There were no significant differences in intelligibility between male and female voices or for gender un/matched pairs. The same listeners also rated the naturalness of each recording on a likert scale from 0 to 5, where 0 was defined as computerized and unnatural to 5, which was defined as natural and human sounding. The average naturalness rating across all voices was 3.5. Voices transformed using a gender matched healthy speaker were judged to be only nominally more natural.

An additional group of 24 monolingual English listeners were recruited for the similarity test. In this task, listeners heard a random sample of 232 pairs of transformed sentences from the eight voice database. For each pair, listeners indicated whether the samples were produced by the same or different speakers. Across all voices, the average listener accuracy was 79.5%.

In summary, VocaliD offers a potentially viable approach to creating a personalized synthetic voice for individuals with severe speech impairment. Harnessing residual phonatory abilities for personalizing speech output on assistive communication devices has the potential to impact technology adoption rates as well as improve satisfaction, social integration and overall quality of life. While the intelligibility scores for the VocaliD system were encouraging, further work is required to achieve high levels of similarity between target and transformed voices. Efforts aimed at developing more sophisticated methods for extracting the phonatory characteristics of the target speaker and using hybrid (model-based and concatenative) approaches to speech generation are currently underway.

5. ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation Grant # 0712821. We thank Matthew Fortier, Jesse Winans and Jamie Becker for their contribution to the software development and usability testing respectively.

6. REFERENCES

[1] Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *J. of Speech, Language & Hearing Research.* 35(2), 296-308.

[2] Bunnell, H. T., Gray, J., Pennington, C. and Yarrington, D. 2005. A system for creating personalized synthetic voices. Presented at ASSETS, Baltimore, MD.

[3] IEEE Recommended Practices for Speech Quality Measurements. (1969). *IEEE Transactions on Audio and Electroacoustics*, 17, 227-46.

[4] Kain, A. and Macon, M. (1998). Personalizing a speech synthesizer by voice adaptation. Proceedings of the Third ESCA/COCOSDA International Speech Synthesis Workshop, 225-230.

[5] Matas, J., Mathy-laikko, P., Beukelman, D., and Legresley, K. (1985). Identifying the non-speaking population: A demographic study. Augmentative and Alternative Communication, 1, 17-31.

[6] Matsumoto, H., Hiki, S., Sone, T. and Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics,* 21(5), 428- 436.

[7] Patel, R. (2003). Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy. *J. of Speech, Language & Hearing Research*, 46, 1401-1415.

[8] Patel, R., and Campellone, P. (2009). Production and identification of contrastive stress in dysarthria. *J. of Speech Language & Hearing Research*, 56, 206-222.

[9] SUSGEN (Semantically Unpredictable Sentence Generator) software available at http://www.asel.udel.edu/speech/download/susgen.tgz

[10] Toda, T., Black, A., and Tokuda, K. (2005). Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter ICASSP, Philadelphia, Pennsylvania.