Evaluation of speaker mimic technology for personalizing SGD voices

Esther Klabbers[1], *Alexander Kain*[1], *and Jan P.H. van Santen*[2]

[1]Biospeech, Inc., Lake Oswego, OR
[2]Center for Spoken Language Understanding at OHSU, Beaverton, OR
klabbers,kain@biospeech.com, vansanten@cslu.ogi.edu

Abstract

In this paper, we demonstrate the use of state-of-the-art speech technology to transform speech from a source speaker to mimic a particular target speaker with the intention of providng personalized voices to users of Speech Generating Devices (SGDs). This speaker mimicry (SM) capability allows us to use high-quality acoustic inventories from professional speakers and transform them to a different target speaker using a very limited set of sentences from that speaker. This technology targets future SGD users who still have a limited vocabulary or available previous recordings. The results of a perceptual study show that listeners can identify which SM voices most resemble their respective target voices.[1]

Index Terms: speech synthesis, voice transformation, prosody modeling, perceptual evaluation

1. Introduction

A wide range of individuals cannot communicate by voice. Voice enabled Augmentative and Alternative Communication (AAC) devices, also called Speech Generating Devices (SGDs), are often the only means available through which these individuals can communicate. While many SGDs are currently available, most of them lack the important ability to generate customized speech that mimics aspects of the user's past or intermittently available speech. Modern concatenative speech synthesis technology can mimic a given speaker's voice, by excising speech fragments from a recorded acoustic inventory and recombining these into output speech using sophisticated algorithms.

The only system currently available that allows SGD users with sufficient speaking ability to record speech for the purpose of creating their own synthetic voices is the ModelTalker system [1, 2]. The recording tool that is part of this system prompts the user to record sentences with a total of about 1650 words and phrases, the first 80 of which include words and phrases that are likely to be of need to users of SGD devices in a variety of daily contexts. The remaining words and phrases were selected to cover all diphones of American English in a broad range of phonetic and prosodic contexts that the synthesizer can use to produce an unlimited number of sentences. Another set of 200 "expressive" sentences is recorded as well. The recording process is said to take approximately 68 hours over 3 or 4 days. The ModelTalker system was shown to have a better word error rate in an intelligibility task than 5 out of 6 other concatenative synthesizers (which were not named) [2]. However, that version of ModelTalker used speech data from the free Arctic data collection initiative, which contains a larger set of sentences than the ModelTalker recording tool offers, and it uses semi-professional speakers [8].

The most important factor in creating high-quality concatenative synthesis output is the ability of the speaker whose recordings are used to create the acoustic inventory to speak consistently over a long period of time. For this reason, the consistent speech of professional speakers is used for the recordings. However, even professional speakers show inconsistencies in their voice quality, average pitch, and speaking rate, when recording during long sessions, or when recording multiple sessions [7].

Many SGD users cannot meet these requirements because they have already lost the capability to speak or because they cannot speak with adequate consistency of pronunciation. We therefore propose a Speaker Mimicry (SM) system in which only a small set of recordings is required from the SGD user ($<$ 50 sentences) to adapt a high-quality synthetic voice based on a professional speaker to sound like their own. In order for the resulting speech to sound like the target speaker, both the spectral characteristics and the prosodic characteristics (the duration and intonation) of the source speaker have to be adapted to the characteristics of the target speaker.

We use Voice Transformation (VT) technology to adapt the spectral characteristics and Prosody Mimicry (PM) to adapt the prosodic characteristics. Using VT technology, we can transform speech spoken by a "source" speaker into speech that is perceived as spoken by a specific "target" speaker. The amount of training recordings is far less than what is needed for a high-quality acoustic inventory. The basis of our PM method is provided by concise quantitative models for capturing prosodic signatures [10, 16, 17, 19, 20]. These models go well beyond characterizing a signature in terms of such coarse parameters as F_0 range or average speaking rate, and capture such highly speaker-dependent features as the amount of lengthening of stressed syllables or the magnitude of the drop in F_0 at the end of a declarative utterance, which, in combination, define the prosodic signature.

2. Research

2.1. Creating a Library of Limited Speech Recordings

As a proof-of-concept we recorded limited acoustic diphone inventories for multiple speakers. We recorded 12 typical,

[1] This project was funded by NIH STTR grant 1R41DC008712-01: User Adaptation of AAC Voices and the Nancy Lurie Marks foundation project "In Your Own Voice: Personal Augmentative and Alternative Communication Voices for Minimally Verbal Children with Autism Spectrum Disorders" and was a collaboration between Biospeech, Inc. and OHSU. OHSU and the authors have a significant financial interest in BioSpeech, Inc., a company that may have a commercial interest in the results of this research and technology. This potential conflict was reviewed and managed by OHSU and the Integrity Program Oversight Council.

healthy speakers: 4 male adult speakers, 4 female adult speakers, 2 male children and 2 female children. The sentences that were recorded were determined as follows. First, we selected 20 sentences from a larger set of phonetically rich sentences (IEEE/Harvard Sentences [11]). These 20 sentences served as the target set (Data Set 1). The target set is used in the evaluation, as a reference of how well a natural voice can produce those sentences. Then, we constructed 276 sentences that included 257 diphones and 19 triphones (Data Set 2). These diphones and triphones were necessary to produce the 20 test sentences using a concatenative TTS system. We included triphones for vowels followed by /l/ or /r/ because these consonants have a large impact on the pronunciation of the preceding vowel.

Finally, we selected a set of sentences for training the voice transformation (Data Set 3). We used a greedy algorithm [18] to search a transcription of Switchboard conversations [3]. The transcription text was transformed into diphone and triphone strings so that the greedy algorithm could search for the smallest set of sentences to cover the concatenative units in Data Set 2. The Switchboard conversations represent a casual speaking style that is in line with our goal of being able to perform Speaker Mimicry using spontaneous speech. A set of 46 sentences was found to provide complete coverage. In this project each speaker served both as a target and a source for the VT so each speaker uttered 342 sentences (Data Sets 1, 2, and 3 combined). The ultimate application will only require a small training set to be recorded by the target speaker.

The recordings were made in a recording studio using a data-collection computer with a hard drive and three flat-panel monitors, a high-quality condenser microphone in a special enclosure Whisperroom, a high-quality A/D sound module, and a mixing board. A special-purpose data-collection program was run on the computer. As sentences were read aloud, the recordings were stored in 16 KHz, 16-bit waveforms. One monitor was used by a technician to administer the program for data collection, and the other two monitors displayed the sentences to be read for recording. One of these was placed inside the enclosure for the speaker to see. Each sentence was preceded by a spoken example so that the speakers knew exactly what to say. Our ultimate goal is to be able to use arbitrary recordings of target speakers, such as old video recordings, but for this project we used only studio-quality recordings.

2.2. Creating Synthetic Voices with a Restricted Set of Diphones

We used the recordings from Data Set 2 to create 12 synthetic voices that consisted of the diphone and triphone characteristics of the source speaker's voices. The set of diphones and triphones in Data Set 2 was sufficient to synthesize the 20 test sentences of Data Set 1. The locations of phoneme boundaries in the recordings were determined using an automated forced-alignment system [5]. These boundary labels were manually corrected to optimize their accuracy. It was imperative to the VT process for these labels to reflect what was actually being said. We normalized the volume using the Snack toolkit [13] to ensure that differences in loudness did not influence the subjective ratings in the perceptual tests. We used an automatic cut-point detector [12] to determine the start and end boundaries of each diphone. We computed pitch marks using a Matlab script based on research by Stylianou [14].

2.3. Training the synthetic voices to mimic target voices

2.3.1. Voice Transformation

Voice transformation is the process of changing the spectral characteristics of a source speaker's voice into that of a target speaker's voice [15]. The first step in transforming the voices is to compute feature vectors describing the spectral characteristics from the source and target voice and to automatically time-align them so that a one-to-one mapping can be made from a source to a target feature vector. The feature vectors describing the spectral characteristics are usually based on a representation of the source-filter speech model [9]. Often, the filter is represented by Linear Prediction Coding (LPC) coefficients. These LPC parameters are usually converted to a number of alternative representations with more desirable properties, such as the ability to interpolate between parameters.

Gaussian mixture models (GMMs) provide a probabilistic approach to VT in which a continuous transformation function is trained [15]. Kain extended the GMM method to perform a joint density estimation of the GMMs in which both the source and target feature vectors were used for estimating the transformation function [6]. We have developed two VT algorithms, the **maximal VT** method (VTmax)and the **minimal VT** method (VTmin):

VTmax uses an LSF vocoder speech model to train the transformation and generate new speech files. For the LSF vocoder, a sample rate of 16 kHz was used and the LSF order was 18. Each LSF vector was computed on asynchronous, 50% overlapping, Hanning-windowed, 25-ms frames. The mapping function between source and target speaker was implemented using a joint-density, Gaussian Mixture Model (GMM) regression function using 24 mixture components. The program consists of three stages: the analysis stage, the learning stage and the synthesis stage. In the analysis stage, we use Data Set 3 for each speaker. We select one speaker to be the source speaker and another speaker to be the target speaker. The source and target speakers are always selected from the same gender and age range. Thus we have 28 possible transformations from source to target speaker (as we do not transform voices into themselves). The learning stage takes the LSF parameters produced in the analysis stage and learns a mapping from the LSF parameters of the source speaker to the LSF vectors of the target speaker. During the synthesis stage, spectral features selected as a result of a unit search on the acoustic inventory, are transformed in-place, and are then subsequently synthesized, thus integrating the VT technology directly in the TTS engine.

VTmin is useful for cases in which only low-quality existing voice recordings are available. VTmin learns the correspondences of the formants between two speakers and performs a piece-wise linear warping function in the frequency domain to map the formants from the source speaker to those of the target speaker. The locations of formants vary due to the shape and size of the vocal tract. Thus they are phoneme-specific and speaker-specific. VTmin captures speaker-specific vocal tract differences. Because it relies on formant frequencies only, the quality of the recordings is not crucial since these parameters can be extracted reliably even from noisy speech waveforms. As such it is irrelevant whether one has access to high quality recordings or lower quality recordings such as from a video camera. An additional benefit of VTmin is that it can be applied to people whose pronunciation abilities are not complete, i. e. they are able to produce some sounds but have problems with other sounds. VTmin uses a global piece-wise linear warping function in the frequency domain to transform all phonemes,

Condition	Spectrum 1	Prosody 1	Spectrum 2	Prosody 2	Effect on discriminability
A	Natural_A	Natural_A	Natural_B	Natural_B	Upper Limit
B	Diphones_A	Natural_A	Diphones_B	Natural_B	Effect of Diphones
C	DIphones_A	Synth_A	Diphones_A	Synth_B	Effect of Prosody Only
D	Diphones_A	Synth_Generic	VTmax	Synth_Generic	Effect of VTmax Only
E	Diphones_A	Synth_A	VTmax	Synth_B	Effect of VTmax + Prosody
F	Diphones_A	Synth_Generic	VTmin	Synth_Generic	Effect of VTmin Only
G	Diphones_A	Synth_A	VTmin	Synth_B	Effect of VTmin + Prosody

Table 1: Stimulus conditions

A few missing data points for phonemes not present in the target speaker's vocabulary will not adversely affect the outcome.

2.3.2. *Prosody Mimicry*

Speaker mimicry relies largely on prosody as each speaker has unique prosodic signatures. For durational characteristics we decided to only control the average speaking rate. In future research, we will explicitly model specific speakers' durational patterns the Sums-of-Products model [16, 17] with speaker adaptation [4]. The pitch is controlled by setting parameters for phrase curve start, mid and end values and accent amplitudes for initial, medial, and final accents. These parameters are used in the OGI Festival TTS system to generate synthetic prosody patterns and are based on the Generalized Linear Alignment Model of intonation (GLAM) [10, 19, 20]. These parameters provide more detailed information about the pitch contours generated than global parameters such as the F_0 range. For this study, the parameters were estimated by manually inspecting F_0 contours for the sentences in Data Set 1 for each speaker and averaging them over all sentences. In future research, we intend to develop automatic parameter extraction methods.

3. Perception experiment

A perception experiment was performed testing speaker identification of both VTmax and VTmin. To reduce the number of stimuli we performed a pre-test to determine which source voice was best suited for transformation to a particular target voice. We found that not every source voice was equally suitable to be transformed to a given target voice. However, the optimal source-target combinations were the same regardless of whether we used VTmax or VTmin. Since there were only two children for each gender, their source-target combinations were fixed. We synthesized the sentences using the TTS system with integrated VT capabilities developed for this purpose. The phoneme durations and intonation contours were obtained separately by running scripts with SABLE markup in Festival. The scripts used a generic Sums-of-Products duration model for generating phoneme durations [16, 17], and a global speaking rate value was added for each speaker. The intonation was generated using an implementation of the GLAM [19, 20, 10]. The SABLE markup language was modified to allow for GLAM parameters to be passed on in a SABLE script. Table 1 lists the 6 stimulus conditions that were presented to the listeners. Condition A presents two natural sentences to give the upper limit of speaker identification. Condition B presents diphone synthesis with natural prosody to show the effect of creating diphone synthesis for the speakers. Condition C presents diphone synthesis with diphones from the same speaker and synthetic prosody from different speakers to show the effect of prosody only on speaker identification. Condition D shows the effect of VTmax with generic synthetic prosody and condition E shows the effect of VTmax with personalized synthetic prosody. Condition F shows the effect of VTmin with generic prosody and condition G shows the effect of VTmin with personalized synthetic prosody.

The evaluation experiment was administered to 12 listeners. They listened to the stimuli using a desktop computer and headphones in a quiet office environment. All listeners had self-declared normal hearing. Since the task was to distinguish speaker identity, non-native speakers of American English were allowed to participate. The listening test was administered using a computerized test controlled via the mouse. Listeners listened to sentence pairs and then indicated whether they thought the sentences were spoken by the same person or two different people on a 4-point scale (1: Certainly Different; 2: Probably Different; 3: Probably Same; 4: Certainly Same). Both the stimulus order and the order in the sentence pair were randomized. In each sentence pair, the sentence content for the sentences was different. The experiment included 7 different conditions (A-G) as displayed in Table 1. The list of 196 stimuli contained an equal number of pairs where the predicted outcome was the same or different.

4. Analysis results

For each listener and each condition (A–G) we computed the percentage of correct responses. The four-point scale was collapsed to a binary scale where 1 indicated same and 0 indicated different. The results support two conclusions. First, referring to Figure 1, we conclude that in all but one condition (F) performance was significantly above chance (using a two-tailed t-test criterion at $p < 0.01$). However, since condition F uses generic prosody (i. e. not speaker specific), it is not relevant for the project since we would always customize the prosody to the target speaker in our product. Second, planned t-tests showed that each component of the proposed solution (i. e. VTmax, VTmin, and PM) had significant impacts: (1) With PM, VTmax was significantly better than generic diphones (comparing conditions E and C, $t(11) = 2.27, p < 0.05$, one-tailed); (2) PM produced significantly better results (comparing conditions E and D, as well as G and F, $t(11) > 3.0, p < 0.05$, one-tailed, in both cases); (3) With PM, VTmin also produced significant results ($t(11) = 1.81, p < 0.05$, one-tailed).

In summary, these results show that VTmax worked well, that VTmin (in combination with prosody mimicry) improves performance compared to generic diphones, and that prosody mimicry is a particularly critical component both by itself and in its ability to amplify the effects of VT.

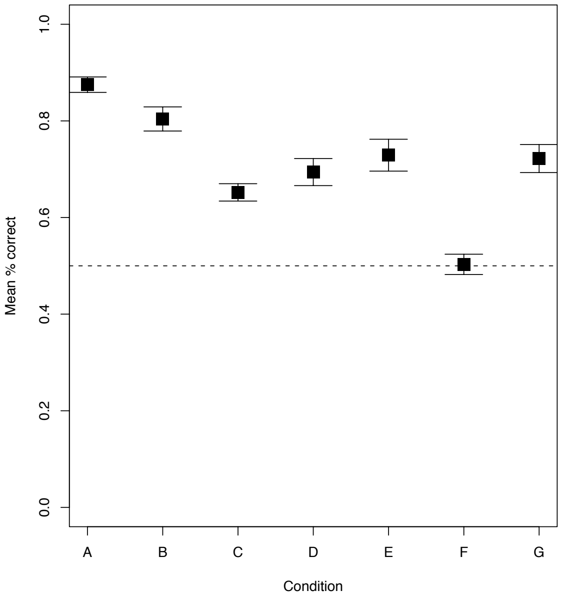

Figure 1: Results of the listening test. The solid blocks represent the average percentage correct per condition, and the solid lines the standard errors computed over the 12 listeners.

5. Discussion

This project has succeeded in several key areas. We have shown that we can record speech from different speakers and create custom synthesized voices for them. We have developed algorithms for state-of-the-art voice transformation to create new synthetic voices. While the quality of the transformation is better for some speakers than for others, we believe that the key concept is valid and can be developed into a product in the future. One consequence of this finding is that we need to record several canonical speakers as the source speakers to find an optimal match for any target voice.

In the future we will further improve the VT algorithms to produce more natural sounding speech. We will also implement a duration adaptation algorithm that adapts the duration model trained for a canonical speaker to fit the durational patterns of the target speaker using the same small set of sentences that will be used for SM training. We will develop parameter estimation algorithms for automatically extracting phrase and accent curve parameters from the pitch contour using GLAM. And we will develop an SGD interface on a netbook. Actual SGD users will provide feedback on the design choices we make as well as the canonical voices we will record. In addition we will evaluate the quality of the SM capability not only in terms of speaker discriminability but also in terms of intelligibility and naturalness.

6. References

[1] H. Bunnell and J. Gray and C. Pennington and D. Yarrington. Automatic construction of concatenative speech synthesis databases for AAC. In *American Speech Language and Hearing Association Conference*, Philadelphia, PA, 2004.

[2] H. Bunnell and C. Pennington and D. Yarrington and J. Gray. Automatic personal synthetic voice construction. In *Proceedings of EUROSPEECH*, Lisbon, Por tugal, 2005.

[3] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 517–520, 1992.

[4] W. Gu, C. Shih, and J. van Santen. An Efficient Speaker Adaptation Method for TTS Duration Model. In *Proceedings of EUROSPEECH*, Budapest, Hungary, 1999.

[5] J.P. Hosom. *Automatic time alignment of phonemes using acoustic-phonetic information*. PhD thesis, Oregon Graduate Institute, Beaverton, OR, 2000.

[6] A. Kain. *High Resolution Voice Transformation*. PhD thesis, OGI School of Science & Engineering at Oregon Health & Science University, Beaverton, OR, 2001.

[7] H. Kawai and M. Tsuzaki. Voice quality variation in a long-term recording of a single speaker corpus. In S. Narayanan and A. Alwan, editors, *Text to Speech Synthesis: New Paradigms and Advances*. Prentice Hall, 2005.

[8] J. Kominek and A. Black. The CMU ARCTIC speech databases for speech synthesis research. *Tech. Rep. CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, 2003.

[9] J. Markel and A. Gray. *Linear Prediction of Speech*. Springer Verlag, New York, NY, 1976.

[10] T. Mishra and J. van Santen and E. Klabbers. Decomposition of pitch curves in the general super positional model. In *Proceedings of Speech Prosody*, Dresden, Germany, 2006.

[11] Institute of Electrical and Electronic Engineers. IEEE recommended practice for speech quality measurements, 1969.

[12] J. Olive, J. van Santen, B. Möbius, and C. Shih. Synthesis. In R. Sproat, editor, *Multilingual Text-to-Speech Synthesis: The Bell-Labs Approach*. Kluwer, Boston, MA, 1998.

[13] K. Sjölander and J. Beskow. Wavesurfer - an open source speech tool. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 2000.

[14] Y. Stylianou. Removing linear mismatches in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(3), 2001.

[15] Y. Stylianou. Voice Transformation In *Springer Handbook of Speech Processing*, T. Benesti, M. Sondhi, and Y. Huang (eds). Springer Verlag, Berlin Heidelberg, 2008, pp. 489–502.

[16] J. van Santen. Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546, 1992.

[17] J. van Santen. Deriving text-to-speech durations from natural speech. In G. Bailly, C. Benoît, and T. Sawallis, editors, *Talking machines: Theories, models and designs*. Elsevier, 1992.

[18] J. van Santen and A. Buchsbaum. Methods for optimal text selection. In *Proceedings of EUROSPEECH*, pages 553–556, Rhodes, Greece, 1997.

[19] J. van Santen and B. Möbius. A model of fundamental frequency contour alignment. In A. Botinis, editor, *Intonation: Analysis, Modeling, and Technology*. Cambridge University Press, 1999.

[20] J. van Santen and T. Mishra and E. Klabbers. Estimating phrase cur ves in the general super positional intonation model. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004.