Personalizing synthetic voices for people with progressive speech disorders: judging voice similarity

S.M. Creer¹, S.P. Cunningham¹, P.D. Green¹ and K. Fatema²

¹University of Sheffield, UK
²University of Kent, UK

S.Creer@dcs.shef.ac.uk, S.Cunningham@sheffield.ac.uk,
P.Green@dcs.shef.ac.uk, kf66@kent.ac.uk

Abstract
In building personalized synthetic voices for people with speech disorders, the output should capture the individual’s vocal identity. This paper reports a listener judgment experiment on the similarity of Hidden Markov Model based synthetic voices using varying amounts of adaptation data to two non-impaired speakers. We conclude that around 100 sentences of data is needed to build a voice that retains the characteristics of the target speaker but using more data improves the voice. Experiments using Multi-Layer Perceptrons (MLPs) are conducted to find which acoustic features contribute to the similarity judgments. Results show that mel-cepstral distortion and fraction of voicing agreement contribute most to replicating the similarity judgment but the combination of all features is required for accurate prediction. Ongoing work applies the findings to voice building for people with impaired speech.

Index Terms: HMM-based speech synthesis, voice similarity measure, voice banking

1. Introduction

Diminishing neurological function due to conditions such as Parkinson’s disease or motor neurone disease (MND) leads to a progressive deterioration in an individual’s ability to produce speech. As motor control is lost, the severity of impairment increases and intelligibility of the speech decreases.

Speech technology can be used to try and help the individual to communicate more easily and many individuals with this severe speech impairment use voice output communication aids (VOCAs). Current VOCAs generally rely on a switch, keyboard or touch screen for input, or in the case of the VIVOCA (voice input voice output communication aid) [1], spoken commands. The VIVOCA, which this paper is related to, is designed as a communication aid for individuals who may find it difficult to use conventional input mechanisms. For output, VOCAs either produce speech that has been pre-stored or, as in this work, synthesize speech to order. When an individual’s condition has progressed to the point where their speech is completely unintelligible to unfamiliar communication partners, the VOCA output effectively becomes the voice of the user.

The synthetic voice of a VOCA must be able to communicate the message intelligibly and it must be sufficiently natural to motivate the user to use it frequently. Increasing the ease of interaction for listeners reduces the potential for social withdrawal. The voice output should also represent the user’s vocal identity appropriately. Preserving the identity of an individual overcomes social distance imposed by using such a device, re-associating the output with the user more closely. It preserves the voice as an identifier of the individual and matches listeners’ expectations of the output, again facilitating ease of interaction. Using the speaker’s own vocal characteristics also allows social bonds to form through associations with features realized in that voice rather than using features which do not match with the individual user and could provoke negative attitudes or disassociations from group membership.

To build personalized voices for people with progressive speech loss, a method must be used which requires only a small amount of data. This is because people with these conditions find it difficult to produce a large volume of speech. Additionally, their speech may already have begun to deteriorate at the point of intervention. HMM (Hidden Markov Model) based speech synthesis has the potential to facilitate construction of voices using a relatively small amount of speech data. It is reported that the HTS (‘H Triple S’ - HMM-based speech synthesis system) tool kit [2] requires approximately 100 sentences, or 6-7 minutes of continuous speech data to produce intelligible, natural-sounding speech output which captures the vocal identity of the person who recorded the data when adapting from speaker-independent or ‘average voice’ models towards those of the target speaker [3].

As a preliminary study to building voices using impaired speech this paper reports on how to ascertain how acceptable a voice would be to listeners, in terms of both naturalness and the match to the original speaker. We report on experiments which seek to substantiate the claims for HTS and techniques to model the judgments of listeners about new synthetic voices.

In section 2 we describe an experiment in which listeners were asked to make similarity judgments, comparing voices built using varying amounts of non-impaired speech data with the original speech of an individual speaker. This replicates a ‘voice banking’ situation where speech has been recorded prior to deterioration in order to use the data to build a voice prosthesis. In order to gain an insight into which auditory features correlate with listeners’ judgments of voice similarity, we trained neural networks to replicate these judgments using a range of features. This is discussed in section 3.

2. Listening experiment

An experiment was performed to investigate how much speech data is required to build a synthetic voice that listeners judge to be sufficiently similar to the original speaker.
2.1. Method

2.1.1. Stimuli

Synthetic voices were constructed for two normal male speakers. Speaker 1 was a professional broadcaster from Barnsley, UK, aged 51 at the time of recording. His voice is distinctive with a definite accent indicative of the Barnsley area and well-known through his broadcasting work. Speaker 2 was a university lecturer from north west England, aged 32 at the time of recording and has a general Northern British English accent.

Recordings were made in a single-walled IAC acoustically-isolated chamber, using a Bruel & Kjaer (B&K) type 4190 0.5 inch microphone located approximately 30 cm in front of the speaker. The signal was pre-amplified using a B&K Nexus model 2690 conditioning amplifier and digitized at 16 kHz using a Tucker Davis Technologies System 3 RP2.1 processor.

Both speakers recorded all the sentences from set A of the Arctic database [4], which is comprised of 593 sentences between 5 and 15 words in length and covers all dialects of US English. A set of 25 sentences that had 8 or fewer words were extracted from the dataset and were used as stimuli in the listening experiment. The remaining sentences were used to construct the sets used to build the different synthetic voices.

The synthetic voices were built using HTS version 2.1 (internal) using a 138-dimensional feature vector containing: 40 mel-cepstral coefficients (including the zeroth coefficient), log F0, 5 band aperiodicity values and the deltas and delta-deltas of each feature. The starting point for adaptation is a speaker-independent ‘average voice’ model trained on multiple speakers with a large amount of data. Adapting from this leads to fewer errors in estimating observations that can occur due to lack of data from the target speaker. The average voice model was constructed using the full Arctic data sets (1132 sentences) as spoken by 6 male speakers: 4 US English speakers, 1 Canadian English speaker and 1 Scottish English speaker. This was the only average voice available at the time of conducting the experiments.

The speaker-specific adapted voices were built using the first 10, 100 and 500 sentences from the data set. These were distinct from the test sentences. Average voice sentences and vocoded (analysed and synthesized using STRAIGHT [5]) original speech sentences from speaker 1 and speaker 2 were also included to allow the participants to continually calibrate their rating by presenting stimuli that should represent the extremes of the rating scale. The vocoding step was necessary to normalize for differences introduced by the sound reproduction process. Each listener heard a total of 375 stimuli: 25 sentences of each of the stimuli in 5 conditions (average voice, voices adapted with 10, 100 and 500 sentences and the resynthesized original - the target). The stimuli were presented as two separate experiments, with one speaker tested per experiment.

2.1.2. Participants

The participants were all native British English speakers aged between 23 and 48 with no reported hearing, language or speech impairment. The speaker 1 experiment had 7 participants (6 male, 1 female) and the speaker 2 experiment had 10 participants (8 male, 2 female). There was some overlap between the participants for both experiments, although the two experiments took place with an interval of two months.

2.1.3. Procedure

The experiments were conducted in a single-walled IAC acoustically-isolated chamber and the stimuli were presented over Sennheiser HD 515 headphones. Participants were initially presented with a set of four examples of the original speech from the target speaker, randomly chosen from the recorded set but different from the test set. They were asked to listen to all four examples and were told that this was the target speaker to which they would be comparing the experimental stimuli. They could listen to each example an unlimited number of times. Participants could listen to the examples any point during the experiment if they wished.

Participants were then presented with a stimulus sentence and asked to rate it on a scale of 1 (sounds like a totally different speaker) to 7 (sounds like the same speaker). This scale was chosen to encourage a discriminating set of responses across the different voices. The rating scale was visible on screen at all times. Participants indicated their response to each presentation using either the keyboard or the mouse. The next stimuli followed immediately once the previous rating had been entered. All of the stimuli were presented to each participant and were split into three blocks of 125 sentences with an opportunity to pause between blocks. Each experiment took a total of approximately 30 minutes to complete.

2.2. Results

The results of the two experiments are presented in Figure 1, panels a and b. The probability of listeners to rate a synthetic voice as more similar to the target increases with the amount of data used to adapt the voice. The responses of listeners were significantly different for the voices under test for both speakers (Speaker 1: \( \chi^2(4) = 27.47, p < 0.001 \); speaker 2: \( \chi^2(4) = 38.73, p < 0.001 \)). Wilcoxon tests corrected for multiple comparisons show that for speaker 1 responses for the average voice differ significantly from those for the 100, 500 and target voices (\( p < 0.0125 \)). For speaker 2 the only significant difference in responses was found between the average voice and the target voice (\( p < 0.0125 \)).

2.3. Discussion

The results of these experiments partially support previous work that, with this technique, 100 sentences is the minimum amount of adaptation data to achieve close similarity to the target speaker. For speaker 1 listeners tended to rate the 100 voice as being more similar to the target than the average voice, however for speaker 2 ratings were less conclusive.

The average voice used in this experiment is an influencing factor in the similarity to the speaker as it sounds North American in accent. Any influence of that accent on the voices built introduces more distance from the target speaker. Using a more appropriate average voice for the target speakers’ voices is likely to reduce the amount of data needed to produce a voice with high similarity. Speaker 1’s voice has a more distinctive and therefore more recognizable voice than that of speaker 2 which could account for the difference in results between the two speakers. Speaker 2 was also known to some of the participants and some commented on the difference between the reference sentences and their recollection of the individual’s actual voice. This difference may be that speaker 1’s professional speaking experience means that he sounds more natural when reading prompts and more like the voice that participants may have been exposed to on television and radio.
3. Acoustic features influencing listener judgments

What acoustic features influence listeners’ judgments of speaker similarity? To address this question we considered the 18 features in table 1, comparing their values for each target utterance $T$ and corresponding synthesized approximation $S$ presented to listeners in the experiment of section 2. The mel-cepstral distortion (feature 1) was obtained from the global cost of an asymmetric Dynamic Time Warping (DTW) alignment between 12 cepstral coefficients for $T$ and $S$, not using overall energy. The remaining features for each $T$ and $S$ were extracted using the PRAAT toolkit [6] and compared by averaging point-by-point absolute differences along the DTW alignment path. Feature 4 is the fraction of frames in which (according to the PRAAT pitch extractor) $T$ and $S$ are both voiced or both unvoiced. Features 5 to 18 were measured only in sections where both $T$ and $S$ were voiced.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mel-cepstral distortion</td>
</tr>
<tr>
<td>2</td>
<td>Intensity mean</td>
</tr>
<tr>
<td>3</td>
<td>Intensity variance</td>
</tr>
<tr>
<td>4</td>
<td>Fraction of voicing agreement</td>
</tr>
<tr>
<td>5</td>
<td>F0 mean</td>
</tr>
<tr>
<td>6</td>
<td>F0 variance</td>
</tr>
<tr>
<td>7</td>
<td>F1 mean</td>
</tr>
<tr>
<td>8</td>
<td>F1 variance</td>
</tr>
<tr>
<td>9</td>
<td>F1 bandwidth mean</td>
</tr>
<tr>
<td>10</td>
<td>F1 bandwidth variance</td>
</tr>
<tr>
<td>11</td>
<td>F2 mean</td>
</tr>
<tr>
<td>12</td>
<td>F2 variance</td>
</tr>
<tr>
<td>13</td>
<td>F2 bandwidth mean</td>
</tr>
<tr>
<td>14</td>
<td>F2 bandwidth variance</td>
</tr>
<tr>
<td>15</td>
<td>F3 mean</td>
</tr>
<tr>
<td>16</td>
<td>F3 variance</td>
</tr>
<tr>
<td>17</td>
<td>F3 bandwidth mean</td>
</tr>
<tr>
<td>18</td>
<td>F3 bandwidth variance</td>
</tr>
</tbody>
</table>

Table 1: Description of the features used to train the MLP

We trained Multi-Layer Perceptrons (MLPs) to model listeners’ judgments given the extracted features (or a subset of them). Each $T$-$S$ pair, each listener and each presentation generated an MLP dataset item in which the input layer received the feature values and there were 7 output nodes corresponding to the listener’s judgment $j$. The output node target value was 1 for the $j$th output node and 0 for the remainder. The trained MLP therefore estimated the probability that a listener’s rating of the similarity between the synthetic utterance $S$ and the target utterance $T$ would take each of the 7 allowed scores. 50% of the dataset, randomly chosen, was used for training and the remainder for testing.

Figure 1, (c) and (d) show that the MLP can reproduce listeners’ judgments ((a) and (b)) accurately, given all 18 features. There were 10 hidden nodes here but that figure is not critical. Listener and MLP results can be compared quantitatively using the mean squared point-by-point error in the $5 \times 7$ condition/judgment matrices (Frobenius norm). Previously, Kominek, Schultz and Black [7] have used mel-cepstral distortion to calibrate synthesized voice quality and Yamagishi and Kobayashi [3] have advocated this feature together with F0 error and vowel duration error. Our comparison, however, shows a more complex picture. While the most important features for the MLP are indeed mel-cepstral distortion and fraction of voicing agreement, (similar to vowel duration error in [3]), and the formant bandwidth features have least influence, the non-linear modelling of many features is required for accurate prediction. This is illustrated in table 2, which gives the MLP prediction error for various feature combinations.

<table>
<thead>
<tr>
<th>Input features</th>
<th>speaker 1</th>
<th>speaker 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0031</td>
<td>0.0046</td>
</tr>
<tr>
<td>4</td>
<td>0.0041</td>
<td>0.0038</td>
</tr>
<tr>
<td>1,4</td>
<td>0.0016</td>
<td>0.0016</td>
</tr>
<tr>
<td>1,2,4</td>
<td>0.0013</td>
<td>0.0012</td>
</tr>
<tr>
<td>1,2,4,5</td>
<td>0.0011</td>
<td>0.0011</td>
</tr>
<tr>
<td>1,2,4,5,7,11,15</td>
<td>0.0008</td>
<td>0.0009</td>
</tr>
<tr>
<td>1,2,3,4,5,6,7,8,11,12,15,16</td>
<td>0.0007</td>
<td>0.0005</td>
</tr>
<tr>
<td>all</td>
<td>0.0006</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Table 2: MLP prediction error for different input feature combinations. Mean values over 10 trials are given. Differences greater than 0.0001 are statistically significant.

4. Conclusions

The listening experiment confirmed that around 100 sentences are needed to build a suitably similar voice to a target speaker, although using an average voice which is closer to the target is likely to produce a better result and may require less adaptation data. Differences were found between the two target speakers which may be related to the proficiency of the speaker to read naturally as one speaker was a professional broadcaster.

The experiments with MLPs show that listeners’ judgments can be modelled closely by comparing classical acoustic features. These comparisons were possible because we had synthetic and target versions of the same sentences. MLPs trained on data from one of our speakers perform reasonably well as listener response predictors for the other speaker (MLP prediction error $\sim 0.01$). Thus it may be possible to use a pre-trained MLP to replace listening tests when assessing voice similarity for a new speaker.

More generally, we have shown that mel-cepstral distortion and voicing agreement have the largest influence on similarity judgment. However, experimentation with the MLP shows that explicit formant information contributes to accurate modelling of listener responses.

5. Work in progress

These experiments can inform work on building voices for speakers with speech impairments. The ‘voice banking’ approach as detailed above has to be altered where the adaptation data is affected by the individual’s condition. Ideally, speaker characteristics should be captured without replicating the impairment in the synthesis. HTS adapts speaker models towards a target speaker and stores F0, spectral information, signal aperiodicity and duration as separate models. This structure allows information from the average voice model to be selectively substituted for that of the speaker to compensate for the effects of the impairment. Results from the above experiments suggest that spectral information, F0 and voiced passage duration contribute significantly to perceptual similarity judgments. Where the speaker’s spectral information is disordered, using selected intelligible sections for adaptation from the speech-impaired...
data allows the retention of the speaker’s spectral characteristics in the model rather than substituting this information from the average voice model. Ideally, the speaker’s F0 and duration information should also be retained, although disordered timing of articulators occurs frequently in these conditions. Preliminary results, not reported here, show that output synthesized using average voice durations with target speaker spectral and F0 information, taken from models adapted with data selected for intelligibility, retains some of the speaker similarity. Using an average voice which shares speaker characteristics with the target, such as accent, may increase the voice similarity further.

6. Acknowledgements

HTS 2.1 internal version and average voice models were provided by Junichi Yamagishi. HTS uses STRAIGHT with permission from Hideki Kawahara. Sarah Creer was funded by EPSRC. The VIVOCA project was funded by the NIHR’s Innovation for Innovation programme.

7. References


