Using HMM-based Speech Synthesis to Reconstruct the Voice of Individuals with Degenerative Speech Disorders

Christophe Veaux, Junichi Yamagishi, Simon King

Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
{cveaux, jyamagis}@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

When individuals lose the ability to produce their own speech, due to degenerative diseases such as motor neuron disease (MND) or Parkinson’s, they lose not only a functional means of communication but also a display of their individual and group identity. In order to build personalized synthetic voices, attempts have been made to capture the voice before it is lost, using a process known as voice banking. But, for some patients, the speech deterioration frequently coincides or quickly follows diagnosis. Using HMM-based speech synthesis, it is now possible to build personalized synthetic voices with minimal data recordings and even disordered speech. In this approach, the patient’s recordings are used to adapt an average voice model pre-trained on many speakers. The structure of the voice model allows some reconstruction of the voice by substituting some components from the average voice in order to compensate for the disorders found in the patient’s speech. In this paper, we compare different substitution strategies and introduce a context-dependent model substitution to improve the intelligibility of the synthetic speech while retaining the vocal identity of the patient. A subjective evaluation of the reconstructed voice for a patient with MND shows promising results for this strategy.

Index Terms: HTS, Voice Cloning, Voice Reconstruction, Assistive Technologies

1. Introduction

Degenerative speech disorders have a variety of causes that include Multiple Sclerosis, Parkinson’s, and Motor Neuron Disease (MND). MND, also known in the USA as Amyotrophic Lateral Sclerosis (ALS), primarily affects the motor neurons in the brain and spinal cord. This causes a worsening muscle weakness that leads to a loss of mobility and difficulties with swallowing, breathing and speech production. As speech becomes disordered, these patients often use a voice output communication aid (VOCA) that embeds a text-to-speech synthesizer (TTS) as speech output function. Currently, most of these devices use a unit-selection synthesizer [1]. The resulting synthetic speech is highly intelligible and can approach human-level naturalness but there are limited opportunities to personalize it to more closely match the vocal identity of the patient. However, for VOCA users, speech synthesis is not an optional extra for reading out text, but a critical function for social communication and identity display. Therefore a personalized VOCA where the synthetic voice has the characteristics of the user is a long requested feature. In order to build personalized VOCA's, several attempts have been made to capture the voice before it is lost, using a process known as voice banking. One example of this approach is ModelTalker [2], a voice building service that can be used on any home computer to build a concatenative synthesis voice. Wants Inc. in Japan also provides a commercial voice building service for individuals called “Polluxstar”. This is based on a hybrid speech synthesis system [3] using both unit selection and statistical parametric speech synthesis [4] to achieve a natural speech quality. However, all these speech synthesis techniques require a large amount of recorded speech in order to build a good quality voice. Moreover, the recorded speech data must be as intelligible as possible, since the data recorded is either used directly or partly as the voice output. This requirement makes such techniques more problematic for those patients whose voices have started to deteriorate. Therefore, there is a strong motivation to improve the voice banking and voice building techniques, so that patients can use their own synthetic voices, even if their speech is already disordered at the time of recordings. A first approach is to try to separate out the disorders from the recorded speech. In this way, Rudzicz [5] has proposed a combination of several speech processing techniques. In its system, a high-pass filter is first used to remove voicing errors in consonants, then TD-PSOLA is applied to stretch out irregular duration, and finally STRAIGHT spectral morphing enables to separate out confusable formants. However, some disorders cannot be simply filtered out by signal processing techniques and a model-based approach seems more appropriate. Kain [6] has proposed a voice conversion framework for the restoration of disordered speech. In its approach, the low-frequency spectrum of the voiced speech segment is modified according to a mapping defined by a Gaussian mixture model (GMM) learned in advance from a parallel dataset of disordered and target speech. The modified voiced segments are then concatenated with the original unvoiced speech segments to reconstruct the speech. This approach can be seen as a first attempt of model-based voice reconstruction although it relies only on a partial modeling of the voice components. Recently, the HMM-based speech synthesis technique has been investigated to create personalized VOCA's [7] [8]. HMM-based speech synthesis statistically represents the acoustic parameters of a source-filter model of the speech. This approach has two major advantages over existing methods for voice banking and voice building. First, it is possible to use existing speaker-independent voice models pre-trained over a number of speakers and to adapt them towards a target speaker. This process known as speaker adaptation [9] requires only a very small amount of speech data. The second advantage of this approach is that we can control and modify various components of the adapted voice model in order to compensate for the disorders found in the patient’s speech. We call this process “voice reconstruction”. In this paper, we compare different strategies of voice reconstruction using the HMM-based synthesis framework and introduce a new technique of context-dependent model substitution that helps to reduce the articulation disorders of the synthetic speech without altering its vocal identity.
2. Acoustic Modeling

In this section, we present shortly the structure of the acoustic models used in HMM-based synthesis. This structure will allow some manipulations of the synthetic voices. We use the state-of-the-art HMM-based speech synthesizer, referred to as HTS. The acoustic models used in HTS are context-dependent hidden semi-Markov models (HSMMs), which are HMMs with explicit state duration distributions. The state output distributions represent three separate streams of acoustic parameters that are required for driving a vocoder. The vocoder used here is STRAIGHT and the acoustic streams correspond respectively to the fundamental frequency (log-F0), the band aperiodicities and the mel-cepstrum, including their dynamics. For each stream, additional information is added to further describe the temporal trajectories of the acoustic parameters, such as their global variances over the learning data. Finally, separate decision trees are used to cluster the state durations probabilities and the state output probabilities.

3. Speaker Adaptation

The speaker adaptation process allows the creation of a synthetic voice clone of any patient’s speech with a limited amount of data. This method starts with a speaker-independent model, or “average voice model”, learned over multiple speakers and uses model adaptation techniques drawn from speech recognition such as maximum likelihood linear regression (MLLR), to adapt the speaker independent model to a new speaker. Speaker adaptive HMM-based synthesis requires as little as 5-7 minutes of recorded speech from a target speaker in order to generate a personalized synthetic voice. This provides a much more practical way for patients to voice-bank their speech. This approach has already been applied to construct a synthetic voice for a patient prior to a laryngectomy operation [8]. A similar approach can also be used for patients with degenerative diseases before the diseases affect their speech. However, we do not want to reproduce the symptoms of a vocal problem if the speech has already been disordered at the time of the recording. That’s the aim of the voice reconstruction introduced in the next section.

4. Voice Reconstruction

If the speech is already disordered at the time of recording, speaker adaptation techniques will clone not only the vocal identity of the patient but also the symptoms of its vocal problem, resulting in a disordered synthetic voice. Therefore we need to remove speech disorders from a synthetic voice, so that it sounds more natural and more intelligible. Repairing synthetic voices is conceptually similar to the restoration of disordered speech mentioned in Section 1, but we can now exploit the acoustic models learned during the training and the adaptation processes. The structure of HTS means that the acoustic models generate the sequence of parameters for each stream separately: duration, log-F0, band aperiodicity and mel-cepstrum. This sows some reconstruction of the voice by substituting models or information from the average voice to compensate for any disorder that occurs in the patient's data. This is illustrated in Figure 1. Although disordered speech perceptually deviates considerably from normal speech in many ways, it is known that its articulatory errors are consistent [10] and hence relatively predictable [11]. Therefore we can pre-define a substitution strategy for a given condition, to some extent. For example, patients with MND often have a disordered speaking rate, contributing to loss of intelligibility of the speech. The substitution of the state duration models (as shown in Figure 1) enables the timing disruptions to be regulated at the phoneme, word, and utterance levels. Furthermore, MND speakers often have breathy or hoarse speech, in which excessive breath through the glottis produces unwanted turbulent noise. In such cases, we can substitute the band aperiodicity models to produce a less breathy or hoarse output. We present different substitution strategies in the following part of this section.

Figure 1: The structure of the acoustic models means that there can be a substitution of state output or state duration models between the average voice model and the patient voice model to compensate for any deterioration in the patient’s speech.
4.1. Baseline model substitution

The first application of the voice reconstruction using HMM-based synthesis was conducted for a patient with Parkinson’s in Sheffield [7]. In this experimentation, the following models and information are substituted:

- Duration and aperiodicity models
- Global variances of log-F0, aperiodicity and mel-cepstrum

These parameters are the less correlated with the speaker identity and their substitution can fix some disorders such as slow speaking rate and excessive hoarseness. However, this substitution strategy cannot correct articulation disorders. It will be referred to as the baseline strategy in this paper.

4.2. Component-wise model substitution

Since the state output distributions have diagonal covariance matrix, we can substitute a component independently from the others. This component-wise substitution strategy allows to substitute the parts of the mel-cepstrum and log-F0 streams that are the less correlated with the speaker identity. In this way, we can further reduce some disorders without altering the voice identity. In particular, we substitute the mean and variance for the following components:

- 1st coefficient of the mel-cepstrum (energy)
- High-order coefficients of the mel-cepstrum
- Dynamics coefficients of the mel-cepstrum and log-F0
- Voiced/Unvoiced weights

The substitution of the high order static coefficients and the dynamics coefficients of the mel-cepstrum will help to reduce the articulation disorders without altering the timbre. In our implementation, we replace all static coefficients of order N>60. The substitution of the dynamics coefficients of the log-F0 will help to regulate the prosodic disorders such as monotonic F0. Finally the replacement of the voiced/unvoiced weights will fix the breathiness disorders. The duration models, aperiodicity models, and global variances are also substituted as in the baseline strategy. We will refer to this method as the component-wise strategy.

4.3. Context-dependent model substitution

In the two previous strategies, the model substitutions are independent of the context. However, in HTS, the acoustic models are clustered after their context by separate decision trees. We can use this contextual information to further refine the model substitution. For example, some MND patients cannot pronounce correctly the plosives, the approximants and the diphthongs. In these contexts, it is preferable to substitute all the mel-cepstrum coefficients in order to enhance the intelligibility of the speech. Therefore, we have defined an extension of the component-wise strategy, in which the mel-cepstrum models are entirely substituted for the diphthongs, plosives, and approximants. We will refer to this method as the context-dependent strategy.

5. Average Voice Models

Ideally, the average voice used for the speaker adaptation and the voice reconstruction should be close to the vocal identity of the patient. On the other hand, a minimum number of speakers are necessary to train robust average voice models. Therefore, we have created a database of about 300 healthy voice donors with various accents (Scottish, Irish, Other UK). Each speaker recorded about one hour of speech (400 sentences). The speakers are clustered according to their gender and their regional accent in order to train specific average voice models. It has been shown that gender dependent average voice models make a better starting point for speaker adaptation [9]. Furthermore, we conducted a listening test of speaker similarity and the results showed that the speakers have generally been rated to be similar to one another according to their accent [12].

6. Experiment

The three substitution strategies presented in Section 4 were evaluated for the case of a MND patient. This patient was a 45 years old Scottish male that we recorded twice. A first recording of one hour (500 sentences) has been made just after diagnosis when he was at the very onset of the disease. At that time, his voice did not show any disorders and could still be considered as “healthy”. A second recording of 15 minutes (50 sentences) has been made 10 months later. He has then acquired some speech disorders typically associated with MND, such as excessive hoarseness and breathiness, disruption of speech fluency, reduced articulation and monotonic prosody. For this experiment, the same male-Scottish average voice model was used to create all the synthetic voices. This average voice was trained on 17 male Scottish speakers using 400 sentences each giving a total of 6800 sentences. The synthetic voices used in this experiment are shown in Table 1. The synthetic voice created from the first recording of the patient (“healthy” speech) was used as the reference voice for the subjective evaluations. This choice of a synthetic voice as reference instead of the natural recordings was done to avoid any bias due to the loss of quality inherent to the synthesis. Three different reconstructed voices were created from the second recording of the patient (“impaired” speech). Each of these voices corresponds to a different substitution strategy.

<table>
<thead>
<tr>
<th>Voice</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV</td>
<td>Average Voice used for speaker adaptation</td>
</tr>
<tr>
<td>HC</td>
<td>Voice Clone of the “healthy” speech (1st recording)</td>
</tr>
<tr>
<td>IC</td>
<td>Voice Clone of the “impaired” speech (2nd recording)</td>
</tr>
<tr>
<td>IR_v0</td>
<td>Reconstructed voice using baseline strategy</td>
</tr>
<tr>
<td>IR_v1</td>
<td>Reconstructed voice using component-wise strategy</td>
</tr>
<tr>
<td>IR_v2</td>
<td>Reconstructed voice using context-dependent strategy</td>
</tr>
</tbody>
</table>

Table 1: Voices compared in the evaluation tests

In order to evaluate the substitution strategies presented in Section 4, two subjective tests were conducted. The first one assesses the voice repair and the second, the speaker similarity.
Figure 3: Similarity to the reference voice HC on a MOS-scale (mean and standard deviation)

6.1. Listening Intelligibility Test

The same 40 semantically unpredictable sentences [13] were synthesized for each of the 5 voices created from the patient’s recordings (see Table 1). The resulting 200 synthesized samples were divided into 4 groups such that each voice is represented by 10 samples in a group. A total of 40 native English participants were asked to transcribe the synthesized samples, with 10 participants for each group. Within each group, the samples were presented in random order for each participant. The participants performed the test with headphones. The transcriptions were evaluated by measuring the word error rate (WER).

6.2. Speaker Similarity Test

The same test sentence “People look, but no one ever finds it.” was synthesized for each of the 6 voices in Table 1. Participants were asked to listen alternatively to the reference voice (HC) and to the same sentence synthesized with one of the other voices including the average voice model (AV). The presentation order of the voices being tested was randomized. Participants should rate the similarity between the tested voice and the reference (HC) on a 5-point scale: 1: Very dissimilar, 2: Dissimilar, 3: Quite Similar, 4: Very similar; and 5: Identical). However, the participants were not given further instruction in order to avoid biasing towards rating any specific form of similarity. A total of 40 native English speakers performed the test using headphones.

7. Results and Discussion

The resulting average WERs for the intelligibility test are shown in Figure 2. We are not interested here in the absolute values of the WER but in their relative values compared to the reference voice HC. As expected, the synthetic voice IC created from the “impaired” speech has a high WER. More surprisingly, the baseline strategy for voice reconstruction increases the WER. One explanation could be that the substitution of the durations without replacing any other temporal parameters (such as the dynamics) could create some artifacts. However, the important results here are that the component-wise strategy succeeds in removing some articulation disorders from the synthetic speech and that the context-dependent strategy performs even better. The results of the similarity test are shown in Figure 3. A first interesting result is that the voice clone IC created by speaker adaptation from the “impaired” speech is more similar to the healthy clone HC than the average voice AV. In the case of this patient, this validates an implicit assumption of the voice reconstruction process: some valuable information about the original vocal identity should remain in the impaired speech. The other important result is the improvement of the mean similarity scores when the voice reconstruction strategies are applied. Between IC and IR_v2, there is a mean improvement of 0.5 MOS that turns out to be significant (p-value << 1.e-5), and between IR_v2 and AV, the mean improvement is about 1 MOS (with a p-value << 1.e-6). One explanation of this trend could be that the similarity of vocal identity is better perceived once the disorders have been regulated.

8. Conclusions

HMM-based speech synthesis has two clear advantages for the creation of personalized voices for people with disordered speech: speaker adaptation and improved control. Speaker adaptation allows the creation of a voice clone with a limited amount of data. Then the structure of the acoustic models can be modified to repair the synthetic speech. We investigated different substitution strategies where a model is replaced by the corresponding model from the average voice model. The evaluation of these methods demonstrates that: a) the voice clone can still bring valuable information about the vocal identity of the patient; b) it is possible to improve the intelligibility of a disordered synthetic speech while retaining its vocal identity. The reconstruction strategies presented here have been designed for MND patients, but their principle could be easily generalized to any other degenerative or acquired speech disorder.

9. References