A Comparison of Manual and Automatic Voice Repair for Individual with Vocal Disabilities

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 neurone 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. The power of this approach is that it is possible to use the patient's recordings to adapt existing voice models pre-trained on many speakers. When the speech has begun to deteriorate, the adapted voice model can be further modified in order to compensate for the disordered characteristics found in the patient's speech, we call this process "voice repair". In this paper we compare two methods of voice repair. The first method follows a trial and error approach and requires the expertise of a speech therapist. The second method is entirely automatic and based on some a priori statistical knowledge. A subjective evaluation shows that the automatic method achieves similar results than the manually controlled method.

Index Terms: HTS, Speech Synthesis, Voice Banking, Voice Reconstruction, Voice Output Communication Aids, MND.

1. Introduction

Degenerative speech disorders have a variety of causes that include Multiple Sclerosis, Parkinson's, and Motor Neurone Disease (MND) also known in the USA as Amyotrophic Lateral Sclerosis (ALS). MND primarily affects the motor neurones 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. Initial symptoms may be limited to a reduction in speaking rate, an increase of the voice's hoarseness, or an imprecise articulation. However, at some point in the disease progression, 80 to 95% of patients are unable to meet their daily communication needs using their speech [1]. As speech becomes difficult to understand, these individuals may use a voice output communication aid (VOCA). These devices consist of a text entry interface such as a keyboard, a touch screen or an eye-tracker, and a text-to-speech synthesizer that generates the corresponding speech. However, when individuals lose the ability to produce their own speech, they lose not only a functional means of communication but also a display of their individual and social identity through their vocal characteristics.

Current VOCAs are not ideal as they are often restricted to a limited set of impersonal voices that are not matched to the age or accent of each individual. Feedback from patients, careers and patient societies has indicated that there is a great unmet need for personalized VOCAs as the provision of personalized voice is associated with greater dignity and improved self-identity for the individual and their family [2]. In order to build personalized VOCAs, 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 [3], a free voice building service that can be used from any home computer in order to build a synthetic voice based on diphone concatenation, a technology developed in the 1980s. The user of this service has to record around 1800 utterances in order to fully cover the set of diphones and the naturalness of the synthetic speech is rather low. Cereproc [4] has provided a voice building service for individuals, at a relatively high cost, which uses unit selection synthesis, and is able to generate synthetic speech of increased naturalness. However, 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 used directly 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. 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. A voice building process using the hidden Markov model (HMM)-based speech synthesis technique has been investigated to create personalized VOCAs [7-10]. This approach has been shown to produce high quality output and offers 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 [11] requires only a very

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 130–133,
Dresden, Germany, 11 September, 2015.

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 repair". In this paper, we compare different strategies of voice repair using the HMM-based synthesis framework. The first method follows a trial and error approach and requires the expertise of a speech therapist. The second method is entirely automatic and based on some a priori statistical knowledge.

2. HMM-Based Speech Synthesis

Our voice building process is based on the state-of-the-art HMM-based speech synthesizer, known as HTS [12]. As opposed to diphone or unit-selection synthesis, the HMM-based speech synthesizer does not use the recorded speech data directly as the voice output. Instead it is based on a vocoder model of the speech and the acoustic parameters required to drive this vocoder are represented by a set of statistical models. The vocoder used in HTS is STRAIGHT and the statistical models are context-dependent hidden semi-Markov models (HSMMs), which are HMMs with explicit state duration distributions. The state output distributions of the HSMMs represent three separate streams of acoustic parameters that correspond respectively to the fundamental frequency (logF0), 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 using symbolic context information at the phoneme, syllable, word, and utterance level. In order to synthesize a sentence, a linguistic analyser is used to convert the sequence of words into a sequence of symbolic contexts and the trained HSMMs are invoked for each context. A parameter-generation algorithm is then used to estimate the most likely trajectory of each acoustic parameter given the sequence of models. Finally the speech is generated by the STRAIGHT vocoder driven by the estimated acoustic parameters.

3. Speaker Adaptation

One advantage of the HMM-based speech synthesis for voice building is that the statistical models can be estimated from a very limited amount of speech data thanks to speaker adaptation. This method [9] 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. It has been shown that using 100 sentences or approximately 6-7 minutes of speech data is sufficient to generate a speaker-adapted voice that sounds similar to the target speech [7]. In the following of this paper we refer the speaker-adapted voices as "**voice clones**". This provides a much more practical way to build a personalized voices for patients. For instance, it is now possible to construct a synthetic voice for a patient prior to a laryngectomy operation, by quickly recording samples of their speech [8]. A similar approach can also be used for patients with neurodegenerative diseases such as MND. 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. This is the aim of the voice repair methods introduced in the section 5 of this paper.

4. Database of Voice Donors

Ideally, the average voice model used for the speaker adaptation 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 more than 900 healthy voice donors with various accents (Scottish, Irish, Other UK). Each speaker recorded about one hour of speech (400 sentences). This database of healthy voices is first used to create the average voice models used for speaker adaptation. Ideally, the average voice model should be close to the vocal identity of the patient and it has been shown that gender and regional accent are the most influent factors in speaker similarity perception [13]. Therefore, the speakers are clustered according to their gender and their regional accent in order to train specific average voice models. A minimum of 10 speakers is required in order to get robust average voice models. Furthermore, the database is also used to select a reference donor for the voice repair procedures described in section 5. The voice repair is most successful when the reference donor is as close as possible to the patient in terms of vocal identity.

5. Voice Repair

Some individuals with neurodegenerative disease may already have speech symptoms at the time of the recording. In that case, the speaker adaptation process will also replicate these symptoms in the speaker-adapted voice. Therefore we need to remove speech disorders from the 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 in order to control and modify various speech features. This is the second major advantage of using HMM-based speech synthesis. In particular, HTS has statistically independent models for duration, log-F0, band aperiodicity and mel-cepstrum. This allows the substitution of some models in the patient's speaker-adapted voice by that of a well-matched healthy voice or an average of multiple healthy voices. For example, patients with MND often have a disordered speaking rate, contributing to a loss of the speech intelligibility. The substitution of the state duration models 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. In the following part of this section, we present two different methods of model substitution. The first one is manually controlled whereas the second one is automatic.

5.1. Manual voice repair

In the manual approach, a speech therapist first selects a reference voice among all the available voices with same accent, gender and age range than the patient. Then the models of this reference voice are used to correct some of the patient's voice models. This correction is based on mean and variance interpolation between models. A graphical interface allows the speech therapist to control the amount of interpolation between the patient's voice models and the reference voice models as illustrated in Figure 1.

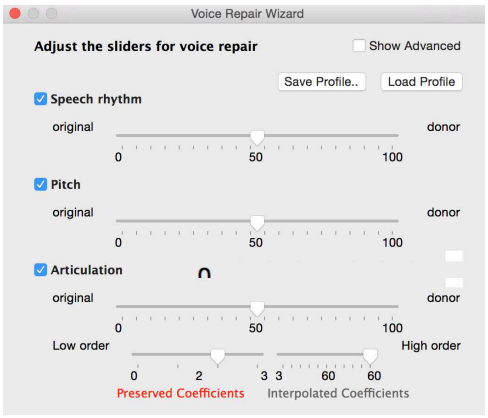

Figure 1: Graphical interface for model interpolation.

The following models and information can be interpolated:

- Duration
- Dynamics coefficients of the log-F0
- Dynamics coefficients of the mel-cepstrum
- Low-order coefficients of the mel-cepstrum
- High-order coefficients of the mel-cepstrum

The voiced/unvoiced weights and aperiodicity models are simply substituted since their impact on voice identity is rather limited and their replacement of will fix the breathiness disorders. The interpolation 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. The interpolation of the dynamics coefficients of the log-F0 will help to regulate the prosodic disorders such as monotonic F0. Finally the global variances of all the parameters are also simply substituted. We will refer to this method as the **manual repair**.

5.2. Automatic voice repair

The manual voice repair requires a lot of expertise from the speech therapist, as it is a trial and error approach. Therefore, we aim to replace it by a fully automated voice repair procedure. We measure the Kullback-Leibler distance (KLD) between the models of the patient voice and the models of the reference voice as illustrated in Figure 2. Then the likelihood of each of the measured distance is evaluated given the statistical distribution of KLD distances between healthy voice models of similar accent, gender and age band. The likelihood values are used to control the interpolation between the patient and reference voice models. For instance, if the likelihood of the KLD distance for a given model of the patient voice is very low, the corresponding model of the reference voice is used to replace it in the patient voice. The reference voice model is also selected automatically as the one that maximizes the likelihood of the patient recording data.

6. Experiment

The manual and automatic voice repair methods presented in Section 5 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.

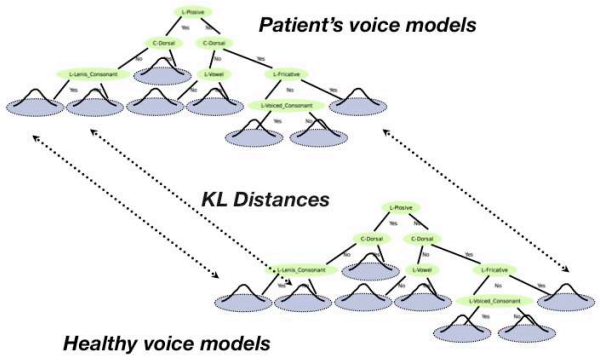

Figure 2: Graphical interface for model interpolation.

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. These two recordings were used separately as adaptation data in order to create two speaker-adapted voices from the same male-Scottish average voice model. 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. Two different reconstructed voices were created from the second recording of the patient ("impaired" speech) using the manual and the automatic voice repair methods respectively. In order to evaluate the voice repair methods, two subjective tests were conducted. The first one assesses the intelligibility of the reconstructed voices whereas the second one measures their similarity with synthetic voice created from "healthy" speech of the patient. We also included the synthetic voices of the donors selected for the manual and the automatic voice repair in the similarity test. All the synthetic voices used in the experiment are summarized in Table 1.

Voice	*Description*
MD	Voice of donor used in manual voice repair
AD	Voice of donor used in automatic voice repair
HC	Voice clone of the "**healthy**" speech (1st recording)
IC	Voice clone of the "**impaired**" speech (2nd recording)
IR_v1	Reconstructed voice using **manual voice repair**
IR_v2	Reconstructed voice using **automatic voice repair**

Table 1: Voices compared in the evaluation tests.

6.1. Listening Intelligibility Test

The same 50 semantically unpredictable sentences were synthesized for each of the 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).

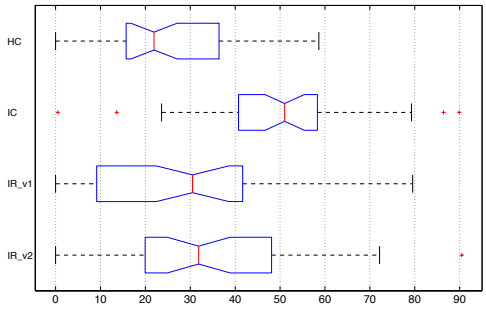

Figure 3: *Word Error Rate (mean and standard deviation)*

6.2. Speaker Similarity Test

The same test sentence "People look, but no one ever finds it." was synthesized for each of the 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. The presentation order of the voice samples was randomized. The participants have been asked to rate the similarity in terms of speaker identity 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). A total of 40 native English speakers performed the test using headphones.

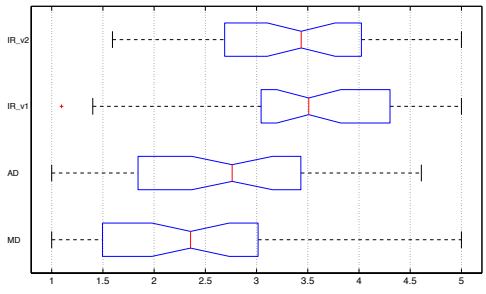

Figure 4: *Similarity to the reference voice HC on a MOS-scale (mean and standard deviation)*

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 healthy voice HC. As expected, the synthetic voice IC created from the "impaired" speech has a high WER. Both manual and automatic voice repair succeeds in removing some articulation disorders from the synthetic speech as we can see a significant decrease of WER. The manual voice repair yields to slightly lower WER than the automatic voice repair although the difference is not significant. The results of the similarity test are shown in Figure 3. The first important result is that the reconstructed voices are still considered more similar to the patient's voice than the closest voice donors (MD and AD) used for the voice repair. This means that both voice repair methods manage to preserve the voice identity to a certain extent. The manual voice repair is performing slightly better than the automatic method but the difference is not significant (p-value ~ 1.e-2).

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 have presented here two different strategies for voice reconstruction. The first one is manual and requires the expertise of a speech therapist whereas the second one is fully automated. The evaluation of these methods demonstrates that: a) it is possible to improve the intelligibility of a disordered synthetic speech while retaining its vocal identity; b) the automatic voice repair performs almost as well as the manual voice repair. 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

[1] Doyle, M. and Phillips, B. (2001), "Trends in augmentative and alternative communication use by individuals with amyotrophic lateral sclerosis," *Augmentative and Alternative Communication* 17 (3): pp.167–178.

[2] Murphy, J. (2004), "I prefer this close': Perceptions of AAC by people with motor neurone disease and their communication partners. *Augmentative and Alternative Communication*, 20, 259-271.

[3] Yarrington, D., Pennington, C., Gray, J., & Bunnell, H. T. (2005), "A system for creating personalized synthetic voices," *Proc. of ASSETS*.

[4] http://www.cereproc.com/

[5] Rudzicz, F. (2011) "Production knowledge in the recognition of dysarthric speech", PhD thesis, University of Toronto.

[6] Kain, A.B., Hosom, J.P. Niu X., van Santen J.P.H., Fried-Oken, M., and Staehely, J., (2007) "Improving the intelligibility of dysarthric speech," Speech Communication, 49(9), pp743–759.

[7] Creer, S., Green, P., Cunningham, S., & Yamagishi, J. (2010) "Building personalized synthesized voices for individuals with dysarthia using the HTS toolkit," IGI Global Press, Jan. 2010.

[8] Khan, Z. A., Green P., Creer, S., & Cunningham, S. (2011) "Reconstructing the Voice of an Individual Following Laryngectomy," Augmentative and Alternative Communication.

[9] Veaux, C., Yamagishi, J., King, S. (2011) "Voice Banking and Voice Reconstruction for MND patients," *Proceedings of ASSETS*.

[10] Veaux, C., Yamagishi, J., King, S. (2012) "Using HMM-based Speech Synthesis to Reconstruct the Voice of Individuals with Degenerative Speech Disorders," *Interspeech*, Portland, USA.

[11] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. & Isogai, J. 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. on ASL, 17, 66-83.

[12] Zen, H., Tokuda, K., & Black, A. (2009) "Statistical parametric speech synthesis, Speech Communication," 51, pp.1039-1064.

[13] Dall, R., Veaux, C., Yamagishi, J. & King, S. (2012) "Analysis of speaker clustering strategies for HMM-based speech synthesis," *Proc. Interspeech*, Portland, USA.