解説

Voice banking and reconstruction

— Speech synthesis technologies for individuals with vocal disabilities —

Junichi Yamagishi, Christophe Veaux, Simon King, Steve Renals
(The Centre for speech Technology Research, University of Edinburgh, U.K.)

ここにPACS No. をご記入ください

1. Introduction

In this invited paper, we overview the clinical applications of speech synthesis technologies and discuss research trends in Japan, U.K. and USA. We also introduce the University of Edinburgh's new project "Voice Banking and reconstruction" for patients with degenerative diseases, such as motor neurone disease and Parkinson's disease and show how speech synthesis technologies can improve the quality of life for the patients.

1.1 Speech Disorder and Impairment

Speech disorders have a variety of causes. The major neurological degenerative diseases resulting in severe speech disorders include Motor Neurone Disease (MND), Parkinson's, and Multiple Sclerosis; non-progressive conditions resulting in such disorders include and Stroke and Cerebral Palsy. Speech impairment may also be caused by conditions such as cancer of the vocal cords, resulting in laryngectomy.

MND, also known in the USA as Amyotrophic Lateral Sclerosis (ALS) or Lou Gehrig's disease, is a debilitating neurological disease that results from the degeneration and death of the large motor neurones in the brain and spinal cord. This causes a worsening muscle weakness that leads to loss of mobility and difficulties with swallowing, breathing and speech production. Approximately 75% of MND patients are unable to speak by the time of their death [1]. MND usually strikes at random, during middle age (There are four types of MNDs. Most cases of ALS and Progressive Bulbar Palsy are recorded in over 60's. Age of onset of Primary Lateral Sclerosis is fifty years. Age of onset of Progressive Muscular Atrophy is usually under fifty years). About 100 new cases of MND are diagnosed in Scotland, U.K. each year.

1.2 AAC and VOCA

When individuals lose the ability to produce their own speech for neurological or other reasons, alternative augmentative communication (AAC) devices [1] can be used. An AAC device with a speech synthesis capability is referred to as a "Voice Output Communication Aid" or VOCA. The standard text-to-speech (TTS) synthesizers such as the Klatt formant synthesizer [2] or unit-selection synthesizers [3] are usually embedded as speech output functions in such devices. Probably Prof. Stephen Hawking, who also has MND, is the most famous person who utilizes a VOCA device in his daily life.

People with such speech disorders lose not only a functional means of communication, but also the vocal expression of individual and social identity. The voice is such an integral part of identity that, when damage occurs, sufferers may withdraw from social interaction, and even from interaction with their own family.

Using a VOCA that sounds like someone from a different geographical or social background, or someone of a different age, can cause embarrassment and a lack of motivation to interact socially [4]. However, current voice communication aids compound this effect. Existing VOCAs on the market typically provide a small and limited range of voices, sometimes of the wrong sex, and almost always of the wrong accent (US-accented English is the default, even in the UK, because of market size).

Currently it is not easy, and certainly not cost effective, for manufacturers to create synthetic voices that clearly belong to the user. This is something long requested by the AAC/VOCA users: for them, speech synthesis is not just an "optional extra" for reading out text, but a critical function for social communication and personal identity.

Therefore a personalized VOCA where the synthetic voice includes characteristics of an individual user could reduce the social distance imposed by this mode of communication by re-associating the output content with the user through vocal identity. In order to build a personalized VOCA, several attempts have been made to mainly capture the voice for the later use before it is lost, using a process known as voice banking [5]. In the next chapter, we first introduce such voice banking approaches in the USA, U.K. and Japan.

2. Commercial Voice Banking Service

ModelTalker, developed by the Nemours Speech Research Laboratory at the Alfred I. duPont Hospital for Children in Wilmington, Delaware, USA, provides a free TTS voice building service (www.modeltalker.com). A user must record around 1800 utterances that fully cover a set of diphones (which is the second half of one phone plus the first half of the following phone) and can upload the recordings to the developers. Within a short time, a personalized TTS voice, based on diphone concatenation synthesis, which was originally developed in the 1980s, becomes available to use [6].

Cereproc (http://www.cereproc.com/) based in Edinburgh, Scotland, U.K. also 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. Roger Ebert, the movie critic and Chicago Sun-Times columnist, used this service in 2010 and he emphasized the importance of the use of his own computer voice "Roger 2.0"[1].

OKI Electric Industry Co., Ltd. in Japan also provides a commercial voice building service for individuals called "Polluxstar"[2]. This is based on a hybrid speech synthesis system [7] using both unit selection and statistical parametric speech synthesis [8] to achieve a natural speech quality.

However, all the speech synthesis techniques used in building the above services require a large amount of recorded speech in order to build a good quality voice. This requirement stems from the fact that in diphone or unit selection synthesis, the speech recordings are segmented into small phonetic units that can be recombined to make new utterances. Moreover it requires the recorded speech data to be as intelligible as possible, since the data recorded is used directly as the voice output. These requirements make this technique problematic, especially for those individuals whose voices have already started to deteriorate.

Unfortunately, our colleagues in speech and language therapy find that some patients only begin treatment once their vocal problems become moderate to severe. Therefore, there is a strong motivation to improve the voice banking and voice building techniques for individuals, so that patients can use their own synthetic voices, even if their speech is already disordered at the time of recordings. In the next section, we

[1] http://www.ted.com/talks/roger_ebert_remaking_my_voice.html
[2] http://www.oki.com/en/press/2008/07/z08050e.html

discuss some promising results on restoring disordered speech and building natural sounding TTS voices.

3. Research Trends

3.1 Restoration of Disordered Speech

Kain et al. have used the Klatt formant synthesizer with voice conversion techniques to repair disordered speech [9]. Their system generated speech by concatenating the original unvoiced speech segments with re-synthesized voiced segments. The re-synthesis combined the original high-frequency spectrum with a re-synthesized and manipulated low-frequency spectrum using modified energy and formant parameters that were mapped using a Gaussian mixture model (GMM). In a similar way to voice conversion techniques, the GMM was learned in advance from a parallel dataset of disordered and target speech to transform the output closer to the correct acoustics.

In contrast, Rudzicz [10] used several speech generation techniques simultaneously. The system first used high-pass filtering to remove voicing errors in consonants, then used TD-PSOLA to stretch out irregular duration, and finally used STRAIGHT spectral morphing to separate out confusable formants.

Nakamura et al. have also proposed a voice conversion approach based on GMMs for the restoration of speech. They have used this voice conversion framework for people with laryngectomies using non-audible murmur (NAM) microphone [11].

3.2 Creating Personalized a Text-to-Speech Synthesis System from Disordered Speech

A) HMM-based speech synthesis

Recently the state-of-the-art text-to-speech synthesis technique, known as statistical parametric speech synthesis, hidden Markov model (HMM)-based speech synthesis, or sometimes “HTS” [8] has been investigated to create personalized VOCAs by the Clinical Application of Speech Technology (CAST) group at the University of Sheffield and the Centre for Speech Technology Research (CSTR) at the University of Edinburgh in U.K [12] [13]. In the following, we elaborate this.

The HMM-based speech synthesis approach developed by Tokuda et al. [9] statistically models several acoustic parameters required for driving a vocoder. The acoustic parameters correspond to both the source and the excitation and include fundamental frequency (F0) and spectral representations such as mel-cepstrum and line spectrum pairs. The vocoders used typically are STRAIGHT and MELP.

B) Speaker adaptation

The use of the HMM-based speech synthesis method has two clear advantages over existing speech synthesis methods for voice banking and voice building for people with disordered speech: speaker adaptation and improved control.

Speaker adaptation may be employed to transform existing speaker-independent acoustic models to match a target speaker using a very small amount of speech data [14]. This method starts with an “average voice model” and uses model adaptation techniques drawn from speech recognition such as maximum likelihood linear regression (MLLR) [15], to adapt the speaker independent HMMs to a new speaker. This adaptation allows text-to-speech synthesizers to build a target voice using much smaller amounts of training data than previously required.

Prior to this, the development of a new voice required many hours of carefully annotated speech recordings from a single 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 [14]. This provides a much more straightforward and

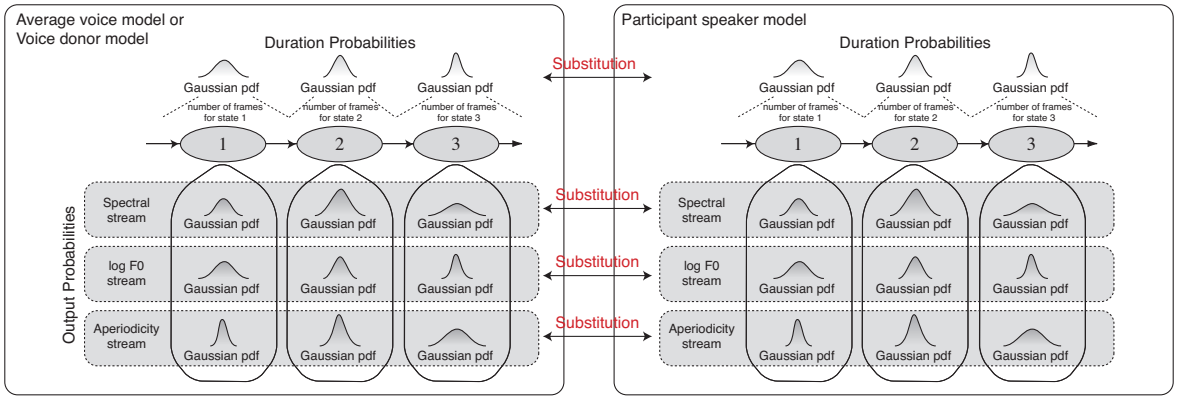

Figure 1 Model-level substitution into average voice or donor voice

practical way for patients to voice-bank their speech. For instance, it is now possible to construct a synthetic voice for a patient prior to a larynjectomy operation, by quickly recording samples of their speech [13]. A similar approach could also be used for patients with degenerative diseases before the diseases affect their speech.

C) Voice reconstruction

If the speech is already disordered at the time of recording, speaker adaptation techniques will clone not only speaker identity but also the symptoms of the 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 3.1.

Since the HMM speech synthesis method is based on the source-filter theory, and the statistical models may be regarded as “abstracting” various speech components, we can control and modify various speech components. This is the second, novel advantage of using HMM speech synthesis. Some examples are given below.

Speakers with MND typically have a highly variable speaking rate. We may model duration explicitly in hidden semi-Markov models (HSMMs) adapted to disordered speech, using duration distributions with parameters (mean and variance) estimated from several healthy normal voices. This model substitution process (Figure 1) enables the timing disruptions to be regulated at the phoneme, word, and utterance levels.

Furthermore, MND speakers often have monotonic or excessive prosody. In such cases, we can substitute and modify the parameters of F0 distributions in a similar way. If an MND patient has breathy or hoarse speech, in which excessive breath through the glottis produces unwanted turbulent noise, we can substitute a statistical aperiodicity model to produce a less breathy or hoarse output.

In summary, the HMM speech synthesis has statistically independent structures of spectral and excitation parameters including their dynamics and temporal information. This allows the substitution of a model for any disorder that occurs in the patient's speech data by that of a well-matched healthy voice or an average of multiple healthy voices. A surgical analogue to this is transplantation of a component of the voice model.

Although disordered speech perceptually deviates considerably from normal speech in many ways, it is known that its articulatory errors are highly consistent [16] and hence relatively predictable [17]. This is one of the reasons this simple substitution works effectively. For example, patients with Parkinson’s typically have

Figure 2 The reconstructed synthetic voice was installed onto patient's VOCA

a weaker voice, very fast speaking rate, and difficulties to get speech started. Patients with Multiple Sclerosis frequently have imprecise articulation with unexpected F0 changes, nasality, hoarseness and poor loudness control. Therefore we can pre-define the substitution strategy for a given condition, to some extent.

To verify this, the construction of personal synthetic voices were conducted for a patient with Parkinson's in Sheffield [12] and a patient with MND in Edinburgh. The patient in Edinburgh had had MND for three years at that time of recording, which consisted of five minutes of an informal interview in the quite office environment. We created his personalized synthetic voice from this short recording. Various components of this model were substituted by components from a model constructed from his brother's voice. Feedback obtained from the patient's family after we installed his reconstructed voice into his VOCA device (Fig. 2) was very positive. The samples of repaired voice are available online[3].

D) Voice donors

In the case study above, we used the patient's brother's models for model substitution. We call such volunteers (who donate their speech data, from which we can create acoustic models to be

[3] http://www.smart-mnd.org/voicebank/

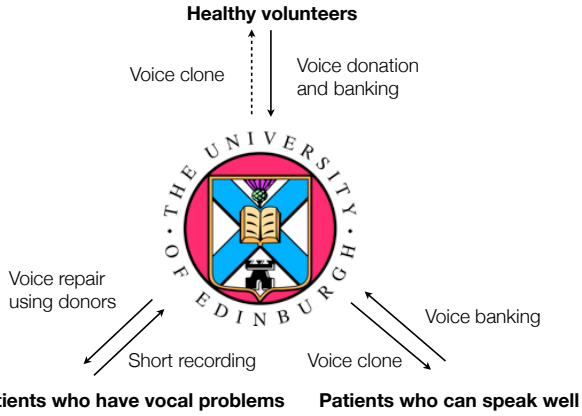

Figure 3 Vision of the Edinburgh voice banking project.

partially substituted for patient's disordered model) "voice donors". This also comes from the analogy to surgical transplantation.

We believe that the voice donor framework has the potential to have a significant social impact because (a) voice donors can help people with vocal problems simply by speaking and (b) a donor's speech data is also automatically voice-banked so they could have a personalized VOCA in case of any speech problems in the future. We hope that this will raise awareness amongst the public of more general issues surrounding vocal health, which is important in itself, but can also be an early indicator of other problems.

At the same time we can collect appropriate data for use in average voice models from the voice banked speech data, so that we can conduct more precise speaker adaptation.

4. Edinburgh Voice Banking Project

We are currently conducting voice banking and reconstruction clinical trials at both the Euan McDonald Centre for Motor Neurone Disease Research and the Anne Rowling Regenerative Neurology Clinic, aiming for 50 new patients (about half of the new MND cases in Scotland) and 150 voice donors annually. The Anne Rowling Clinic (founded by donations by J.K. Rowling and opening in December 2011) will have a recording facility specialized for voice

baking purposes.

In this project, we carried out voice banking, voice donation, and voice repair at the same time (Fig. 3). So far we have completed high quality recordings of about a hundred UK-wide healthy voice donors with various accents and 10 patients with various conditions of MND in the semi-anechoic chamber at CSTR. The recordings will be continued throughout the project and this will become the largest UK speech database available to researchers. At this point, the database is already larger than the WSJCAM0 database, a standard UK speech research database.

5. Conclusions

In this invited paper, we have overviewed the clinical applications of speech synthesis technologies and have introduced the University of Edinburgh's new project "Voice Banking and Reconstruction" for patients with neurological degenerative diseases.

Due to limited space, we have not mentioned other important research in the area, such as perceptual studies and the analysis of social factors for VOCAs. For a review of such work, we refer the reader to [18].

6. Acknowledgements

The authors would like to thank Mr. Euan MacDonald, Prof. Siddharthan Chandran of Euan MacDonald Centre, Mrs. Phillipa Jane Rewaj, and Mrs. Shuna Colville for the support of Edinburgh voice banking project. The authors would like to thank Prof. Phil Green and Dr. Sarah Creer of the University of Sheffield for their useful comments. Finally the authors would like to gratefully thank "voice donors" for their cooperation to this project. This research was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology), and by the Euan MacDonald Centre for Motor Neurone Disease Research.

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] Klatt, D. H. (1980), "Software for a cascade/ parallel formant synthesizer," *The Journal of the Acoustical Society of America*, *67* (3), pp.971–995

[3] Black, A. and Cambpbell, N. (1995), "Optimising selection of units from speech database for concatenative synthesis," *Proc. EUROSPEECH-95*, pp.581–584

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

[5] Stern, S.E. (2008), "Computer-Synthesized Speech and Perceptions of the Social Influence of Disabled Users." *Journal of Language and Social Psychology* 27 (3): pp.254

[6] Yarrington, D., Pennington, C., Gray, J., and Bunnell, H. T. (2005), "A system for creating personalized synthetic voices," *Proc. ASSETS*, pp.196–197

[7] Kawai, H., Toda, T., Yamagishi, J., Hirai, T., Ni, J., Nishizawa, N., Tsuzaki, M., and Tokuda, K. (2006) "XIMERA: a concatenative speech synthesis system with large scale corpora," *IEICE Trans. Information and Systems*, J89-D-II (12), pp.2688–2698

[8] Zen, H., Tokuda, K., & Black, A. 2009. Statistical parametric speech synthesis, *Speech Communication*, 51, pp.1039-1064

[9] 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.

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

[11] K Nakamura, T Toda, H Saruwatari, K Shikano (2006), "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech", Proc Interspeech 2006

[12] Creer, S., Green, P., Cunningham, S. and Yamagishi, J. (2010) "Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit," *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, John W. Mullennix and Steven E. Stern (Eds), IGI Global press

[13] Khan, Z. A., Green P., Creer, S., & Cunningham, S. (2011) "Reconstructing the Voice of an Individual Following Laryngectomy," *Augmentative and Alternative Communication*, 27, pp.61–66

[14] 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. Speech, Audio, & Lang. Proces.*, 17, pp.66-83

[15] Woodland, P. C. (2001) "Speaker adaptation for continuous density HMMs: A review". Proc. the ISCA workshop on adaptation methods for speech recognition, pp.11–19

[16] Yorkston, K. M., Beukelman, D. R. and Bell, K. R. (1998) "*Clinical management of dysarthric speakers,"* College-Hill Press.

[17] Mengistu, K.T. and Rudzicz, F., (2011) "Adapting acoustic and lexical models to dysarthric speech," Proc. ICASSP 2011

[18] Mullennix, J.W. and Stern, S.E. (2010) "Computer Synthesized Speech Technologies: Tools for Aiding Impairment," IGI Global press.