Voice Conservation: Towards Creating a Speech-Aid System for Total Laryngectomees

Zdeněk Hanzlíček and Jindřich Matoušek

Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 306 14 Plzeň, Czech Republic
{zhanzlic,jmatouse}@kky.zcu.cz

Abstract. This paper describes the initial experiments on voice conservation of patients with laryngeal cancer in an advanced stage. The final aim is to create a speech-aid device which is able to "speak" with their former voices. Our initial work is focused on applicability of speech data from patients with an impaired vocal tract for the purposes of speech synthesis. Preliminary results indicate that appropriately selected synthesis method can successfully learn a new voice, even from speech data which is of a lower quality.

Keywords: speech aid, total laryngectomy, voice conservation, speech synthesis

1 Introduction

Laryngectomees are people who underwent a laryngectomy surgery. This medical intervention is performed on patients with laryngeal cancer when other types of treatments (e.g. radiation or chemotherapy) fail or are not possible. According to the extent of the carcinoma various sections of larynx are removed. In the case of total laryngectomy, the removal of the whole larynx together with the vocal folds is performed. Then, respiratory and digestive tracts are separated from each other. The laryngectomee breathes through a stoma – an opening in the trachea.

A significant consequence of this surgery is the inability to produce speech in the common manner. However, there are several alternatives for producing speech sounds:

- **tracheo-esophageal speech** – a special voice prosthesis need to be surgically placed between trachea and esophagus, it contains a one-way valve that allows the air to flow from lungs into the oral cavity. Tracheostoma has to be plugged during the speech production.

- **esophageal speech** – the air is swallowed into the esophagus and then it is pushed back into the oral cavity for articulation. This method is very exacting due to the low capacity of esophagus.
- **electrolaryngeal speech** – the function of vocal folds is substituted by an external device (electrolarynx) which is put to the neck where produces mechanical vibrations while the speaker articulates.

All the aforementioned kinds of speech (also called alaryngeal speech) suffers from lack of naturalness and speaker identity.

Recently, a new important task – alaryngeal speech enhancement – is solved in the field of computer speech processing [1, 2]. The objective is to improve the sound quality and recover the speaker identity. The final aim is to create a speech-aid device which allows the laryngectomee to communicate with a more natural voice.

From the practical point of view, the source alaryngeal speech should not disturb and coincide with the final enhanced speech. This could be done by rotating stages of alaryngeal and enhanced speech.

Another solution is using the NAM (non-audible murmur) microphone that is able to detect so-called body conducted speech. It is caused by vibrations of air in the vocal tract passed on the soft tissues of the head or neck. The standard electrolarynx excitation is swapped with a source of small-power vibrations. Thus the produced alaryngeal speech is nearly inaudible.

In fact, there are two basic approaches to design a speech-aid system for laryngectomees:

- Using **speech recognition**. Produced speech is first recognized by a specially designed ASR (automatic speech recognition) system. A text of the utterance is extracted this way. Then this text is synthesised by a new voice. The knowledge of the utterance content allows to add some higher speech properties, e.g. the course of the fundamental frequency. An important disadvantage of such a speech-aid system is a delayed response caused by speech recognition process. The resulting delay could be similar as in the case of human simultaneous translation.
- **Speech signal transformation** without speech recognition. Spectral characteristics of alaryngeal speech are converted and the enhanced (i.e. more human like) speech is reconstructed from those characteristics.

2 Voice Conservation

A natural requirement for any speech-aid system is to produce a voice which is close to the former voice of each laryngectomee. Most speech synthesis

methods are able to produce voice with required speaker identity. However, a quite huge amount of training speech data is necessary to learn this voice. In the case of a laryngectomee, to obtain a sufficient amount of speech could be a substantial problem.

Naturally, speech recorded by the healthy voice (before the disease has broken out) would be preferred. Some people could have recordings related to their job, e.g. various speeches, performances, presentations, or also some personal recordings, e.g. family events, reading fairy tales to their children, etc. Unfortunately, the acoustic conditions of such audio data are often not optimal and it is not suitable for the purposes of speech synthesis. Moreover, most people do not have any usable recordings at all.

Another solution is to record the speech data after the diagnosis before surgery. However, in those stages of disease, the vocal tract is usually significantly damaged which causes various speech problems. The overall voice quality is poor or unstable and the speaking could be very exhausting for the patients. The voice could be also affected by the mental condition of patients because they are often significantly stressed by their diagnosis and expected surgery. The last chance is to acquire recordings from another willing person with a similar voice.

The process of obtaining and storing speech data of the patients can be called voice conservation. It is only one fundamental step in solving the complex problem of developing a speech-aid device for laryngectomees.

3 Speech Synthesis

Modern speech synthesis methods (e.g. unit selection or statistical parametric speech synthesis) need a large amount of training data (several hours of speech) to create a new system producing the desired voice in a high quality. The standard recording procedure is demanding even for a healthy professional speaker because high-quality data are required for the purposes of speech synthesis. Thus each utterance is repeated until it is perfectly pronounced. The overall recording process lasts for several weeks.

It would be nearly impossible to record such a huge amount of speech data by a patient with advanced laryngeal cancer. Fortunately, there are several alternatives with lower demands on speech data. The most promising one is using adaptation methods [3] within the statistical parametric speech synthesis (also known as HMM-based speech synthesis [4]). This synthesis method employs statistical models (hidden Markov models, HMMs) to represent the statistical acoustic features of speech. These models are trained by

using speech data of the desired speaker. During synthesis, speech is generated from those trained models.

Model adaptation is a transformation of models from one voice to another. Source models are trained from speech data of a professional speaker, or data from more speakers can be used to trained so-called average models. For the model adaptation significantly less speech data is needed. Moreover, a great advantage of the adaptation is the lower sensitivity to the quality of the source speech data.

4 Preliminary Experiments

For our first experiments we apply an experience from our previous work in the domain of statistical parametric speech synthesis [5, 6].

In cooperation with the Motol University Hospital one voluntary female patient with laryngeal cancer diagnosis was selected and her speech recorded. Recording conditions (similar to those described in [7]) was acoustically perfect, an anechoic chamber for acoustical measuring and experiments was used.

The recording process had to be adjusted to the specific condition of the patient. Several concessions had to be done, otherwise it would be unfeasible to record required amount of speech during one session. Consequently, the recorded utterances contains various stumbles, unexpected pauses and voice failures. About 500 utterances (approx. 1 hour of speech) were recorded during this session.

In our first experiments we used about 30 minutes of speech to adapt models trained from 5 hours of speech from a professional female speaker. The speech produced by using adapted models was definitely identified as the female patient. Considering the utilised data, the quality was also acceptable.

5 Conclusion

Our first experiments are promising. Voice conservation of patients with laryngeal cancer diagnosis, even when their speech is of lower quality, opens the possibility to create a speech-aid device producing the former personal voice. Although the way to develop such a device is still long and a lot of research work has to be done, the current results can be already practically utilised. Laryngectomees can run the speech synthesizer with their own voices on their computers. This could be helpful in the post-operative stage when the possibilities of inter-personal communication are very limited which could be frustrating.

References

1. Doi, H., Nakamura, K., Toda, T., Saruwatari, H. and Shikano, K.: An Evaluation of Alaryngeal Speech Enhancement Methods nased on Voice Conversion Techniques. In: Proceedings of ICASSP 2011, pp. 5136–5139 (2011)
2. Nakamura, K., Toda, T., Saruwatari, H. and Shikano, K.: Speaking Aid System for Total Laryngectomees using Voice Conversion of Body Transmitted Artificial Speech. Proceedings of Interspeech 2006, pp. 1395-1398 (2006)
3. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. and Isogai, J.: Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. In: IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, pp. 66–83 (2009)
4. Zen, H., Tokuda, K. and Black, A. W.: Review: Statistical parametric speech synthesis. *Speech Communication*, vol. 51, 1039–1064 (2009)
5. Hanzlíček, Z.: Czech HMM-Based Speech Synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) *Text, Speech and Dialogue, LNCS (LNAI)*, vol. 6231, 291–298, Springer Berlin/Heidelberg (2010)
6. Hanzlíček, Z.: Czech HMM-Based Speech Synthesis: Experiments with Model Adaptation. In: Habernal, I. and Matoušek, V. (eds.) *Text, Speech and Dialogue, LNCS (LNAI)*, vol. 6836, 107–114, Springer Berlin/Heidelberg (2011)
7. Matoušek, J., Romportl, J.: Recording and Annotation of Speech Corpus for Czech Unit Selection Speech Synthesis. In: Matoušek, V. and Mautner, P. (eds.) *Text, Speech and Dialogue, LNCS (LNAI)*, vol. 4629, 326–333, Springer Berlin/Heidelberg (2007)