INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

Voice Conservation and TTS System for People Facing Total Laryngectomy

Markéta Jůzová[12], *Daniel Tihelka*[1], *Jindřich Matoušek*[12], *Zdeněk Hanzlíček*[1]

[1]New Technologies for the Information Society (NTIS), [2]Department of Cybernetics
Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

juzova@kky.zcu.cz, dtihelka@ntis.zcu.cz, jmatouse@kky.zcu.cz, zhanzlic@ntis.zcu.cz

Abstract

The presented paper is focused on the building of personalized text-to-speech (TTS) synthesis for people who are losing their voices due to fatal diseases. The special conditions of this issue make the process different from preparing professional synthetic voices for commercial TTS systems and make it also more difficult. The whole process is described in this paper and the first results of the personalized voice building are presented here as well.

Index Terms: speech synthesis, text corpus, voice loss, voice conservation, speech impairment, laryngectomy

1. Introduction

A human voice is very important in our daily life communication and it is also an essential part of our human identity. This is a reason why the voice loss affects us deeply, all the more so it comes suddenly and unexpectedly. There is e.g. a considerable amount of people who are about to lose their ability to speak because of a diagnosed laryngeal cancer, leading to total laryngectomy – a surgical removal of the larynx with vocal chords. After the cancer diagnosis, the patients are going to lose their voice in a very short period, usually few weeks, or even few days in more serious cases.

Although there are several alternative solutions how to communicate after the total voice loss (electrolarynx, voice prosthesis, esophageal voice and, nowadays, generic-voice speech synthesis), none of them can replace the original human voice of the particular patient; not to mention that the first three options are very demanding for the patients, do not allow longer talk without any rest, and not all patients are able to learn some of these. On the other hand, the use of a general, anonymous synthetic voices on portable devices does lack the personal feeling to the user and may even increase the depression of identity loss [1].

On that account, we have started the recordings of such patients 6 years ago [2] to preserve this important part of their own personality – process called as "voice conservation" or "voice banking". In addition, from these recordings we also build the voice database usable within our TTS system on a PC or a smartphone [3, 4]. We offer the TTS system with their personal voice to the patients as a part of study testing the procedure of speech synthesis building focused on end-users who are going to depend on it due to the voice loss in a short time.

2. Creating personalized synthetic voices

The creation of a personalized synthetic voice significantly differs in several aspects from the well-studied process of building a new professional synthetic voice (a voice built for a commercial TTS system), which is usually expensive and time consuming and requires a certain amount of manual post-production and regular maintenance work [5].

In case of voice conservation, the patients have no experience with any recording; they are mostly elderly people with poorer abilities to use computer devices. Moreover, at the time of a diagnosis, the voice can already be partially damaged and speaking for a longer time may be painful for these people.

2.1. Text corpus building

Facing the unpredictable ability of non-professional speakers to read and pronounce the sentences to be recorded in a consistent style, we have collected a specially designed set of sentences (described in Section 2.1) allowing us to balance the use of statistical parametric synthesis (SPS) in cases of lower amount of recordings (or recordings of a poor voice quality), or the use of unit selection-based speech synthesis (USEL) in cases of adequate amount of sufficiently clear recordings (see Section 3), similarly to [6].

Our pilot experience with such non-professional speakers showed that they are not able to fluently read longer, complex or compound sentences, not to mention most of longer foreign-like words included for rare units or contexts. Therefore, we filtered our source text data (a large collection of Czech texts containing more than half a million sentences from various domains) to remove the problematic ones.

Contrary to the professional voice building [5], we do not know in advance how many sentences the patient is willing and able to record, since the period between the diagnosis of the fatal disease and the radical surgery is very short. This fact leads us to change the paradigm of text-corpus building – instead of maximizing a criterion through the whole corpus (e.g. 10,000 sentences with uniformly balanced units, see [7]), the text corpus building process must be split into several levels. In each level, the selection criterion focuses on different speech units, starting with simple phones and ending with diphones in different prosodic contexts. The higher levels increase the quality of the created synthetic voice not only by adding more and more sentences, but by increasing the richness of speech units coverage. The higher levels of the algorithm also suppose more cooperative speakers, thus, the sentences are intentionally selected for increasing the final quality of speech synthesis.

Thanks to this algorithm, no matter the number of sentences a speaker is able to record, we can guarantee the highest coverage reached by the last recorded sentence at the level the sentence belongs to, as well as the fulfilment of all the lower-level criteria [8]. This allows us to use SPS-based speech synthesis without the need of voice conversion when low amount (few hundreds) of sentences are recorded, as well as the use of unit selection method (from less than 1000 phrases) when each unit is guaranteed to have several candidates to select from. Let us point out that these numbers are significantly lower than those reported by [6].

2.2. Voice banking process

Currently, the voice banking of the patients takes place in a special soundproof room and it is carried out in a supervised mode, i.e. a trained person supervises the recording to help the speakers. During the pilot study, a new recording application has been being created to make the recording easier for the speakers not accustomed to the computer work.

Currently, we are working on the full automation of the whole process of voice banking and TTS voice database building. Firstly, it comprises the recording in the unsupervised mode, allowing to record in the comfort of speakers' home. The human supervision is being replaced by quality check modules ranging from acoustic surroundings checkers to our well-tuned ASR system [9], all of them providing input for the automatic decision on whenever accept the recorded phrase, or to repeat the recording, or to use another phrase from a pool of phrases. The framework must ensure the expected coverage described in Section 2.1, since some of the required units may be lost due to defects in the recording. Secondly, the whole TTS voice building process is also being automated, starting from speech corpus mastering, segmentation, statistical model training and unit selection features building – based on the decisions about the size and quality of the recorded corpus, as provided by the recording-related modules. Moreover, as our experience shows, since the patients' voices are sometimes damaged to the level that the use of ASR starts to be problematic, there still is a possibility for the speaker (or another delegated person) to manually check the recordings and fix their annotations, which can, in turn, be used to dynamically adapt the ASR models.

3. Conclusion

Our previous research served as a *proof of concept* of the viability of the idea to offer the patients the possibility to conserve their voices and to use their own "personalized" synthetic voice as one of the means of (not physically exhaustive) communication. In the recent years, we have created 20 personalized synthetic voices for laryngectomized people, and some of these patients have started to use it in their daily life, both for personal and job-related communication [1]. The samples of several patients' synthetic voices can be found at `https://goo.gl/XoUPqA` and the informations on patients who has undergone the described procedure so far are listed in Table 1.

Table 1: *The base statistics of speech corpora recorded by* 20 *patients whose voices were conserved before the surgery. The duration of speech in the corpus is in minutes excluding pauses.*

	sent.	dur.	method		sent.	dur.	method
P1	3,500	227.6	USEL,SPC	P11	684	38.4	USEL,SPC
P2	2,016	138.9	USEL,SPC	P12	683	52.6	USEL,SPC
P3	2,014	116.6	USEL,SPC	P13	555	34.7	USEL,SPC
P4	1,800	123.1	USEL,SPC	P14	473	31.5	SPC
P5	1,431	99.0	USEL,SPC	P15	469	46.2	USEL,SPC
P6	1,049	211.2	USEL,SPC	P16	403	26.2	USEL,SPC
P7	1,038	62.8	USEL,SPC	P17	350	16.0	SPC
P8	856	79.5	USEL,SPC	P18	300	17.2	SPC
P9	769	41.3	USEL,SPC	P19	230	15.3	SPC
P10	700	73.3	SPC	P20	210	15.1	SPC

As can be seen from the table, the unit selection method was successfully used from several hundreds sentences. The SPS method was successfully performed from 200 sentences without using any adaptation techniques, since the SPS-based speech synthesis method can produce satisfactory results still keeping the basic characteristics of the original patient's voice, which is not easy to be guaranteed by any voice conversion method.

However, due to the fact that the build of each voice required a lot of manual work (starting from annotation, through all the steps of TTS voice build), especially for cancer-demaged voices, works on full automation of the voice building process have been started. The system will be available online to offer voice banking for anyone interested in, not only for to-be-laryngectomized patients but also for people suffering from neurodegenerative diseases or simply for anyone who desires to make a backup of his/her voice.

4. Acknowledgements

This research was supported by the Technology Agency of the Czech Republic, project No. TH02010307, and by the grant of the University of West Bohemia, project No. SGS-2016-039. We would like to thank to anyone participated in the previous research of voice conservation, especially to our colleagues from Motol University Hospital in Prague and from New Technologies Research Centre (NTC), University of West Bohemia in Pilsen.

5. References

[1] J. Mertl, E. Žáčková, and B. Řepová, "Quality of life of patients after total laryngectomy: Struggle against stigmatization and social exclusion using speech synthesis," *Disability and Rehabilitation: Assistive Technology*, pp. 1–11, 2017. [Online]. Available: http://dx.doi.org/10.1080/17483107.2017.1319428

[2] Z. Hanzlíček and J. Matoušek, "Voice conservation: Towards creating a speech-aid system for total laryngectomees," in *Beyond AI: Interdisciplinary Aspects of Artificial Intelligence*. University of West Bohemia, Pilsen, 2011, pp. 55–59.

[3] D. Tihelka and P. Stanislav, "ARTIC for assistive technologies: Transformation to resource–limited hardware," in *Proceedings of the World Congress on Engineering and Computer Science 2011*, San Francisco, USA, 2011, pp. 581–584.

[4] Z. Hanzlíček, J. Romportl, and J. Matoušek, "Voice conservation: Towards creating a speech-aid system for total laryngectomees," in *Beyond Artificial Intelligence: Contemplations, Expectations, Applications*, J. Kelemen, J. Romportl, and E. Zackova, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 203–212.

[5] J. Matoušek, D. Tihelka, and J. Romportl, "Building of a speech corpus optimised for unit selection TTS synthesis," in *LREC 2008, proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco: ELRA, 2008, pp. 1296–1299.

[6] F. Malfrère, O. Deroo, E. Franques, J. Hourez, N. Mazars, V. Pagel, and G. Wilfart, "My-own-voice: A web service that allows you to create a text-to-speech voice from your own voice," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA*, 2016, pp. 1968–1969.

[7] J. Matoušek and J. Romportl, "On building phonetically and prosodically rich speech corpus for text-to-speech synthesis," in *Proceedings of the 2nd IASTED international conference on Computational intelligence*. San Francisco, USA: ACTA Press, 2006, pp. 442–447.

[8] M. Jůzová, D. Tihelka, and J. Matoušek, "Designing high-coverage multi-level text corpus for non-professional-voice conservation," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary*, A. Ronzhin, R. Potapova, and G. Németh, Eds. Cham: Springer International Publishing, 2016, pp. 207–215.

[9] P. Ircing, J. Psutka, and J. V. Psutka, "Using morphological information for robust language modeling in Czech ASR system," *IEEE Transactions on Audio Speech and Language Processing*, vol. 17, pp. 840–847, 2009.