IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

RESTORE Project: REpair, STOrage and REhabilitation of speech

Inma Hernáez[1], *Eva Navas*[1], *Jose Antonio Municio*[2], *Javier Gómez*[2]

[1]UPV/EHU
[2]HUC-Biocruces

inma.hernaez@ehu.eus, eva.navas@ehu.eus
joseantonio.municiomartin@osakidetza.eus, jgomezsuarez@seorl.net

Abstract

RESTORE is a project aimed to improve the quality of communication for people with difficulties producing speech, providing them with tools and alternative communication services. At the same time, progress will be made at the research of techniques for restoration and rehabilitation of disordered speech. The ultimate goal of the project is to offer new possibilities in the rehabilitation and reintegration into society of patients with speech pathologies, especially those laryngectomised, by designing new intervention strategies aimed to favour their communication with the environment and ultimately increase their quality of life.

Index Terms: alaryngeal voice, oesophageal speech, speaking aids, voice rehabilitation, statistical parametric speech synthesis, voice bank

1. Introduction

According to the Spanish Statistical Institute (Instituto Nacional de Estadística) (data from 2012), there are more than 410 000 people with 'disability to produce spoken messages' in Spain. This kind of disability, if severe, produces social isolation since the communication with the environment is seriously affected so as to hamper personal relationships and generate problems of integration at the workplace. The origin and displaying forms of the oral disabilities is varied. They might be caused by traumatic events such as stroke or surgery like Total Laryngectomy (TL) or also by degenerative diseases (such as ELA or Parkinson) which deteriorate the functioning of the motor speech system. This project is concerned with two specific groups of affected persons: on the one hand, people with a laryngectomy and on the other hand, people with dysarthria suffering from poor articulation of phonemes due to neurological injury of the motor speech system.

In the Rehabilitation Service of the Cruces University Hospital (the main public Hospital of the area of Bilbao), almost 6000 speech therapy sessions are performed annually. Among them, around 50 patients are treated after a total laryngectomy. Even though nowadays laryngeal cancer is treated with chemotherapy and radiotherapy to preserve the laryngeal function, in some cases total laryngectomy must be performed [1]. In 2014 laryngeal cancer had an incidence in Spain of 3182 per 100 000 inhabitants with a mortality of 1.3% and a 5-year prevalence of 112 per 1000 inhabitants [2]. After total laryngectomy patients must attend a number of speech therapy sessions to acquire a new voice, generally called alaryngeal voice. In the Cruces University Hospital all patients start the rehabilitation treatment 2 or 3 weeks after hospital discharge.

People affected by ELA are frequent users of Text to Speech Conversion systems, usually integrated in AAC systems (Alternative and Augmentative Communication). These AAC systems are designed to help the user to quickly build messages to be spoken out loud (with a keyboard or more sophisticated input devices). Commercial AAC systems have usually a limited choice of the synthetic voice: they are usually high quality voices sounding like a very healthy young person. Nevertheless, statistics show a reality in which a great majority of the people affected are elderly people, whose real voice would not match with the prosthetic one. Similarly, there is a lack of children's voices, and according to the data in Spain there are 32 700 children between 6 and 15 years affected by this disability.

RESTORE is a project aimed to improve the quality of communication for people with difficulties producing speech, providing them with tools and alternative communication services. At the same time, progress will be made at the research of techniques for restoration and rehabilitation of disordered speech. The following goals were proposed for the project:

1. Implementation of a donors voice bank and creation of the service of personalised synthetic speech for people with oral disabilities in general and people who have been laryngectomised, in particular at Cruces University Hospital.

2. Design and development of a Serious Game as a multimedia tool to support the speech rehabilitation process for people with voice disorders.

3. Improvement of the intelligibility of alaryngeal voices (voices from people who have been laryngectomised) and dysartric voices.

The Project coordinates the research work of two groups, one from the field of Signal Processing and Engineering and the other from the fields of Otorhinolaryngology and Speech Therapy. The Voice Banking service proposed is a groundbreaking service for the country, since there are only a few experiences with the same aim in USA (VocalID[1], ModelTalker[2]) and Europe (SpeakUnique[3]) [3]. This kind of service offers personalised synthetic speech that can be used in existing AAC devices. In our proposal the communication is achieved with a personalised voice that might be similar to the patients own voice in cases of progressive diseases or laryngectomies. Besides, the design and development of a voice training tool is pursued in order to incentive and facilitate the speech rehabilitation process, a real hard and complex task depending on the specific pathology.

In this paper we describe the main tasks of the project and briefly sketch the main results up to date.

[1] www.vocalid.co
[2] www.modeltalker.com
[3] www.speakunique.org

10.21437/IberSPEECH.2018-34

2. Tools to support the speech therapist

Total Laryngectomy surgery completely removes the larynx of the patient while separating the airway from the mouth, nose and oesophagus. Consequently, patients who undergo a TL can not produce speech sounds in a conventional manner because their vocal cords have been removed. The rehabilitation process of a patient starts immediately upon confirmation of the surgery. Through a pre-surgical interview an orientation framework is offered both to the patient and his or her family where they will receive information about:

- The anatomical and physiological changes that result from surgery
- The way to communicate during the period immediately following surgery
- The speech therapy sessions that will follow surgery

The main objective of the rehabilitation after TL is to return to the patients the possibility of oral communication for reintegration into their social, work and personal life. After surgical intervention, additionally to medical treatment and the importance of tracheostomy protection and care of the tracheal cannula the patient will be informed about the therapy process to acquire alaryngeal voice. There are three possibilities for voice rehabilitation:

- Oesophageal speech
- Tracheoesophageal speech
- Use of an electrolarynx

Oesophageal speech (ES) is preferred by medical doctors because it does not require a voice prosthesis, but it is also most effortful and difficult to acquire. Tracheoesophageal speech is the most successful method and also produces the most understandable speech, but requires a voice prosthesis placed during total laryngectomy or later in a secondary puncture. Finally, the electrolarynx is an external vibrating handheld device which is placed to the neck or the face. The vibrating sound is modulated by the movements of the articulators to produce understandable speech. It produces a robotic voice and it is sometimes used also as a backup secondary method.

In Cruces University Hospital laryngectomised patients start rehabilitation after 2 or 3 weeks after hospital discharge with the aim of learning to produce oesophageal speech. The patient attends around 50 rehabilitation sessions during a period of 4 months. If the final speaking method is tracheoesophageal, the average learning period is only 5 days.

In RESTORE we have developed an interactive video aimed at helping the patient during and after this rehabilitation period. This video considers the main difficulties faced by the laryngectomised patient and proposes exercises and advices to overcome them. Using a comic style representation of a food market, the main character represents the patient itself, going through the different market stalls, in each of which he or she will practice a new rehabilitation exercise. Video recordings of real sessions with a speech therapist are also included, as well as short interviews with laryngectomised persons that have succeed in the rehabilitation process and share their own feelings and experiences.

Clinical evaluation of the developed tool is currently taking place with laryngectomised patients.

3. Personalisation of the synthetic voice

One of the goals of RESTORE project is to improve the already existing ZureTTS voice bank web portal[4] making it more flexible and allowing to provide a personalised TTS service to people with oral disabilities.

3.1. Voice bank and web portal

In the previous version of the voice bank web portal, each voice donor had to record 100 sentences to get his or her personalised voice. This is not a problem for healthy people, but many patients are not able to record such a long corpus. Therefore the original 100 sentences corpus has been divided into three corpora of 33, 33 and 34 sentences. Each donor can choose how many corpora to record and the personalised voice will be produced with the available speech material. Besides, two new languages have been incorporated to the portal: Gaelic and the Navarro-Lapurdian dialect of Basque.

3.2. Recording protocol for pre-laryngectomised people

To be able to generate a personalised voice for laryngectomees, the ideal situation is to have recordings of the patient made prior to the surgery. If this is not the case, recordings made by close family members can be used to produce a personalised synthetic voice for the patient, as voices are usually similar among family members of the same gender [4] [5].

The first step to get these recordings is to introduce the recording procedure in the hospital protocol. This protocol has established the following criteria to select patients that can take part in the project:

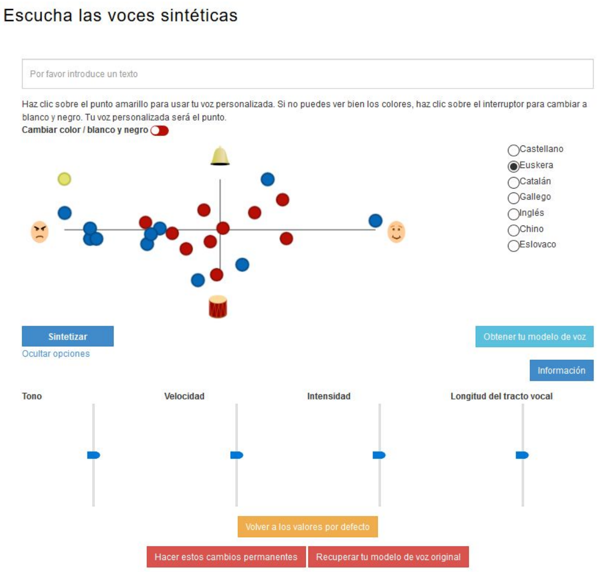
1. Patients older than 18 years old with a TL programmed.
2. Close family members with voices similar to the patient.
3. Any patient older than 18 years old without any speech pathology who comes to the otorhinolaryngology service of the Cruces University Hospital for any reason.

Once these criteria have been fixed, the protocol has been sent to the Basque Ethics Committee for approval.

3.3. Personalisation for dysarthric voices

One person with ELA make use of the ZureTTS portal to obtain a synthetic voice. However, the voice was already affected by the disease and resulting synthetic voice also showed the same problems of the original voice. The main issue was the rhythm, which was very slow, with very long vowels and very frequent long pauses, thus confirming other studies [6][7]. Also, some consonants were poorly realised. The slow prosody provoked the malfunctioning of the automatic alignment algorithm thus contributing to the low quality of the synthetic voice. Nevertheless, even when a new specific alignment method adapted to the multiple pauses was applied, the synthetic voice was not of the desired quality, mainly because of the long phones. To overcome it, prosody transplantation from a healthy voice was performed with good results. This voice was offered to the user, instead of the one automatically provided by the system. We also tried to palliate the pronunciation problems applying the adaptation techniques using only vowels described in [8][9], but the new synthetic voice, although with an improved pronunciation lost the personality of the speaker. We are also experimenting with model surgery on some phones [10], but the improvements are subtle.

[4] aholab.ehu.eus/zureTTS



Escucha las voces sintéticas

Por favor introduce un texto

Haz clic sobre el punto amarillo para usar tu voz personalizada. Si no puedes ver bien los colores, haz clic sobre el interruptor para cambiar a blanco y negro. Tu voz personalizada será el punto.

Cambiar color / blanco y negro

○Castellano
●Euskera
○Catalán
○Gallego
○Inglés
○Chino
○Eslovaco

Sintetizar
Ocultar opciones
Obtener tu modelo de voz
Información
Tono
Velocidad
Intensidad
Longitud del tracto vocal
Volver a los valores por defecto
Hacer estos cambios permanentes
Recuperar tu modelo de voz original

Figure 1: *Catalogue of voices in ZureTTS voice bank portal.*

3.4. Synthetic voices catalogue

If the person with oral disabilities is not able to record the corpus and there is no family member with a similar voice, he or she is still able to get a customised synthetic voice, by selecting the one of his or her preference among all the donated voices.

To make the selection of a personalised voice easier, a catalogue with the available voices has been included in the WEB portal. A subjective evaluation where listeners qualified all the synthetic voices according to several attributes was developed. These attributes were: white - hoarse voice, sweet - dominant voice, warm - high-pitched voice, clean - nasal voice and monotonous - expressive voice. A bidimensional representation of the voices using the two most discriminative dimensions (sweet-dominant and white-hoarse) has been integrated in the portal, as shown in Figure 1.

The voices can be easily modified in tone, rhythm, intensity and vocal tract length to get a synthetic voice that pleases the user. The customised voices can be obtained from the web portal in standard format for Android OS, iOS and Windows, so they can be directly used by Augmentative and Alternative Communication Devices.

4. Voice conversion

In the production of oesophageal speech the pharyngo-oesophageal segment is used as a substitutive vibrating element for the vocal folds. Due to the nature of the intervention, the air used to create the vibration of the oesophagus can not come from the lungs and the trachea as happens during normal speech production. Instead, the air is swallowed from the mouth and introduced in the oesophagus, being then expelled in a controlled way while producing the vibration. These huge differences in the production mechanisms lead to a diminution of naturalness and intelligibility [11][12][13]. As a consequence, the communication with others is hindered. Moreover, these less intelligible voices are an added problem for the automatic speech recognition algorithms that are becoming ubiquitous in the human computer interaction technologies. One of the goals of this project is the development of techniques and algorithms aimed at modifying oesophageal speech in such a way as to improve the performance of a state of the art ASR system with these modified signals as input. To achieve this goal we decided to experiment with voice conversion (VC) algorithms. In this section we summarize the efforts done within the project in this direction.

4.1. ASLABI Database

To our knowledge, there are not publicly available databases for oesophageal Spanish speech. There are studies published concerning the quality and characteristics of ES, some or them for Spanish, but they use their own recording data, mostly developed for the purpose of the specific research [14][15][16][17]. Additionally, most VC algorithms make use of parallel databases. The availability of recorded voices for the 100 phonetically balanced sentences of ZureTTS for healthy speech made us decide the recording of the same set of sentences for oesophageal speakers. To make the recordings, we contacted the local Association *ASociación de LAringectomizados de BIzkaia* who showed an enormous interest in the project and collaborated with enthusiasm. A total of 32 persons went to make the recordings to the acoustically isolated room in our Faculty. The database also includes one Basque session.

4.2. Improving the intelligibility of the oesophageal voice

We have tried several strategies to improve the intelligibility of the oesophageal speech. First, we tried a classical GMM based voice conversion, using parallel data. This was followed by DNN based approaches, using LSTM and more recently also including a WaveNet vocoder.

There are several specific problems that must be faced when processing and converting oesophageal signals:

- Oesophageal signals lack the regular periodicity (intonation) typical of laryngeal signals. Although they have a certain periodicity at certain segments, the fundamental frequency is very low (around 80Hz) and very irregular. Usual F0 calculation algorithms generate many errors and do not result in a realistic measure of the local periodic segments. Thus a specific F0 detection algorithm has been used.
- The rhythm in general, the duration of syllables and the duration of the phones inside them, vary significantly in relation to healthy speech. Additionally, noises and pauses are inserted in between words and syllables even inside the syllable. Therefore, the alignment of healthy and oesophageal parallel sentences becomes a tricky task.
- Many phones (mainly corresponding to consonants) in the sounds stream are not present in the signal or they are realised in a completely different way. This fact also complicates the parallel alignment task and generates many recognition errors.
- In general, a fundamental frequency curve must be estimated for the converted spectrum (except for the case of using WaveNet). Simple conversion of the source F0 values as usually done in the VC field is not feasible, which opens a wide range of research possibilities.

To evaluate the intelligibility of the resulting converted signals we have used a Kaldi-based ASR system [18] trained with material described in [19]. This approach was selected because it allows us to control the exact processing operations followed during the recognition (such us the use of transformations like fMLLR) as well as basic aspects of the recognition process such as the lexicon and the language model. This turned out to be very important with our reduced set of 100 phonetically rich sentences, containing many very unlikely words, proper names etc. The procedure and results of the different ASR tests are described in [20][21].

5. Discussion and future work

This project represents an effort to promote modern technological advances in the area of speech processing among the group of people with oral disabilities. In particular we have put special emphasis to introduce the benefits of the advances in speech synthesis in the rehabilitation process of the laryngectomised people.

In relation to the intelligibility of oesophageal speech, we plan to improve the techniques to evaluate not only human intelligibility but also the effort employed by the listener ('listening effort').

At the time of writing this paper, only two laryngectomised persons have been recorded for the voice bank previous to TL. It must be taken into account that the elapsed time between the communication of the need to do the TL surgery and the surgery itself is very short (usually of a few days). So there is not the necessary time for the medical team to explain the patient the future benefits of making the recordings. Additionally, these patients have their voice already very harsh and speak with difficulties (it is usually the reason why they have contacted the unit). This is why we consider the catalogue of synthetic voices, possibly including voices of the patient's relatives as a very good alternative. We hope that this voice bank pilot experiment will continue growing and expanding the service to other hospitals after the end of the present project.

6. Acknowledgements

This project has been founded by the Spanish Ministry of Economy and Competitiveness with FEDER support (RESTORE project, TEC2015-67163-C2-1-R and TEC2015-67163-C2-2-R).

7. References

[1] J. P. Rodrigo, F. López, J. L. Llorente, C. Álvarez-Marcos, and C. E. Suarez, "Results of total laryngectomy as treatment for locally advanced laryngeal cancer in the organ-preservation era." *Acta otorrinolaringologica espanola*, vol. 66 3, pp. 132–8, 2015.

[2] S. Sociedad Española de Oncología Médica, "Las cifras del cáncer en España 2014," 2014.

[3] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Nov 2013, pp. 1–4.

[4] H. S. Feiser and F. Kleber, "Voice similarity among brothers: evidence from a perception experiment," in *Proc. 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, 2012.

[5] I.-S. Ahn and M.-J. Bae, "On a similarity analysis to family voice," *Advanced Science Letters*, vol. 24, no. 1, pp. 744–746, 2018.

[6] W. G, J. JY, L. JS, K. RD, and K. JF, "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," *Folia Phoniatrica et Logopaedica*, vol. 53, no. 1, pp. 1–18, 2001.

[7] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in als: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, no. 7-8, pp. 494–500, 2013.

[8] A. Alonso, D. Erro, E. Navas, and I. Hernaez, "Speaker Adaptation using only Vocalic Segments via Frequency Warping," in *INTERSPEECH 2015*, 2015, pp. 2764–2768.

[9] ——, "Study of the effect of reducing training data in speech synthesis adaptation based on frequency warping," in *Advances in Speech and Language Technologies for Iberian Languages*, A. Abad, A. Ortega, A. Teixeira, C. García Mateo, C. D. Martínez Hinarejos, F. Perdigão, F. Batista, and N. Mamede, Eds. Cham: Springer International Publishing, 2016, pp. 3–13.

[10] A. Pierard, D. Erro, I. Hernaez, E. Navas, and T. Dutoit, "Surgery of speech synthesis models to overcome the scarcity of training data," vol. 10077 LNAI, pp. 73–83.

[11] B. Weinberg, "Acoustical properties of esophageal and tracheoesophageal speech," *Laryngectomee rehabilitation*, pp. 113–127, 1986.

[12] T. Most, Y. Tobin, and R. C. Mimran, "Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production," *Journal of communication disorders*, vol. 33, no. 2, pp. 165–181, 2000.

[13] T. Drugman, M. Rijckaert, C. Janssens, and M. Remacle, "Tracheoesophageal speech: A dedicated objective acoustic assessment," *Computer Speech & Language*, vol. 30, no. 1, pp. 16–31, 2015.

[14] R. Ishaq and B. G. Zapirain, "Esophageal speech enhancement using modified voicing source," in *Signal Processing and Information Technology (ISSPIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 000 210–000 214.

[15] A. Mantilla, H. Pérez-Meana, D. Mata, C. Angeles, J. Alvarado, and L. Cabrera, "Recognition of vowel segments in spanish esophageal speech using hidden markov models," in *Computing, 2006. CIC'06. 15th International Conference on*. IEEE, 2006, pp. 115–120.

[16] A. Mantilla, H. Perez-Meana, D. Mata, C. Angeles, J. Alvarado, and L. Cabrera, "Analysis and recognition of voiced segments of esophageal speech," in *Electronics, Robotics and Automotive Mechanics Conference, 2006*, vol. 2. IEEE, 2006, pp. 236–244.

[17] A. Mantilla-Caeiros, M. Nakano-Miyatake, and H. Perez-Meana, "A pattern recognition based esophageal speech enhancement system," *Journal of applied research and technology*, vol. 8, no. 1, pp. 56–70, 2010.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[19] L. Serrano, D. Tavarez, I. Odriozola, I. Hernaez, and I. Saratxaga, "Aholab system for albayzin 2016 search-on-speech evaluation," in *IberSPEECH*, 2016, pp. 33–42.

[20] L. Serrano, D. Tavarez, X. Sarasola, S. Raman, I. Saratxaga, E. Navas, and I. Hernaez, "LSTM based voice conversion for laryngectomees," in *Proc. IberSPEECH 2018*, 2018.

[21] S. Raman, I. Hernaez, E. Navas, and L. Serrano, "Listening to laryngectomees: A study of intelligibility and self-reported listening effort of spanish oesophageal speech," in *Proc. IberSPEECH 2018*, 2018.