PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

A multifaceted enrichment of oesophageal speech

Sneha Raman, Inma Hernaez, Eva Navas, Luis Serrano

University of the Basque Country (UPV/EHU), Spain, sneha.raman@ehu.eus

Abstract

Post a laryngectomy, oesophageal speakers struggle with communication owing to limitations in producing speech. Oesophageal speech (ES), therefore, is of poor intelligibility and demands increased listening effort. With the aim of improving intelligibility for an Automatic Speech Recognition (ASR) system, we performed three enrichment tasks on oesophageal speech: one heavy-weight strategy (voice conversion) and two light-weight strategies (cleaning up of undesired silences and synthesis with Wavenet vocoder that generates high quality natural sounding speech). While voice conversion (targeted at a healthy speaker) is the conventional approach to improve oesophageal speech, we were motivated to try the aforementioned light-weight strategies, as they are simple and straightforward, and any improvement with these methods will add to the benefits gained from other methods. Voice conversion was the most useful enrichment technique in improving ASR scores. Synthesizing with Wavenet improved the voice quality of some parts of ES but it also added some unwanted glitches. As a result, it did not improve the ASR scores for ES. Removing undesired silences did not have any improvement in ASR for a well-trained ES speaker as there were few instances of undesired silences. However, for a poorly intelligible speaker it showed marginal improvement.

Keywords: Pathological speech, Laryngectomy, Enrichment, Automatic Speech Recognition

1 INTRODUCTION

A large number of people (0.4% of the total population of Europe) have speech impairments (1). People with speech disorders have unsatisfying social lives, as they are difficult to understand and communicate with. Voice rehabilitation for pathological speech is a way to alleviate this issue by providing restored voices to people with speech disorders.

Oesophageal Speech (ES) is a pathological speech resulting from the removal of the larynx. It is known to have poorer intelligibility and demands more effort to listen to compared to healthy laryngeal speech (6,7,8).

In this age, machines are beginning to take the top spot as participants of speech communication. This study focuses on improving machine communication for oesophageal speakers by enhancing the intelligibility of their speech using several speech modifications techniques.

2 THE PROBLEM AND OBJECTIVES

Laryngectomy involves the removal of the source or fundamental frequency of the speech. Therefore, the primary enrichment to make it sound like natural speech involves the addition or reconstruction of the fundamental frequency (2). Voice conversion has been used for restoration of ES using statistical methods (3).

Another way of enrichment is by removing the unwanted artefacts introduced while speaking, such as swallowing air to speak. These pauses of swallowing that are necessary for ES speech production, disturb the durations and hence, the rhythm of the speech. Smearing of rhythm of speech is known to affect recognition (10). Moreover, the swallowing sounds are unpleasant to listen to, affecting the process of listening.

The objective of this study is to implement some light-weight strategies directly on the speech, such as improving the rhythmic structure of the speech by eliminating undesired silences and improving spectral characteristics by resythesizing with a high quality natural sounding vocoder. We have used one high intelligibility ES (HIES) speaker and one low intelligibility ES (LIES) speaker in our study.

3 METHODS

3.1 Alignment and removal of undesired silences

Segmentation and labelling of ES is a tricky process. The forced alignment feature built into our generic Spanish Automatic Speech Recognition (ASR) systems was unsuitable for ES. Therefore, using the Montreal Forced Alignment tool (5), new models were made for ES. Segmentation with this forced aligner gave us the labels of pauses in the signal. We removed portions in the speech that corresponded to these pauses or silences. A comparison of the number of pauses and the their duration for HIES and LIES is presented in the results.

3.2 Wavenet synthesis

Wavenet (4) can be used as a vocoder that is known to generate high quality natural sounding speech (11). We used this ability of wavenet by extracting features (Mel filter bank parameters) from ES and synthesising samples from wavenet using these extracted features as local conditioning.

4 RESULTS AND DISCUSSION

In our previous work, we performed voice conversion for HIES using GMM and LSTM (9) and the ASR performance improved by nearly 20% (Table 1).

Removing silences showed improvement for LIES (4.17%, see Table 1) but not HIES. This is possibly because HIES had fewer pauses. The mean number of pauses per utterance and mean duration of pauses for LIES (13.99±4.18, 280±53 ms) was higher than that for the HIES (7.56±2.29, 169±60 ms).

Wavenet was not useful in improving ASR, although during informal listening tests, it is reported to be preferred over other enrichment methods.

Speech Type	ASR Scores: Word Error Rate in %	
	High Intelligibility ES	Low intelligibility ES
Unprocessed Oesophageal Speech	57.00	93.19
Voice Conversion GMM	38.99	NA
Voice Conversion LSTM	40.35	NA
Removal of undesired silences	57.99	89.02
Wavenet synthesis	57.46	94.55

Table 1. ASR scores for unprocessed and enriched oesophageal speech for high and low intelligibility speech. Green text shows numbers where improvement was observed

Our next step is to retain the silences that are desired in the speech (inter-word silences for instance). Also, we will evaluate these systems from a Human Speech Recognition and listening effort perspective too.

5 CONCLUSIONS

While voice conversion remains the greatest contributor in improving ASR performance, it was observed that undesired silence removal was beneficial for low intelligibility ES. Synthesis with a rich vocoder did not help in ASR improvement but it has scope for positive responses in perceptual evaluations.

ACKNOWLEDGEMENTS

This project was supported by funding from the EUs H2020 research and innovation programme under the MSCA GA 67532*4 (the ENRICH network: www.enrich-etn.eu)

REFERENCES

[1] Dupre D.; Karjalainen A. Employment of disabled people in Europe in 2002, Eurostat-Your key to European statistics, 2003, https://ec.europa.eu/eurostat/documents/3433488/5542140/KS-NK-03-026-EN.PDF/0b806b41-0898-45d2-ac99-0f085e983887 (Accessed: 07 Nov 2018).

[2] Sharifzadeh HR, McLoughlin IV, Ahmadi F. Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. IEEE Transactions on Biomedical Engineering. 2010 Oct;57(10):2448-58.

[3] Doi H, Nakamura K, Toda T, Saruwatari H, Shikano K. Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models. IEICE TRANSACTIONS on Information and Systems. 2010 Sep 1;93(9):2472-82.

[4] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior AW, Kavukcuoglu K. WaveNet: A generative model for raw audio. SSW. 2016 Sep 13;125.

[5] McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. InInterspeech 2017 (pp. 498-502).

[6] Most T, Tobin Y, Mimran RC. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. Journal of communication disorders. 2000 Mar 1;33(2):165-81.

[7] Weinberg B. Acoustical properties of esophageal and tracheoesophageal speech. Laryngectomee rehabilitation. 1986:113-27.

[8] Raman S, Hernaez I, Navas E, Serrano L. Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech. Proc. IberSPEECH 2018. 2018:107-11.

[9] Serrano L, Tavarez D, Sarasola X, Raman S, Saratxaga I, Navas E, Hernaez I. LSTM based voice conversion for laryngectomees. Proc. IberSPEECH 2018. 2018:122-6.

[10] Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech reception. The Journal of the Acoustical Society of America. 1994 Feb;95(2):1053-64.

[11] Tamamori A, Hayashi T, Kobayashi K, Takeda K, Toda T. Speaker-Dependent WaveNet Vocoder. InINTERSPEECH 2017 Aug (pp. 1118-1122).