IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech.

Sneha Raman, Inma Hernaez, Eva Navas, Luis Serrano

AHOLAB Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Spain
sneha.raman@ehu.eus, inma.hernaez@ehu.eus, eva.navas@ehu.eus, lserrano@aholab.ehu.eus

Abstract

Oesophageal speakers face a multitude of challenges, such as difficulty in basic everyday communication and inability to interact with digital voice assistants. We aim to quantify the difficulty involved in understanding oesophageal speech (in human-human and human-machine interactions) by measuring intelligibility and listening effort. We conducted a web-based listening test to collect these metrics. Participants were asked to transcribe and then rate the sentences for listening effort on a 5-point Likert scale. Intelligibility, calculated as Word Error Rate (WER), showed significant correlation with user rated effort. Speaker type (healthy or oesophageal) had a major effect on intelligibility and effort. Listeners familiar with oesophageal speech did not have any advantage over non familiar listeners in correctly understanding oesophageal speech. However, they reported lesser effort in listening to oesophageal speech compared to non familiar listeners. Additionally, we calculated speaker-wise mean WERs and they were significantly lower when compared to an automatic speech recognition system.
Index Terms:Spoken language understanding, Speech intelligibility, Speech and voice disorders, Pathological speech and language, Speech perception

1. Introduction

Oesophageal speech is an equipment-free speech production method used by people whose larynx has been surgically removed (laryngectomees). In spite of the absence of vocal folds, they can still utter intelligible speech using alternative vibrating elements. In the production of oesophageal speech, the pharyngo-oesophageal segment is used as a substitutive vibrating element. Air is swallowed from the mouth and is introduced into the oesophagus, after which it is expelled in a controlled way, thereby producing vibration. This generation mechanism introduces acoustic artefacts and makes oesophageal speech difficult and effortful to understand [1] [2], which greatly affects communication, interpersonal relationships and hence, quality of life. Moreover, these less intelligible voices are problematic for Automatic Speech Recognition (ASR) systems that are becoming ubiquitous in human-computer interaction technologies. The aim of the work presented in this paper is to quantify the difficulty in understanding oesophageal speech by measuring intelligibility and listening effort. Intelligibility is quantified in both Human Speech Recognition (HSR) and ASR contexts.

Several studies have devised systems for the analysis of pathological speech which also includes intelligibility measurements [3] [4] [5]. In [3] the authors use a Hidden Markov Model based Automatic Speech Recognition (ASR) system to measure objective intelligibility of laryngectomees, and also cleft lip and palate speech. A similar tool for Dutch pathological speech intelligibility calculation is proposed in [5]. In [4] some intelligibility measurement techniques that do not use ASR are described. The main advantage of measures based on ASR is that it is an objective measure and therefore easy to replicate and to implement. However it only evaluates machine intelligibility, and not how intelligible it is to humans. It also does not consider other important factors like pleasantness, acceptability or listening effort.

Some studies have been conducted in measuring the intelligibility of Spanish oesophageal speech. In [6] the voice intelligibility characteristics for Spanish oesophageal and tracheoesophageal speech is reported. This study was conducted for two syllable words. Another study [7] showed how the formant frequencies were higher and the duration of vowels was longer for laryngectomees as compared to healthy speech. The work in [8] describes a real time recognition system for vowel segments of Spanish oesophageal voice.

The above mentioned studies focus on the micro level of words and vowels. Sentence level HSR studies on the intelligibility of Spanish oesophageal voice is a less traversed area of investigation. In this study, sentences are used as our stimuli.

The downside of intelligibility measurements is that they indicate only how many words have been correctly identified but not how difficult it was to identify them. This does not do justice to the problem of how effortful the listening was, especially in adverse listening conditions such as pathological speech. A review of listening effort and various methods of measuring listening effort is presented in [9].

Some research focussed on measuring the processing load associated with oesophageal speech. In [10] the authors measured the acceptability of oesophageal, electro-laryngeal and healthy speech. They found that healthy speech was the most acceptable, followed by superior oesophageal speech and then artificial larynx speech. In [11] high intelligibility tracheoesophageal speech was played to listeners and they were asked to rate the effort of listening as well as acceptability for each sample and found an inverse correlation between listening effort and acceptability. Another observation from this study was that even highly intelligible speech can have varying listener effort. In this study we attempt to explore this processing load phenomenon in addition to the ASR and HSR intelligibility measurements. Firstly, we investigate whether the intelligibility (both ASR and HSR) of healthy and disordered speech is comparable. Secondly, we are interested in seeing if the intelligibility and listening effort are correlated.

In [12], the idea of intelligibility differences between experienced and inexperienced listeners of oesophageal speech was explored and the findings stated that oesophageal speech was ranked similarly for intelligibility by both experienced and inexperienced listeners. Following this thread, we were interested in investigating if the same result was observed for our dataset for listeners that are familiar and unfamiliar with oesophageal speech. We consider friends, family and close relatives of oesophageal speakers as familiar listeners.

10.21437/IberSPEECH.2018-23

In short, the hypotheses of the experiment described in this paper are:

- Intelligibility or Word Error Rate measurement is correlated with user rated listening effort.
- Healthy voices are more intelligible and less effortful, compared to oesophageal voices.
- Listeners familiar with oesophageal speech find it less effortful to process, compared to listeners that are not.
- ASR performs worse than HSR for healthy voices, but even more so for oesophageal voices.

We begin by describing the methodology, corpus and details of the listening test. This will be followed by analysis methods used and results. Finally, the conclusions and future work are presented.

2. Methodology

2.1. Experimental Design

The main task for this experiment was the word recall and transcription task: Participants listened to a sentence and then wrote what they have understood. According to [13] the strengths of sentence repetition tasks are that they are "fairly simple cognitive tasks" and that they are "consistent throughout the age span" in the area of neurophysiological tests. Moreover, sentence transcription tasks have been widely used for subjective intelligibility measurements. The work in [14] reports the agreement of sentence transcription tasks with a wide range of intelligibility quantification techniques and in [15] the method is described as "human speech recognition". Therefore, we chose this approach to calculate WER and consequently the intelligibility.

We were also interested in knowing the listening effort of these utterances. We had the participants rate the sentences for listening effort on a 5 point Likert scale. The options were 'very little','a little', 'some', 'quite' and 'a lot'.

To avoid priming and sentence order bias, the sentences were played only once and in a random order.

2.2. Corpus and Stimuli

The parallel data used for this task is 100 phonetically balanced sentences selected from a bigger corpus [16], recorded by 35 healthy speakers and 32 oesophageal speakers. The recordings of oesophageal speakers were done in an acoustically isolated room with a studio microphone (Neumann TLM 103). The recordings of the healthy speakers have variable sources because they have been acquired through an online platform [17]. However, some of them were made in the aforementioned acoustically isolated room although with a different microphone.

2.2.1. Selection of Speakers

For this experiment we chose oesophageal speakers based on two criteria: proficiency and accessibility. Proficient speakers were those who underwent laryngectomy, and had begun training to speak for at least two years prior to the recording. Additionally, an oesophageal voice quality assessment tool [18], based on the factors (speaking rate, regularity etc.) of the A4S scale of [19], was used as a guide to assess proficiency. Accessibility of speakers was considered because the opportunity to obtain follow-up recordings could be useful for future research. Based on these criteria, we chose 4 speakers, three male and one female, making it gender inclusive (there are only 4 women in the whole database and only 2 of them fulfilled the two criteria of proficiency and accessibility).

The criteria for choosing healthy speakers was quality of recording as well as gender balance. One male and one female healthy speaker was chosen.

2.2.2. Selection of Sentences

A pilot listening test was conducted within the lab to assess the feasibility of this corpus for the sentence transcription task. The participants chosen for this pilot study were unfamiliar with the sentences of the corpus and thus not subject to priming. After the pilot test, the participants reported that some sentences were too long to remember and hence, effortful to transcribe. Additionally, although semantically and syntactically correct, the sentences were rich in content and contained words that are difficult to guess, often containing proper names, dates etc.

This led us to reconsider the length of sentences and we decided to choose a subset of shorter sentences, which would make them suitable for sentence transcription. We used the CorpusCRT tool [20] which generates a phonetically balanced subset of sentences based on the provided phonetic criteria. In this case, the criteria we used was a maximum of 40 phonemes and this gave us a set of 30 sentences, each of which had a maximum of 10 words. Some examples of the sentences are the following: '¿Qué diferencia hay entre el caucho y la hevea?' *What is the difference between rubber and hevea?*, 'Unos días de euforia y meses de atonía.' *A few days of euphoria and months of atony.*

All the selected sentences (both from oesophageal and healthy speakers) were normalised to a common peak value (0.8) to achieve a homogeneous and comfortable level of loudness.

2.3. The Listening Test

We created six mutually exclusive sets of sentences such that each set contained 30 different sentences and exactly 5 sentences from each speaker. As a result, all 180 sentences (30 sentences from six speakers) was covered after every sixth participant. This ensured equal coverage of all sentences and speakers. Each participant was assigned one of these sets and they listened to the sentences in a random order.

The participants were asked to use headphones for the study unless impossible. They were assured that it was not a test of hearing and that the test was being conducted to obtain their honest and uninhibited response. They were told that the sentence could be played only once and that they should pay close attention and type what they hear. If they missed some portions or were unsure of what they heard, they could put three dots (...) in that place. Additionally, they were asked to mark a response for the amount of effort they experienced for that sample on the aforementioned Likert scale. The first couple of sentences that were presented were practice sentences (one healthy and one oesophageal), to familiarise the participant with the task. These sentences were sampled from the same corpus [21] but different from the ones that appear in the actual test.

We took the following information from the participants: age, presence of hearing impairment, the kind of audio equipment used (good quality headphones, normal quality headphones, good loudspeakers and bad equipment) and whether the listener had close contact with laryngectomees.

The listening test[1] was web based and it was possible to

[1] https://aholab.ehu.eus/users/sneha/Listening_test.php

reach out to a wide range of participants. However, this also meant differences in audio equipment and the effects of this on the responses are reported in the results section.

2.4. Automatic Speech Recognition

To have an objective measure of the intelligibility (WER) we prepared an ASR system for Spanish using the Kaldi toolkit [22]. This approach was chosen as it allowed us to control the processing operations followed during the recognition as well as basic aspects of the recognition process such as the lexicon and the language model. It is implemented following the recipe s5 for the Wall Street Journal database. The acoustic features used are 13 Mel-Frequency Cepstral Coefficients (MFCCs) to which a process of mean and variance normalization (CMVN) is applied to mitigate the effects of the channel. The details of the training procedures are described in [23].

The audio material used to train the Spanish recogniser was healthy laryngeal speech as described in [24]. However, due to the characteristics of the sentences used for the evaluation, some modifications were made in this ASR system. Although the acoustic models were maintained, a new lexicon was created from the 100 sentences corpus used in the experiment (701 words). This was done because using the original lexicon (with $37,632$ entries) as much as 23% of the words were out of vocabulary (OOV) words. This is due to the fact that the sentences are phonetically balanced and many sentences containing proper names and many unusual words were chosen to maximize the variability of the phonetic content. Together with this reduced lexicon, a unigram language model with equally probable words was used.

Although the final WER numbers obtained in this way are not comparable to a realistic ASR situation, the procedure serves our purpose of evaluating the intelligibility of the sentences, comparing the performance of healthy and oesophageal speakers, and establishing a baseline reference for future developments in the field (such as evaluating the improvements of speech modification algorithms).

3. Analysis and Results

We had 57 native Spanish participants in this test, out of which 15 of them had close contact with laryngectomees and hence were familiar with oesophageal speech. The age of listeners ranged from 21 to 70 and mean age was 36.6.

Prior to calculating WER, an initial clean-up was performed on the data. This included removing any punctuations or special characters, and some typing errors (accented vowels, use of upper and lower case, spelling of proper or foreign names etc.). The WER was obtained after correcting these transcription errors.

WER was calculated [25] using the Levenshtein distance between the reference sentence and the hypothesis sentence (the sentence transcribed by the listener). This method calculates the distance by quantifying the insertions, deletions and substitutions that are observed in the hypothesis sentence when compared to the reference sentence.

Self reported listening effort responses were assigned numeric values that ranged from 1 to 5 with 1 corresponding to 'very little' and 5 to 'a lot'.

We performed ANOVA analysis on the dataset using the JASP tool [26] to quantify the effects of speaker type and familiarity of the listener with oesophageal speech. The audio device used by the listener had no effect on the HSR WER results (F(3,1256)=0.707, p=0.548) and on listening effort (F(3,1256)=0.705, p=0.549).

In addition, we present the WER results from the ASR system for all speakers.

Table 1: *Mean WER and Effort*

		Oesophageal	**Healthy**
Effort	Familiar	2.61	1.25
	Not familiar	3.54	1.26
	Total mean effort	**3.07**	**1.255**
WER (in %)	Familiar	17.39	7.42
	Not familiar	18.35	4.85
	Total mean WER	**17.87**	**6.16**

3.1. Word Error Rates from HSR

Table 1 presents mean WERs and Figure 1 shows the speaker-wise WERs for familiar and unfamiliar listeners. OM, OF, HM, HF are acronyms for Oesophageal Male, Oesophageal Female, Healthy Male and Healthy Female respectively. Mean WER is always higher for oesophageal speech compared to healthy speech, as expected. There is no major difference in the WER for familiar and unfamiliar listeners in the case of oesophageal speech. This result corroborates the conclusions in [12]. For healthy speech there is slight difference of around 3 points in the mean WER, but as can be seen in Figure 1 the difference is not meaningful.

The ANOVA results show that familiarity with oesophageal speech had no effect on WER (F(1,1590)=0.360,0.548). On the other hand, speaker-type has a strong effect on WER (F(1,1590)=129.552, p<0.001).

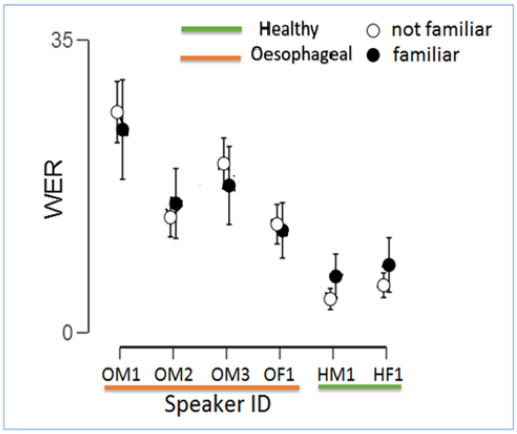

Figure 1: *Mean speaker-wise Word Error Rates for oesophageal (OM1, OM2, OM3, OF1) and healthy (HM1, HF1) speakers. Error bars show 95% confidence intervals*

3.2. Self-reported Listening Effort

Mean self-reported listening effort values are stated in Table 1 and Figure 2 shows the speaker-wise values. As expected, it is higher for oesophageal speech compared to healthy speech. However, when listening to oesophageal speech the perceived effort is significantly lower for familiar listeners than for not familiar listeners. Indeed, ANOVA analysis shows

that familiarity with oesophageal speech has an effect on effort ($F(1,1590)=84.94$, $p<0.001$) and Speaker-type has a strong effect on effort ($F(1,1590)=1243.94$, $p<0.001$).

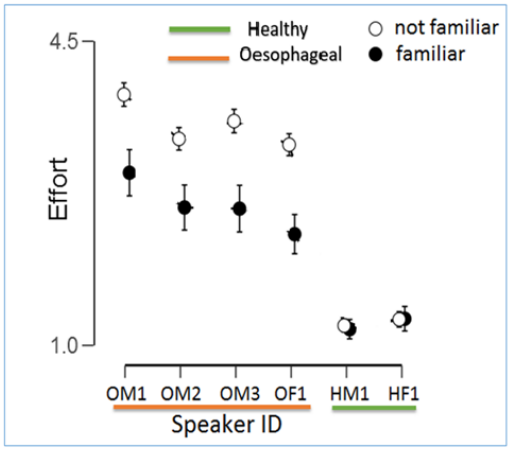

Figure 2: *Mean speaker-wise self reported effort values for oesophageal (OM1, OM2, OM3, OF1) and healthy (HM1, HF1) speakers. 1 corresponds to least effortful and 5 to most effortful. Error bars show 95% confidence intervals*

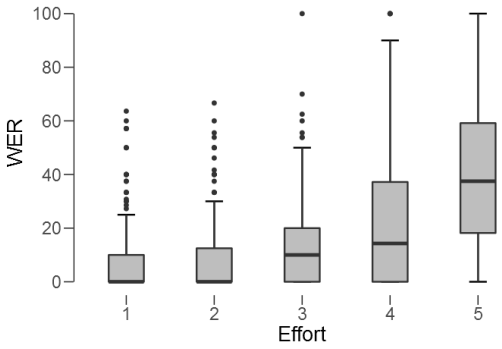

Figure 3: *Correlation between WER and user rated listening effort*

3.3. Correlation of Intelligibility and Listening Effort

Correlation between intelligibility (WER) and self reported effort is 0.479 (Pearson's r, $p < 0.001$). This is a weak but significant correlation that indicates that sentences with more transcription errors are perceived as more effortful. This relationship between WER and self-reported effort is illustrated as a box-plot in Figure 3.

3.4. Word Error Rates from ASR

The ASR experiment was performed using all the 100 available sentences for each speaker and not only the subset used for human intelligibility measurements. This was convenient in order to obtain a reliable WER measure. It can be observed from Figure 4 that the ASR performs poorly for both healthy and oesophageal speech. The fact that the system is using a unigram language model contributes greatly to this poor performance. As expected, WER for oesophageal speakers is significantly higher than healthy speakers.

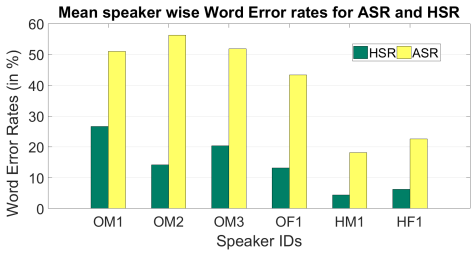

Figure 4: *Word Error Rates for HSR and ASR*

Figure 4 also shows the HSR results. We can observe HSR and ASR perform differently for the different speakers. However, the number of speakers in this experiment is small to draw any reliable conclusion about the variation of ASR and HSR across speakers.

4. Conclusions and Future Work

Healthy voices are on an average three times as intelligible as oesophageal voices. The mean self reported effort was also three times larger for oesophageal speech compared to healthy voices. There was significant correlation between intelligibility and effort. Speaker type had an effect on both intelligibility and effort. Listeners familiar with oesophageal fared the same for intelligibility as people who were not. However, they reported less effort in listening to oesophageal speech than the not familiar listeners. The ASR system we chose for this task had poorer WER for oesophageal voice compared to healthy voice.

The listening effort obtained through this study is based on the listener's own interpretation of 'effort involved in listening'. This will provide us with a reference for comparison when we perform objective listening effort measurements in the future using physiological methods such as EEG and pupillometry. If these subjective measures are found to be correlated with the physiological measurements, then that opens the possibility of using the less cumbersome self report strategy to achieve our purpose of evaluation.

Both HSR intelligibility and ASR intelligibility play different but important roles in oesophageal speech evaluation. While improved HSR would enable better human-human interactions, an improved ASR performance would enable better human-machine interactions (eg. digital voice assistants). Lower listening effort would also contribute towards better communication with humans.

Our main future work is to build an oesophageal voice restoration system aimed at better ASR and HSR intelligibility and low listening effort.

5. Acknowledgements

This project has received funding from the European Unions H2020 research and innovation programme under the Marie Curie European Training Network ENRICH (675324).

The work has been partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER support (MINECO/FEDER, UE) (RESTORE project, TEC2015-67163-C2-1-R).

6. References

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

[2] 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.

[3] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "Peaks–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[4] C. Middag, T. Bocklet, J.-P. Martens, and E. Nöth, "Combining phonological and acoustic asr-free features for pathological speech intelligibility assessment," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Dia: a tool for objective intelligibility assessment of pathological speech," in *6th International workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*. Firenze University Press, 2009, pp. 165–167.

[6] J. L. Miralles and T. Cervera, "Voice intelligibility in patients who have undergone laryngectomies," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 3, pp. 564–571, 1995.

[7] T. Cervera, J. L. Miralles, and J. González-Àlvarez, "Acoustical analysis of spanish vowels produced by laryngectomized subjects," *Journal of speech, language, and hearing research*, vol. 44, no. 5, pp. 988–996, 2001.

[8] 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.

[9] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay, "Listening effort and fatigue: What exactly are we measuring? a british society of audiology cognition in hearing special interest group white paper," *International journal of audiology*, 2014.

[10] S. Bennett and B. Weinberg, "Acceptability ratings of normal, esophageal, and artificial larynx speech," *Journal of Speech, Language, and Hearing Research*, vol. 16, no. 4, pp. 608–615, 1973.

[11] K. F. Nagle and T. L. Eadie, "Listener effort for highly intelligible tracheoesophageal speech," *Journal of Communication Disorders*, vol. 45, no. 3, pp. 235–245, 2012.

[12] W. L. Cullinan, C. S. Brown, and P. D. Blalock, "Ratings of intelligibility of esophageal and tracheoesophageal speech," *Journal of communication disorders*, vol. 19, no. 3, pp. 185–195, 1986.

[13] J. Meyers, K. Volkert, and A. Diep, "Sentence repetition test: Updated norms and clinical utility," vol. 7, pp. 154–9, 02 2000.

[14] K. M. Yorkston and D. R. Beukelman, "A comparison of techniques for measuring intelligibility of dysarthric speech," *Journal of communication disorders*, vol. 11, no. 6, pp. 499–512, 1978.

[15] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.

[16] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sanchez, I. Saratxaga, and I. Odriozola, "Versatile Speech Databases for High Quality Synthesis for Basque," in *8th international conference on Language Resources and Evaluation (LREC)*, 2012, pp. 3308–3312. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/126_Paper.pdf

[17] D. Erro, I. Hernáez, A. Alonso, D. García-Lorenzo, E. Navas, J. Ye, H. Arzelus, I. Jauk, N. Hy, C. Magariños, R. Pérez-Ramón, M. Sulír, X. Tian, and X. Wang, "Personalized synthetic voices for speaking impaired: Website and app," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[18] N. Tits, "Exploring the parameters describing the quality and intelligibility of alaryngeal voices," University of Mons, 2017.

[19] 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.

[20] A. Sesma and A. Moreno, "Corpuscrt 1.0: Diseno de corpus orales equilibrados," *UPC, Tech. Rep., Dec.2000*.

[21] D. Erro, I. Hernáez, E. Navas, A. Alonso, H. Arzelus, I. Jauk, N. Q. Hy, C. Magarinos, R. Pérez-Ramón, M. Sulír *et al.*, "Zuretts: online platform for obtaining personalized synthetic voices," *Proceedings of eNTERFACE*, pp. 1178–1193, 2014.

[22] 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.

[23] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.

[24] 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.

[25] E. Polityko. Word error rate. [Online]. Available: https://www.mathworks.com/examples/matlab/community/19873-word-error-rate , access date: 20th February 2018

[26] JASP Team, "JASP (Version 0.8.6)[Computer software]," 2018. [Online]. Available: https://jasp-stats.org/ ,access date: 20th February 2018