applied sciences

MDPI

Article

Enrichment of Oesophageal Speech: Voice Conversion with Duration–Matched Synthetic Speech as Target

Sneha Raman *, Xabier Sarasola, Eva Navas and Inma Hernaez *

HiTZ-Aholab, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain; xabier.sarasola@ehu.eus (X.S.); eva.navas@ehu.eus (E.N.)

* Correspondence: sneha.raman@ehu.eus (S.R.), inma.hernaez@ehu.eus (I.H.)

Abstract: Pathological speech such as Oesophageal Speech (OS) is difficult to understand due to the presence of undesired artefacts and lack of normal healthy speech characteristics. Modern speech technologies and machine learning enable us to transform pathological speech to improve intelligibility and quality. We have used a neural network based voice conversion method with the aim of improving the intelligibility and reducing the listening effort (LE) of four OS speakers of varying speaking proficiency. The novelty of this method is the use of synthetic speech matched in duration with the source OS as the target, instead of parallel aligned healthy speech. We evaluated the converted samples from this system using a collection of Automatic Speech Recognition systems (ASR), an objective intelligibility metric (STOI) and a subjective test. ASR evaluation shows that the proposed system had significantly better word recognition accuracy compared to unprocessed OS, and baseline systems which used aligned healthy speech as the target. There was an improvement of at least 15% on STOI scores indicating a higher intelligibility for the proposed system compared to unprocessed OS, and a higher target similarity in the proposed system compared to baseline systems. The subjective test reveals a significant preference for the proposed system compared to unprocessed OS for all OS speakers, except one who was the least proficient OS speaker in the data set.

Keywords: pathological speech; voice conversion; intelligibility; speech recognition

Citation: Raman, S.; Sarasola, X.; Navas, E.; Hernaez, I. Enrichment of Oesophageal Speech: Voice Conversion with Duration–Matched Synthetic Speech as Target. *Appl. Sci.* **2021**, *11*, 5940. https://doi.org/10.3390/app11135940

Academic Editor: Francesc Alías

Received: 9 April 2021
Accepted: 18 June 2021
Published: 26 June 2021

1. Introduction

Laryngectomy is the surgical procedure of removing the larynx [1]. In addition to several functional disorders and lifestyle changes [2], this results in the loss of vocal folds and the patient's pre-surgery speech [3]. One of the several alternative ways that a laryngectomee can communicate [3,4] is to speak using the vibrations of the pharyngoesophageal segment [5], known as Oesophageal Speech (OS). Generating OS introduces acoustic artefacts [6] and makes OS less intelligible [7,8], which affects communication, social activities and quality of life [2,9].

OS is less intelligible and more effortful to listen to compared to healthy speech (HS). This is evident from previous listening experiments [10,11] as well as acoustic characteristics and challenges of OS [12]. Prolonged exposure to effortful speech causes fatigue in listeners [13]. Therefore, there is a strong motivation to make OS more intelligible and pleasant to listen to. We aim to enrich OS by closing the OS-HS gaps in intelligibility, quality and listening effort (LE).

Modern speech technologies and machine learning have great potential for use in the healthcare sector, be it for improvement of healthcare services [14] or to aid patients with speech impairments [15]. One such application is transforming pathological speech with the aim of making it more intelligible, pleasant and easier to process. This can reduce the load on the listeners and improve communication for people with speech pathologies.

One of the possible approaches to enrich OS is to use a voice conversion (VC) system. The goal of a VC system is to convert the utterances of a source speaker to sound like those

of a target speaker [16]. In the OS enrichment context, utterances of an OS speaker can be mapped to those of a healthy speaker, thereby having OS acquire characteristics of HS.

Some OS enrichment has been done using statistical VC methods such as Gaussian Mixture models (GMMs) [17–19]. In these methods, OS and HS are modelled by a linear combination of Gaussian distributions. In the training process, the Gaussian distributions of OS are mapped to those of HS. The output of such a training session is a conversion function mapping OS to HS. This conversion function can then be used to convert new OS samples, thereby getting OS speech that has characteristics of HS. In recent times, Deep Neural Networks (DNN) are more popular and effective compared to GMM based methods for enhancement of alaryngeal speech [20–23] and other types of pathological speech [24,25]. Another attempt to enrich OS was by using the eigenvoices concept [26], which was inspired by the eigenfaces concept [27]. Some studies have used filtering approaches [28], formant synthesis [29] and increasing the harmonics to noise ratio of OS [30].

Like our previous approaches [31,32], the proposed method is also based on VC. The bidirectional long short-term memory (BLSTM) based transformation [31] had better Automatic Speech Recognition (ASR) scores compared to OS. The method used by Serrano et al. [32] was inspired by a Phonetic Posteriorgrams (PPG) based system [33] which had good results for HS-HS VC. When applied for OS-HS VC, there was no improvement in ASR. Mel Cepstral Distortion (MCD) was reduced by both systems. Unprocessed OS was preferred over both of the systems in preference tests.

VC systems may be parallel (requires temporally aligned source–target utterance pairs) or non-parallel (requires many hours of speech data). Due to data limitations (100 sentences per speaker), parallel VC is best suited for our purposes. A parallel VC requires the parallel source and target sentences to be aligned for training. This is primarily done by Dynamic Time Warping (DTW) alignment which finds an optimal match based on similarities in the two sequences.

Helander et al. [34] describe some challenges of DTW in the context of VC. One of them is the presence of silences or extra sounds in the source and not in the target. Another one is the poor estimation of end points of silences and phonemes. A third case is the many-to-one and one-to-many nature of the DTW mapping. For example, if the source contains a phoneme with a longer duration compared to the target, then a single frame of the target may be mapped to several frames of the source. OS has undesired silences and artefacts and longer and varying durations of phonemes. These qualities make DTW challenging in the OS-HS VC task.

As a workaround, in our previous attempt [31], we performed alignment at two stages: first aligning the phone boundaries and then applying DTW, anchoring the phone boundaries. In this paper, we took advantage of the available phone labels and the possibility of generating synthetic speech (SS) with explicit phone durations. This resulted in SS that matches in duration with the source OS utterances, and would be a perfectly aligned target. This eliminated the need for DTW and its limitations. We hypothesise that this DTW-free VC would improve the intelligibility and quality of the enriched OS compared to our previous methods which required source–target alignment.

A robust enrichment system should ideally work with OS speakers of varying speaking proficiency. Therefore, we performed enrichments for OS speakers ranging from very low to very high intelligibility. As the enrichment system is built to improve communications for the OS speaker, it is important that the output of the enrichment system is preferred by listeners over the unprocessed OS. Moreover, given that voice interactions with machines are becoming more and more common, the enriched outputs should be intelligible to machines. Taking these points into consideration, we evaluated the subjective preference of the enriched system amongst human listeners as well as objective intelligibility and ASR performance.

To sum up, we present a novel, DTW-free, parallel VC system for OS enrichment which includes an SS target. We evaluate its outputs for ASR performance, an objective intelligibility metric and a preference test in comparison with unprocessed OS.

2. Data

We chose four OS speakers (3 male, 1 female) with a wide range of intelligibility from an OS database that contains over 30 OS speakers [12]. In the original database, the four speakers were identified as '02M3', '04M3', '16M3' and '25F3', and we continue to use these IDs. Some details of the four speakers such as age, sex, time passed since the laryngectomy operation, stimulus duration and speaking rates are presented in Table 1 [35]. The age and the time since laryngectomy were collected on the day of the recording. Speakers 04M3 and 16M3 are relatively recent laryngectomees and hence, are less proficient than the other two speakers.

For each speaker, we used a parallel dataset of 100 phonetically-balanced Spanish sentences (described in detail in [12]). The sentences were syntactically and semantically predictable but had some low frequency words. The number of words in each sentence ranged between 9 and 18 words (mean = 13.19, SD = 3.66).

Table 1. Speaker characteristics.

Speaker IDs	Sex	Age	Time since Laryngectomy	Duration Per Stimulus Mean ± SD (Seconds)	Speaking Rate Mean ± SD (Syllables Per Second)
02M3	Male	75 years 5 months	8 years 1 month	7.48 ± 1.67	4.32 ± 1.80
04M3	Male	59 years 4 months	1 year 7 months	9.27 ± 2.36	3.84 ± 1.71
16M3	Male	66 years 4 months	1 year 10 months	12.52 ± 3.61	2.59 ± 1.19
25F3	Female	59 years 3 months	11 years 11 months	7.85 ± 2.02	4.24 ± 1.86

3. Proposed VC System

The proposed VC system, BLSTM with SS as target (BLSTMSS), is a DNN based system with OS as source and SS with matching durations as target (see Figure 1). The procedure is described in detail in the following steps.

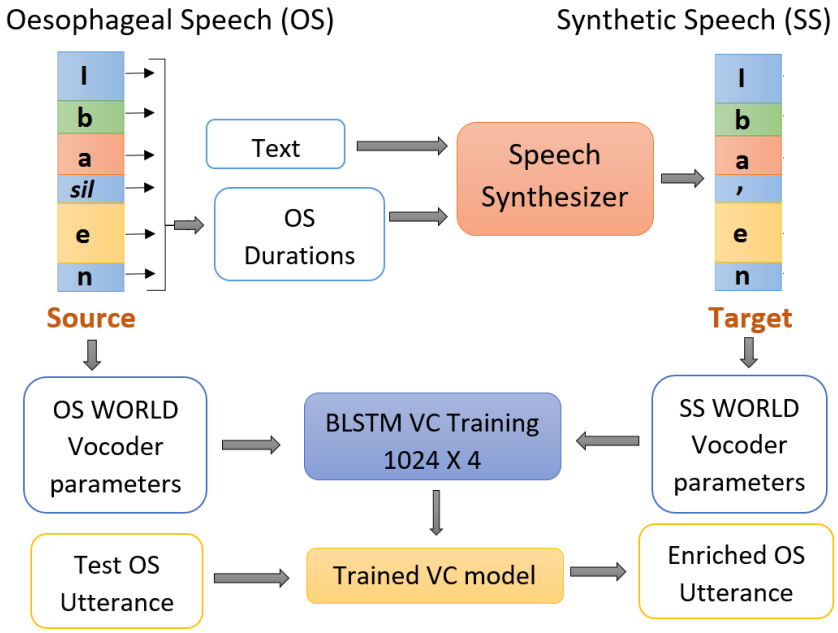


Figure 1. The proposed OS-HS VC system: BLSTMSS.

3.1. Labelling of Oesophageal Speech

Segmentation and labelling of OS is a tricky process owing to undesired artefacts, incorrect pronunciations of some consonants and unstable fundamental frequency. The forced alignment feature built into generic Spanish ASR systems such as Kaldi [36] was unsuitable for OS. Therefore, using the Montreal Forced Alignment tool [37], and with the aid of a manually labelled set for one speaker (speaker 02M3), new models were created by using OS as the training material. Performing segmentation with this forced aligner gave us the phone labels and their durations for the source OS utterances.

3.2. Generating Target Synthetic Speech

Using the labels, their durations and the utterance text, SS was generated by explicitly assigning these durations to the phones. We used an HMM based text-to-speech system [38] which was originally developed for the Basque language. The Spanish version is described in [39]. This process gave us equal-sized frame-by-frame aligned pairs of OS and SS.

Due to constant swallowing of air to produce speech, OS contains several pauses with artefacts within utterances. During the SS generation, these pauses were replaced with silences.

3.3. Voice Conversion Neural Network

Voice conversion was performed with the VC recipe of the Merlin toolkit [40]. Parametrisation and resynthesis was done using the WORLD Vocoder [41]. The extracted parameters included 60 Mel Cepstral Coefficients (MCCs), 1 excitation parameter (log F0), 1 Band Aperiodicity Parameter (BAP), the deltas and the delta deltas of the MCCs, log F0 and BAP and a voiced/unvoiced binary parameter. In all, there were 187 parameters extracted every 5 milliseconds.

A matrix of size 187 $\times$ (number of 5 ms frames) of OS and SS utterances were the source and target inputs, respectively. We split the 100 source–target pairs into 90 train and 10 test pairs. As the source and the target had the same number of frames, we skipped the alignment step in the training process. The train parameters were normalised to 0 mean and unit variance and then fed into a 4 layered BLSTM (4 $\times$ 1024) training network. After training, the source test utterance parameters were converted using the trained model. A denormalisation of the mean and the variance was applied to the output parameters, followed by a Maximum Likelihood Parameter Generation using the variances from the training data. The resulting converted parameters were fed into the vocoder to synthesise the converted speech. A cross validation was performed 10 times, so that all the 100 sentences were available as test sentences.

4. Evaluations and Results

Evaluations involved comparing BLSTMSS outputs to unprocessed OS using three ASR systems, STOI scores and a preference test. In addition, we compared ASR and STOI scores of BLSTMSS with those of our previous systems.

4.1. ASR

We evaluated the outputs of our proposed enrichment system using three ASR systems: the speech-to-text system from Microsoft Azure using the python azure-cognitiveservices-speech library (ASR 1) [42], the Elhuyar speech recognition system (ASR 2) [43] and a Kaldi [36] based system (ASR 3) developed in our laboratory and described in [44]. The input files to these ASR systems were the 100 single channel wav files sampled at 16,000 Hz. The outputs were text files containing the transcriptions.

The reason for using three ASR systems was to have a diverse set of evaluations. ASR 1 is a well known commercial ASR system used world wide and therefore easier for comparisons in future studies elsewhere. ASR 2 is a commercial system built locally in Spain, and therefore, better adapted to the speech style and vocabulary of the speakers involved in this study. ASR 3 is a customised Kaldi based ASR with full control of all the

components such as the language model, dictionary etc. We presume that the amount of audio used to train ASR 3 (approximately 5 h of audio) was smaller in comparison to the other two commercial systems. It uses a lexicon limited to the vocabulary of the corpus used in this study. It also uses a unigram language model and was used in our previous studies [31,32]. The advantage of ASR 3 is that it is not prone to updates as is the case of commercial ASRs. This allows us to make fair and accurate comparisons of our ongoing work with our previous work.

We calculated two metrics from the ASR transcriptions: Word Error Rates (WER) and Percentage Words Correct (PWC). For WER, we calculated the Levenshtein distance [45] between the reference sentence (original recording utterance text) and the hypothesis sentence (ASR transcription output) using the Word Error Rate Matlab toolbox [46]. The WER formula is shown in Equation (1). The Levenshtein distance WER takes into account the insertions, deletions, and substitutions that are observed in the transcribed output. Please note that the WER can be more than 100% if the total insertions, substitutions and deletions exceed the total number of words in the reference sentence.

$$\text{WER} = \frac{\text{Substitutions + Insertions + Deletions}}{\text{Total number of words in reference sentence}} \times 100. \quad (1)$$

PWC is the percentage of words from the reference sentence correctly identified in the transcribed sentence. The PWC formula is shown in Equation (2).

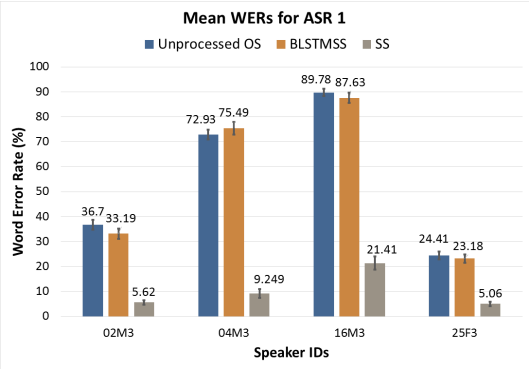
$$\text{PWC} = \frac{\text{Words correctly identified in transcription}}{\text{Total number of words in reference sentence}} \times 100. \quad (2)$$

Figures 2–4 show mean WER and PWC scores for the 100 sentences obtained from the transcriptions of ASR 1, 2 and 3, respectively. WER scores were lower (i.e., higher intelligibility) for BLSTMSS compared to unprocessed OS for all ASRs and speakers with 2 exceptions—speaker 04M3 in ASR 1 and speaker 16M3 in ASR 2. In the case of PWC scores, a higher PWC score (i.e., higher intelligibility) was observed for the BLSTMSS samples compared to unprocessed OS samples for all speakers and ASRs.

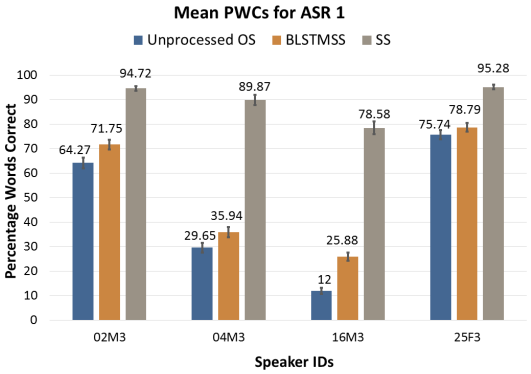
When comparing the different ASR systems, the best WER and PWC scores for unprocessed OS were obtained by ASR 1, followed by ASR 3 and ASR 2. In addition, there were fewer differences between OS and enriched OS in ASR 1 compared to the other two systems. Amongst all the ASRs, ASR 3 had the best WER and PWC scores for enriched OS.

We did correlation analysis of WERs and PWCs of OS obtained from ASR 1. There was a significant negative correlation (Pearson's $r = -0.959$, $p = 0.041$) between WER and the number of months since laryngectomy and a significant positive correlation (Pearson's $r = 0.952, p = 0.048$) between PWC and the number of months since laryngectomy. A similar correlation was observed with speaking rate, but it did not reach significance.

In our previous studies [31,32], we worked with speaker 02M3 and ASR 3. Figure 5 shows the WER scores of BLSTMSS in comparison to our previous methods, PPG [32] and BLSTMHS [31]. It can be observed that the proposed system was able to significantly reduce ASR errors in comparison to previous methods.

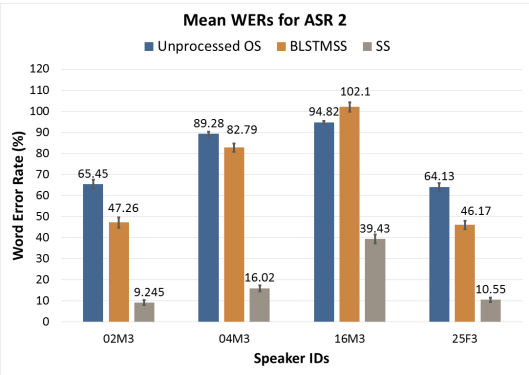


(**a**) Word Error Rates

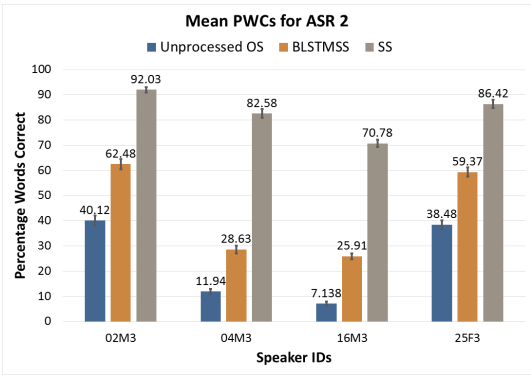


(**b**) Percentage Words Correct

Figure 2. ASR 1 WER and PWC scores for unprocessed OS (source), the BLSTMSS converted outputs and target SS (target). Error bars show standard errors.

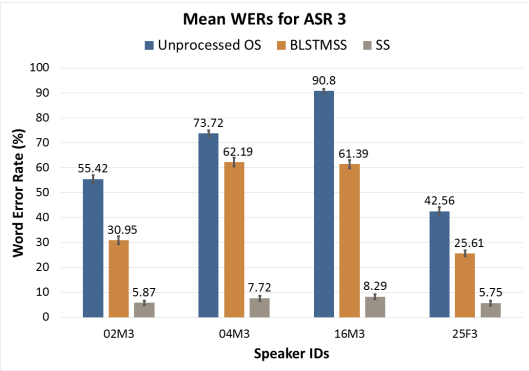


(**a**) Word Error Rates

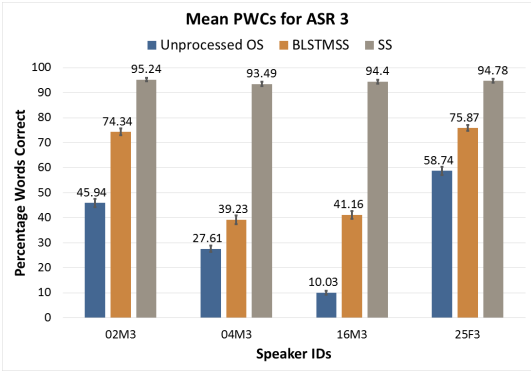


(**b**) Percentage Words Correct

Figure 3. ASR 2 WER and PWC scores for unprocessed OS (source), the BLSTMSS converted outputs and target SS (target). Error bars show standard errors.



(**a**) Word Error Rates



(**b**) Percentage Words Correct

Figure 4. ASR 3 WER and PWC scores for unprocessed OS (source), the BLSTMSS converted outputs and target SS (target). Error bars show standard errors.

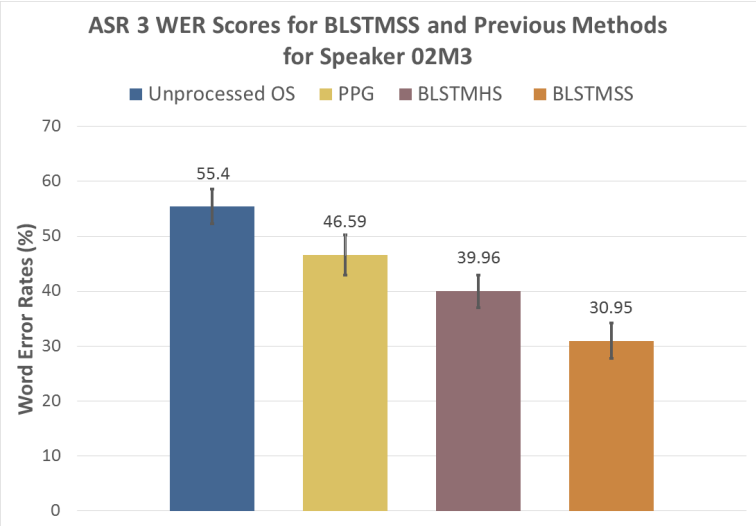


Figure 5. WER scores for Unprocessed OS, previous systems (PPG and BLSTMHS) and the proposed BLSTMSS system as calculated by ASR 3 for speaker 02M3. Error bars show standard errors.

4.2. STOI Scores

STOI [47] is an intrusive objective intelligibility measure which is known to be correlated with subjective intelligibility scores for noisy speech. An intrusive intelligibility measurement requires a degraded signal and an aligned reference signal. We calculated STOI for unprocessed OS samples and converted BLSTMSS samples for the four OS speakers using the already aligned duration–matched SS (target signal) as the reference signal. We used the SS as reference because they were the best possible clean aligned signals available. Calculating STOI with aligned healthy laryngeal speech would have resulted in alignment errors and hence, in an inaccurate STOI measurement. The STOI results are shown in Figure 6.

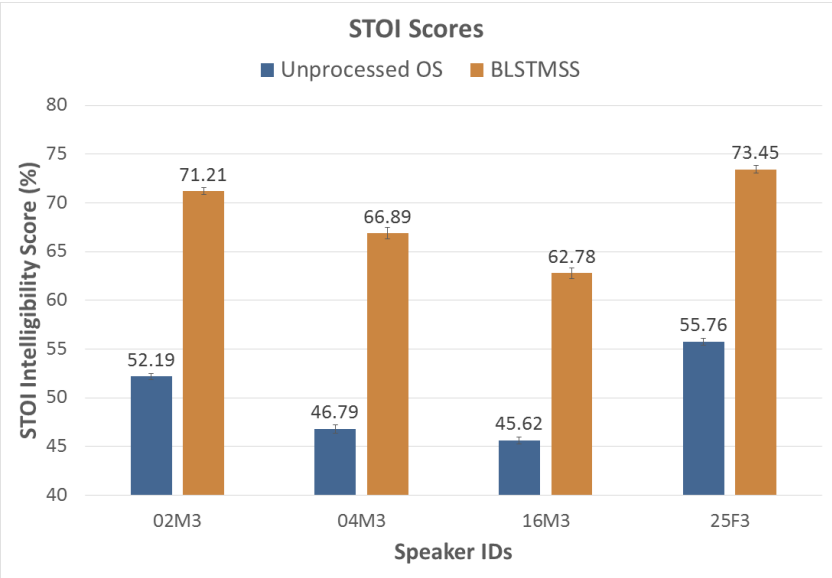


Figure 6. STOI scores for the four OS speakers and the enriched versions. Reference signal for STOI is duration–matched SS. Error bars show standard errors.

We can observe that the STOI scores have improved considerably (at least 15 percentage points) from OS to BLSTMSS for all four speakers. A high STOI score of over 60% was observed for all the BLSTMSS samples with intelligible synthetic speech (>70% ASR accuracy) as reference.

Like the ASR, we compared the STOI scores of the proposed system with those of our previous methods (see Figure 7). The references used to calculate STOI were the same duration–matched SS signals. The proposed system has higher STOI scores (about 5%) compared to previous systems and unprocessed OS.

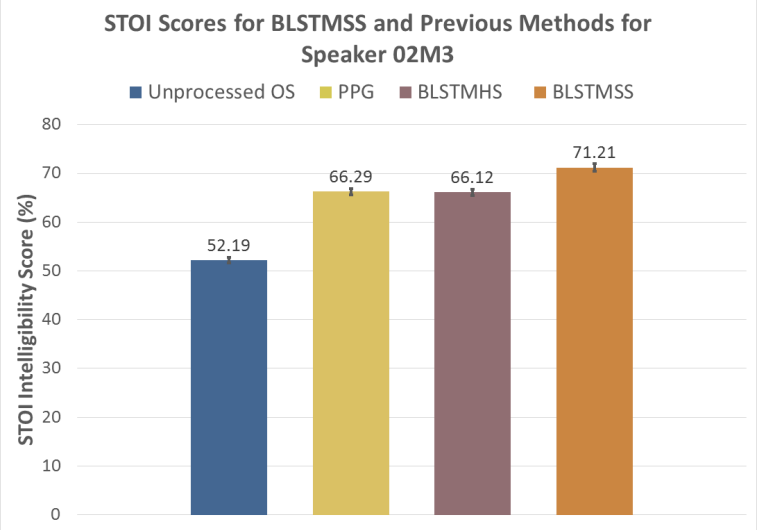


Figure 7. STOI scores for Unprocessed OS, previous systems (PPG and BLSTMHS) and proposed BLSTMSS system for speaker 02M3. Reference signal for STOI is duration–matched SS. Error bars show standard errors.

5. Subjective Test

While unprocessed OS has several undesired artefacts and lacks a natural fundamental frequency, it is natural speech. On the other hand, although the BLSTMSS outputs are much clearer sounding, they are synthetically produced and may have some limitations because of that. The success of the enrichment depends majorly on whether listeners prefer to listen to the enriched version more than the unprocessed OS. Therefore, we performed a preference test to collect listeners' opinion on whether they prefer listening to the outputs of the proposed system or the unprocessed OS.

The preference test was a 5-point Comparison Mean Opinion Scores (CMOS) test conducted using a web-based interface (https://aholab.ehu.eus/users/sneha/BLSTMSS_evaluation/preference_test.php (accessed on 25 June 2021)). A web-based test was considered more appropriate owing to COVID restrictions. Participants were sourced by sending emails to speech technology networks in Spain and other local networks. The participants were instructed to perform the test with headphones. They were informed that there are no correct or incorrect answers in the test and that they should state their opinions with full liberty.

The participants listened to 10 pairs of sentences from each of the four speakers—a total of 40 pairs of sentences. Each pair contained one unprocessed OS sentence and the corresponding BLSTMSS output of the same sentence. The chosen 10 pairs were the shortest sentences in the set, as that allowed us to have the maximum number of evaluations while keeping the test under 20 min. The presentation order of all the pairs, as well as the order of the BLSTMSS and OS sentences within each pair was randomised to avoid order bias. After listening to the two stimuli in each pair, the participants were asked to mark the preferred stimulus. To do so, they were given the following options: 'Prefiero claramente la primera' (I clearly prefer the first one), 'Prefiero la primera' (I prefer the first one), 'No percibo diferencia/Ninguna suena mejor' (I do not perceive any difference/Neither one sounds better), 'Prefiero la segunda' (I prefer the second one), 'Prefiero claramente la segunda' (I clearly prefer the second one).

Apart from the 40 test pairs, there were 4 pairs (presented at regular intervals) where both the samples were the same file, which was a sentence spoken by a healthy

speaker. As both the files in these 4 control pairs were the same exact file, we expected the participants to mark the third option ('I do not perceive any difference/Neither one sounds better'). Only those participants who correctly marked this option for at least 3 of these 4 control pairs, were included in the analysis. This ensured reliability of the participants' responses.

We asked the participants to describe the audio equipment they used during the test. This was to ensure that they were not using any bad equipment. The options were: good headphones, normal headphones, good loudspeakers, normal loudspeakers and bad equipment. We also asked whether the participants had any experience with using speech technologies. The options were: no experience, experts, sporadic users and through perception tests. This was not to study the effect of speech expertise on the evaluations, but to ensure a good mix of all kinds of listeners.

A total of 32 native Spanish participants performed the listening test. Two of them were rejected because they failed the control test. One other participant described their audio equipment as 'bad equipment' and was excluded too. 16 out of the chosen 29 listeners had no experience with using speech technologies. Five of them were speech technology experts, 4 were sporadic users of speech technology and 4 stated that their experience of speech technologies was through perception tests.

Overall, the most chosen option was 'Preference for BLSTMSS' as can be observed in Figure 8c. There were more responses in the 'Clear preference for BLSTMSS' and 'Preference for BLSTMSS' categories compared to 'Clear preference for OS' and 'Preference for OS' categories, respectively. 'Clear preference for OS' was the least chosen option.

When looking at speakers separately we observed that speaker 16M3 (Figure 8d) has a different trend compared to other speakers. For speakers 02M3 (Figure 8a), 04M3 (Figure 8b) and 25F3 (Figure 8e), the most preferred option was 'Preference for BLSTMSS'. However for speaker 16M3, the least intelligible speaker in the dataset, most responses were in the 'No preference for either' or the undecided category. The next most preferred option was 'Preference for BLSTMSS'. Additionally, for the more proficient speakers (25F3 and 02M3), there were less instances of the 'No preference for either' category compared to the non-proficient speakers.

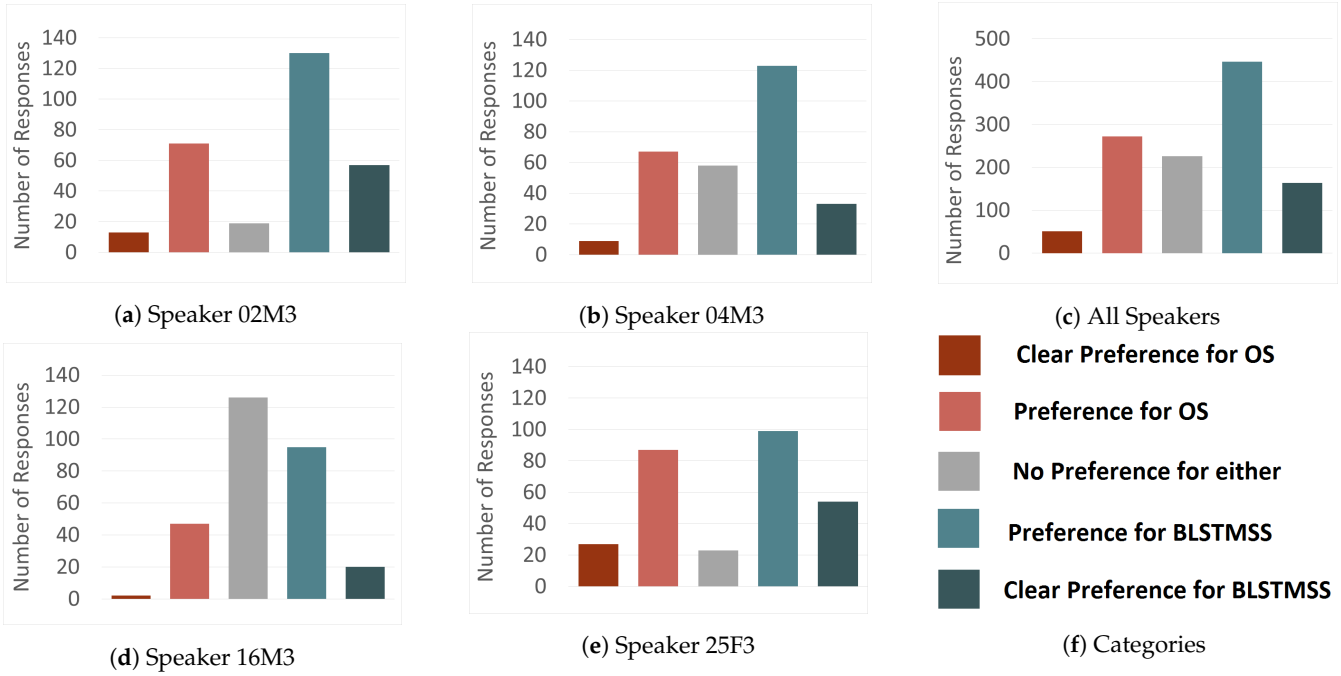


Figure 8. Histogram plots for the preference scores of the four speakers separately and all together.

6. Discussion

In this study, we have employed and evaluated a novel DNN-based voice conversion system aimed at enriching OS. The evaluations involved subjective (preference test) as well as objective (STOI, ASR) aspects. The evaluations were performed on unprocessed-enriched pairs of samples from four OS speakers. Additionally, objective results of the proposed system were compared with those of our previous experiments.

In the ASR evaluations, the results of WER improvement for the proposed system was not unanimous. However, for all the 3 ASR systems and all 4 speakers, our proposed system had better PWC scores compared to unprocessed OS. This means that our enrichment resulted in the ASR systems recognising more number of words in comparison to unprocessed OS. Correlation analysis suggests that more errors were found in speakers who underwent the laryngectomy more recently. This is expected, as these speakers have had less time to train and practice the techniques of OS production. This explains the higher WER and lower PWC for speakers 04M3 and 16M3 compared to the other two more proficient speakers.

STOI as an objective intelligibility measure is usually applied in cases where already available clean signals are degraded with noise. In the case of OS, the original signal itself is degraded. Although the duration–matched SS is clean and aligns with the original OS and the enriched outputs, it cannot be considered as a clean reference in the true sense. Moreover, the reference SS signal was also the target of the VC process. Therefore, STOI scores in this case may be interpreted as a measure of similarity with the target SS (which is the goal of a VC task) rather than objective intelligibility. Nonetheless, for all the four speakers, a higher STOI value for the proposed system (over 62% to 73%) compared to unprocessed OS (45% to 55%) indicates that there was an improvement in intelligibility. Comparisons with previous methods also revealed improved intelligibility with the proposed method.

The preference test revealed a preference for the proposed method for three of the four OS speakers. For the remaining fourth and the least proficient speaker, there was no preference for the proposed system, but no preference for unprocessed OS either. The listeners were mostly undecided. This is possibly because the speaker's intelligibility is poor, and although the conversion system helped in improving some spectral characteristics, some other characteristics of the speaker such as the long duration of the stimuli, phones and silences, and the resulting slow and unnatural rhythm, were present in the converted version too.

In a previous experiment of enriching OS using a DNN-based VC system [31], there was an improvement in ASR scores. However, the listeners preferred a system that performed a simple fundamental frequency transformation, which did not improve ASR or intelligibility scores. Similarly, in another experiment [32], we did not achieve intelligibility improvement or a preference for the proposed method. Both the above experiments involved only one OS speaker (02M3), who also was one of the most intelligible speakers of the dataset. In the current study, with a novel strategy and more speakers, we have shown that our OS enrichment method improves intelligibility in addition to being preferred by listeners. The source–target alignment process is particularly problematic for OS and that was a limitation in the aforementioned previous systems. Using a duration–matched target and skipping the error–prone alignment process helped in overcoming that limitation and improving results.

The proposed system can possibly be improved by using newer DNN VC technologies and newer speech synthesis systems. The speech synthesis method used in this study is relatively old and was specifically chosen for its ability to generate speech with forced durations. Although newer DNN based speech synthesis systems are of better quality, they do not have this ability. If a speech synthesis system in the future can generate forced-duration SS of better quality, it can possibly improve results. Using a more modern vocoder, increasing the amount of training data or including more diverse training data (utterances in noise, more OS speakers) are also possible ways of improving the results.

7. Conclusions and Future Work

Due to the anatomical alteration post-laryngectomy, it is difficult for OS speakers to produce intelligible speech. Our study was an attempt to enhance OS with the aim of making it more intelligible. We performed voice conversion using a BLSTM network and a novel training data selection approach (using a target that is matched in duration with source) which eliminated the need for the source–target alignment process.

The proposed system showed significant improvements in objective evaluations of intelligibility in comparison to our previous systems. Compared to the unprocessed OS utterances, the proposed system's outputs were recognised with more accuracy by three different ASR systems. In recent times, communication with digital assistants and other devices is on the rise. Therefore, an improvement in this direction is desirable for efficient communication with digital devices and dialogue systems. Additionally, for the three most intelligible OS speakers out of the four, the proposed system was preferred by listeners over unprocessed OS. While this is encouraging, more effort is needed to have similar results for the low intelligibility OS speaker.

An extension of the system described in this paper is under development: a multi-speaker system with many OS speakers' utterances as source and corresponding duration–matched SS as target. This will be a generic OS enrichment system and will enable the enrichment of more OS speakers. In addition to ASR scores, STOI scores and preference tests, we are interested in investigating subjective and physiological listening effort for unprocessed OS, enriched OS and HS. This is because, while intelligibility reveals what percentage of the speech was understood correctly, it does not tell us how difficult it was to understand it. Listening effort provides useful additional information about whether enriched OS is easier to perceive and process compared to unprocessed OS. Therefore, our future studies will focus on LE in addition to intelligibility and listener preferences.

Our aim for the future is to install this enrichment system as a face-to-face communication aid in a stand-alone device or a smartphone. The device will take the unprocessed OS input and play out the enriched version in real time or with negligible delay. Another possible practical application of the proposed system could be in the form of a software plugin coupled with the microphone of the devices used by an OS speaker. This would convert any microphone input (Unprocessed OS) to an enriched version of the speech in real time or with minimum delay. Any app which requires a microphone input will use this modified speech instead. In this way, the OS speaker would be able to use the benefits of the enriched speech for telephonic conversations, zoom calls, voice commands to digital assistants and other voice based apps. Generating real time outputs will require solving latency problems when generating BLSTM outputs (see [48] for example). Once we tackle the problems of making the enrichment system work in real time and embedding it into a device, our research efforts will be able to translate into real world applications.

Author Contributions: conceptualization, S.R. and X.S.; methodology, S.R. and X.S.; software, S.R. and X.S.; validation, S.R., I.H. and E.N.; formal analysis, S.R.; investigation, S.R.; resources, S.R., I.H. and E.N.; data curation, S.R.; writing—original draft preparation, S.R.; writing—review and editing, S.R., I.H. and E.N.; visualization, S.R.; supervision, I.H. and E.N.; project administration, I.H. and E.N.; funding acquisition, I.H. and E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by funding from the European Union's H2020 research and innovation programme under the MSCA GA 67532*4 (the ENRICH network: www.enrich-etn.eu (accessed on 25 June 2021)), and the Basque Government (PIBA_2018_1_0035 and IT355-19).

Data Availability Statement: The oesophageal speech data used in this study is available at http://catalog.elra.info/en-us/repository/browse/ELRA-S0413/.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

OS	Oesophageal Speech
HS	Healthy Speech
SS	Synthetic Speech
LE	Listening Effort
ASR	Automatic Speech Recognition
VC	Voice Conversion
SD	Standard Deviation
GMM	Gaussian Mixture Models
DNN	Deep Neural Networks
BLSTM	Bidirectional Long Short Term Memory
BLSTMSS	Bidirectional Long Short Term Memory with Synthetic Speech as target
BLSTMHS	Bidirectional Long Short Term Memory with Healthy Speech as target
PPG	Phonetic Posteriorgrams
DTW	Dynamic Time Warping
MCD	Mel Cepstral Distortion
HMM	Hidden Markov Models
MCC	Mel Cepstral Coefficients
BAP	Band Aperiodicity Parameter
WER	Word Error Rate
PWC	Percentage Words Correct
STOI	Short Term Objective Intelligibility
CMOS	Comparison Mean Opinion Scores

References

1. Ward, E.C.; van As-Brooks, C.J. *Head and Neck Cancer: Treatment, Rehabilitation, and Outcomes*; Plural Publishing: San Diego, CA, USA, 2014.
2. Ackerstaff, A.; Hilgers, F.; Aaronson, N.; Balm, A. Communication, functional disorders and lifestyle changes after total laryngectomy. *Clin. Otolaryngol. Allied Sci.* **1994**, *19*, 295–300. [CrossRef] [PubMed]
3. van Sluis, K.E.; van der Molen, L.; van Son, R.J.; Hilgers, F.J.; Bhairosing, P.A.; van den Brekel, M.W. Objective and subjective voice outcomes after total laryngectomy: A systematic review. *Eur. Arch. Oto-Rhino* **2018**, *275*, 11–26. [CrossRef] [PubMed]
4. Koike, M.; Kobayashi, N.; Hirose, H.; Hara, Y. Speech rehabilitation after total laryngectomy. *Acta Oto-Laryngol.* **2002**, *122*, 107–112. [CrossRef]
5. Štajner-katušić, S.; Horga, D.; Mušura, M.; Globlek, D. Voice and speech after laryngectomy. *Clin. Linguist. Phon.* **2006**, *20*, 195–203. [CrossRef]
6. Weinberg, B. Acoustical properties of esophageal and tracheoesophageal speech. *Laryngectomee Rehabil.* **1986**, 113–127.
7. Most, T.; Tobin, Y.; Mimran, R.C. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *J. Commun. Disord.* **2000**, *33*, 165–181. [CrossRef]
8. Drugman, T.; Rijckaert, M.; Janssens, C.; Remacle, M. Tracheoesophageal speech: A dedicated objective acoustic assessment. *Comput. Speech Lang.* **2015**, *30*, 16–31. [CrossRef]
9. Mohide, E.A.; Archibald, S.D.; Tew, M.; Young, J.E.; Haines, T. Postlaryngectomy quality-of-life dimensions identified by patients and health care professionals. *Am. J. Surg.* **1992**, *164*, 619–622. [CrossRef]
10. Raman, S.; Hernáez, I.; Navas, E.; Serrano, L. Listening to Laryngectomees: A study of Intelligibility and Self-Reported Listening Effort of Spanish Oesophageal Speech. IberSPEECH. 2018. Available online: https://www.isca-speech.org/archive/IberSPEECH_2018/abstracts/IberS18_O3-1_Raman.html (accessed on 25 June 2021)
11. Raman, S.; Serrano, L.; Winneke, A.; Navas, E.; Hernaez, I. Intelligibility and Listening Effort of Spanish Oesophageal Speech. *Appl. Sci.* **2019**, *9*, 3233. [CrossRef]
12. García, L.S.; Raman, S.; Rioja, I.H.; Cordón, E.N.; Sanchez, J.; Saratxaga, I. A Spanish Multispeaker Database of Esophageal Speech. *Comput. Speech Lang.* **2020**, *66*, 101168. [CrossRef]
13. McGarrigle, R.; Munro, K.J.; Dawes, P.; Stewart, A.J.; Moore, D.R.; Barry, J.G.; Amitay, S. Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *Int. J. Audiol.* **2014**, *53*, 433–445. [CrossRef]
14. Latif, S.; Qadir, J.; Qayyum, A.; Usama, M.; Younis, S. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Rev. Biomed. Eng.* **2020**, *14*, 342–356. [CrossRef]
15. Hawley, M.S.; Green, P.; Enderby, P.; Cunningham, S.; Moore, R.K. Speech technology for e-inclusion of people with physical disabilities and disordered speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.

16. Mohammadi, S.H.; Kain, A. An overview of voice conversion systems. *Speech Commun.* **2017**, *88*, 65–82. [CrossRef]
17. Doi, H.; Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Statistical approach to enhancing esophageal speech based on Gaussian mixture models. In Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, TX, USA, 14–19 March 2010; pp. 4250–4253.
18. Doi, H.; Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models. *IEICE Trans. Inform. Syst.* **2010**, *93*, 2472–2482. [CrossRef]
19. Doi, H.; Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Enhancement of Esophageal Speech Using Statistical Voice Conversion. 2009. Available online: https://www.semanticscholar.org/paper/Enhancement-of-Esophageal-Speech-Using-Statistical-Doi-Nakamura/bd88fe19deb4ed4991b64daf164d27af0d1197d4 (accessed on 25 June 2021).
20. Othmane, I.B.; Di Martino, J.; Ouni, K. Enhancement of esophageal speech obtained by a voice conversion technique using time dilated Fourier cepstra. *Int. J. Speech Technol.* **2019**, *22*, 99–110. [CrossRef]
21. Dinh, T.; Kain, A.; Samlan, R.; Cao, B.; Wang, J. Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency. *Proc. Interspeech* **2020**, *2020*, 4781–4785. [CrossRef]
22. Urabe, E.; Hirakawa, R.; Kawano, H.; Nakashi, K.; Nakatoh, Y. Enhancement of Electrolarynx speech based on WaveRNN. In Proceedings of the 7th ACIS International Conference on Applied Computing and Information Technology, Honolulu, HI, USA, 29 May 2019; pp. 1–6.
23. Urabe, E.; Hirakawa, R.; Kawano, H.; Nakashi, K.; Nakatoh, Y. Electrolarynx System Using Voice Conversion Based on WaveRNN. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 4–6 January 2020; pp. 1–2. [CrossRef]
24. Chen, C.Y.; Zheng, W.Z.; Wang, S.S.; Tsao, Y.; Li, P.C.; Li, Y. Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-based Voice Conversion System. In Proceedings of the IEEE Interspeech, Shanghai, China, 25–29 October 2020.
25. Sudro, P.N.; Kumar Das, R.; Sinha, R.; Mahadeva Prasanna, S.R. Enhancing the Intelligibility of Cleft Lip and Palate Speech Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 720–727. [CrossRef]
26. Doi, H.; Toda, T.; Nakamura, K.; Saruwatari, H.; Shikano, K. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 172–183. [CrossRef]
27. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef]
28. Garcia, B.; Ruiz, I.; Méndez, A. Oesophageal speech enhancement using poles stabilization and Kalman filtering. In Proceedings of the 2008 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Las Vegas, NV, USA, 31 March–4 April 2008; pp. 1597–1600.
29. Matsui, K.; Hara, N.; Kobayashi, N.; Hirose, H. Enhancement of esophageal speech using formant synthesis. *Acoust. Sci. Technol.* **2002**, *23*, 69–76. [CrossRef]
30. Oleagordia-Ruiz, I.; Garcia-Zapirain, B. Harmonic to noise ratio improvement in oesophageal speech. *Technol. Health Care* **2015**, *23*, 359–368. [CrossRef]
31. Serrano, L.; Tavarez, D.; Sarasola, X.; Raman, S.; Saratxaga, I.; Navas, E.; Hernaez, I. LSTM based voice conversion for laryngectomees. In *IberSPEECH*; International Speech Communication Association: Barcelona, Spain, 2018; pp. 122–126.
32. Serrano, L.; Raman, S.; Tavarez, D.; Navas, E.; Hernaez, I. Parallel vs. Non-Parallel Voice Conversion for Esophageal Speech. In Proceedings of the Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 4549–4553.
33. Sun, L.; Li, K.; Wang, H.; Kang, S.; Meng, H. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
34. Helander, E.; Schwarz, J.; Nurminen, J.; Silen, H.; Gabbouj, M. On the impact of alignment on voice conversion performance. In Proceedings of the 9th Annual Conference of the International Speech Communication Associationa, Interspeech 2008, Brisbane, Australia, 22–26 September 2008; pp. 1453–1456.
35. Serrano, L. Técnicas Para la Mejora de la Inteligibilidad en Voces Patológicas. Ph.D. Thesis, University of the Basque Country (UPV/EHU), Biscay, Spain, 2019.
36. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, USA, 3 February 2011.
37. Ling, Z.H.; Kang, S.Y.; Zen, H.; Senior, A.; Schuster, M.; Qian, X.J.; Meng, H.M.; Deng, L. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Process. Mag.* **2015**, *32*, 35–52. [CrossRef]
38. Erro, D.; Sainz, I.; Luengo, I.; Odriozola, I.; Sánchez, J.; Saratxaga, I.; Navas, E.; Hernáez, I. HMM-based speech synthesis in Basque language using HTS. *Proc. FALA* **2010**, 67–70. Available online: http://lorien.die.upm.es/~lapiz/rtth/JORNADAS/VI/pdfs/0012.pdf (accessed on 25 June 2021).
39. Sainz, I.; Erro, D.; Navas, E.; Hernáez, I.; Sánchez, J.; Saratxaga, I.; Odriozola, I.; Luengo, I. Aholab speech synthesizers for Albayzin 2010. *Proc. FALA* **2010**, *2010*, 343–348.

40. Wu, Z.; Watts, O.; King, S. Merlin: An Open Source Neural Network Speech Synthesis System. In *Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 Sepember 2016*; International Speech Communication Association: Barcelona, Spain, 2016; pp. 202–207.
41. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inform. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
42. Microsoft. Microsoft Azure Cognitive Services Speech-to-Text. Available online: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/get-started-speech-to-text (accessed on 10 October 2020).
43. Elhuyar. Aditu—El Reconocedor del Habla de Elhuyar Basado en Inteligencia Artificial y Redes Neuronales. Available online: https://aditu.eus/ (accessed on 10 October 2020).
44. Aholab Speaker Diarization System for Albayzin 2016 Evaluation Campaign. 2016. Available online: https://iberspeech2016.inesc-id.pt/wp-content/uploads/2017/01/OnlineProceedings_IberSPEECH2016.pdf (accessed on 25 June 2021).
45. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
46. Polityko, E. Word Error Rate. MATLAB Central File Exchange. Available online: https://ch.mathworks.com/matlabcentral/fileexchange/55825-word-error-rate (accessed on 25 June 2021).
47. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, TX, USA, 14–19 March 2010; pp. 4214–4217.
48. Xue, S.; Yan, Z. Improving latency-controlled BLSTM acoustic models for online speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5340–5344. [CrossRef]