A Spanish Multispeaker Database of Esophageal Speech

Luis Serrano García, Sneha Raman, Inma Hernáez Rioja, Eva Navas Cordón, Jon Sanchez, Ibon Saratxaga

HiTZ Basque Center for Language Technology, University of the Basque Country (UPV/EHU), Bilbao, Spain

Abstract

A laryngectomee is a person whose larynx has been removed by surgery, usually due to laryngeal cancer. After surgery, most laryngectomees are able to speak again, using techniques that are learned with the help of a speech therapist. This is termed as alaryngeal speech, and esophageal speech (ES) is one of the several alaryngeal speech production modes. A considerable amount of research has been dedicated to the study of alaryngeal speech, with a wide range of aims such as helping speech therapists with evaluation and diagnosis, and improving its quality and intelligibility using digital signal processing techniques. We present to you a database of Spanish ES voices, named AhoSLABI, which is designed to allow the development of new support technologies for this speech impairment. The database primarily consists of recordings of 31 laryngectomees (27 males and 4 females) pronouncing phonetically balanced sentences. Additionally, it includes parallel recordings of the sentences by 9 healthy speakers (6 males and 3 females) to facilitate speech processing tasks that require small parallel corpora, such as voice conversion or synthetic speech adaptation. Apart from the sentences, the database includes sustained vowels and a small set of isolated words, which can be valuable for research on ES analysis, diagnosis and evaluation. The paper describes the main contents of the database, the recording protocols and procedure, as well as the labeling process. The main acoustic

[1]Current Address: Communications Engineering Dept., Faculty of Engineering of Bilbao, University of the Basque Country (UPV/EHU) Spain

characteristics of the voices, such as speaking rate, durations of the recordings, phones and silences, and other such characteristics are compared with those of a reduced set of healthy voices. In addition, we describe an experiment using the database to improve the performance of an ASR system for ES speakers. This new resource will be made available to the scientific community with the hope that it will be used to improve the quality of life of the laryngectomees.

Keywords: esophageal speech, voice conversion, speech databases, speech intelligibility, speech analysis

2010 MSC: 00-01, 99-00

---

1. Introduction

Esophageal speech (ES) is a type of speech produced by laryngectomees, which are people whose larynx has been surgically removed. The larynx is a fundamental organ in the speech production mechanism. It contains the vocal folds, which are responsible for generating the air vibrations that are perceived as a sound. In addition to the removal of the larynx, the laryngectomy separates the nasal cavity and the vocal tract. As a result, the laryngectomees breathe through a hole (called the *stoma*) which lets the outside air directly into the trachea. Despite the removal of the vocal folds, it is still possible for people who have undergone a total laryngectomy to produce intelligible speech via one of the three main types of alaryngeal speech: using an electro-larynx (EL Speech), tracheoesophageal speech (TES) and ES.

EL speech uses an external vibration device which is placed in contact with the throat. This device generates an acoustic buzz that can be modulated by the movement of the articulators. Intelligible speech is obtained with this method, but the quality is poor, mainly due to the dominant buzzing. The main and perhaps the only advantage of this method is that no learning is required.

In Healthy Laryngeal Speech (HS), the air that flows through the lungs and the trachea vibrates the vocal folds to create sound. This is not anatomically possible for a laryngectomee. Therefore, airflow is produced using other strate-

gies, the effectiveness of which depends on the characteristics and anatomy of each person. A surgical solution is to create a fistula allowing air to pass between the trachea and the esophagus. A valve is placed in the fistula so that no food or liquid can pass to the trachea. The airflow aided by this valve produces vibrations of the esophageal sphincter, which generates TES. TES is more intelligible and requires less effort from the speaker than other methods of obtaining the air [1][2][3][4][5][6]. However, the valves must be changed periodically (requiring surgery) and there are other possible medical complications associated with the implant [7][8].

Unlike TES speakers, ES speakers do not have the valve that allows the controlled entry of air. ES speakers achieve this function by swallowing air and expelling it, very much like the production of a burp. Like TES, the pharyngoesophageal segment is used as a substitute vibratory element instead of the vocal folds. Learning to produce speech in this manner requires long periods of training (usually months) with the assistance of a speech therapist. Due to the difficulties in the production method [9], some individuals never manage to learn ES. However, despite the long periods of learning, ES has the advantage of not requiring a device or periodic surgeries. Therefore, we consider there is a clear advantage in promoting the learning of ES.

As ES and HS production mechanisms are very different, their speech signal characteristics differ greatly too [4][10]. The main consequence for ES is a dramatic reduction in naturalness and intelligibility [11][5][12]. A considerable amount of research effort has been devoted to overcome these limitations of ES, some of which involve artificially modifying its characteristics.

There have been several approaches to enhance the quality and intelligibility of alaryngeal voices. Some studies use source-filter analysis of the pathological signal and focuses on modifying the source, the filter, or both. An example of this approach can be found in [13], where an adaptive gain equalizer algorithm was used to modify the ES source; or in [14], where the reconstruction of normal sounding speech for laryngectomy patients was attempted through a modified CELP (Code Excited Linear Prediction) codec. In [15], different

manipulations of both source and filter were evaluated. Another approach to improve intelligibility and quality is to work with the prosodic elements. In [16], the pitch information extracted from an electroglottograph (EGG) was used to create a synthetic glottal signal which reduced jitter and shimmer. Additionally, spectral smoothing and tilt correction were applied. These modifications reduced the harshness and breathiness of the TES speech. The same authors describe a method for rectifying the duration of pathological phones in [17]. Along the same lines, [18] presents a system where concatenation of randomly chosen healthy reference patterns replaces the pathological excitation, adjusting the short, medium and long-term variability of the pitch.

A different approach to the problem is to use Voice Conversion (VC) techniques. VC aims to modify the characteristics of the voice of an input speaker, making them sound like those of a target speaker. In the classical approach [19], a conversion function is trained using data from both source and target speaker. Although non-parallel training is also possible [20], in VC, a set of parallel source-target sentences is desirable. A set of 50 phonemically balanced sentences in Japanese were used to evaluate the performance and capability of the different VC strategies that were aimed at improving the quality and intelligibility of alaryngeal voices [21][22][23][24][25][26][27].

In this paper, we present an acoustic speech database specifically designed for the research of speech conversion techniques, applied to ES. The purpose of developing the AhoSLABI database[2] was to compile acoustic data which would allow us to investigate the use of VC techniques in improving the intelligibility and quality of ES. Some of our previous work in the field of personalized synthetic voices [28][29] has revealed that laryngectomees are a highly interested user group of the technological developments in speech synthesis, speech recognition and VC techniques. Although no official statistics have been published, in 2018, an estimated 1200 laryngectomies were performed in Spain [8]. We

[2]The name is a combination of the laboratory name of the authors (Aholab), and the name of the Biscayan Association of Laryngectomees ASLABI

aim to provide some useful tools for these laryngectomees, and for social and geographical reasons, we have developed the database in Spanish. To promote research and the comparison of techniques and results, we have provided an open access phonetically labeled database.

First, we present an overview of the existing databases in the following section. Thereafter, we describe the contents of the database and the processes performed. Section "Design of the AhoSLABI database" describes the corpus contents and characteristics of the speakers, as well as the recording setup. The "Results" section presents some metrics of the database and provides some linguistic and acoustic statistics about its contents. In this section, we also report the process of extraction of phonetic labels and the evaluation of the automatic labeling procedure. In addition, we give some preliminary results of ASR and VC experiments performed with the database. The final section presents the conclusions and discusses possible future uses of the database.

2. Existing Related Material

In some types of pathological speech such as dysarthric speech, certain databases have been extensively used and have become a de facto standard [30][31][32]. The same cannot be said for alaryngeal speech databases. For ES, many different recordings of varied characteristics have been performed, each adapted to the purpose of the study. In this section, we review the research publications in the field and give an overview of the existing recordings and their characteristics.

Research on alaryngeal speech has traditionally focused on the production of sustained vowels. Vowels allow easy measurement of fundamental frequency, harmonic properties, and intensity and duration of phonation, which are basic features in assessing the speaker's voice quality and speaking proficiency. Vowels based analysis were performed in a number of studies [33][4][34][35][36][37][38][39]. Some studies used recordings of words and sentences to measure the speaking rate [40][2][41], to study pauses [42] or both [43]. Recordings of words and sen-

tences have also been used in perceptual evaluations [44][45][46][47][48][49][50], and to evaluate synthetic manipulations [15].

Automatic speech recognition (ASR) is also problematic for alaryngeal voices. Some ASR experiments use only vowels [51][52]. Typically, hundreds of sentences are used to train such ASR systems. In [53], a parallel database of 500 sentences pronounced by seven EL and seven HS German speakers were used to evaluate an ASR designed for HS speakers. In [54] 480 sentences produced by one French ES speaker were recorded with the purpose of improving the performance of an existing ASR system.

The statistical VC experiments described in [21][22][55][23][24][25][26][27] use 50 parallel HS-ES sentences, but in Japanese. In order to facilitate the alignment procedure, the HS speaker tried to imitate the rhythm of the ES speakers' utterances. Such a parallel HS-ES database is desirable for VC.

In conclusion, to the best of our knowledge, no standard database exists to perform comparable research of Spanish ES, let alone to carry out VC experiments. We hope to fill this void with the database described in this paper.

3. Design of the AhoSLABI database

3.1. Text content

We selected the Spanish text corpus called ZureTTS described in [28] for the recordings. This corpus contains 100 phonetically balanced sentences encompassing all the phonemes in Castilian Spanish. The phoneme frequency distribution is shown in Table 1. The phoneme codes follow the Spanish SAMPA convention[3]. The total number of phones is 5625. This distribution is consistent with other previous Spanish corpora (see for example [56]). The sentences in the corpus are semantically relatively complex. As we already have HS recordings of this corpus, it made sense to record the ES database with the same corpus.

[3]https://www.phon.ucl.ac.uk/home/sampa/spanish.htm

This allowed us to have a parallel ES-HS corpus which is useful for tasks such as parallel VC.

For a healthy speaker, the recording process usually takes between 30 and 40 minutes. For an ES speaker, the same task takes longer (see subsection Recorded Material and Durations) and for novice ES speakers, it can be quite exhausting. This is why the 100 sentences recorded were further divided in three blocks of 33, 33 and 34 sentences respectively. Each one of these blocks was phonetically balanced within itself. Therefore, if a speaker was tired and decided to not continue with the recording process after the first or the second block, the collected material would still be useful.

Table 1: **Percentage of phonemes in the AhoSLABI corpus.**

Phoneme	**Occurrences (%)**	**Phoneme**	**Occurrences (%)**
a	12.71	b	2.83
e	13.17	d	4.98
i	8.69	g	1.44
o	9.76	p	1.92
u	4.43	t	4.48
m	2.52	k	3.47
n	7.13	f	1.08
J	0.30	s	5.99
l	4.96	T	1.99
L	0.69	x	0.82
jj	0.41	tS	0.44
r	4.75	rr	1.03

In addition to the 100 sentences, each ES speaker recorded 4 instances of the sustained articulation of all five Spanish vowels. Four words containing diphthongs were also recorded (*murciélago*, *acuífero*, *ayuntamiento*, *aceituno*). Ten isolated words, which are also present in the ZureTTS corpus, were included in the recordings, to enable future evaluations of spoken term detection tasks

and the like.

3.2. Characteristics of the Speakers

All the ES speakers who participated in the recording process are members of the Association of Laryngectomees of Biscay (AhoSLABI). The speakers underwent speech therapy sessions after the laryngectomy to learn ES production techniques.

Most candidates performed the recordings months after having finished the speech therapy sessions. We call these speakers 'proficient' ES speakers. On the other hand, 4 of them were still attending the therapy sessions and their speech had very low intelligibility. We call these speakers 'non-proficient' speakers. Out of the 4 non-proficient speakers, 2 returned after finishing the therapy and performed the recordings again. We have kept all these sessions in the database.

The database contains recordings from 31 speakers (27 male and 4 female). It is composed of 34 different sessions as follows:

- 26 proficient ES speakers with one recording session each
- 2 non-proficient ES speakers with one recording session each
- 2 ES speakers with one recording session each when they were non-proficient and one when they were proficient (in total 4 sessions)
- 1 speaker's recordings in both TES and ES (in total 2 sessions)

In summary, out of the 34 sessions, 29 correspond to proficient ES speakers, one to a proficient TES speaker and the remaining four to non-proficient ES speakers.

The mean age of the speakers was 65 years and 4 months, but with large variation. The youngest was 51 years and 4 months old at the time of recording, and the oldest was 82 years and 5 months old.

In order to identify each session, a four character code is used:

- The first two numbers identify the speaker (01 to 32)[4]
- One character specifies the speaker's gender M or F.
- One character specifies the kind of speaker: "3" for the proficient speakers and "2" for the non-proficient speakers. For the TES speaker a "T" has been used.

The majority of sessions (25) feature proficient male speakers. Table 2 lists all the session identifiers.

Table 2: **Session identifiers.**

	Session identifier
Non-proficient, male	13M2, 14M2, 16M2
Non-proficient, female	15F2
Proficient, female	11F3, 15F3, 25F3, 28F3
Tracheoesophageal speaker, male	09MT
Proficient, male	All the others

In addition to the ES speakers, recordings of the 100 sentences from 9 healthy speakers (6 males, 3 females, average age: 36 years and 3 months) are provided. These speakers were selected because of their availability and willingness to be part of the public database, and no criteria of age balance was considered.

3.3. Recording protocol

The database recording protocol and procedures were approved by the ethics committee of the University of the Basque Country (UPV/EHU) (signed on 26th February 2017). The recordings were made in the soundproofed recording room at the Faculty of Engineering (UPV/EHU). Four different microphones (studio microphone - Neumann TLM 103, instrumentation microphone

[4]Recordings from speaker number 27 are not included in the database

-Behringer ECM8000, headphone microphone -DPA 4066-F and condenser microphone -AKG C542BL) were connected by means of an audio acquisition interface (Fireface 400) to a PC (Dell Latitude E4200) with a Firewire cable as shown in Figure 1. With the four microphones, we had four different recordings of each utterance. The recording sampling frequency was 48kHz which was downsampled to 16kHz.

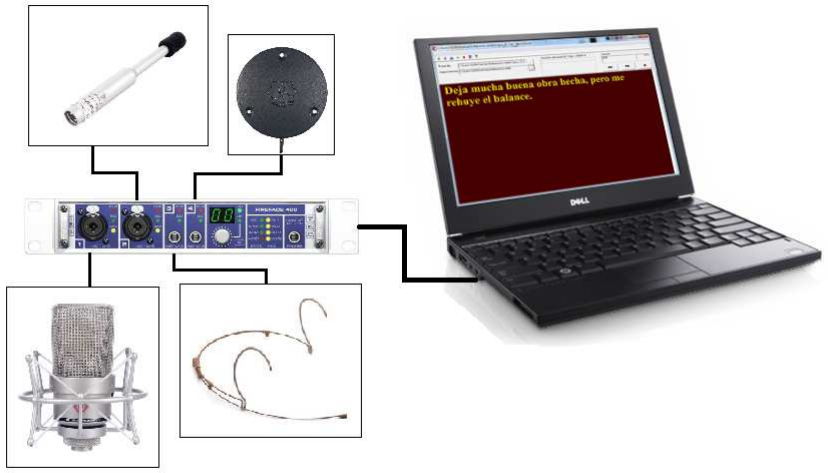

Figure 1: **Scheme of the recording system.**

During the recording process, the speaker was accompanied by one technician. The sentence to be recorded was presented to the ES speaker on a computer screen. First, the assistant read the sentence aloud, in order to demonstrate to the speaker how it should be read. Then, the ES speaker read the sentence while being recorded. The demonstration by the technician has drastically reduced pronunciation errors, i.e. unintentional word substitutions, omissions or insertions. It can be argued that this procedure may modify the natural speaking style of the speaker. However, the prosody of ES is very poor by nature. We think that the advantages of avoiding most spelling errors was preferable over the downsides of possible prosody mimicking. Nonetheless, some

errors still occurred and were annotated by the technician. In the utterances with errors, the corresponding text transcriptions were modified to match what was said in the recorded audio. The main issues observed were the repetition of syllables, the omission of part of a word and the mispronunciations of some phonemes. More details on utterance errors can be found in section 4.2.

4. Results

4.1. Recorded Material and Durations

Thirty out of the 34 sessions contain recordings of all the 100 sentences. Out of the 4 sessions where it was not possible to record all the sentences, one contains 91 sentences produced by one non-proficient speaker. Another session corresponds to the TES speaker pronouncing 33 sentences without the TE valve. The other two incomplete sessions contain 33 sentences from two non-proficient speakers.

The sustained vowels were recorded in all the sessions. The 14 isolated words (including the 4 words with diphthongs) were also recorded in all the sessions, except for the ES recordings of the TES speaker.

Table 5 contains a summary of the content and duration of each session. In total, 9 hours and 31 minutes of audio was recorded. The duration of the sentences is 8 hours and 49 minutes. The duration of the isolated words is 17 minutes and that of vowels is about 25 minutes.

Fig 2 shows a comparison between the duration of the 30 ES and the 9 HS sessions. The differences between both groups of speakers are clear. The average duration of an ES speaker is 773.25 seconds (with a standard deviation of 231.96 s), while for a healthy speaker this value is 405.11 seconds (and a standard deviation of only 27.3 s).

4.2. Ortographic transcription

Although the recording process was designed so that the speakers would faithfully reproduce the sentences, the recordings were not free of mistakes.

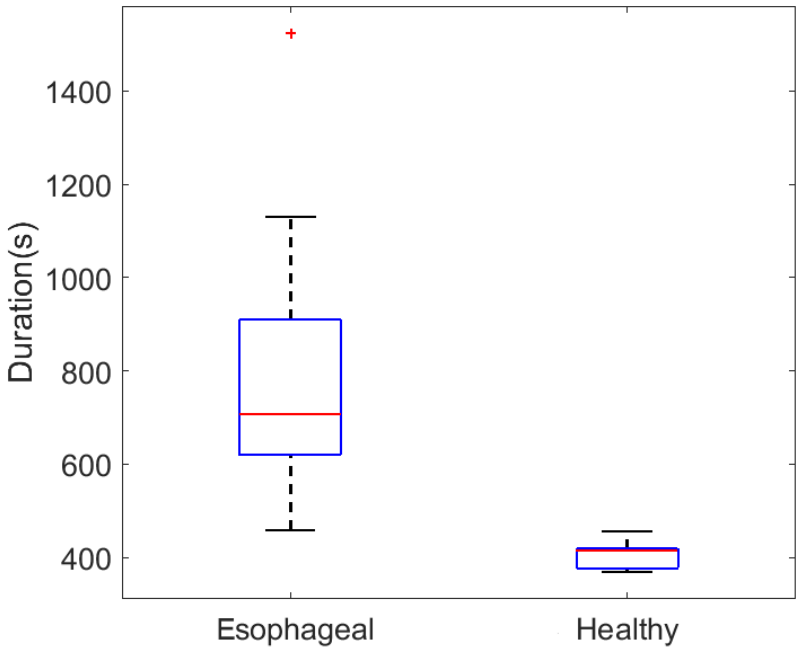

Figure 2: Time taken to utter 100 sentences by 30 esophageal speakers and by 9 healthy speakers.

The transcriptions of each session were later corrected taking into account the annotations made by the technician. The errors were classified as one of the following types:

- Substitution: Instead of the desired word, another word is pronounced. It may be due to a mistake, the failure to pronounce all the phonemes, or a mispronunciation due to lack of air.

- Insertion: A word that does not exist in the original prompt text is uttered. It is most often caused by hesitations or repetitions when trying to pronounce a word, although sometimes they also appear due to reading errors.

- Deletion: A word is not uttered. The most common cause is that the speaker runs out of air and the word is totally inaudible, but it can also

be caused by reading errors.

The errors made in each session can be seen in more detail in Table 6. In 398 out of the 3190 recorded sentences, at least one error was annotated (around 13% of the sentences). The total number of annotated errors is 636, from which 53.4% are substitutions, 36.1% insertions and 10.4% deletions.

4.3. Temporal Analysis

One of the characteristics of ES is that its rhythm is considerably altered. On one hand, laryngectomees need to pause more often for air intake, as the amount of phonatory air is quite reduced compared to that available for the production of normal speech. On the other hand, the control of the vibratory pharyngoesophageal segment is not as precise as the control of the vocal folds, leading to more erratic and less regular rhythm than in healthy speech. Measurements such as speaking rate, phrase length and pause timing were used as an indication of the laryngectomee's speech proficiency.

In this section we present the results of the temporal analysis of the recordings in the database. To calculate the duration of the sounds it was necessary to label the database at phone level. In the following part, we explain the procedure followed to assign these labels. Then we describe the results of the analysis in terms of duration of phones and silences, and speaking rate.

4.3.1. Phonetic labeling

The most common way to segment and label a database at phone level is by using forced alignment that comes as a subproduct of an ASR process. We performed an initial segmentation trial using a Spanish ASR developed in the lab (the one described in section 4.4), with models based on HS. The results showed that this approach was not valid for segmenting ES, due to the enormous differences in voice quality, pauses, rhythm, duration of the sounds etc. Hence, new acoustic models had to be developed, either by adapting the healthy speech models or by creating new models from scratch using the available recordings.

We opted for the latter and used the Montreal Forced Alignment tool [57] to create the new models to perform segmentation.

To provide a reference and assess the accuracy of the automatic labeling procedure, the segmentation of one recorded session (05M3) was manually corrected. Speaker 05 was chosen because he was perceived as having mid to high ES proficiency. Table 3 shows the average manual correction of the time marks. Only phone labels were considered in the table (silences were not taken into account). As shown in the table, 83% of the marks had a difference of less than 5 ms and 97% differed less than 50 ms from the manual reference segmentation. Thus the incidence of large errors was low for this speaker. For more proficient speakers, similar accuracy was expected. However, more segmenting errors may occur with the recordings of less proficient speakers.

Table 3: **Segmentation correction (%).Percentage of marks that were inserted at time instants more than 5, 10, 20 or 50 ms away from the reference mark.**

Session	**Error**			
	< 5 ms	< 10 ms	< 20 ms	< 50 ms
05M3 (5740 marks)	83.03	84.43	89.37	97.14

Fig. 3 shows statistics (median, 25^{th} and 75^{th} percentiles and extreme values) for the duration of phones and silences between words, both for ES (30) and for HS (9) speakers. As expected, both phones and silences were clearly longer in ES compared to HS. We also calculated the average number of inter-word silences per sentence for HS and ES speakers. The results for the average number of silences and the durations of silences and phones are shown in Table 4. We can see that while for HS the average number of silences per sentence was 1.49, for the ES speakers this number increased to 6.28. Also, the average duration of the ES inter-word silences was more than twice the value of the HS speakers' group, with greater variability. Therefore it is evident that the the utterance style of ES speakers is very different to HS speakers.

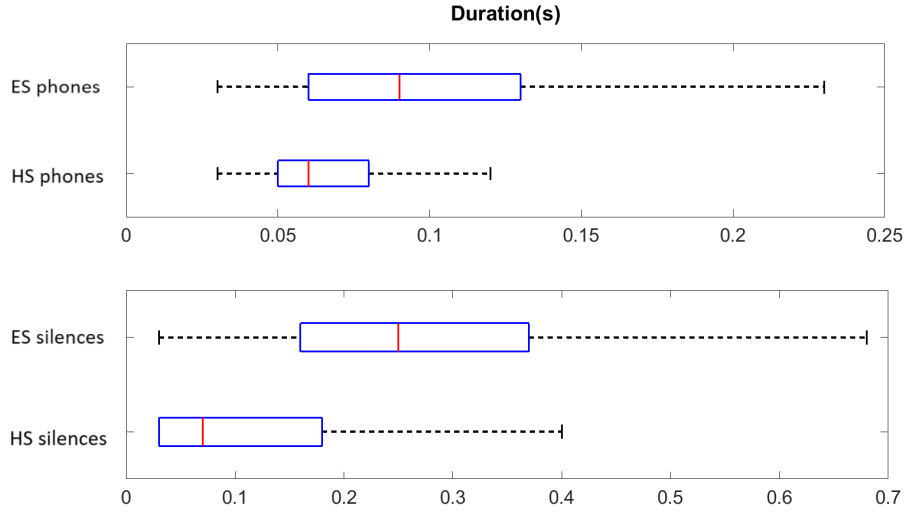

Figure 3: **Duration of phones and silences.** Duration of phones and silences for 30 ES speakers and 9 healthy speakers. In each box, the center line is the median, the edges of the box represent the 25^{th} and 75^{th} percentiles, the whiskers extend to the most extreme values that are not considered outliers. The outliers are not shown for clarity reasons.

Table 4: **Average number of silences per sentence and duration of silences and phones**

	Healthy	**Esophageal**
Average # of silences per sentence	1.49	6.28
Average duration of silences$\pm\sigma$ (ms)	128 $\pm$ 16	299 $\pm$ 57
Average duration of phones $\pm\sigma$ (ms)	68.2 $\pm$ 1.01	98.7 $\pm$ 3.59

4.3.2. Speaking rate

Another way to characterize the dysfluency of an ES speaker is to calculate the *Speaking rate*. Usually this is expressed as the number of phonations per unit of time, such as syllables per second or words per minute. Using the corrected orthographic transcriptions of the sentences, the number of syllables per sentence was calculated. The duration of each sentence was calculated without the initial and final silences.

Fig. 4 shows the resulting median and percentile values for the number of

syllables per second obtained for each session, ordered by mean. The results for the set of 9 HS speakers are also shown. As expected, HS showed a higher speaking rate than ES. It can also be seen that the TES speaker (session 09MT) achieved a speaking rate which does not differ from that of a healthy speaker which corroborates previous analysis on TES and ES [44]. Moreover, when the same speaker did not use the valve (session 09M3), his speaking rate slowed considerably. Another interesting result is that 3 out of the 4 non-proficient speakers had the slowest speaking rates. Two of these non-proficient speakers repeated the recordings three months later, after gaining more control and speech proficiency. While speaker 15F increased her speed, speaker 16M was speaking even slower. However, based on only these two speakers, we cannot generalise these observations.

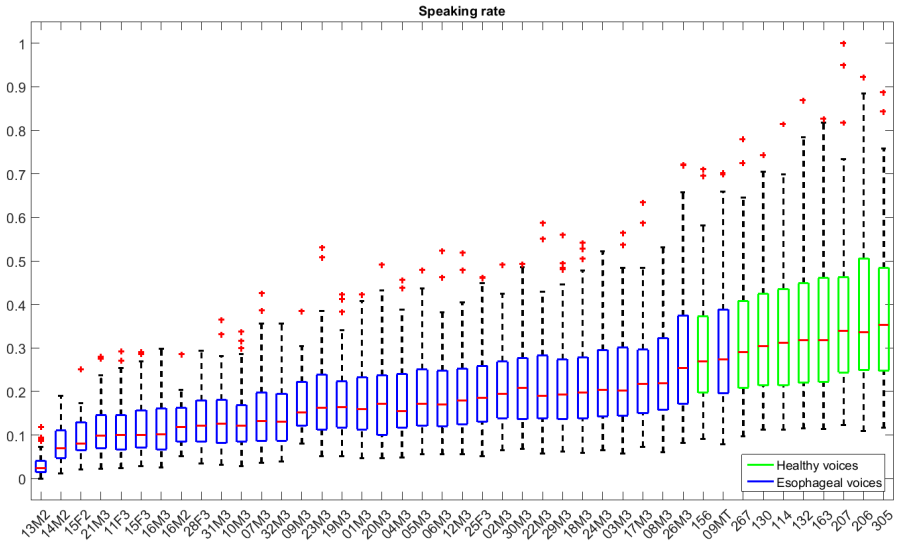

Figure 4: **Speaking rate.** Speaking rate calculated for 34 sessions of esophageal speakers (blue) and 9 of healthy speakers (green). In each box, the center line is the median, the edges of the box represent the 25^{th} and 75^{th} percentiles, the whiskers extend to the most extreme values not considered outliers, and the outliers are shown individually with a red cross.

4.4. ASR experiments

Standard ASR systems normally use healthy speech as training material and therefore perform poorly for ES. In this subsection, we describe an experiment where we compare the results of two ASR systems, one trained with HS and the other one with ES from the AhoSLABI database.

The starting point for both ASR is a standard Spanish ASR built using the Kaldi toolkit [58]. The specific implementation for Spanish is described in [59] and it is implemented following the recipe s5 for the Wall Street Journal database. The training begins with a flat-start initialization of context-independent phonetic Hidden Markov Models (HMM), and then a series of accumulative trainings are done. For the final step of the recognizer, a neural network is trained. The input features to the neural network consist of a series of 40-dimensional features. The network sees a window of these features, with 4 frames on each side of the central frame. The features are derived by processing the conventional 13-dimensional 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 necessary steps are described in [60] and basically consist in applying a series of transformations to the normalized cepstra: first linear discriminant analysis (LDA), then maximum likelihood linear transform (MLLT) and global feature-space maximum likelihood linear regression (fMLLR). At the recognition stage, the same transformations are applied to the test data, handling them as a block.

The main corpus used for the training of the acoustic models is the Spanish section of a subset of the Basque Parliament database. This subset contains the recordings of 47 parliamentary sessions of the Basque Parliament in both Basque and Spanish, together with their correspondent transcriptions [5]. Some preliminary work has been done to separate the Spanish interventions from the Basque ones. As a result, there are more than 124 hours of speech in Spanish

[5] This database is presently being developed by the GTTS research group of the UPV/EHU, contact german.bordel@ehu.eus

uttered by 84 different speakers, 45 male and 39 female. Additionally to the Basque Parliament database, about 4 hours of speech extracted from 5 audio files in Spanish extracted from the Spanish MAVIR workshops held in 2006, 2007 and 2008 was also used to train the acoustic models (see [61] for more details).

To avoid the effects of Out Of Vocabulary (OOV) words, the lexicon for both ASR systems has been reduced to the vocabulary of the 100 sentences of the database and unigram models are used. For the ASR trained with HS, the healthy speakers of the database had a mean WER score of 15.8 ± 3.9, while the ES speakers had a mean WER score of 68.7 ± 16.9. These results show how problematic generic ASR trained with HS can be for ES.

To train the system with ES, we used all the ES speakers for which the complete set of 100 sentences was available. The speakers were divided into 3 blocks of 10 speakers each. The sentences were divided into 10 blocks. A two level cross validation was performed, one at the speaker level and the other at the sentence level. In total 10 (sentence blocks) times 3 (speaker blocks) i.e., 30 cross-validations were performed to ensure that the test data was not seen in the training phase. In each of these cross-validations, 90 sentences from all the speakers of 2 blocks were used as training material and the 10 test sentences of the 3rd block of speakers were evaluated. When done 30 times, all the sentences from all the speakers were covered.

The ASR scores for the 29 proficient speakers from both systems (ASR trained with HS and ASR trained with ES) are presented in figure 5. The non-proficient speaker (14M2) has been removed from the global results due to their poor performance (WER higher than 100%). The WER scores from the ASR trained with HS were significantly higher than the ASR trained with ES ($t(28)=16.14$, $p<0.001$). As can be observed, some speakers benefit more than others from the ES training. The mean improvement in WER is 23.2 ± 7.7.

This result demonstrates that generic Spanish ASR sytems can be made more ES inclusive by using the AhoSLABI database.

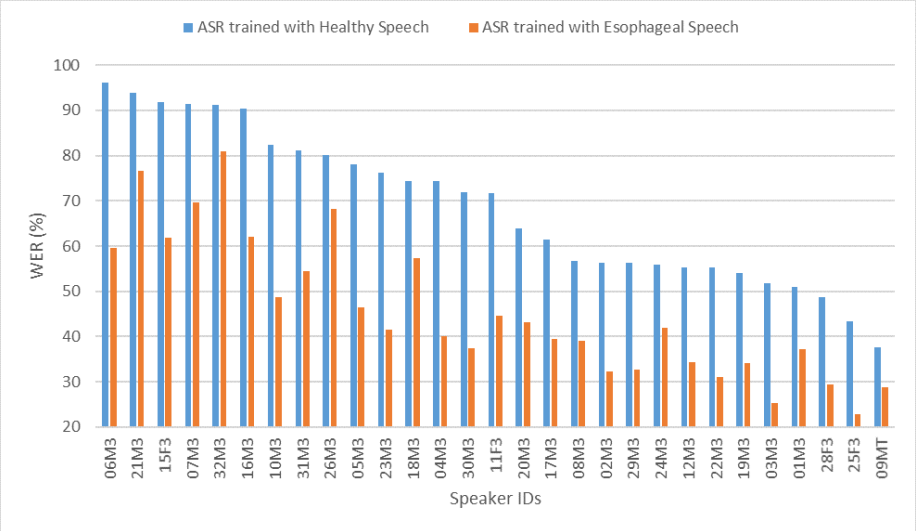

Figure 5: **ASR Results.** Mean speaker-wise Word Error Rates (in %) for ASR trained with HS and ASR trained with ES

5. Conclusions

In this article we have described a database of Spanish ES called AhoSLABI. The database comprises mainly male ES, although it also contains the recordings of four female ES speakers and one male TES speaker. The main content of the database is the recordings of a set of 100 phonetically balanced sentences. The database also contains parallel recordings of 9 healthy speakers. We performed segmentation and labeling on the data. We have described the main aspects of the experimental setup, speaker characteristics and the acoustic properties of the recordings.

The primary motivation for creating this database was the authors' desire to have the laryngectomees benefit from recent advances in speech technologies, specifically in the field of VC. In particular, as reported in section 1, VC techniques have been proposed in the literature to improve the intelligibility of these voices. This was the main reason to record the sentences, as most VC systems need parallel source-target utterances to train the conversion function. Some of our VC work ([62] and [63]) demonstrates how ES can be made more intelligible

or more preferable to listeners using VC techniques.

Although VC was our main intended application, there are many other areas of study where these recordings could be of interest. The sustained vowels recordings are helpful in the evaluation of fundamental frequency, shimmer, jitter, and intensity and duration of phonation. The signals can be used to train and test the performance of ASR systems with ES as shown in section 4.4 of this paper. Additionally, a small set of isolated words is also available which can be useful to test ASR systems in a spoken term detection task.

Another research area is related to the loss of identity in the laryngectomees voices. One's voice is a very important personality trait which is lost with laryngectomy. The recordings available could be useful in the emulation of pre-laryngectomy speech characteristics. Investigating ways to restore this identity could be more feasible if pre-surgery recordings were available. In the future, the authors intend to also record voices of pre-laryngectomy patients.

Subjective evaluation of the quality and intelligibility of alaryngeal speech to improve diagnosis and therapy is also possible with these recordings, because the number and variety of individuals is considerably high. A preliminary study of the intelligibility and listening effort for AhoSLABI was conducted in [64].

We believe that it is not only speech engineers but also researchers in speech therapy who can benefit from this database [6].

6. Acknowledgments

This work was partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER support (RESTORE project, TEC2015-67163-C2-1-R), the Basque Government (PIBA-018-0035) and by the European Union's H2020 research and innovation program under the Marie Curie European Training Network ENRICH (675324).

[6] The database is available for researchers through the European Language Resources Agency repository.

The authors want to thank the Asociación Bizkaina de Laringectomizados for their valuable collaboration and all the laryngectomees for their voice donations. We also would like to thank the reviewers for their fruitful comments that have contributed greatly to the value of the paper.

References

[1] S. E. Williams, J. B. Watson, Speaking proficiency variations according to method of alaryngeal voicing, Laryngoscope 97 (1987) 737–739.

[2] R. H. Pindzola, B. H. Cain, Acceptability ratings of tracheoesophageal speech, Laryngoscope 98 (1988) 394–397.

[3] W. Ainsworth, S. W., Perceptual comparison of neoglottal, oesophageal and normal speech., Folia Phoniatr (Basel) 44 (6) (1992) 297–307.

[4] F. Debruyne, P. Delaere, J. Wouters, P. Uwents, Acoustic analysis of tracheo-oesophageal versus oesophageal speech, The Journal of Laryngology & Otology 108 (4) (1994) 325–328.

[5] T. Most, Y. Tobin, R. C. Mimran, Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production, Journal of communication disorders 33 (2) (2000) 165–181.

[6] L. Širić, D. Šoš, M. Rosso, S. Stevanović, Objective assessment of tracheoesophageal and esophageal speech using acoustic analysis of voice, Collegium antropologicum 36 (2) (2013) 111–114.

[7] B. M. Op de Coul, F. J. Hilgers, a. J. Balm, I. B. Tan, F. J. van den Hoogen, H. van Tinteren, A decade of postlaryngectomy vocal rehabilitation in 318 patients: a single Institution's experience with consistent application of provox indwelling voice prostheses., Archives of otolaryngology–head & neck surgery 126 (11) (2000) 1320–8. doi:10.1001/archotol.126.11.1320.
URL http://www.ncbi.nlm.nih.gov/pubmed/11074828

[8] P. Díaz de Cerio Canduela, I. Arán González, R. Barberá Durban, A. Sistiaga Suárez, M. Tobed Secall, P. L. Parente Arias, Rehabilitation of the laryngectomised patient. Recommendations of the Spanish Society of Otolaryngology and Head and Neck Surgery, Acta Otorrinolaringológica Española (2018) 1–6doi:10.1016/j.otorri.2018.01.003.
URL https://doi.org/10.1016/j.otorri.2018.01.003

[9] E. Lundström, Voice Function and Quality of Life in Laryngectomees, in: Studies in Logopedics and Phoniatrics, 13, Karolinska Institutet, Stockholm, 2009.

[10] W. Wszolek, M. Modrzejewski, M. Przysiezny, Acoustic analysis of esophageal speech in patients after total laryngectomy, Archives of Acoustics 32 (4 (Supplement)) (2007) 151–158.

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

[12] T. Drugman, M. Rijckaert, C. Janssens, M. Remacle, Tracheoesophageal speech: A dedicated objective acoustic assessment, Computer Speech & Language 30 (1) (2015) 16–31.

[13] R. Ishaq, 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. 000210–000214.

[14] H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec, IEEE Transactions on Biomedical Engineering 57 (10) (2010) 2448–2458.

[15] R. van Son, I. Jacobi, F. J. Hilgers, et al., Manipulating tracheoesophageal speech., in: Interspeech, 2010, pp. 274–277.

[16] A. Del Pozo, S. Young, Continuous tracheoesophageal speech repair, in: Signal Processing Conference, 2006 14th European, Citeseer, 2006, pp. 1–5.

[17] A. Del Pozo, S. Young, Repairing tracheoesophageal speech duration, in: Proc Speech Prosody, Citeseer, 2008, pp. 187–190.

[18] O. Schleusing, R. Vetter, P. Renevey, J.-M. Vesin, V. Schweizer, Prosodic speech restoration device: Glottal excitation restoration using a multi-resolution approach, in: International Joint Conference on Biomedical Engineering Systems and Technologies, Springer, 2010, pp. 177–188.

[19] Y. Stylianou, O. Cappé, E. Moulines, Continuous probabilistic transform for voice conversion, IEEE Transactions on Speech and Audio Processing 6 (2) (1998) 131–142. doi:10.1109/89.661472.

[20] D. Erro, A. Moreno, A. Bonafonte, Inca algorithm for training voice conversion systems from nonparallel corpora, IEEE Transactions on Audio, Speech, and Language Processing 18 (5) (2009) 944–953.

[21] M. Kishimoto, T. Toda, H. Doi, S. Sakti, S. Nakamura, Model training using parallel data with mismatched pause positions in statistical esophageal speech enhancement, in: Signal Processing (ICSP), 2012 IEEE 11th International Conference on, Vol. 1, IEEE, 2012, pp. 590–594.

[22] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models, IEICE TRANSACTIONS on Information and Systems 93 (9) (2010) 2472–2482.

[23] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Statistical approach to enhancing esophageal speech based on gaussian mixture models, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010, pp. 4250–4253.

[24] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Speaking-aid systems based on one-to-many eigenvoice conversion for total laryngectomees, APSIPA ASC 2010 - Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.

[25] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 5136–5139.

[26] K. Yamamoto, T. Toda, H. Doi, H. Saruwatari, K. Shikano, Statistical approach to voice quality control in esophageal speech enhancement, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4497–4500.

[27] H. Doi, Augmented speech production beyond physical constraints using statistical voice conversion – alaryngeal speech enhancement and singing voice quality control, Ph.D. thesis, Nara Institute of Science and Technology (2013).

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

[29] D. Erro, I. Hernaez, A. Alonso, D. García-Lorenzo, E. Navas, J. Ye, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños, R. Pérez-Ramón, M. Sulír, X. Tian, X. Wang, Personalized synthetic voices for speaking impaired: Website and app., in: Interspeech, 2015, pp. 1251–1254.

[30] M. Eye, E. Infirmary, Voice disorders database, version. 1.03 (cd-rom) (1994).

[31] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, H. T. Bunnell, The nemours database of dysarthric speech, in: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, Vol. 3, IEEE, 1996, pp. 1962–1965.

[32] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang,

K. Watkin, S. Frame, Dysarthric speech database for universal access research, in: Proceedings of Interspeech, 2008, pp. 1741–1744.

[33] M. Kinishi, M. Amatsu, Pitch perturbation measures of voice production of laryngectomees after the amatsu tracheoesophageal shunt operation, Auris Nasus Larynx 13 (1) (1986) 53–62.

[34] M. R. Arias, J. L. Ramón, M. Campos, J. J. Cervantes, Acoustic analysis of the voice in phonatory fistuloplasty after total laryngectomy, Otolaryngology—Head and Neck Surgery 122 (5) (2000) 743–747.

[35] C. J. van As-Brooks, F. J. Koopmans-van Beinum, L. C. Pols, F. J. Hilgers, Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech, Journal of Voice 20 (3) (2006) 355–368.

[36] M. Carello, M. Magnano, A first comparative study of oesophageal and voice prosthesis speech production, EURASIP Journal on Advances in Signal Processing 2009 (1) (2009) 821304.

[37] J. K. MacCallum, L. Cai, L. Zhou, Y. Zhang, J. J. Jiang, Acoustic analysis of aperiodic voice: perturbation and nonlinear dynamic properties in esophageal phonation, Journal of Voice 23 (3) (2009) 283–290.

[38] N. Deore, S. Datta, R. Dwivedi, R. Palav, R. Shah, S. Sayed, M. Jagde, R. Kazi, Acoustic analysis of tracheo-oesophageal voice in male total laryngectomy patients, The Annals of The Royal College of Surgeons of England 93 (7) (2011) 523–527.

[39] H.-J. Shim, H. R. Jang, H. B. Shin, D.-H. Ko, Cepstral, spectral and time-based analysis of voices of esophageal speakers, Folia Phoniatrica et Logopaedica 67 (2) (2015) 90–96.

[40] J. Robbins, H. B. Fisher, E. C. Blom, M. I. Singer, A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production, Journal of Speech and Hearing disorders 49 (2) (1984) 202–210.

[41] C. Finizia, H. Dotevall, E. Lundström, J. Lindström, Acoustic and perceptual evaluation of voice and speech quality: a study of patients with laryngeal cancer treated with laryngectomy vs irradiation, Archives of Otolaryngology–Head & Neck Surgery 125 (2) (1999) 157–163.

[42] J. Merol, F. Swierkosz, O. Urwald, T. Nasser, M. Legros, Acoustic comparison of esophageal versus tracheoesophageal speech, Revue de laryngologie-otologie-rhinologie 120 (4) (1999) 249–252.

[43] M. H. Bellandese, J. W. Lerman, H. R. Gilbert, An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers, Journal of speech, language, and hearing research 44 (6) (2001) 1315–1320.

[44] S. E. Williams, J. B. Watson, Speaking proficiency variations according to method of alaryngeal voicing, The Laryngoscope 97 (6) (1987) 737–739.

[45] J. L. Miralles, T. Cervera, Voice intelligibility in patients who have undergone laryngectomies, Journal of Speech, Language, and Hearing Research 38 (3) (1995) 564–571.

[46] M. L. Ng, C.-L. I. Kwok, S.-F. W. Chow, Speech performance of adult cantonese-speaking laryngectomees using different types of alaryngeal phonation, Journal of Voice 11 (3) (1997) 338–344.

[47] R. McDonald, V. Parsa, P. Doyle, G. Chen, On the prediction of speech quality ratings of tracheoesophageal speech using an auditory model, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, 2008, pp. 4517–4520.

[48] A. Huang, T. H. Falk, W.-Y. Chan, V. Parsa, P. Doyle, Reference-free automatic quality assessment of tracheoesophageal speech, in: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE, 2009, pp. 6210–6213.

[49] I. K.-Y. Law, E. P.-M. Ma, E. M.-L. Yiu, Speech intelligibility, acceptability, and communication-related quality of life in chinese alaryngeal speakers, Archives of Otolaryngology–Head & Neck Surgery 135 (7) (2009) 704–711.

[50] C. M. G. Membiela, M. J. F. Gutiérrez, S. M. Andrés, L. S. Rabanal, P. S. Rodríguez, C. Á. Marcos, La voz del laringectomizado: incapacidad, percepción y análisis acústico, Revista de Logopedia, Foniatría y Audiología 36 (3) (2016) 127–134.

[51] R. Pietruch, A. Grzanka, Combining acoustic and visual modalities in vowel recognition system for laryngectomees, in: Neural Network Applications in Electrical Engineering (NEUREL), 2010 10th Symposium on, IEEE, 2010, pp. 175–179.

[52] R. W. Pietruch, A. D. Grzanka, Vowel recognition of patients after total laryngectomy using mel frequency cepstral coefficients and mouth contour, Journal of Automatic Control 20 (1) (2010) 33–38.

[53] A. K. Fuchs, J. A. Morales-Cordovilla, M. Hagmüller, Asr for electrolaryngeal speech., in: ASRU, 2013, pp. 234–238.

[54] O. Lachhab, J. Di Martino, E. I. Elhaj, A. Hammouch, A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion, SpringerPlus 4 (1) (2015) 644.

[55] H. DOI, K. NAKAMURA, T. TODA, H. SARUWATARI, K. SHIKANO, Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models, IEICE Transactions on Information and Systems E93.D (9) (2010) 2472–2482. doi:10.1587/transinf.E93.D.2472.

[56] V. Aubanel, M. L. G. Lecumberri, M. Cooke, The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology, International Journal of Audiology 53 (9) (2014) 633–638.

[57] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Montreal forced aligner: Trainable text-speech alignment using kaldi, in: Pro-

ceedings of Interspeech, 2017, pp. 498–502. doi:10.21437/Interspeech.2017-1386.

[58] 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.

[59] L. Serrano, D. Tavárez, I. Odriozola, I. Hernaez, I. Saratxaga, Aholab system for albayzin 2016 search-on-speech evaluation, in: Proceedings of IberSPEECH, 2016, pp. 33–31.

[60] S. P. Rath, D. Povey, K. Veselỳ, J. Cernockỳ, Improved feature processing for deep neural networks., in: Proceedings of Interspeech, 2013, pp. 109–113.

[61] J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, L. Serrano, I. Hernaez, A. Coucheiro-Limeres, J. Ferreiros, J. Olcoz, J. Llombart, Albayzin 2016 spoken term detection evaluation: an international open competitive evaluation in spanish, EURASIP Journal on Audio, Speech, and Music Processing 22. doi:10.1186/s13636-017-0119-z.

[62] L. Serrano, D. Tavarez, X. Sarasola, S. Raman, I. Saratxaga, E. Navas, I. Hernaez, LSTM based voice conversion for laryngectomees, in: Proceedings of IberSPEECH, 2018, pp. 122–126. doi:10.21437/IberSPEECH.2018-26.
URL http://dx.doi.org/10.21437/IberSPEECH.2018-26

[63] L. Serrano, S. Raman, D. Tavarez, E. Navas, I. Hernaez, Parallel vs. Non-Parallel Voice Conversion for Esophageal Speech, in: Proc. Interspeech 2019, 2019, pp. 4549–4553. doi:10.21437/Interspeech.2019-2194.
URL http://dx.doi.org/10.21437/Interspeech.2019-2194

[64] S. Raman, I. Hernaez, E. Navas, L. Serrano, Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech, in: Proceedings of IberSPEECH, 2018, pp. 107–111. doi:10.21437/IberSPEECH.2018-23.
URL http://dx.doi.org/10.21437/IberSPEECH.2018-23

Table 5: **Contents and duration of each session.**

Session ID	Sentences	Words	Sustained vowel repetitions	Duration (min s)
01M3	100	14	4	16' 34,6"
02M3	100	14	4	13' 46,6"
03M3	100	14	4	13' 13,5"
04M3	100	14	4	16' 21,5"
05M3	100	14	4	14' 34,5"
06M3	100	14	4	14' 54,9"
07M3	100	14	4	19' 14,6"
08M3	100	14	4	12' 53,4"
09M3	33	0	4	05' 48,0"
09MT	100	14	4	13' 17,1"
10M3	100	14	4	20' 37,2"
11F3	100	14	4	22' 17,5"
12M3	100	14	4	13' 50,2"
13M2	91	14	4	52' 19,8"
14M2	100	14	4	30' 08,1"
15F2	33	14	4	08' 08,3"
15F3	100	14	5	20' 41,6"
16M2	33	14	4	07' 14,6"
16M3	100	14	5	22' 17,0"
17M3	100	14	4	12' 55,1"
18M3	100	14	4	13' 45,3"
19M3	100	14	4	16' 19,7"
20M3	100	14	4	17' 03,7"
21M3	100	14	4	22' 00,3"
22M3	100	14	4	12' 41,5"
23M3	100	14	4	15' 56,5"
24M3	100	14	4	13' 04,2"
25F3	100	14	4	13' 52,9"
26M3	100	14	4	09' 59,7"
28F3	100	14	4	19' 40,8"
29M3	100	14	4	13' 45,2"
30M3	100	14	4	12' 45,2"
31M3	100	14	4	20' 24,4"
32M3	100	14	4	18' 21,4"

Table 6: **Errors committed in each session.**

Session	Recorded sentences	Sentences with errors	Total number of errors	Sub	Ins	Del
01M3	100	1	1	1	0	0
02M3	100	0	0	0	0	0
03M3	100	5	6	4	2	0
04M3	100	4	5	3	2	0
05M3	100	1	1	0	1	0
06M3	100	4	5	1	4	0
07M3	100	10	11	6	3	2
08M3	100	15	18	7	6	5
09M3	33	3	4	0	3	1
09MT	100	6	7	5	1	1
10M3	100	15	22	12	10	0
11F3	100	27	43	26	13	4
12M3	100	7	9	4	4	1
13M2	91	10	15	7	1	7
14M2	100	8	14	4	9	1
15F2	33	13	22	5	17	0
15F3	100	18	29	14	14	1
16M2	33	3	3	1	1	1
16M3	100	13	13	9	2	2
17M3	100	4	7	1	6	0
18M3	100	20	32	18	12	2
19M3	100	3	9	3	6	0
20M3	100	44	81	43	31	7
21M3	100	7	8	2	2	4
22M3	100	10	13	6	6	1
23M3	100	23	33	23	8	2
24M3	100	17	34	21	13	0
25F3	100	6	8	6	2	0
26M3	100	41	72	36	27	9
28F3	100	3	3	3	0	0
29M3	100	3	3	2	1	0
30M3	100	10	13	8	5	0
31M3	100	16	17	15	1	1
32M3	100	28	75	44	17	14