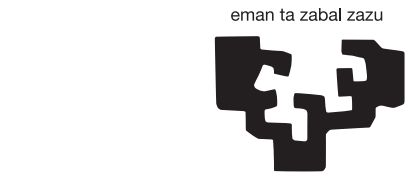


Oesophageal Speech: Enrichment and Evaluations

Doctoral thesis presented by Sneha Raman within the Language Analysis and Processing programme under the supervision of Prof. Inmaculada Hernáez Rioja and Dr. Eva Navas

Sneha Raman

Doctor of Philosophy
University of the Basque Country (UPV/EHU)
15 November 2021

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Sneha Raman*)

Dedicated to my late father for showing me through words and actions what unconditional support really means. You were one of a kind. May more people have fathers like you.

Acknowledgements

A PhD came to me like an unexpected visitor. I was not prepared for it, nor was I seeking it. However, it has left a very profound impact in my life, unlike anything else I have experienced before. Through this process I have gained many life skills and lessons. To manage and take care of this unexpected visitor, I was also sent a host of people and organisations to whom I express my sincerest gratitude. With your presence and participation, you have elevated my doctoral experience.

First and foremost, I thank you Prof. Inma Hernaez for taking me under your wing and teaching me to fly, even at times when I was not willing to. Your direction and dedication has made me a more confident person and researcher. Thank you for pushing me whenever I have been lazy, and for patting my back whenever I have done well. Long lab lunches, being stranded at airports and a celebration of scotch whisky post a rejected conference paper are all incidents I will remember vividly for years. They have surely helped in extending our relationship outside of work and in adding some sense of humour in an otherwise very serious profession.

I thank you Dr. Eva Navas for your invaluable contributions towards my PhD. The one semester course you gave us early on in the PhD was one of my finest experiences as a student. Your efficient and crystal clear advice has helped me tackle many conflicts and difficult situations throughout the PhD. I have and will always admire the elegance with which you carry yourself and your work.

Thank you Dr. Axel Winneke and Dr. Clara Martin for giving me your knowledge and expertise in neuroimaging which has improved my research tremendously. Working with you has opened my eyes to a whole new world of research and research methodologies.

I thank my colleagues Xabier Sarasola, Luis Serrano, Itxasne Diez, Ibon Saratxaga, Jon Sanchez, David Tavarez, Victor Garcia, Inge Salomons and Eder del Blanco for being the salt and pepper of my PhD. Without you guys this PhD would not have been so interesting and balanced. Thank you Xabier and Luis for some very important contributions to my PhD. Some of your ideas have been strong foundations for my research. Ibon, your effervescent personality

and readiness to help with all my roadblocks is much appreciated. Jon, thanks for being the go to person for any technical difficulties and ensuring that the office pantry is up to date! Itxasne, you have advised and helped me like a big sister. You never lose a chance to have a good big laugh, a reminder to all of us to take things with a good spirit! Inge, although you appeared in my life in the final stages of my PhD, your impact has been huge. Thanks for your wise company and making the thesis writing stage more enjoyable.

Thank you Peio for being a strong pillar throughout this process and for keeping me physically and mentally healthy. You declared me a doctor long before I finished my PhD and brought me back into the game whenever I had thoughts of quitting. You believed that I will finish this more than myself. You pushed me to achieve goals whether it was a journal paper, 50 burpees or climbing the last hundred steps to reach the mountain top. This PhD is a result of the endurance I built with you. Thanks also to Garbiñe and Javi, your lovely parents who gave me so much love and affection and rushed to help me whenever I needed it.

A big thank you to the ENRICH family for providing me the warmth and nourishment in this journey. All the regular update meetings and exchange of ideas helped me be on track and organised and actually finish this PhD. I will fondly remember the getting together and venting of frustrations, fun outings and fancy dinners that I shared with the ENRICH members.

Amidst all the gains, I experienced a big loss too. I lost my father who was my mentor and advisor and the person from whom I have received the most affection. But he has left me with all of his positive energy which has helped me finish this thesis and will help me in my future endeavours too. I want to thank my sister Shreya for coming into my life and for being by my side to experience all the fun and tragic moments of life. Life would be so bland without you. And finally, but most importantly, I want to thank my mother for being my constant cheerleader whether I realised it or not. Thank you for all the love.

Abstract

After a laryngectomy (i.e. removal of the larynx) a patient can no more speak in a healthy laryngeal voice. Therefore, they need to adopt alternative methods of speaking such as oesophageal speech. In this method, speech is produced using swallowed air and the vibrations of the pharyngo-oesophageal segment, which introduces several undesired artefacts and an abnormal fundamental frequency. This makes oesophageal speech processing difficult compared to healthy speech, both auditory processing and signal processing. The aim of this thesis is to find solutions to make oesophageal speech signals easier to process, and to evaluate these solutions by exploring a wide range of evaluation metrics.

First, some preliminary studies were performed to compare oesophageal speech and healthy speech. This revealed significantly lower intelligibility and higher listening effort for oesophageal speech compared to healthy speech. Intelligibility scores were comparable for familiar and non-familiar listeners of oesophageal speech. However, listeners familiar with oesophageal speech reported less effort compared to non-familiar listeners. In another experiment, oesophageal speech was reported to have more listening effort compared to healthy speech even though its intelligibility was comparable to healthy speech. On investigating neural correlates of listening effort (i.e. alpha power) using electroencephalography, a higher alpha power was observed for oesophageal speech compared to healthy speech, indicating higher listening effort. Additionally, participants with poorer cognitive abilities (i.e. working memory capacity) showed higher alpha power.

Next, using several algorithms (preexisting as well as novel approaches), oesophageal speech was transformed with the aim of making it more intelligible and less effortful. The novel approach consisted of a deep neural network based voice conversion system where the source was oesophageal speech and the target was synthetic speech matched in duration with the source oesophageal speech. This helped in eliminating the source-target alignment process which is particularly prone to errors for disordered speech such as oesophageal speech. Both speaker dependent and speaker independent versions of this system were implemented. The outputs

of the speaker dependent system had better short term objective intelligibility scores, automatic speech recognition performance and listener preference scores compared to unprocessed oesophageal speech. The speaker independent system had improvement in short term objective intelligibility scores but not in automatic speech recognition performance. Some other signal transformations were also performed to enhance oesophageal speech. These included removal of undesired artefacts and methods to improve fundamental frequency. Out of these methods, only removal of undesired silences had success to some degree (1.44 % points improvement in automatic speech recognition performance), and that too only for low intelligibility oesophageal speech.

Lastly, the output of these transformations were evaluated and compared with previous systems using an ensemble of evaluation metrics such as short term objective intelligibility, automatic speech recognition, subjective listening tests and neural measures obtained using electroencephalography. Results reveal that the proposed neural network based system outperformed previous systems in improving the objective intelligibility and automatic speech recognition performance of oesophageal speech. In the case of subjective evaluations, the results were mixed - some positive improvement in preference scores and no improvement in speech intelligibility and listening effort scores. Overall, the results demonstrate several possibilities and new paths to enrich oesophageal speech using modern machine learning algorithms. The outcomes would be beneficial to the disordered speech community.

Resumen

Después de una laringectomía (es decir, extirpación de la laringe), el paciente ya no puede hablar con una voz laríngea sana. Por lo tanto, estos pacientes deben adoptar métodos alternativos de habla, como el habla esofágica. En este método, el habla se produce utilizando aire tragado y las vibraciones del segmento faringoesofágico, que introduce ruidos y efectos no deseados y produce una frecuencia fundamental anormal. Esto dificulta el procesamiento del habla esofágica en comparación con el habla sana, tanto el procesamiento auditivo como el procesamiento automático de señales. El objetivo de esta tesis es encontrar soluciones para lograr que las señales de habla esofágica sean más fáciles de procesar tanto por parte de los humanos como de los ordenadores y evaluar dichas soluciones explorando una amplia gama de métricas de evaluación para encontrar la más adecuada para este problema.

En primer lugar, se realizaron algunos estudios preliminares para comparar el habla esofágica y el habla sana. Esto reveló una inteligibilidad significativamente menor y un mayor esfuerzo de escucha para el habla esofágica en comparación con el habla sana. Las puntuaciones de inteligibilidad fueron comparables para los oyentes familiarizados y no familiarizados con el habla esofágica. Sin embargo, los oyentes familiarizados con el habla esofágica expresaron menos esfuerzo en comparación con los oyentes no familiares. En otro experimento, se concluyó que el habla esofágica supone un mayor esfuerzo de escucha en comparación con el habla sana, aun para locutores con un nivel comparable de inteligibilidad. Al investigar los correlatos neuronales del esfuerzo de escucha (es decir, la potencia de la banda alfa) mediante electroencefalografía, se observó una potencia alfa más alta para el habla esofágica en comparación con el habla sana, lo que indica un mayor esfuerzo de escucha. Además, los participantes con peores habilidades cognitivas (es decir, menor capacidad de memoria de trabajo) mostraron mayores valores de potencia en esta banda alfa.

A continuación, utilizando varios algoritmos (con enfoques tanto preexistentes como novedosos), se transformó el habla esofágica con el objetivo de hacerla más inteligible y con menor exigencia de esfuerzo de escucha. El enfoque novedoso consistió en un sistema de conversión de

voz basado en una red neuronal profunda donde la fuente era el habla esofágica y el objetivo era el habla sintética con una duración coincidente con la del habla esofágica de origen. Esto ayudó a eliminar el proceso de alineación fuente-objetivo que es particularmente propenso a errores cuando se aplica al habla esofágica. Se implementaron versiones de este sistema tanto dependientes como independientes del locutor. Las salidas del sistema dependiente del hablante obtuvieron mejores puntuaciones de inteligibilidad objetiva a corto plazo, mayor rendimiento de reconocimiento automático del habla y mejores puntuaciones de preferencia del oyente en comparación con el habla esofágica no procesada. El sistema independiente del hablante tuvo una mejora en las puntuaciones de inteligibilidad objetiva, pero no en el rendimiento del reconocimiento automático del habla. También se probaron otras transformaciones de señal para mejorar el habla esofágica como la eliminación de pausas no necesarias y la disminución de ruidos y efectos no deseados para mejorar la frecuencia fundamental. De estos métodos, solo la eliminación de silencios no deseados tuvo éxito hasta cierto punto (1,44 % puntos de mejora en el rendimiento del reconocimiento automático del habla, únicamente para el habla esofágica de baja inteligibilidad).

Por último, el resultado de estas transformaciones se evaluó y comparó con sistemas anteriores utilizando un conjunto de métricas de evaluación como la inteligibilidad objetiva a corto plazo, el reconocimiento automático del habla, las pruebas de escucha subjetiva y medidas neuronales obtenidas mediante electroencefalografía. Los resultados revelan que el sistema propuesto basado en la red neuronal superó a los sistemas anteriores en la mejora de la inteligibilidad objetiva y del rendimiento del reconocimiento automático del habla esofágica. En el caso de las evaluaciones subjetivas, los resultados fueron mixtos: alguna mejora positiva en las puntuaciones de preferencia y ninguna mejora en la inteligibilidad del habla y las puntuaciones de esfuerzo auditivo.

En general, los resultados de la tesis muestran varias posibilidades y nuevos caminos para enriquecer el habla esofágica utilizando algoritmos modernos de aprendizaje automático. Los resultados de este trabajo serán de utilidad para la comunidad del habla patológica.

Laburpena

Laringektomia baten ondoren (hau da, laringea erauzi ondoren), pazienteak ezin du laringeahots osasuntsu batekin hitz egin. Beraz, paziente horiek hizketa-metodo alternatiboak erabili behar dituzte, hala nola ahots esofagikoa. Metodo honetan, irentsitako airea eta faringoesofagikoaren segmentuaren bibrazioak erabiliz sortzen da hizketa. Metodo honek nahi ez diren zaratak eta efektuak sartzen ditu, eta oinarrizko maiztasun anormala sortzen du. Horrek zaildu egiten du ahots esofagikoaren prozesamendua hizketa osasuntsuarekin konparatuz, bai entzumenaren ikuspuntutik bai seinaleen prozesamendu automatikoaren ikuspuntutik. Tesi honen helburua ahots esofagiko seinaleak bai gizakiek bai ordenagailuek errazago prozesa ditzaten teknikak garatzea da eta baita irtenbide horiek ebaluatzea, ebaluazio-metrika ugari aztertuz, arazo horretarako egokiena dena aurkitzeko.

Lehenik eta behin, zenbait azterketa egin ziren, ahots esofagikoa eta hizketa osasuntsua konparatzeko. Azterketa honek agerian utzi zuen ahots esofagikoaren ulergarritasuna nabarmen txikiagoa zela, eta entzuteko ahalegin handiagoa egin behar zela hizketa osasuntsuarekin alderatuta. Ulergarritasunaren puntuazioak hizketa esofagikoarekin familiartasuna zeukaten eta ez zeukaten entzuleentzat konparagarriak izan ziren. Hala ere, ahots esofagikoa ezagutzen zituzten entzuleek ahalegin gutxiago behar zutela jakinarazi zuten. Beste esperimentu batean ondorioztatu zen ahots esofagikoak entzuteko ahalegin handiagoa eskatzen duela hizketa osasuntsuarekin alderatuta, baita hizlarien ulergarritasun-maila konparagarria denean ere. Elektroentzefalografiaren bidez entzuteko esfortzuaren korrelatu neuronalak (hau da, alfa bandaren potentzia) aztertzean, alfa potentzia handiagoa ikusi zen ahots esofagikorako hizketa osasuntsuarekin alderatuta, eta horrek entzuteko ahalegin handiagoa adierazten du. Gainera, gaitasun kognitibo okerrenak (hau da, lan-oroimenerako gaitasun txikiagoa) zituzten parte-hartzaileek potentzia-balio handiagoak erakutsi zituzten alfa banda horretan.

Ondoren, zenbait algoritmo erabiliz (lehendik eskuragarriak zeuden batzuk nahiz beste berri batzuk), ahots esofagikoa bihurtu egin zen, ulergarriagoa eta entzuteko ahalegin txikiagokoa egiteko. Ikuspegi berria ahotsa bihurtzeko sistema berri bat izan zen, sare neuronal sakon

batean oinarritua, non iturria ahots esofagikoa baitzen, eta helburua, berriz, hizketa sintetikoa, jatorrizko ahots esofagikoaren iraupen berarekin. Horrek iturriaren eta helburuaren arteko lerrokatze-prozesuaren beharra ezabatzen zuen eta hau oso komenigarria da lerrokatzen prozesu honek bereziki akatsetarako joera baitu ahots esofagikoari aplikatzen denean. Sistema horren zenbait bertsio eraiki ziren, bai esatariaren mendekoak, bai esatariaren menpekotasunik ez dutenak. Hiztunaren mendeko sistemaren irteerek puntuazio hobeak izan zituzten: epe laburreko ulergarritasun objektibo hobea, hizketa automatikoki ezagutzeko errendimendu handiagoa eta entzulearen lehentasunezko puntuazio hobeak, prozesatu gabeko ahots esofagikoarekin konparatuta. Hiztunaren sistema independenteak hobetu egin zituen ulergarritasun objektiboaren puntuazioak, baina ez hizketa automatikoki ezagutzearen errendimendua. Ahots esofagikoa hobetzeko beste seinale-eraldaketa batzuk ere probatu ziren, hala nola beharrezkoak ez ziren etenak kentzea eta oinarrizko maiztasuna hobetzeko nahi ez ziren zaratak eta efektuak gutxitzea. Metodo horietatik, nahi ez ziren isiluneak ezabatzeak bakarrik izan zuen arrakastarik puntu jakin bateraino (hizketa automatikoki ezagutzeko errendimendua % 1,44hobetzea, ulergarritasun gutxiko ahots esofagikoarentzat soilik).

Azkenik, transformazio horien emaitza ebaluatu zen aurreko sistemekin lortzen zutenarekin konparatuz, ebaluazioko metrika-multzo bat erabiliz, hala nola epe laburreko ulergarritasun objektiboa, hizketaren ezagutza automatikoa, entzute subjektiboko probak eta elektroentzefalografiaren bidez lortutako neurri neuronalak. Emaitzen arabera, sare neuronalean oinarritutako proposatutako sistemak aurreko sistemak gainditu zituen, hizketa esofagikoaren ulergarritasun objektiboaren eta ezagutza automatikoaren errendimenduaren hobekuntzan. Ebaluazio subjektiboen kasuan, emaitzak mistoak izan ziren: lehentasun-puntuazioetan hobekuntza positiboren bat, eta hizketaren ulergarritasunean eta entzumen-ahaleginaren puntuazioetan hobekuntzarik ez.

Oro har, tesiaren emaitzek zenbait aukera eta bide berri erakusten dituzte ahots esofagikoa aberasteko, ikaskuntza automatikoko algoritmo modernoak erabiliz. Lan honen emaitzak erabilgarriak izango dira hizketa patologikoaren komunitatearentzat.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "*For millions of years, mankind lived just like the animals. Then something happened which unleashed the power of our imagination. We learned to talk and we learned to listen. Speech has allowed the communication of ideas, enabling human beings to work together to build the impossible. Mankind's greatest achievements have come about by talking, and its greatest failures by not talking. It doesn't have to be like this. Our greatest hopes could become reality in the future. With the technology at our disposal, the possibilities are unbounded. All we need to do is make sure we keep talking.*"
>
> — Stephen Hawking

I would like to introduce you to this thesis by bringing some questions to your attention: What is the role of verbal communication in our lives? How important is smooth and efficient verbal communication? What is the effect of our speech on its receivers? What importance does our voice play in forming our identity? What if some key characteristics of our speech are lost? How can science and technology restore these lost characteristics? Are these restorations helpful?

These questions are the starting points to the problems dealt with in this thesis. In this introductory chapter, you will read about the importance of speech communication and difficulties of disordered speech communication. I will also explain the reason for choosing to work on this problem and the possible ways to address it.

1.1 Speech Communication

We communicate for several reasons: to get help, to express our emotions, for leisure, to have and maintain relationships and so on. These are some of the key aspects of our life and makes our life wholesome. Being able to speak well aids you in achieving all the key objectives of communication.

There are many layers in speech communication. Listening, decoding the message, interpreting the message, forming a response, and then articulating it. Each of these layers is a field of study in itself. The focus in this thesis is on the articulation, listening and decoding of the message - the top level interactions. We also delve a little bit deeper by discussing our brains' response to listening to speech.

1.2 Disordered Speech Communication

For the most of us, our voice is the main medium to communicate. We do it without giving it much thought, just like the way we walk with our two legs and see with our two eyes. It is only when we lose the ability to speak properly that we realise its importance. We may have briefly experienced it when we have a bad cold or when we have to speak in a really noisy room or over a poor telephone connection. But what if this disability is permanent? We all know someone who has difficulties speaking. Some are born with these difficulties while others get them at a later stage in life.

A large number of people have speech impediments (0.4% of the total population of Europe) [20]. Speech disorders can be structural or anatomical (cleft lip and palate speech), neurological (dysarthria,apraxia), articulation disorders, stuttering and several other types. Oftentimes, people with speech disorders have unsatisfying social lives, as they are difficult to understand and communicate with.

Voice rehabilitation is a way to alleviate communication problems of disordered speech. This task is primarily performed by speech pathologists who train and help people with disordered speech towards producing intelligible speech. The improvement obtained via speech training is limited due to the physical or neurological constraints. Another possibility of voice rehabilitation is by way of providing restored voices to people with speech disorders. Restoration of voices is done using software algorithms and with the aim of improving the intelligibility and comprehension ability of speech. Modern speech technologies have been shown to help in restoring of disordered voices.

Every speech disorder has its own peculiar characteristics and aetiology. Therefore voice

rehabilitation is also tailored to that particular speech disorder and that particular individual. In this thesis, we focus on one kind of disordered speech known as Oesophageal Speech (OS), which is the speech acquired by a person who has had their larynx removed. The current chapter briefly describes the problems of this type of disordered speech and the aims of this thesis towards tackling some of those problems. In the following chapter, the characteristics of this disorder will be explained in detail.

Across literature, varying terms have been used to denote speech spoken by people with speech disorders and that by healthy individuals with a normally functioning speech apparatus. Some of these terms are typical and atypical/non-typical speech [81], impaired and normal speech [106], degraded and non degraded speech [42] etc. In this thesis, we use the terms disordered speech and healthy speech (HS) as is used in many studies dealing with post-laryngectomy speech [83, 94, 5]. The term alaryngeal speech is used to denote all kinds of speech modalities spoken by a laryngectomee.

1.3 Problem Statement

Due to physical alterations resulting from a laryngectomy i.e absence of a larynx, verbal communication needs to be performed using alternate methods. OS is one such method and it involves speaking with the pharyngo-oesophageal segment, which is a physically taxing process. Moreover the resulting speech is difficult to understand and is unpleasant to listen to. This inability to produce intelligible speech in addition to the added physical effort of producing speech are barriers for the OS speaker in conducting smooth verbal communication. This in turn affects their quality of life and their ability to integrate in the society and conduct day-to-day tasks independently. On the other hand, there are challenges when listening and understanding to OS be it during face-to-face conversations, video or telephone calls and digital devices. This is challenging especially for people who need to interact closely with OS speakers (family, friends etc.) for long periods of time. A more detailed description of the characteristics and challenges of OS can be found in Section 2.1.

1.4 Aim

The general aim of this thesis is to improve the communication of OS speakers and listeners in all possible modalities: telephonic or remote video conversation, face-to-face conversation and interacting with digital devices. To achieve these aims, we need to break them down into two channels: OS speaker improvement and OS listening improvement. Firstly, to improve the

communication on the speaker side, the aim is to explore techniques to enrich OS signals by making it more intelligible and less effortful to process. It must be noted that the said techniques would focus on modifying the digital speech signals of an OS speaker and not on training the OS speaker to produce more intelligible speech. Secondly, to understand and overcome the difficulties in listening to OS, we need to evaluate OS and the enrichments using a diverse set of metrics.

In the literature so far, disordered speech enrichment is largely a signal processing and machine learning task. The evaluation of these systems have mostly been very uni-dimensional and do not answer complex questions that we usually encounter with speech perception, especially for disordered speech. There has been many advanced algorithms designed to enrich various kinds of speech, but with minimum focus on the evaluation of such outputs. On the other hand, the field of speech perception and evaluation of degraded speech has been limited to speech-in-noise, hearing impairment and other commonly researched scenarios. There is very little focus on an in depth evaluation of disordered speech and disordered speech modified with software-based interventions.

Through this thesis, I aim to create a bridge and open some conversations between these two fields (speech enrichment and speech evaluation). This thesis is an attempt to solve a problem of speech communication using the strengths and the approaches of both these fields.

1.5 The Organisation of the Thesis

This introductory chapter presented a brief idea of the problems to be addressed in this thesis. The following chapters of this thesis will elaborate each and every aspect of the problem and the approaches to solve them.

The very next chapter contains all the background information and literature review related to speech disorders, speech enhancement, speech enrichment and speech evaluation, with particular focus on OS. This will provide a strong foundation for understanding the experiments featured in this thesis. The background chapter will be followed by the description of the OS corpus we have used to work on our problem. Understanding this corpus is key, as all the experiments are designed based on the nature of this corpus. Chapter 4 will describe some preliminary evaluations of OS and how it differs from HS. Chapter 5 is a follow up of Chapter 4 where we delve into brain activity and its relationship with speech perception. Both Chapter 4 and 5 are helpful in setting the stage for the final evaluations of OS and its enrichments. Then we move on to the enrichments chapter (Chapter 6) which presents some OS enrichment

methods and their performance. Chapter 7 describes the final experiment of this thesis where we evaluate OS and the outputs from our enrichment methods using a wide platter of evaluation metrics. Finally, in the concluding chapter the main outcomes and contributions of the thesis, future directions and the list of publications are presented.

Chapter 2

Background

"Start where you are. Use what you have. Do what you can."

— Arthur Ashe

This chapter presents all the material that has laid the foundation for the experiments performed as part of this thesis. There are three aspects of the problem that needed a thorough exploration to be able to enrich and evaluate OS. Firstly, Section 2.1 contains an in-depth understanding of OS and OS speakers which was necessary before delving into the task of enrichment. Secondly, Section 2.3 lists some of the previous and ongoing attempts at enriching OS and other kinds of disordered speech. These were useful references and guides for the enrichment process. Thirdly, Section 2.2 describes a wide range of methods used to evaluate speech and speech perception abilities, and how these methods can help in evaluating OS and its enrichments.

A thesis titled 'Técnicas para la mejora de la Inteligibilidad en voces patológicas' (Techniques for the Improvement of Intelligibility of Pathological Voices) by Luis Serrano [123] contains research methodologies, corpus and evaluations that have laid the foundations for this current thesis.

2.1 Alaryngeal Speech Characteristics

2.1.1 Anatomy

Laryngeal cancer patients are advised to undergo a total laryngectomy when it is not possible to save the larynx by radiation, chemotherapy, or other procedures. This is a surgical procedure where the entire larynx is removed and separate airways are created for the mouth, nose and oesophagus. Along with the larynx, the vocal folds are also removed which leaves the

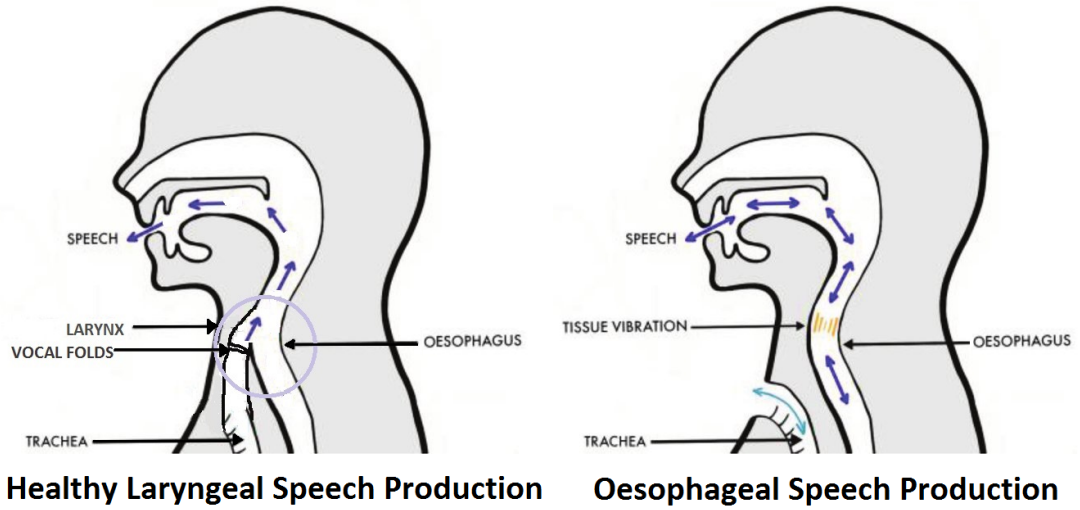


Figure 2.1: Differences in the anatomy of HS and OS

laryngectomee with a disability to speak.

In spite of the absence of vocal folds, laryngectomees can still manage to speak using alternative methods. As mentioned before, OS is one of the alternative methods of speaking after a laryngectomy. The other alternative speaking methods are Electrolaryngeal Speech (ELS) and Tracheoesophageal Speech (TOS) [61]. Both ELS and TOS require external aids—a tracheoesophageal prosthesis in the case of TOS; and electrolarynx (an external electronic substitute for the vocal folds) for ELS. The TOS prosthesis must be changed periodically (around 6 months) by a physician.

In the case of OS, the vibrations in the food passage, or to be more correct, the vibrations in the pharyngo-oesophageal segment are used as a source to generate speech (See Figure 2.1) and that is the reason it is called oesophageal speech. Unlike TOS and ELS, OS does not require any external equipment. It is a skill that is developed with training from a speech therapist and requires several months of practice. It is also of poorer quality compared to TOS or ELS [133]. Nonetheless, OS has the advantage that once the skill is mastered, the laryngectomee is self-sufficient in producing speech and this makes it a promising option post-laryngectomy. Moreover, for ELS or TOS users, getting OS skills is beneficial as it can help them communicate during unexpected situations (lost or broken devices, low battery, etc.).

A comparison of the anatomy of HS and OS is shown in Figure 2.1. We can observe that the HS production system has a larynx and vocal folds with which we produce speech. However, as the anatomy is altered after a laryngectomy, two separate passages are created: one for breathing (light blue line) and the other for food (dark blue lines).

2.1.2 Challenges

Unlike HS, which is produced with vibrations from the vocal folds, OS is produced from the vibrations of the pharyngo-oesophageal segment (See Figure 2.1). Air is swallowed, inhaled, or injected and is introduced into the oesophagus, after which it is expelled with control, thereby producing vibration [133]. This generation mechanism introduces acoustic artefacts and makes OS difficult to understand [155, 92], which greatly affects communication, social activities, and hence, quality of life [90].

Moreover, these less intelligible voices are not well received by machines that are operated by speech input. An increase in the popularity of devices with voice-based interaction means that machine intelligibility is gaining importance too. We know that machine recognition, or Automatic Speech Recognition (ASR) performance, is lower compared to Human Speech Recognition (HSR) performance [73], although better ASR systems are being built increasingly, aiming towards human-like recognition abilities [132]. However, this is for HS and not for disordered speech, like OS.

Figure 2.2 shows the differences in the signal (top) and spectrogram (bottom) characteristics of HS and OS for a voiced phoneme. As we can see in the top figure, HS has clear periodic lines and OS does not. This irregular periodicity in OS affects its fundamental frequency and prosody, which are important characteristics in expressive communication as well as speaker identity. Please note that the OS signal has lines (although unclear) with less frequency compared to the HS signal. This is indicative of the much lower pitch that an OS speaker appears to have. In Figure 2.3, we can see the results of the pitch estimation process by Praat [9], a software used to study and process speech signals. The figure contains the signals (top) and the pitch curves (bottom) for the word 'abrigadero' spoken by a male HS speaker and a male OS speaker. The vertical blue lines on the speech signal (top) represent pulses and the corresponding blue curves (bottom) represent the estimated pitch curve or the f0 curve in Hertz.

For the HS pitch curve, we can see a continuous non-breaking blue string around the 128.9 Hz point, indicating a continuous pitch in a normal range of male fundamental frequency (100 to 150 Hz). On the other hand, for OS, the pulses are few in number and the calculated pitch curve is around 395.3 Hz. This is incorrect as the OS speakers appear to speak in a much lower fundamental frequency and not such a high frequency as calculated. The results on this figure demonstrate the erroneous calculation of fundamental frequency of OS signals by a standard speech software.

The observable differences between OS and HS characteristics in the signal and spectrogram are supported by the following studies. Cervera et. al [11] and Liu et. al [74] observed that

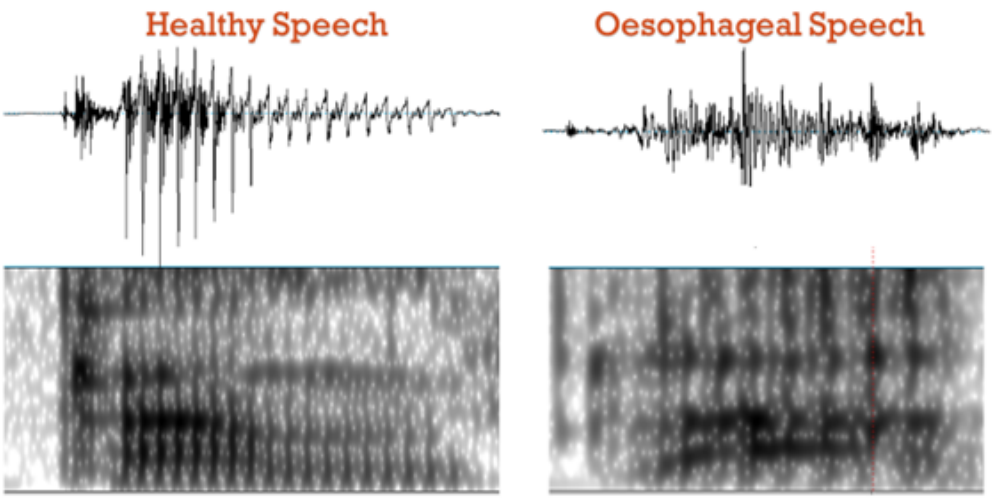


Figure 2.2: Differences in the signal and spectrogram of HS and OS

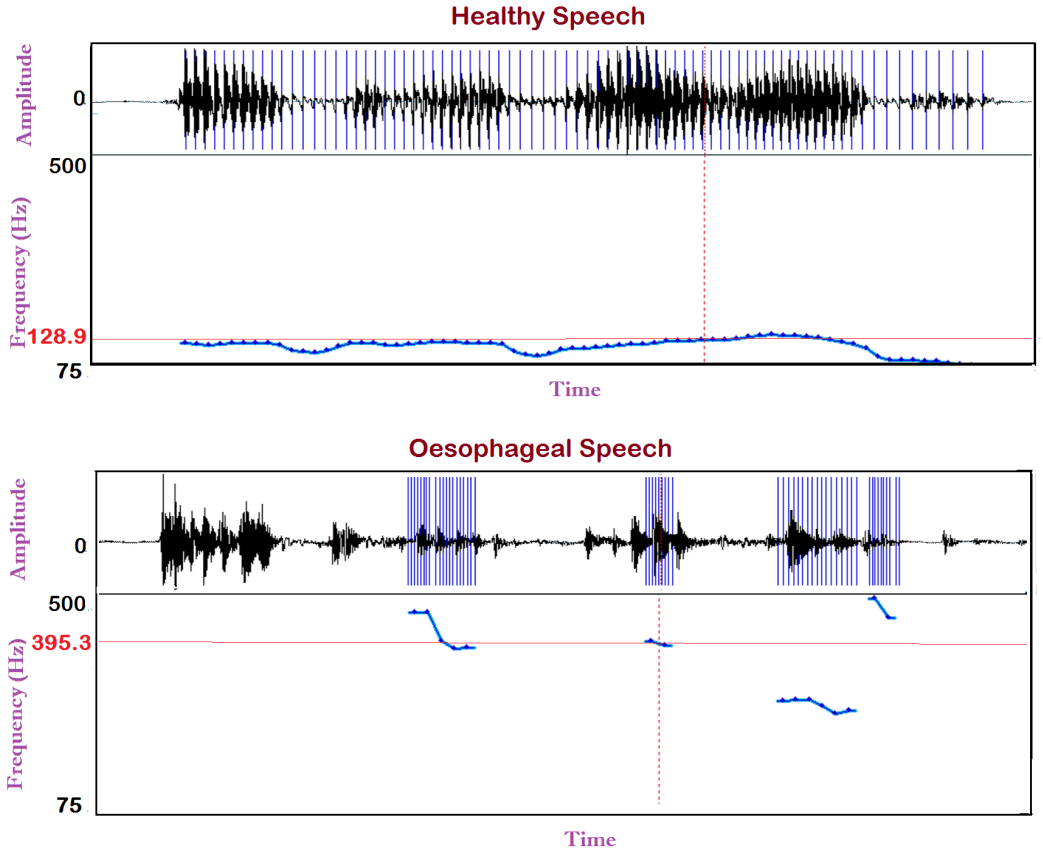


Figure 2.3: Differences in the calculated fundamental frequencies of HS (top) and OS (bottom)

the formant frequencies were higher and the duration of vowels was longer for laryngectomees as compared to HS. Fundamental frequency, intensity, and signal-to-noise ratio of HS were significantly higher than those of OS [74]. Jitter and shimmer were significantly lower for HS compared to OS [123, 74]. On average, OS had about 10 dB less intensity compared to HS [156]. Vowel duration is longer for OS compared to HS [74, 13]. Generally speaking HS has higher speaking rate compared to OS. However, some OS speakers do have speaking rates matching that of HS [123].

2.2 Evaluation Metrics

For any task aimed at making improvements to an existing situation, it is important to conduct evaluations at the beginning and at every stage of progress. This process can tell us clearly whether we are indeed making improvements, or at least progressing in the right direction. In this section, you will find a summary of several evaluation metrics, how they have been used and on what kind of speech they have been used. This knowledge is useful to identify the metrics that are interesting and not yet explored in the context of OS enrichment or the enrichment of any other disordered speech.

2.2.1 Intelligibility

Intelligibility is the most important and the most commonly evaluated property of speech. Speech is useful as a successful communication tool only if it is correctly understood. Speech intelligibility (SI) is a widely researched field, and several intelligibility measurement metrics (subjective and objective) have been explored and analysed. Subjective measures are based on the responses or opinions of the listeners. They may also be derived from listener responses. Objective measures of intelligibility are calculated using a formula, an algorithm or a software program.

Intelligibility may also be categorised as HSR-based or ASR-based. Both involve getting transcriptions of the speech utterance and calculating transcription errors. A review of HSR and ASR methods can be found in [122].

Objective Measurements of Intelligibility

Objective measurements of intelligibility include Speech Transmission Index (STI), Articulation index (AI), Signal-to-Noise Ratio (SNR), Harmonics-to-Noise Ratio (HNR), Short Term Objective Intelligibility (STOI) [134] and ESTOI [58].

STI takes into account the noise and reverberation, and measures the reduction of performance due to speech of varying intelligibility [51]. Measuring AI involves dividing the speech spectrum into several frequency bands and a calculation of the speech-to-noise ratio in each band [66]. For both these measures, a value of 0 means poor intelligibility or no speech was heard and a value of 1 means high intelligibility or speech was perfectly heard.

STOI [139] and ESTOI [58] are intrusive objective intelligibility measures which are known to be correlated with subjective intelligibility scores for noisy speech. An intrusive intelligibility measurement requires a degraded signal and an aligned reference signal. The STOI and ESTOI scores range from 0 to 1 where 0 means the least similarity to the clean reference signal (not intelligible) and 1 means totally similar to the clean reference signal (very intelligible).

On the other hand non-intrusive intelligibility measures do not require a reference signal. Some examples of such measures include the non-intrusive STOI [1], methods based on Deep Neural Networks (DNN) [168, 117] and methods based on auditory-inspired filterbank analysis [35].

The main advantage of objective measurements is that they are easy to replicate and to implement. There is no need for a human subject to evaluate the speech, thereby avoiding hassles usually associated with such perceptual experiments. These metrics measure how a certain speech is received by machines or digital devices. Several of these measurements, including STOI and ESTOI, have high correlations with listening test scores [150] and hence are often used in the place of conducting and actual listening test. Other newer approaches of intelligibility measurement can be found in [1, 130, 150].

Objective measures of intelligibility can also be obtained from an ASR system. An ASR system, as the name suggests, recognises a speech utterance and gives a text output which is the transcription of the utterance. When the transcription matches the message presented in the utterance, then the ASR system is said to have recognised the utterance with 100 percent accuracy, or 0 percent errors. Such an utterance would have 100 percent intelligibility as per that specific ASR system. Currently such a high performance is not obtained with any ASR system. The lowest error (WER) so far is 2 to 3 percent [59] even with clear HS. By running speech utterances through an ASR system and then calculating the errors in transcription, we can assign an objective intelligibility score to the utterance.

The errors in transcription of an ASR system are computed using a metric called Word Error Rate (WER). WER is obtained by calculating the Levenshtein distance [71] between the reference sentence and the hypothesis sentence (the sentence transcribed by the listener). The Levenshtein distance takes into account the insertions, deletions, and substitutions that are

observed in the hypothesis sentence. The calculation was performed with the Word Error Rate Matlab toolbox [105]. The formula used is shown in Equation 2.1.

$$WER = \frac{Substitutions + Insertions + Deletions}{Total\ number\ of\ words\ in\ reference\ sentence}. \tag{2.1}$$

Another way to evaluate the performance of an ASR system is to use the Percentage Words Correct (PWC) method. PWC is the percentage of words from the reference sentence correctly identified in the transcribed sentence. In this case it is not a measure of error, but of accuracy. Higher PWC means higher intelligibility. The PWC formula is shown in Equation 2.2.

$$PWC = \frac{Words\ correctly\ identified\ in\ transcription}{Total\ number\ of\ words\ in\ reference\ sentence} \times 100. \tag{2.2}$$

There are some other objective metrics that do not measure intelligibility, but other important characteristics of intelligible speech. For example, Mel Cepstral Distortion (MCD) evaluates differences in the mel cepstra by calculating the Euclidean cepstral distance of a signal from a reference signal. It is often used to evaluate speech synthesis systems [63] and Voice Conversion (VC) systems [126, 124, 19]. See Section 2.3.1 for details on VC. The lower the MCD, more alike the signal is to the reference signal. On the other hand, Perceptual Evaluation of Speech Quality (PESQ) evaluates quality of the signal. This is done by aligning the reference and degraded signal, performing an auditory transform and extracting distortion parameters from the difference of these two signals. PESQ provides a prediction of a subjective MOS for the degraded signal [116].

Subjective Measurements of Intelligibility

Subjective intelligibility involves ratings or responses from an actual human listener. It needs proper experimental design and setup and well-thought recruitment of participants to obtain reliable measures. Although this is time consuming, there is an obvious advantage with this approach as the evaluations come from an actual person which is the case in face-to-face communications in the real world. Some metrics in subjective intelligibility tests such as Mean Opinion Scores (MOS) and Speech Reception Threshold (SRT) are explained below.

A MOS is a very simple and versatile metric. It is very easy to implement and useful to get a numeric grading of any characteristic that needs to be evaluated. However, MOS values are unidimensional and prone to "misuse and misinterpretation" [137]. Despite of these limitations, MOS is a very popular method of evaluation be it for speech synthesis systems, VC systems or for other speech similarity and naturalness evaluations [154, 75, 152]. A variation of MOS is

the Comparative Mean Opinion Scores (CMOS) where the intelligibility (or any other speech property) is compared for two different signals. This method is useful when comparing the performance of two different systems.

SRT is defined as the minimum hearing level at which 50 percent of the speech material is understood by the listener. This metric is used mainly in hearing impaired research [151].

Another popular method is to conduct a sentence recognition task. Sentence transcription tasks or "human speech recognition" tasks [73] have been widely used for subjective intelligibility measurements [93, 62]. Yorkston et. al [167] reported the agreement of sentence transcription tasks with listeners' estimates and ratings of intelligibility. A variation of the sentence transcription task is the sentence repetition task where instead of writing down the sentence heard by the listener, they simply have to repeat what they heard. This is less taxing for the listener. Moreover, the strengths of sentence repetition tasks are that they are "fairly simple cognitive tasks" and that they are "consistent throughout the age span" in the area of neurophysiological tests [84]. For sentence repetition or sentence transcription tasks as well, WER or PWC that was explained in the previous section can be used as an intelligibility measure. Therefore, although the responses in such tests can be subjective, in the sense that it comes from a subject, the error in transcription or recognition can be measured objectively. This method was most predominantly used to calculate SI in the experiments of this thesis.

Some alternatives to sentence recognition tasks are word [17] or digit recognition tasks [159], sentence last word recognition tasks [67] and keyword recognition tasks [6]. Both last word recognition and word recognition tasks were also used for evaluations in the experiments of this thesis.

Intelligibility of Disordered Speech

Intelligibility measurements for the analysis of disordered speech have been explored in [76, 86, 87]. Some of them are ASR-based [76, 87], while others are not [86]. In an evaluation of disordered speech resulting from cerebral palsy and Amyotrophic Lateral Sclerosis (ALS), a high correlation was reported between STOI and subjective intelligibility ratings [55]. Intelligibility of OS was found to be lower compared to HS and intelligibility improved for OS in the combined auditory-visual mode compared to the auditory only mode [52]. In another study, Holley et. al [50] found higher intelligibility for HS compared to OS as well in the quiet condition.

Some studies have been conducted in measuring the intelligibility of Spanish OS. Authors of [88] studied the voice intelligibility characteristics for Spanish OS and TOS. This HSR study was conducted for two-syllable words, and it was reported that nasal sounds resulted in most

transcription confusions for OS. The work in [77] describes a real time recognition system for vowel segments of Spanish OS.

2.2.2 Listening Effort

While intelligibility is an important property of speech, measures of intelligibility cannot quantify how much effort was required to correctly understand the message. While poorly intelligible speech is usually difficult to understand, sometimes even highly intelligible speech is difficult to understand. For instance, if you are in a noisy room, you can still understand the person next to you, but the process is much more difficult and tiring. This is where LE is helpful as it gives an idea of the effort involved in listening to the message, and the resultant fatigue from prolonged listening to effortful speech.

In literature, LE has been defined as "the mental exertion required to attend to, and understand, an auditory message" [104]. Another related concept is that of cognitive load. The accomplishment of a listening task, especially in adverse conditions, involves the use of cognitive resources of the listener. These resources are present in the listener in the form of working memory. Moreover, each listening task has three aspects that decide the cognitive load or the listening effort of that task: intrinsic load (complex sentences, grammar, poor speech production etc.), the load on the listener for semantic processing, and extraneous load (distracting speakers or images, noise etc.). Therefore the exerted cognitive load depends on the amount of available working memory and how much of it goes into each of these kinds of loads [2]. For example, listening to a very complex sentence in a noisy environment would entail more cognitive load compared to a quieter environment.

Figures 2.4 shows an infographic of the concept of LE. As you can see, the speaker is saying 'I need some water'. Both listeners understand it as 'I am some water'. Therefore, they have both understood 75 percentage of the message correctly (1 word wrongly identified out of 4). However, listener 2 has to exert more effort to decipher the message compared to listener 1, which is indicated by the redness in the head. This might depend on several factors such as the cognitive and the hearing capacities of the listener, the listening environment etc. For example, if listener 1 were to listen to the same speech in noise, he would have needed to exert more effort. The effort might also depend on the speech itself. We aim to explore all these LE factors in the context of OS and normal hearing listeners.

LE has been measured in several ways: Self-reporting (questionnaires, ratings etc.); behavioural measures (performance in single tasks or multiple tasks and deriving LE from them); and physiological measures (electroencephalography, pupillometry etc.). A review of LE and

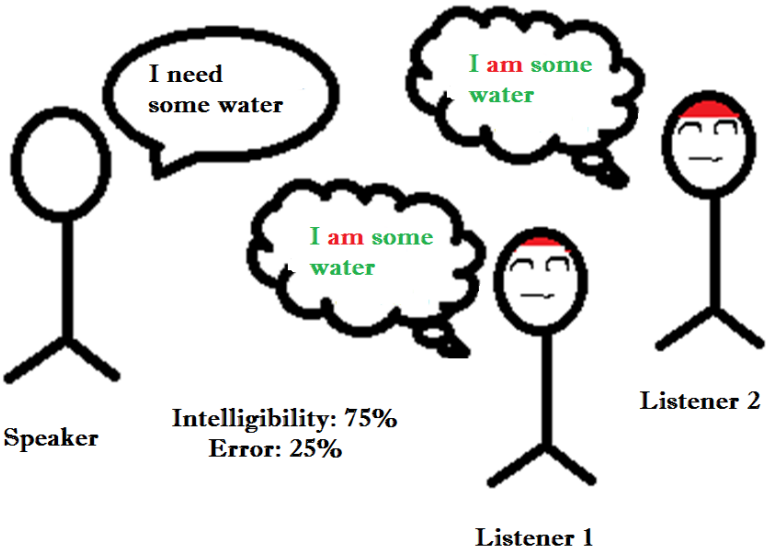


Figure 2.4: An infographic explaining the distinction of intelligibility and LE. The level of redness in the head represents the level of LE.

various methods of measuring LE is presented by McGarrigle et. al [80].

LE has been measured in several contexts where the listeners have to put in extra investment of their neurocognitive resources. This includes understanding of distorted speech signals. Distortion can come from the speaker (e.g., foreign accent, disordered speech), the listener (e.g., hearing impairment) or from the environment or channel (e.g., noise). While there is research on LE in the context of speech in noise [115], non-native or accented speech [10, 149], and hearing impairment [46], LE of disordered speech is a less researched field, and physiological LE measurements, even less so.

Self-reported Listening Effort

Studies on LE often involve subjective ratings where participants are asked to indicate the perceived effort, for example, using Likert scales [112] or visual analogue scales [79, 94].

Self-reporting measures or subjective ratings are based on a set of questionnaires such as 'Do you have to concentrate very much when listening to someone or something? '; 'Can you easily ignore other sounds when trying to listen to something? '; and 'Do you have to put in a lot of effort to hear what is being said in conversation with others?'. Based on the response of participants (say from 0 to 10), the LE can be calculated. While this method is easy and does not require much expertise, there are limitations as the responses are subjective and the threshold for effortful listening can be different for different individuals. For example, what

is effortful for a subject may not be as effortful as for another subject as they may have less tolerance to effort or may base their effort rating on the extent to which they could complete the task. [80]

In the experiments of this thesis, two kinds of self-reported LE are used. The first one is a very simple 5- point Likert scale response to the question 'How effortful was the speech to listen to?'. This method was used in the first preliminary experiment (See Section 4.2). The second method used in this thesis is the 14-point scale known as the Adaptive Categorical Listening Effort Scaling (ACALES) procedure [65]. This scale is used widely in speech-in-noise research [115, 41]. As this scale is designed for speech-in-noise experiments, the top most level is the 'only noise' level. This is the case where the participant hears only noise and no speech. As we are not studying the effect of environmental noise while listening to OS, but the effort associated due to the disordered speech, we have not used this top level. Therefore our modified ACALES scale to measure self-reported LE is a 13-point scale that goes from 'Ningún esfuerzo' (No effort) to 'Muchísimo esfuerzo' (Extreme effort). There are 7 labels in all, but also intermediate labels ('–') that allows participants to choose in-between options. The LE rating labels, their English translations, and their values are presented in Table 2.1. This rating scale was used in the experiments in Chapter 4, 5 and 7.

Table 2.1: LE rating labels, their English translations, and the values assigned.

LE Rating Labels	**English Translations**	**Values Assigned**
Muchísimo esfuerzo	Extreme effort	13
–	–	12
Mucho esfuerzo	A lot of effort	11
–	–	10
Esfuerzo considerable	Considerable effort	9
–	–	8
Esfuerzo moderado	Moderate effort	7
–	–	6
Poco esfuerzo	Little effort	5
–	–	4
Muy poco esfuerzo	Very little effort	3
–	–	2
Ningún esfuerzo	No effort	1

Listening Effort from Behavioural Data

In the case of behavioural measures, the effort is measured by behavioural tasks, which can include single tasks or multi-tasking. In single tasks, the participant is given a listening task and their responses are recorded. In addition to the responses, the response times may also indicate LE as they are known to be slower for challenging listening conditions. In a dual task

paradigm, the participant is overloaded with a task, giving limited resources for the other task. In this way the effort taken by the task in 'limited resources' condition can be measured. [80]

Listening Effort from Physiological Activity

Recent studies have explored the possibility of using physiological measurements that indicate cognitive load. As opposed to subjective ratings which only give a single data point for an entire stimulus, a physiological marker can provide a more online and continuous measure of LE. McGarrigle et al. [80] presented the results of the some physiological LE measures. It was reported that there is scope for more exploration of the relationship between self-report measures and physiological LE measures.

Some of these physiological measures that have been used to measure LE are functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET), electroencephalography, pupillometry, heart rate variability, skin conductance, monitoring hormone (e.g. adrenaline) levels etc. [2]

All of the the aforementioned physiological methods have some disadvantages. For example, sensitive and consistent methods such as fMRI and PET are too intrusive and cumbersome. Measures like heart rate variability cannot detect instantaneous fluctuations in load whereas detecting hormone levels is a slow process and unsuitable for online load detection. Pupillometry and electroencephalography have relatively lesser disadvantages and therefore, are more popular. [2]

Pupillometric measures (changes in pupil size, peak pupil dilation) have been used to show changes as a function of listening task demand in clinical and non-clinical populations [80]. A pupillometry study from 2015 by [160], found that the effort required to process degraded speech was more than that for non-degraded speech even if the intelligibility was 100 percent. It was reported that 'Pupillary responses were a sensitive and highly granular measurement to reveal changes in LE. In a review article, pupillometry was shown to be a promising tool for measuring LE amongst children in acoustically inadequate conditions [40]. Another recent study revealed that pupil size was sensitive to accuracy and engaged effort [14].

Like pupillometry, brain activity is also known to be linked to LE. Research shows that measuring brain activity helps us better understand the underlying neural processes linked to effortful listening as well as subjective LE ratings. Some studies involved fMRI [158] while others involved electroencephalography [142, 147].

An electroencephalogram (EEG) is the recording of electrical signals emanating from the brain's outer layer, or the cerebral cortex. These EEG signals reflect individual thoughts emo-

tions and behaviour thereby providing a "window on the mind". EEG recording may be intra-cranial (invasive, inside the skull) or may be recorded on the scalp (non-invasive). Some medical studies (such as epilepsy) or some brain studies would require an intra-cranial EEG, which can provide a more detailed and accurate estimates of the brain activity. This scale of detail and accuracy is not possible through scalp recordings. However, scalp recordings can still provide important information for cognitive and behavioural scientific studies. Thus, a non-invasive EEG recorded on the scalp is a great tool to understand some cognitive processes [97].

Out of the several neural markers that are studied, a marker called alpha oscillations has been studied with particular focus [135, 161, 98, 82]. While these aforementioned studies observed higher alpha power oscillation for a more difficult listening condition, there have also been evidences for the opposite effect [43, 48]. Wisniewski et. al [162] discuss in their study that among other factors, these diverging evidences are due to the differences in the kind of stimuli. It can be said that alpha oscillations as a measure of LE has not been explored for disordered speech, and especially for OS. Therefore an explorations in this direction is warranted to understand the role of alpha oscillations in OS speech perception and LE.

The disadvantages of pupillometry are that they are not suitable for tasks that involve continuous reading and the sensitivity of pupillary responses diminish with age. On the other hand, the disadvantages of EEG are that it needs a large number of trials to be reliable and is prone to artefacts such as blinks, motion and electrical noise. However the advantages of EEG are that they have good temporal resolution, better ecological validity, are non-invasive, low cost and can be recorded without extensive medical expertise. The good temporal resolution in EEG means that it is possible to monitor the brain activity to the millisecond level, making it a very useful tool. [2]

For the above reasons, and due to the availability of required resources for EEG from our research partners, EEG-based LE became the method of choice to collect physiological measures of LE. See Chapter 5 to read about the experiments on EEG based LE.

Appendix B contains the detailed procedure of EEG data processing: from raw EEG data to extraction of features associated with LE.

Listening Effort of Disordered Speech

Amongst disordered speech, we found some LE research for dysarthric speech. Whitehill and Wong [157] conducted a listening experiment to collect LE ratings, and found segmental features and factors of voice quality to be the predictors of LE ratings. The authors of [68] found that severity of speech impairment and listener familiarity had an effect on 'ease of listening'. In

[15], words spoken by children with dysarthria were presented and listeners performed a word recognition and subjective LE rating task. The response time (time taken to recognise the words) was a significant predictor of the subjective LE ratings. Words with high accuracy in recognition had shorter response times and lower LE ratings.

We found some research that looked into some subjective processing load based measures of OS. In [8] the authors measured the acceptability of OS, ELS, and HS. They found that HS was the most acceptable, followed by superior OS and then ELS. In [95], high-intelligibility TOS was played to listeners and they were asked to rate the effort of listening as well as acceptability for each sample and found an inverse correlation between LE and acceptability. Another observation from this study was that even highly intelligible speech can have varying listener effort.

Therefore, by measuring LE for OS, we hope to expand the literature on processing load-based measurements of disordered speech as well as gain an understanding of OS on a dimension beyond intelligibility.

A previous thesis [123] has measured jitter, shimmer, speaking rate and such other assessments for the OS speakers used in this database. Jitter refers to variations in fundamental frequency and shimmer refers to variations in the amplitude of the signal.

Some metrics such as Voice Handicap Index (VHI) [54] and Voice-related Quality of Life (V-RQOL) [49] are questionnaires to assess the quality of the voice and the life of people with speech pathologies. These methods have been used for alaryngeal speech too [118, 143].

2.3 Enriching Oesophageal Speech

Modern speech technologies and machine learning have great potential for use in the healthcare sector, be it for improvement of healthcare services [69] or to aid patients with speech disorders [44]. One such application is transforming disordered 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 disorders, including OS.

2.3.1 Voice Conversion

As OS lacks some important HS characteristics such as a normal fundamental frequency, prosody and rhythm, one way to enrich OS is to provide the characteristics of HS to OS using VC technology

VC is a process where the voice properties of a speaker are converted into those of another

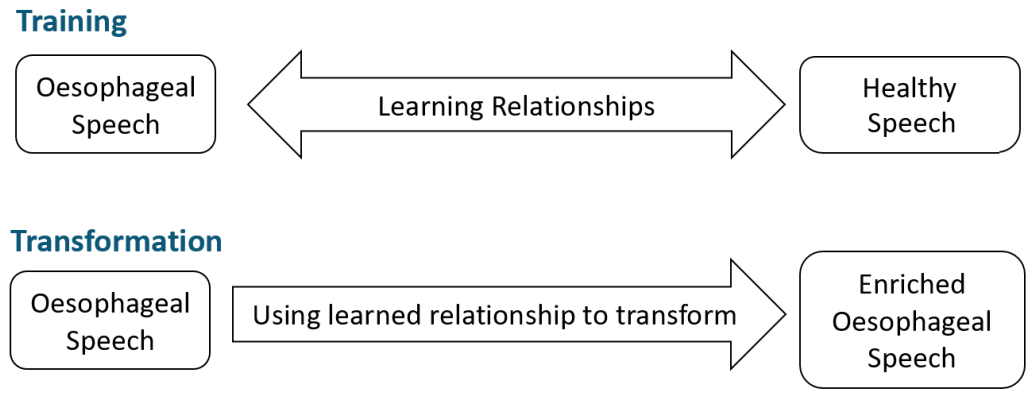


Figure 2.5: A simplified description of the voice conversion process

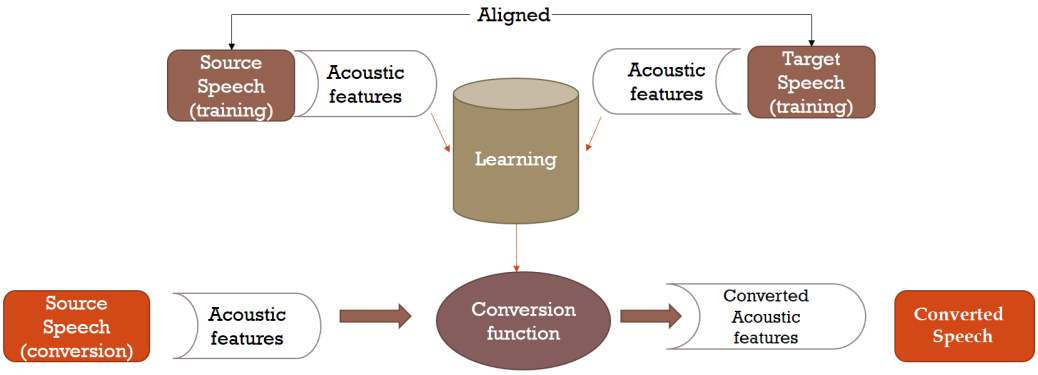


Figure 2.6: Basic building blocks of a voice conversion system

speaker [89]. Figure 2.5 shows how a VC system works in a simplified manner. Figure 2.6 is a more elaborate depiction of a VC process. It represents a parallel VC process in which we have matching spoken utterances of OS and HS (same sentences spoken by OS and HS). We employ a machine learning algorithm to learn the relationship of OS to HS. This learned relationship is then used to transform OS to HS.

Some OS enrichment has been done using such statistical VC methods such as Gaussian Mixture Models (GMM) [24, 23, 22, 123]. 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. The outputs of these systems were evaluated by MCD and MOS and showed improvement in these metrics.

In recent times, DNNs are more popular and effective compared to GMM based methods.

Othmane et.al [102] used DNN-based VC to enhance OS. This system had better intelligibility and naturalness (calculated as MOS) when compared to the GMM-based method. An additional effort (time dilation algorithm) to remove unpleasant OS sounds was found to boost intelligibility and naturalness even more compared to the DNN only method.

Dinh et. al [21] used a DNN-based intonation conversion and a GAN-based cepstrum conversion. CMOS Evaluations were performed for 3 TOS and 1 ELS speakers. Their system improved naturalness for all speakers and improved intelligibility for one of the 3 TOS speakers. Another attempt to enrich OS was by using the eigenvoices concept [25], which was inspired by the eigenfaces concept [146]. This method is based on the idea that there are speaker independent features and speaker dependent features. A source speech is converted to a desired target speech by adding speaker dependent "ingredients" in the right recipe (obtained during training) to the principal speaker independent ingredient(s). One advantage of this method is that VC is possible with a small set of source speakers samples (around 50 phonetically balanced parallel sentences). This is beneficial for OS speakers as they can provide only short amounts of recorded samples owing to their physical limitations and discomfort. For a more detailed and technical description of this method, see Toda et. al [145] and Doi et. al [25] as applied for OS VC.

A Long Short-Term Memory (LSTM) based transformation [126] developed in our lab had better ASR scores than OS. Another method described in [124] was inspired by a Phonetic Posteriorgrams (PPG) based system [138] which had good results for HS-HS VC. When applied for OS-HS VC, there was no improvement in ASR. MCD was reduced by both systems. Unprocessed OS was preferred over either of the systems in preference tests. Both these systems were developed using the OS database used in this thesis too. This database is described in detail in Chapter 3. These systems are also the chosen baseline systems in Chapter 7 where we evaluate the enrichments performed as part of the current thesis.

2.3.2 Other Enrichment Methods

Not all OS enhancement uses VC. Some studies have used a more signal processing based approach. One such approach involved vocal tract transformations with a Kalman filter (a popular speech enhancement algorithm) with the aim of making OS less "noisily corrupted" [38]. Another similar approach involved enhancement of the Spanish 'a' phoneme [101]. Both these methods resulted in improved HNR for enhanced OS compared to unprocessed OS.

Based on the assumption that a laryngectomy does not alter the articulation-based features of speech (mouth movements, position of the tongue etc.), Matsui et. al [78] performed an

enhancement of vocal tract features by formant synthesis by inverse filtering the voice source. Subjective evaluations revealed preference for the enhanced OS compared to unprocessed OS.

The idea of all these approaches focus on removing undesired and abnormal sounds present in OS, removing noise and correcting its abnormal vocal tract characteristics. These signal processing based approaches require a thorough understanding of the nature of OS signals. This is in contrast to VC approaches, where the focus is on mapping of OS characteristics to HS without delving deep into the signal characteristics of OS.

2.4 Conclusions

This chapter described all the necessary background information needed to understand the contents of this thesis hereafter. This included the anatomy and challenges of laryngectomees, the several speech evaluation metrics that are used in literature currently and previously and the methods used to enrich OS so far.

One main shortcoming of the disordered speech research is the lack of exploration of newer machine learning methods for its enrichment. In addition, as it is difficult to perform phonetic alignment, pitch irregularities etc. on OS, newer methods to preprocess OS data need to be explored to facilitate the enrichment process. These concerns are investigated in this thesis.

As observed in the literature review, OS enhancement research usually revolves around some selected few evaluation metrics such as MOS and MCD (See section 2.3.1). On the other hand, general speech perception research has several standard and established metrics and new metrics (such as LE) are still being investigated. However, these metrics are applied mostly to commonly occurring problems such as speech-in-noise and hearing impairment and not as much for disordered speech, let alone OS.

The studies mentioned in Section 2.2.1 focus on the micro-level of words and vowels. Sentence level intelligibility and Listening Effort (LE) studies on the intelligibility of Spanish oesophageal voice are a less traversed area of investigation. In addition, their focus is usually on one aspect, either ASR or HSR. Therefore, in this thesis, we used sentences as our stimuli and a combined approach of ASR and HSR intelligibility measurements.

In the following chapters, I will describe the database that is used to perform the experiments of this thesis followed by the experiments in the later chapters.

Chapter 3

Corpus and Stimuli

> "*Words are the coins making up the currency of sentences, and there are always too many small coins.*"
>
> — Jules Renard

My first introduction to this doctoral research was listening to recorded speech from several OS speakers of varying severities. As a first time listener of OS, I was both taken aback and intrigued. The following few months were spent familiarising with a database of OS speakers that we recorded in our lab. This chapter describes the database in detail and the insights gained from it which facilitated the design of the enrichment and evaluation experiments.

3.1 OS Database - Already Available

The OS database contained sentences, words and sustained vowels recordings from over 30 OS speakers. The speakers are 32 laryngectomised patients who are members of 'Association of Laryngectomees in Bilbao'. There were two categories of speakers in the database: proficient and non-proficient speakers. Proficient speakers are those who finished their speech therapy sessions months before the recording session. Non-proficient speakers on the other hand were still undergoing the speech therapy sessions. Out of these 32 speakers, 26 of them were proficient speakers, 2 of them were non-proficient speakers, 2 speakers were recorded when they were proficient as well as non-proficient and 1 other speaker was recorded in proficient OS and TOS modes. In total there were 32 speakers and 34 sets of recordings. In this thesis, we have only used the proficient speakers.

The data was recorded using 4 different microphones - A studio microphone (Neumann TLM 103), an instrumentation microphone (Behringer ECM8000), a headphone microphone (DPA 4066-F), a condenser microphone (AKG C542BL) in an acoustically isolated room. However

the recordings used in the studies of this thesis were from the studio microphone. The protocols of recordings and the detailed description of the database is available in [39] and [123]. Some important details of the database that are relevant to this thesis are described below.

3.1.1 Sentences

The database contained recordings of 100 phonetically-balanced sentences selected from a bigger corpus [121, 33]. The selection of sentences was performed with a greedy-algorithm-based tool called corpusCRT [128], with the criteria of maximised diphone coverage and a maximum of 15 words per sentence.

These 100 sentences were chosen to be phonetically balanced to ensure maximum phonetic content variability. They were syntactically and semantically predictable, but had some proper nouns and many unusual words that are hard to guess. This is to be kept in mind while considering intelligibility measurements. The difficulty of these sentences make them appropriate for intelligibility and LE experiments. There are short and long sentences in the database which allows the experimenter to choose stimuli as per the need of the experiment. Some examples of the sentences are the following: '¿Qué diferencia hay entre el caucho y la hevea?' *What is the difference between rubber and hevea?*, 'Unos días de euforia y meses de atonía.' *A few days of euphoria and months of despair.* Words such as hevea (specific name of rubber tree) and atonía (despair) are not commonly used words and hence are difficult to guess.

3.1.2 Sustained Vowels

Each OS speaker recorded 4 instances of the sustained articulation of all five Spanish vowels. These data were not used in the experiments described in this thesis. However, they were great tools to understand the voice characteristics of OS.

3.1.3 Words

The database contained 14 isolated words which included four words containing diphthongs. These words are useful for spoken term detection tasks. These words were not used in any of the experiments either.

3.2 HS Database Description - Already Available

HS samples were obtained from an online platform [32] and hence, were recorded in variable environments. However, some of them were recorded in the aforementioned acoustically isolated

room, although with a different microphone. The number of speakers in the HS database keeps growing as it is an open online platform [1]. There were over 35 speakers in the HS database at the time of the initial experiments. Some of the speakers in this database were used in the experiments performed in this thesis.

3.3 Additional OS Data - Newly Recorded

3.3.1 Words

We recorded 150 low frequency difficult words from one OS speaker and one HS speaker. The words were chosen using an online database [2]. These words were used in an experiment in Chapter 7. See the list of words in Appendix D.

3.3.2 Continuous Speech

In addition to the 150 words, five continuous speech passages were recorded for the same one OS and one HS speaker. The passages were chosen from the Cervantes resources for reading for intermediate level Spanish [3]. This was meant to be used in the experiment described in Chapter 7, but eventually was not used. However, these passages are useful for future OS studies with continuous speech.

For one of the passages, a video recording was performed. This passage with the video recording (duration: 2 minutes and 17 seconds) was used to build an interactive demonstration of the outcomes of OS enrichment. More details on this demonstration will be presented in Chapter 6. See the contents of the passage in Appendix C.

3.4 Manual Labelling

In order to create phonetic labels of the speech signals in the database, an automatic forced aligner which is part of an ASR system was used. However, using a forced aligner which was trained using HS was unfit for OS. On manual inspection, several errors were encountered in the labels likely owing to poor spectral and temporal characteristics of OS. Therefore, new acoustic models were made using the available OS recordings with the Montreal Forced Alignment tool [72].

[1] `https://aholab.ehu.eus/ahomytts/`
[2] https://www.bcbl.eu/databases/espal/idxword.php
[3] `https://cvc.cervantes.es/aula/lecturas/intermedio/`

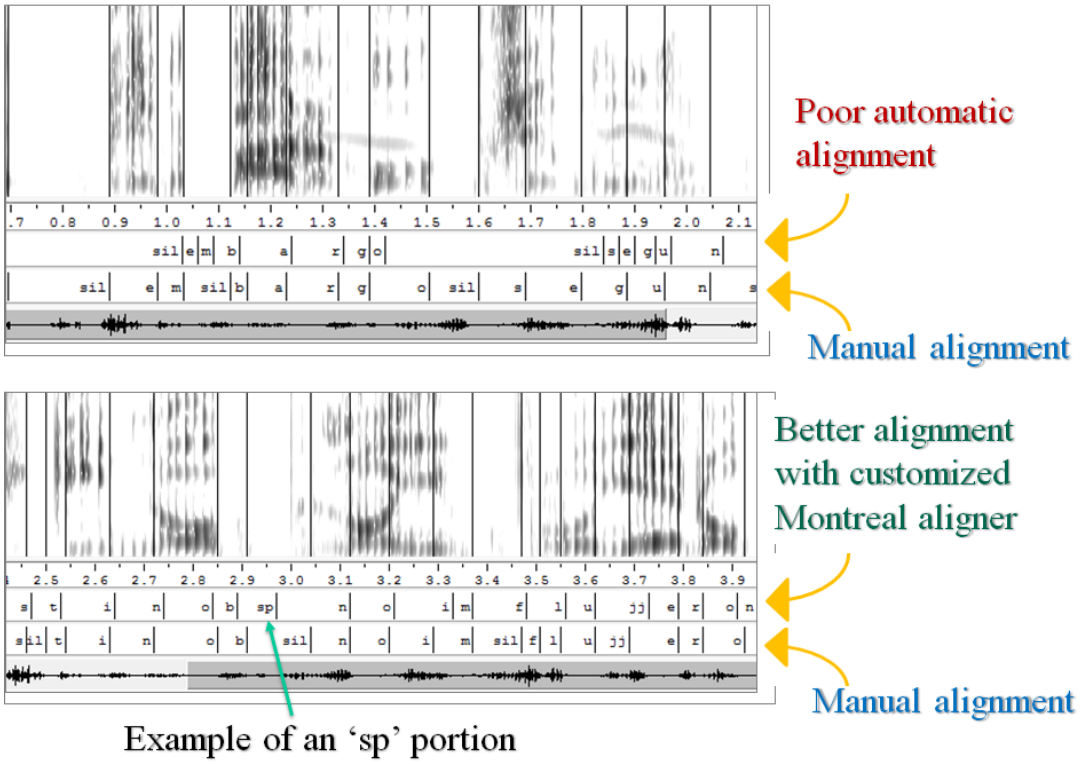


Figure 3.1: Comparison of the automatic labelling, manual labelling and customised automatic labelling of OS

Manual labelling was performed for one speaker (Speaker ID:05M3) so that these manual labels can be used to evaluate the accuracy of the automatic aligner. It was not possible to perform manual labelling for all 32 speakers as it is a time consuming process. The process involved listening to each and every phoneme in the speech samples and correcting the positions of the respective phone boundary. This was done using the wavsurfer software [131].

We assessed the accuracy of the automatic labelling process by comparing the label boundary positions of the automatic labelling process to that of the manual process. Results showed that 97 percent of the errors were less than 50 ms and 83 percent of the errors were less than 5 ms [39]. Figure 3.1 shows how the the default automatic labelling and the customised automatic labelling compare to the manual alignment.

These new labels were used in training an ASR system adapted to OS [39] as well as for generating Synthetic Speech (SS) with durations matching with OS (See Section 6.2).

The duration matched SS was created for all the 32 speakers. This parallel SS database can be useful for future OS enhancement research.

3.5 Intelligibility

The intelligibility of 29 proficient speakers was calculated by calculating ASR WER scores. Two different types of ASRs were used: ASR trained using HS and ASR trained with OS. Figure 3.2 shows the results of the ASR performance. It can be observed that the system trained with OS data has fewer errors compared to the system trained with HS data. There was an improvement of around 16 percentage points when trained with OS data.

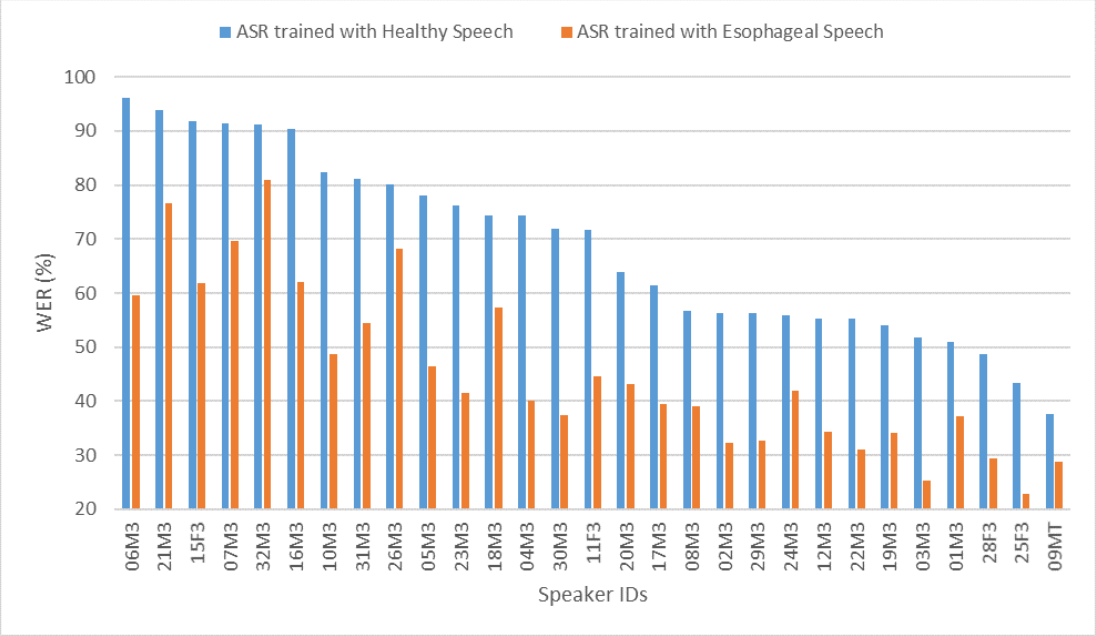


Figure 3.2: **ASR Results.** Mean speaker-wise Word Error Rates (in %) for ASR trained with HS and ASR trained with OS

3.6 Conclusions

This chapter described all the data that was available to us for performing our experiments. However, not all of the data was used in the experiments described in this thesis. The experiments listed in Chapters 4, 5 6 and 7 describe in detail the chosen subsets of the database in their respective methods sections.

The manual and the automatic labelling process described in this chapter form an important step in the enrichment strategies described in Chapter 6.

Additionally, intelligibility of the speakers obtained from an ASR system was described. The system trained with OS data had fewer errors compared to those with HS. These WER scores are used as a criteria when selecting speakers to evaluate in the aforementioned experiments.

A complete and extensive description of the database was published as a paper titled 'A

Spanish Multispeaker Database of Esophageal Speech' [39].

Chapter 4

Preliminary Measures of Intelligibility and Listening Effort

"If you think communication is all talking, you haven't been listening."

—Ashleigh Brilliant

In the previous chapter, I described some key features of the OS database. I listed some differences in the acoustic and linguistic characteristics of OS and HS. In Section 2.1.2 we have seen that OS has a lower speaking rate, higher jitter and shimmer and lower intensity compared to HS. All these are factors that affect the ability to perceive and understand speech. To understand in what ways OS is more difficult to process compared to HS, and how much, it is necessary to conduct well designed listening experiments.

In this chapter, I describe two listening experiments that collected some preliminary intelligibility and LE metrics for a small set of speakers from the database and some control healthy speakers. Both the experiments were designed to evaluate the gaps between OS and HS in intelligibility and ease of processing. Once these gaps are known, appropriate enrichment methods can be designed aimed at closing these gaps i.e. bringing enriched OS closer to the metrics of HS than unprocessed OS.

The contents of this chapter have featured in previously published paper titled 'Intelligibility and Listening Effort of Spanish Oesophageal Speech' [112].

4.1 Introduction

Listening to disordered speech is a challenging task, and it demands a lot of attention and effort. To quantify the challenges of listening to OS we begin by measuring its intelligibility in comparison to HS. Intelligibility measurements are common and are a useful way to quantify what percentage of the spoken message has been correctly understood (See Section 2.2.1 for more details). In this study, we have evaluated intelligibility in human–human (speaker is human, listener is a human) as well as human–machine (speaker is human, listener is a device/software) interactions.

In addition to intelligibility, there is growing interest in research measuring LE and other processing load aspects of speech as it gives an additional dimension for understanding challenges in speech perception in adverse listening conditions. The motivations for measuring LE is described in detail in Section 2.2.2. In this study, we have attempted to explore LE in addition to the intelligibility measurements.

In chapter 2, we presented some research were HS was found to be more intelligible than OS (Section 2.2.1) and HS was found to be more acceptable compared to OS (Section 2.2.2). These studies are the foundations for the hypotheses of this experiment. Significant positive correlations were observed in ASR and HSR for listening in adverse conditions such as age related hearing loss [37] and speech disorders [53]. This led us to hypothesise that like HSR perfomance, ASR performance will be lower for OS compared to HS.

The idea of intelligibility differences between experienced and inexperienced listeners of OS was explored in [16]. The findings were that OS was ranked similarly for intelligibility by both experienced and inexperienced listeners. This was intriguing, and led us to investigate the effect of familiarity with OS on its intelligibility. In addition, as we were collecting LE ratings too, we were interested in seeing if the same was observed for LE ratings, or if they would tell a different story. We consider friends, family (spouse, siblings, children), and caretakers of OS speakers as familiar listeners.

This study contains two experiments. The first experiment was web-based, and was focused on getting preliminary intelligibility and LE metrics for our data. We investigated how intelligibility (both ASR and HSR) and LE differ for the two speech types (OS and HS). We also investigated the effect of familiarity and to what extent intelligibility and LE are correlated. The second experiment (an extension of Experiment 1) was conducted in a laboratory setting, which allowed us better control of the experiment environment. The aim of this experiment was to find out if more LE is reported for OS even if the intelligibility of OS is close to that of HS. Additionally, in this experiment, we also investigated if the participants' performance in

the speech perception tasks depended on their cognitive abilities.

The hypotheses for Experiment 1 are:

- WER is positively correlated with self-reported LE ratings.
- HS is more intelligible and less effortful, compared to OS.
- Listeners familiar with OS find it less effortful to process OS, compared to listeners that are not.
- ASR performs worse for OS than for HS.

Our hypotheses of Experiment 2 are:

- For the case that intelligibility of OS is similar to that of HS, there is still more effort in understanding OS.
- Listeners with better cognitive abilities have better intelligibility scores and report lesser effort.

We begin by describing the materials and methods and the results of Experiment 1 in Section 4.2, followed by those of Experiment 2 in Section 4.3. Finally, a general discussion and conclusions are presented.

4.2 Experiment 1: Preliminary Word Error Rate and Listening Effort Measurements

4.2.1 Materials and Methods

Experimental Design

The main task for this web-based experiment was the sentence recall and transcription task. Participants listened to a sentence and then typed what they had understood. To collect LE rating measures, we asked the participants to rate the sentences for LE on a 5-point Likert scale. The sentences were played only once (to avoid any possible memory effect) and in a random order (to avoid sentence order bias).

Corpus and Stimuli

The corpus we used was a part of the larger database of 32 OS speakers described in detail in Section 3.1. The chosen stimuli were picked from the 100 sentences described in Section 3.1.1.

A subset of 6 speakers (OS speaker IDs: OM1, OM2, OM3, OF1 ; HS speaker IDs: HM1, HF1)[1] and 30 sentences were used as our stimuli (Sentence list provided in Appendix A). The 'M' and the 'F' in the speaker IDs refer to male speakers and female speakers respectively. The criteria and the procedure for these choices are described next.

OS speakers to be evaluated were chosen based on two criteria—proficiency and accessibility. Speakers who practised for at least two years after the laryngectomy qualified as proficient speakers. Additionally, an OS voice quality assessment tool [144] was used as a guide to assess proficiency. This tool was based on the factors used in the A4S scale of [26] such as speaking rate, regularity, etc. We also considered accessibility of speakers as a factor, because their willingness and availability to come for follow-up recordings are useful for future research. Based on these criteria, we chose 4 speakers, three male and one female, making it gender-inclusive (there are only 4 women in the whole database and only 2 of them fulfilled our criteria). The criteria for choosing healthy speakers was quality of recording as well as gender balance. One male and one female healthy speaker were chosen.

We conducted a pilot listening test to check if our stimuli were suitable for the sentence transcription task. We played the stimuli to some pilot participants who were not familiar with the stimuli, and hence, not primed. They reported that some sentences were effortful to transcribe as they were too long to remember. Therefore, we decided to use a subset of shorter sentences (maximum 40 phonemes), extracted using the aforementioned CorpusCRT tool. The result was a phonetically balanced set of 30 sentences with 7 to 10 words in each.

To sum up, we had 30 sentences from each of the 6 speakers (4 OS and 2 HS), a total of 180 stimuli. They were normalised to a common peak value (0.8) to achieve a homogeneous and comfortable level of loudness.

Listening Test

We took the following information from the participants: Age group (21–30, 31–40, etc.), presence of hearing impairment, the kind of audio equipment used ('good quality headphones', 'normal quality headphones', 'good loudspeakers', 'normal loudspeakers', and 'bad equipment') and whether the listener had close contact with laryngectomees.

We had 57 native Spanish participants (from Spain) in this test, out of which 15 had close contact with laryngectomees and hence, were familiar with OS. Out of the 57 participants, 11 listeners were from the age range of 21 - 30, 11 from 31 - 40, 10 from 41 - 50, 8 from 51 - 60 and 2 from 61 - 70. There was no hearing impairment reported for any participant.

[1] (The corresponding speaker IDs in the Ahoslabi database: 01M3 (OM1), 02M3 (OM2), 04M3 (OM3), 25F3 (OF1), 114 (HM1), 207 (HF1))

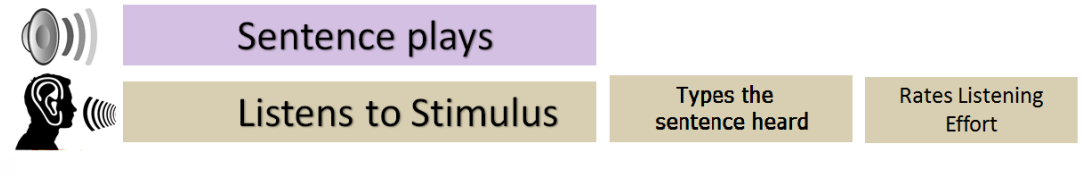


Figure 4.1: Preliminary LE and SI Task Schematic Representation

Participants listened to 5 randomly chosen stimuli from each of the 6 speakers, a total of 30 stimuli. Each of these 30 stimuli was a different sentence. Using a Latin Square design [108], coverage of all the 180 stimuli was ensured.

The participants were instructed to use headphones, pay close attention to the stimuli, and provide the responses honestly and uninhibitedly. There was a sound sample provided along with the instructions for them to ensure they could hear the sound properly. Participants typed what they heard in a text box and provided the LE rating on a drop-down menu Likert scale. The options were 'Very little','A little', 'Some', 'Quite' and 'A lot'. If they missed some portions of the sentence or were unsure of what they heard, they were asked to put three dots (...) in the text box. The first two sentences were presented as practice sentences (one HS and one OS), to familiarise the participant with the task. These sentences were sampled from the same corpus of 100 sentences but were different from the ones presented in the actual test.

The listening test (`https://aholab.ehu.eus/users/sneha/Listening\_test.php`) was web-based, and it was possible to reach out to a wide range of participants. However, this also meant differences in audio equipment, the effects of this on the responses are reported in the Results section.

Automatic Speech Recognition

We conducted automatic transcriptions with a Kaldi-based [107] Spanish ASR system. It is implemented with the s5 recipe for the Wall Street Journal database. We used 13 Mel-Frequency Cepstral Coefficients (MFCC) as acoustic features and a Cepstral Mean and Variance Normalisation (CMVN) to mitigate the effects of the channel. The details of the training procedures are described in [113].

The training material for the ASR system was HS, as described in [125]. Some modifications were made to the ASR system to adapt it to the requirements of our experiment. The sentences we used contained many low-frequency words. About 23% of the words were out of vocabulary (OOV) words in the lexicon of the original ASR system, which contained $37,632$ entries. Therefore, we created a new lexicon with the 701 words present in the 100 sentences of the original corpus. Together with this reduced lexicon, a unigram language model with equally probable

words was used. The acoustic models were unchanged. A unigram language model is a very simple language model and only considers probabilities of single words. A more sophisticated language model would consider the probabilities of single words and also combinations of words (word pairs, word triplets, etc.), which can result in better ASR performance.

Although the final WER numbers obtained with this simple ASR are not comparable to a sophisticated ASR system, the procedure serves our purpose of evaluating the intelligibility, comparing the performance of healthy and oesophageal speakers, and establishing a baseline reference for other parallel research in the field (such as evaluating the improvements of speech modification algorithms [127]).

4.2.2 Analysis and Results

WER was calculated using the Equation 2.1 defined in Section 2.2.1. Prior to calculating WER, an initial clean-up was performed on the data. This included removing any punctuation or special characters and some typing errors (accented vowels, use of upper and lower case, spelling of proper or foreign names, etc.). The WER was obtained after correcting these transcription errors.

For HSR, two different WERs were calculated: One, where all words of the sentence were considered; and two, where only content words were considered. In this case, errors in function words like prepositions, pronouns, conjunctions, etc., were overlooked. This helped us see how much of the content of the conversation was correctly understood. We contrasted these content-words-only WERs with the all-inclusive WERs. This was not done for ASR as when the ASR system transcribes a word, it is present in the dictionary and hence, transcription errors do not occur.

We performed a 2×2 repeated measures ANOVA with the within-subjects factors speaker type (OS vs. HS) and between-subjects factor familiarity (familiar vs. not familiar) to quantify the effects of speaker type and familiarity on WER and LE ratings using the JASP tool [56]. We ran an additional ANOVA to compare the effect of audio device on WER and LE ratings. Sphericity and homogeneity checks were performed on the data with the JASP tool to ensure that assumptions of an ANOVA test are met.

Finally, we present speakerwise average ASR WERs of the 30 sentences and compare them with the HSR WERs.

Word Error Rates from HSR

Figure 4.2 shows all the WER results. The blue bars represent 'All words WER ' from 'not familiar' listeners; and the orange bars represent the same from 'familiar' listeners. Patterned blue and orange bars represent the 'content words only WERs' for the 'familiar' and 'not familiar' categories, respectively. OM, OF, HM, and HF are acronyms for Oesophageal Male, Oesophageal Female, Healthy Male, and Healthy Female, respectively.

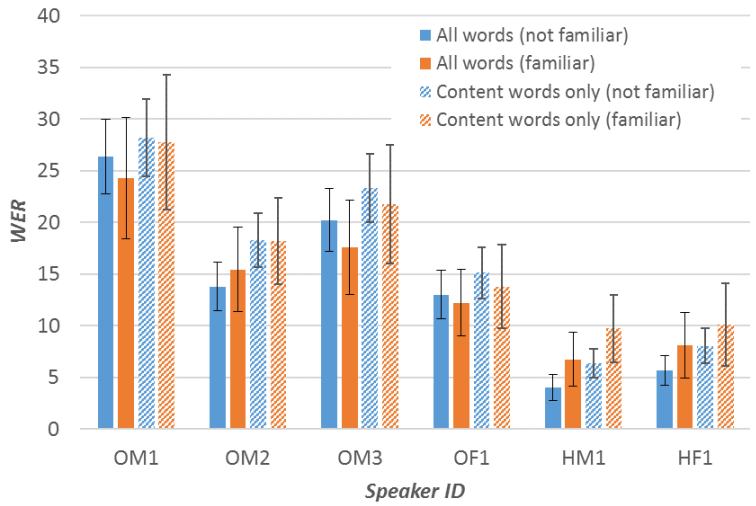


Figure 4.2: Mean speakerwise 'All words' and 'content words only' Word Error Rates (WER) for 'familiar' and 'not familiar' listeners. OM1, OM2, OM3, OF1 are oesophageal speakers; HM1 and HF1 are healthy speakers. Higher WER corresponds to lower intelligibility. Error bars show 95% confidence intervals.

As can be seen in Figure 4.2, the mean 'content words only WERs' and 'All words WERs' differ in value, but only by a small number (mean value of 2.73 percentage points). The 'content words only WERs' were highly correlated with the 'All words WERs' (Pearson's r = 0.889, $p < 0.001$). Therefore, we can see that the trend of the non-patterned bars and the patterned bars is similar.

Now, focusing on the speakerwise WERs for familiar and unfamiliar listeners, we can see that mean WER is always higher for OS compared to HS, as expected. There is no major difference in the WER for familiar and unfamiliar listeners in the case of OS. It is confirmed by a two-sample Kolmogorov–Smirnov (KS) test that the data came from the same continuous distributions (Alpha = 0.05). This result corroborates the conclusions in [16], where intelligibility was scored similarly by expert and novice listeners. For HS, there is a slight difference of around 3 points in the mean WER, but, as can be seen in Figure 4.2, the difference is not meaningful. From a two-sample KS test, we find that the data (HS-familiar and HS-nonfamiliar) came from the same continuous distributions (Alpha = 0.05). Mean WERs and LE values can be found in

Table 4.1.

Table 4.1: Mean Word Error Rate (WER) and Listening Effort (LE) for OS and HS for familiar and not familiar listeners.

		OS	HS
WER (in %)	Familiar	17.39	7.42
	Not familiar	18.35	4.85
	Total mean WER	17.87	6.16
LE	Familiar	2.61	1.25
	Not familiar	3.54	1.26
	Total mean LE	3.07	1.255

The ANOVA results show that familiarity with OS had no effect on WER ($F(1,55) = 0.007$, $p = 0.934$). On the other hand, speaker-type had a strong effect on WER ($F(1,55) = 223.593$, $p < 0.001$, $\eta^2 = 0.788$), with higher WER for OS compares to that of HS. These results can also be observed in Figure 4.2.

The audio device used by the listener had no effect on HSR WER ($F(3,1256) = 0.707$, $p = 0.548$).

Self-reported Listening Effort

Mean LE ratings are stated in Table 4.1. Figure 4.3 shows the speakerwise LE ratings. As expected, it is higher for OS compared to HS. However, when listening to OS, the LE is significantly lower for familiar listeners than for not familiar listeners. Indeed, ANOVA analysis shows that familiarity with OS has an effect on LE ($F(1,55) = 20.22$, $p < 0.001$, $\eta^2 = 0.269$) and Speaker-type has a strong effect on LE ($F(1,55) = 315.00$, $p < 0.001$, $\eta^2 = 0.808$).

The audio device used by the listener had no effect on LE ($F(3,1256) = 0.705$, $p = 0.549$).

Correlation of Intelligibility and Listening Effort

Correlation between intelligibility (WER) and LE ratings was 0.479 (Spearman's rho = 0.475, $p < 0.001$). This is a significant correlation, indicating that sentences with more transcription errors are perceived as more effortful. Spearman's rho correlation was used as LE rating is an ordinal variable.

Word Error Rates from ASR

The mean ASR score for OS was 49.55 ± 3.39 and for HS it was 19.57 ± 1.50. Mean speakerwise ASRs are shown with corresponding HSRs in Figure 4.4. It can be observed from the figure that the ASR performs poorly for both HS and OS. The fact that the system used a unigram

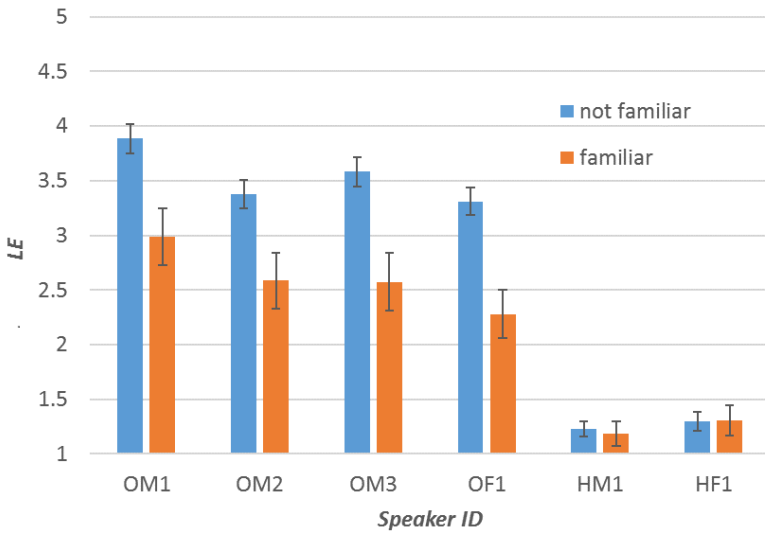


Figure 4.3: Mean speakerwise LE for oesophageal (OM1, OM2, OM3, OF1) and healthy (HM1, HF1) speakers. On the y-axis, 1 corresponds to least effortful and 5 to most effortful. Error bars show 95% confidence intervals.

language model contributes greatly to this poor performance. As expected, WER for OS is significantly higher ($t(4) = 11.42, p < 0.001$) than for HS.

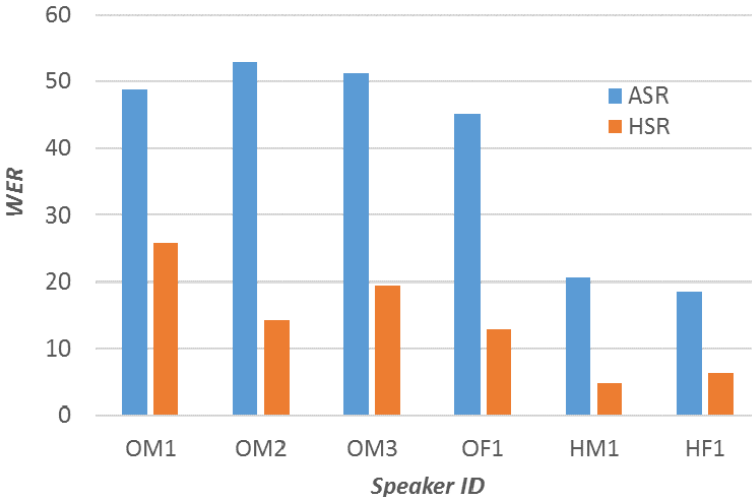


Figure 4.4: Word Error Rates (WER) for Human Speech Recognition (HSR) and Automatic Speech Recognition (ASR) for oesophageal (OM1, OM2, OM3, OF1) and healthy (HM1, HF1) speakers

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

4.3 Experiment 2: Listening Effort for Highly Intelligible Oesophageal Speech

4.3.1 Materials and Methods

Experimental Design

Based on our preliminary intelligibility and self-reported LE experiment (See Experiment 1, Section 4.2), we had the chance to probe further into our data by designing a study specifically aimed at measuring LE. The aim was to investigate the differences in LE for a set of HS and OS speakers that have comparable intelligibility. As pointed out in [95], even highly intelligible OS speech was found to have different LE ratings. Therefore, this is a methodological decision in order to rule out that observed effects are due to differences in intelligibility.

The experiment was designed for an EEG-based LE measurement. We aimed to record EEG data of participants while they listened to OS and HS to investigate if there are differences in the LE correlates of brain activity. Along with measuring the EEG data, we also collected subjective LE ratings from the listeners. As the participants had an EEG cap on, the stimuli were played on a loudspeaker, and not on headphones. Here, we present the LE ratings findings. EEG data acquisition process and findings are presented in Chapter 5.

In addition to the LE experiment, a separate intelligibility test was conducted to replicate the results of Experiment 1 in a laboratory setting. This time we asked the participants to listen to the sentence and repeat out loud what they heard. This is less taxing for the participant as they do not have to type their responses. Also, this resulted in speedier responses and hence less effort from the listeners' side in memorising the sentence. The advantage of oral response is that typing errors can be excluded as a confounding factor for WER. However, this involves post-processing, i.e., the task of transcription of their oral responses to text to calculate WER.

In order to investigate the relationship between LE, SI and the listeners' cognitive capacities, we conducted a Flanker task [31] and a backward digit span task [47] after the behavioural tasks. The Flanker task measures the selective attention ability [31] and the digit span task measures the working memory capacity of the listener [47]. Both of these processes are relevant in speech perception. OS has swallowing sounds and undesired artefacts which need to be ignored by the listener to selectively focus on the speech message. Working memory is also a crucial factor in speech perception (see phonological loop in [4]), especially for OS which spans a longer duration than HS.

Stimuli

We picked a subset of one HS speaker and one OS speaker from our dataset of Experiment 1 based on intelligibility similarity. Speaker OF1 and speaker HF1 were the two speakers that had significantly similar intelligibility based on a two-sample KS test. The null hypothesis that they come from same distributions was accepted with a significance of Alpha of 0.01.

All 100 sentences mentioned in Section 3.1 were used for this experiment. An intelligibility test was performed on the same 30 sentence subset described in Experiment 1. For the LE rating task, we used the other 70 sentences which were longer. This was to ensure that the participants had a sufficiently long stimulus to respond to, and that the EEG recording for each stimulus was sufficiently long to process and analyse. The sentences contained several low frequency words which made them sufficiently difficult for an LE task.

For the 70 LE sentences, the number of words in each sentence ranged between 9 and 18 words (mean = 13.19, SD = 3.66). The mean duration of the OS stimuli was 8.81 seconds (SD = 1.58, min = 6.00, max = 12.55) and that of HS was 5.27 seconds (SD = 1.045, min = 3.31, max = 8.28). The lengths of OS stimuli were significantly longer ($t(69)=1.66$, $p<0.001$) than HS stimuli. Average speaking rates (syllables per second) for HS and OS were 4.32 ± 1.79 and 7.36 ± 3.35 respectively.

Listening Test

Sixteen native Spanish speakers (7 female, 9 male; age range: 19–35, mean = 26.56, SD = 4.50) participated in the study. All participants were native Spanish speakers from South America, except one who was from Spain. They were given monetary compensation for participating in the test. Ethics for conducting the experiment was approved by the local ethics committee of the University of Oldenburg. All participants had normal hearing except one participant with a 55 dB hearing loss in the left ear. The inclusion of this participant did not alter the observations of the study and hence, we chose to keep this participant. The stimuli were presented with a loudspeaker placed at a 0° in front of the participant at distance of 1m at a comfortable listening level of 60 dB SPL.

The test began with the LE task first (Figure 4.5), where 60 (30 OS and 30 HS) out of the 70 available LE sentences were played in 3 blocks of 20 sentences (randomised) each. For 15 sentences of each block, participants were prompted to provide LE ratings as per the 13-point ACALES scale (See Section 2.2.2 for more details). In the other 5 sentences (presented at random intervals), they were asked to repeat the last word of the sentence, which the experimenter scored as correct or incorrect. This was to ensure that they were attentive and actively listen-

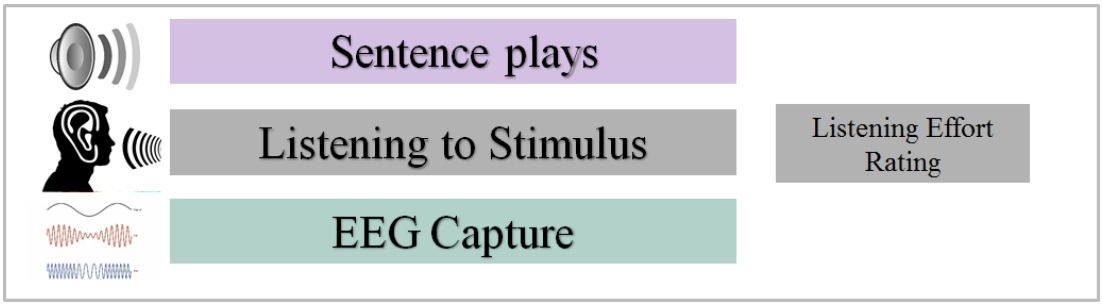


Figure 4.5: LE Task Schematic Representation

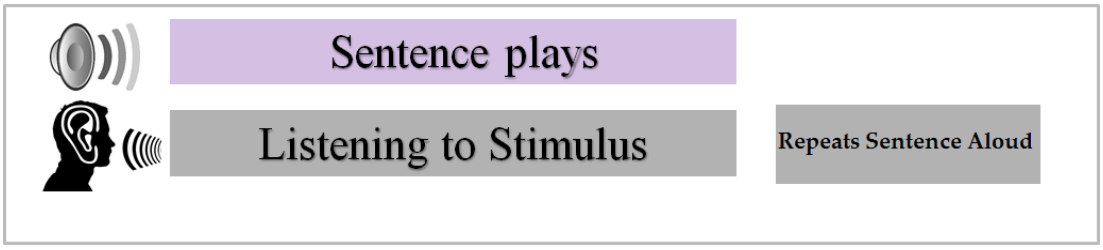


Figure 4.6: SI Task Schematic Representation

ing to the stimuli. The LE task lasted for around 20-25 minutes. The average inter-stimulus interval (time between the response and onset of the next sentence) was 1.59±0.63 seconds.

After the LE task, the participants got a break (approximately 10 minutes) and then they proceeded to the intelligibility task (Figure 4.6). In this task, they listened to a sentence and received a prompt on the screen to repeat the sentence that they heard. They provided oral responses for the 30 sentences. The whole session of the intelligibility test was recorded with a microphone so that it could be transcribed later. This task lasted around 15 minutes.

In all, we had 45 subjective LE ratings and 15 last word recognition scores from the LE task, and 30 SI scores from the SI task for each participant. EEG data was recorded for the entire duration of the LE task and therefore were available for all the 60 LE task stimuli.

Cognitive Tasks

In the Flanker task, participants were presented 24 congruent ("<<<<<"), 24 incongruent (">><<<"), and 24 neutral ("−− < −−") stimuli. They were asked to focus on the middle symbol and correctly identify it by pressing "<" or ">" on a keyboard as quickly as possible. Their response accuracy and reaction times were recorded. The better the performance (i.e. shorter reaction times on accurate responses), the better the participant's selective attention.

In the backward digit span task, the experimenter read a list of digits and the participant was asked to repeat them in reverse (For example, experimenter: '3 2 9 5'; participant: '5 9 2 3'). The digits were read in an even tone at intervals of approximately one second. The

experimenter started with the set of three-digit lists. If the participant was able to correctly recall 5 three-digit lists out of 6, they graduated to the set of four-digit lists. This went on until the participant could no longer recall at least 5 lists in a set or until they reached the final nine-digit list set. The digit span score was the maximum number of digits in the list where the participant could recall 5 lists correctly. The larger the correctly recalled digit span, the larger is the participant's working memory capacity.

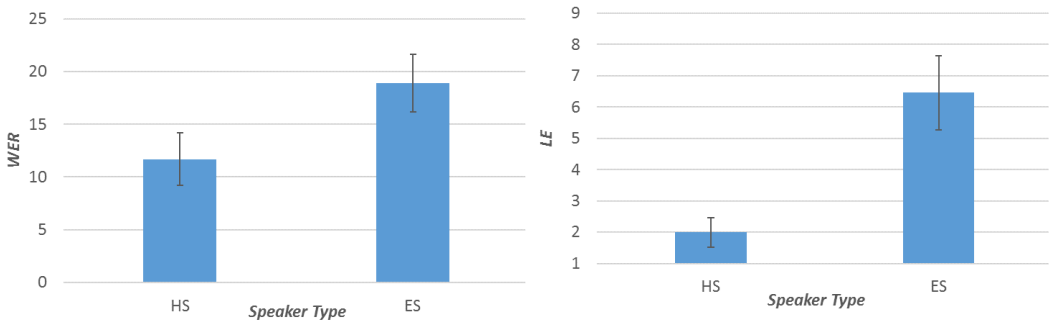
The cognitive tasks lasted 10-15 minutes. No EEG was recorded during the cognitive tests.

4.3.2 Analysis and Results

Out of the 16 participants, transcriptions were available only for 13 participants as we could not record responses of 3 participants due to technical problems. LE ratings were not available for one other participant, also due to a technical problem with saving data. Flanker effect score was not available for a participant. For analysis purposes, these missing data were filled with the mean values of the responses of other participants. Sphericity and homogeneity checks were performed on the data with the JASP tool to ensure that assumptions of an ANOVA test are met.

Intelligibility

The audio responses of the sentence recognition task were transcribed by a native Spanish speaker, who was also a speech expert. WER was calculated using the same methods as elaborated in Section 2.2.1. As the WER was found to be highly correlated with all inclusive WERs in Experiment 1, we decided to proceed with all inclusive WERs only.



(a) Word Error Rates

(b) Self-reported LE ratings

Figure 4.7: WER and LE for oesophageal (OF1) and healthy (HF1) speakers. Error bars show 95% confidence intervals.

Percentage WER scores for OS was 18.88 ± 5.57 and for the healthy speaker it was 11.69 ± 5.07 (Figure 4.7a). ANOVA showed that speaker type had an effect on WER (F(1,15)

$= 27.20$, $p < 0.001$, $\eta^2 = 0.645$).

Listening Effort Ratings

Mean LE (from a 13-point scale) for the OS speaker was 6.457 ± 3.150 and for the healthy speaker it was 1.994 ± 1.611 (Figure 4.7b). There was a difference of 6 points in median LE for OS and HS. The median LE for HS was 1 (no effort) and for OS speaker it was 7 (moderate effort). ANOVA showed that speaker type had an effect on LE ($F(1,15) = 77.55$, $p < 0.001$, $\eta^2 = 0.838$).

There were 15 responses per participant for the task of repeating the last word in the LE task. The average error made in the recognition of the last word was 1.067 response per participant. The total last word recognition error across all the 15 participants was 7 percent. We can tell, therefore, that the participants were attentive during the LE rating task.

LE, WER and Cognitive Tasks

The Flanker effect was calculated as shown in Equation 4.1. RT_incong and RT_neutral are the average reaction times taken to respond to an in-congruent trial (">><<<<") and a neutral trial ("−− < −−"), respectively. These reactions times are calculated for correct trials only.

$$Flanker\ Effect = \frac{log(RT\_incong) - log(RT\_neutral)}{log(RT\_neutral)}. \tag{4.1}$$

The mean Flanker effect score was 0.413 ± 0.019 and the mean digit span score was 4.125 ± 1.258. No significant correlations were found between digit span scores and Flanker scores indicating that they measure separate cognitive abilities.

Correlations between Flanker effect and mean LE ratings were not significant (Spearman's rho = −0.432, $p = 0.096$). Significant negative correlation (Pearson's r = −0.554, $p = 0.049$) was found between digit span scores and mean WERs (average of OS and HS).

4.4 Discussion

In Experiment 1, we were able to show that speaker type (OS or HS) had an effect on both LE and WER. OS speakers had poorer intelligibility compared to HS speakers and also a higher LE. The correlation between LE and WER suggests that more effort was reported as the intelligibility of the speaker worsened. Therefore, a drop in intelligibility caused an increase of LE. A further step in this direction would be to know what aspects of OS contribute more

to LE: Its spectral characteristics, lack of fundamental frequency, poor rhythm in speech, or a combination of these.

Our findings about the effect of familiarity with the listener on intelligibility are in a similar vein to a study investigating the experience of the listener (speech expert vs. novice) on OS intelligibility [16]. However, in this study we were more interested in investigating the experience that comes from constant exposure as family members, close friends, and caretakers. We found that indeed the intelligibility scores were similar for familiar and unfamiliar listeners. However, interestingly, familiar listeners reported less LE. So LE was able to provide additional insight about listening to OS.

As far as ASR is concerned, we found that ASR WER scores were higher for OS compared to HS. We compared WERs from our ASR system with HSR WERs. ASR WERs were higher compared to HSR WERs, but it could be because our ASR system was based on a unigram language model and focused only on acoustic models. The reason to choose such an ASR was to evaluate the drop in intelligibility owing to acoustic degradations, which is the case for OS.

In Experiment 2, the goal was to measure LE when listening to OS and HS with similar intelligibility scores taken from Experiment 1. Although WER data in Experiment 2 indicate a higher intelligibility for HS than OS, the overall intelligibility for both OS and HS can be considered to be very high. Despite this high intelligibility for both speaker types, we observed a considerable gap in LE, and this suggests that LE is a relevant dimension to be considered in OS evaluation.

The negative correlation of the digit span scores with WER suggests that participants with a poorer working memory (denoted by lower digit span scores) made more errors in recognition. This is understandable, as the ability to hold more information helps in correctly recalling and repeating the stimuli. The correlation with Flanker effect was not significant, suggesting that, in this case, selective inhibition plays a minor role to explain differences in LE. Flanker task is a measure of selective inhibition of distracting signals, such as noise added to the signals or signals with distracting speakers. However, the distractions in our stimuli are not of that nature. It is more in the form of undesired pauses and swallowing sounds that appear within and between words in the OS signal. The increase in LE was observed likely due to poorer quality of speech, rather than due to interfering information that has to be suppressed as is the case in noisy environments. On the whole, we cannot tell from these results alone whether better cognitive abilities mean better performance (low LE and low WER) in OS speech perception. Future studies using different cognitive test batteries are necessary to help us answer that question better.

Finally, the familiarity effect on LE, as reported in Experiment 1, could mean that OS speakers might find it easier communicating with family and close friends as opposed to others. Although this was not investigated in this study, it would be interesting to know at what level of familiarisation does this effect show and also whether there is a ceiling effect to this familiarisation. That is, is there a point where, due to familiarity, they find OS as effortful as HS?

4.5 Conclusions

We performed two different experiments to collect intelligibility and LE metrics for OS and HS. The first experiment, a web-based one, was used to collect intelligibility and self-reported LE metrics. The conclusions of this experiment were that speaker type (HS or OS) had an effect on both intelligibility and effort. There was significant correlation between WER and LE. Listeners familiar with OS fared the same for intelligibility as people who were not. However, they reported less effort in listening to OS than the not familiar listeners. The ASR intelligibility was poorer for OS compared to HS.

The second experiment was to measure LE for HS and OS in a laboratory setting. The conclusions were that even if the intelligibility of OS was close to HS, there was a considerable difference in LE.

LE obtained through these experiments is based on the listener's own interpretation of 'effort involved in listening'. In the next chapter, we look deeper into LE by investigating brain activity as a physiological measure of LE, and study its relationship with this self-reported measure of LE.

We have built an OS restoration system aimed at better ASR and HSR intelligibility and low LE (see Chapter 6). The methods used in this study will be used to evaluate the outputs of this system (see Chapter 7).

Both HSR intelligibility and ASR intelligibility play different but important roles in OS evaluation. While improved HSR would enable better human–human interactions, an improved ASR performance would enable better human–machine interactions (e.g., digital voice assistants). Lower LE would also contribute towards improved communication with fellow humans. The evaluation of all these three metrics provides an all-round understanding of OS speech perception.

Experiment 1 from this Chapter was presented at a conference [109] and the combination of both Experiment 1 and 2 was published as a paper [112].

Chapter 5

Listening Effort and Oesophageal Speech: An EEG Study

> *"Not everything that can be counted counts and not everything that counts can be counted"*
>
> —Albert Einstein

Recent advancements in technology has enabled us to gain in-depth understanding of the human body. One such technology is neuroimaging which allows us to have a peek into the functioning of the brain. In this chapter, I describe an experiment that explored the differences in cerebral activity when listening to HS and OS. This experiment is an attempt to answer the questions: Do listeners' brain signals reveal any differences while processing OS and HS? What are the factors that influence these differences in the brain signals?

5.1 Introduction

Speech communication requires a great deal of cognitive processing. It involves a vast network of activities such as acoustical processing, linguistic processing and emotion recognition, all performed in a very short span of time [36]. Listening to speech in (acoustically) challenging conditions increases the cognitive demand [80]. Challenging conditions can be attributed to any of the components of speech communication: sender (disordered speech, foreign accent[149]), receiver (hearing impairment [46], non-native listener [10]) or channel (reverberation, background noise [115], poor telephone connection). Moreover, listening to speech in challenging conditions for prolonged periods causes fatigue [80]. To overcome these additional challenges posed on the sensory-cognitive system of the listener, additional LE is required to understand

the signal of interest. For more background information on LE, refer to Section 2.2.2.

We know from the previous chapter that OS is less intelligible and more effortful to listen to compared to HS. This result of LE was based on subjective ratings from listeners. As it is subjective in nature, this rating varies from person to person. Moreover, the definition of "effortful" is very subjective. What may be effortful to one person may not be as effortful to the other person. Perhaps the other person has better cognitive abilities or they have better cognitive load bearing capacity. Our aim is to answer all these questions and to better understand the neural processes involved in speech processing and effortful listening linked to OS, with the help of EEG.

The EEG activity can be decomposed into different frequency bands such as alpha (8-12Hz), beta (16-31Hz), gamma (>32Hz), theta (4-7Hz) and delta (0.5-4Hz) bands. Investigating these frequency bands has given clues to understanding problems such as speech processing in adverse conditions [153] and sensing imagined speech [28]. Alpha (8-12 Hz) power (See Section 5.2.1 for alpha power calculation procedure), particularly in parietal regions, has been found to be related to LE for speech-in-noise and is suggested to reflect the suppression of task-irrelevant information [135, 161]. Additionally, alpha power is known to increase with increasing acoustical degradation of speech [82] as well as with increasing working memory (WM) demands [98]. As OS is acoustically degraded, less intelligible and requires more LE (as per listeners' subjective ratings), we hypothesise that this will result in a higher alpha power for OS compared to that for HS. Given the importance of cognitive functioning in speech perception, particularly of working memory, we also included individual differences in working memory functioning in the analysis. We assumed that working memory capacity could serve as buffer for LE -i.e. larger working memory capacity is associated with less LE.

Some studies have looked into differences in brain activity while listening to degraded speech and control HS. Theys et al. [142] performed a study based on ERP components (See Appendix B for details on ERP components). They observed increased N100 amplitude and decreased N100 latency while listening to dysarthric speech compared to HS, indicating that the inherent degradation in dysarthric speech influenced early sensory auditory processing, and that degraded speech requires more neurophysiological resources in early processing stages. In a Near-infrared Spectroscopy (NIRS) study [114], 16 typically developing children listened to whispered speech and normally vocalised speech. A higher haemodynamic response was observed in the left ventral sensorimotor cortex (in the frontal-parietal region) for whispered speech compared to normal speech. This indicated increased cognitive effort while processing whispered speech. The degradation present in OS) has certain properties in common with

whispered speech such as low energy and degraded fundamental frequency. As there have not been studies on OS and brain activity, we base our experiment on the aforementioned studies conducted on other kinds of degraded speech.

We expand the research on perception of degraded speech by investigating differences in the neural correlates of LE when listening to HS and OS and how they relate to subjective ratings of LE and behavioural performance (speech intelligibility) as well as the cognitive capacities of the participants.

5.2 Materials and Methods

The materials and methods of this experiment were the same as 'Experiment 2' of Chapter 4 (Section 4.3.1), in which the behavioural data results were presented. The current study presents the EEG data obtained from the same setup. Therefore this section only contains the detailed description of EEG acquisition.

I will briefly summarise the experimental procedure here for the benefit of the reader. Sixteen participants listened to sentences spoken by one HS and one OS speaker. The tasks involved a SI task with 30 short sentences and a LE task with 60 longer sentences. EEG was recorded for all the 60 sentences of the LE task. In addition, cognitive tasks such as the backward digit span task and the Flanker task was conducted. For a more detailed description see Section Section 4.3.1.

5.2.1 EEG Acquisition and Analysis

A continuous EEG was recorded using a 24-channel wireless Smarting EEG system (mBrain-Train, Belgrade, Serbia) at a sampling rate of 500 Hz, with a low-pass filter of 250 Hz. The 24 electrodes were attached to an elastic EEG cap (EasyCap, Herrsching, Germany) according to the International 10/20 system [57]. To record the EEG data the software Lab Streaming Layer [64] and Smarting Streamer 3.1 (mBrainTrain, Belgrade, Serbia) was used. EEGLab v.14.1.1 [18] was used offline to process and analyse the EEG data.

EEG recordings were re-referenced off-line to an average of all electrodes. The EEG data was then filtered with a 0.1Hz to 45Hz bandpass filter. Excessive ocular artefacts, such as eye blinks, and other EEG artefacts were identified and corrected, using an independent component analysis as implemented in EEGLab.

Epochs were extracted from the continuous EEG. The lengths of the extracted epochs were varied and was as long as the length of the entire duration of the stimulus (i.e. sentence length).

As mentioned in Section 4.3.1, there was a significant difference in the lengths of OS and HS samples and therefore we chose to include the EEG data corresponding to the entire stimulus in the epochs for a more accurate and justified analysis. Therefore, each epoch contained the EEG data corresponding to the length of the audio stimuli that was heard and a 500 ms pre-stimulus interval as baseline.

For each epoch, a Welch's Power Spectral Density (PSD) was calculated (pwelch from the Matlab signal processing toolbox) with a window length of 1000 samples (corresponding to 2 seconds), an overlap of 99 percent and 1000 points to calculate the Fourier transform. The mean alpha power for each trial was then calculated as the mean of the power values of the frequencies between 8 and 12 Hz. This was performed for all electrodes, but our analysis was focused on centro-parietal electrodes (P3, P4, PZ, C3, CZ, C4, CP1, CP2, CP5, CP6, CPZ).

For more in-depth details on EEG data acquisition and processing, see Appendix B.

5.3 Results

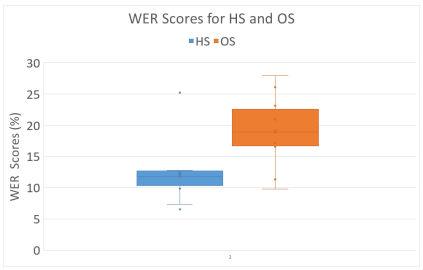
Out of the 16 participants, 4 were excluded from analysis due to poor EEG data quality (i.e., reference mastoid electrodes were either defective or not well prepared resulting in extremely noisy data). The subjective LE ratings from this experiment with all the 16 participants were presented in 4.3.2. We reiterate the results that are relevant to this chapter. Here, for a justified comparison, we present the behavioural results for the 12 participants with good EEG recordings. The overall pattern of results for 16 and 12 participants remain the same.

All statistical analysis were performed using the statistical software program JASP [141].

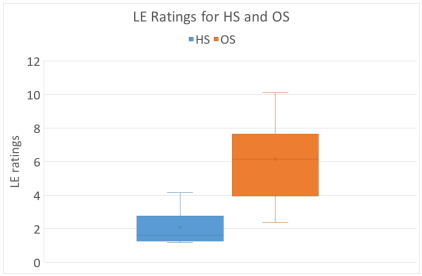
5.3.1 Behavioural Data

On observing histograms, we found that WER scores for the 12 participants were not normally distributed. Therefore we performed a non-parametric paired samples test. A Wilcoxon signed-rank test showed that the WER for HS (Median=11.71) was significantly lower than for OS (Median=18.88), W(1,11)=1.00, $p = 0.002$, Hodges–Lehmann estimate=-6.869, Rank-Biserial correlation=-0.974 (See Figure 5.1a). Nonetheless, both conditions had a high intelligibility of over 80 percent.

Subjective LE ratings followed a non-normal distribution too. Median LE (from a 13-point scale) for the OS speaker was higher than for the HS speaker (see Figure 5.1b). A Wilcoxon signed-rank test showed that this effect was significant, W(1,11)=78, $p < 0.001$, Hodges–Lehmann estimate=4.073, Rank-Biserial correlation=1.00.



(a) WER Scores for HS and OS



(b) Subjective LE Ratings for HS and OS

Figure 5.1: WER scores and subjective LE for HS and OS

5.3.2 EEG Data

We analysed alpha frequency (8-12Hz) power for HS and OS in the centro-parietal region which included central electrodes (C3, C4, CZ), parietal electrodes (P3, P4, PZ) and centro-parietal electrodes (CP1, CP2, CPZ, CP5, CP6). The mean alpha power was the average of the values of all the above mentioned 11 electrodes. For mean alpha power, normality was checked by using the Shapiro-Wilk test, which showed a significant departure from normality (W(11)=0.598, $p < 0.001$). The non-normality was also confirmed by a visual observation of the histograms. A Wilcoxon signed-rank test showed that the alpha power for OS (Median=0.3109) was significantly higher than for HS (Median=0.2786), W(1,11)=61.50, $p = 0.042$, Hodges-Lehmann estimate=0.024, Rank-Biserial correlation=0.577. Figure 5.2 shows a graphical representation of the same data. Figure 5.3 shows a topographic plot of alpha power for HS and OS. The value at each electrode position corresponds to the mean value (across all 12 participants) of alpha at that electrode.

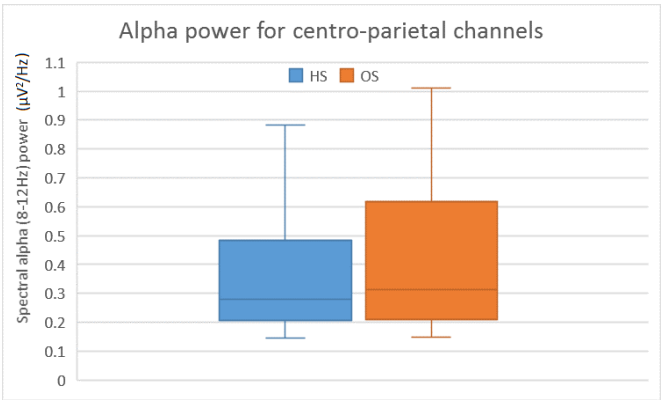


Figure 5.2: Boxplot for average alpha frequency (8-12Hz) power for HS and OS for centro-parietal channels (P3, P4, PZ, C3, CZ, C4, CP1, CP2, CP5, CP6, CPZ)

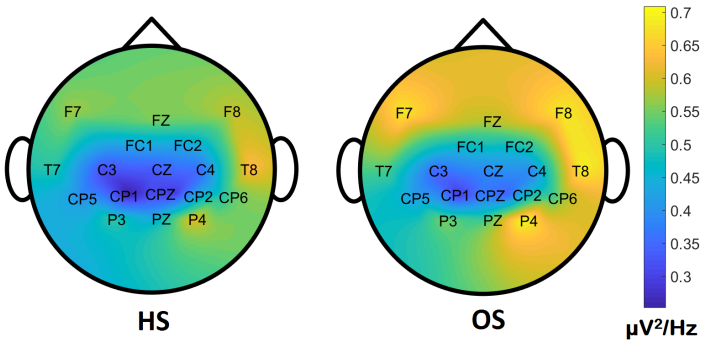


Figure 5.3: Topography plot for average alpha frequency (8-12Hz) power for HS and OS. The channels O1, O2, TP1, TP2, FP1, FP2 were excluded from the analysis due to noise in those channels in the majority of participants

5.3.3 Behavioural Data, Cognitive Tasks and EEG Activity

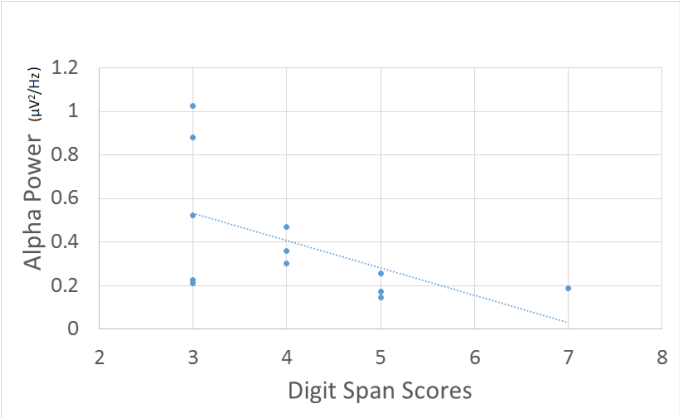
We investigated the relationship of alpha power to subjective LE and cognitive tasks by running correlation analyses. Digit span scores were found to be negatively correlated with alpha power for the HS (Spearman's rho=−0.641, p=0.025) as well as OS conditions (Spearman's rho=−0.670, p=0.017). See Figure 5.4 for scatter plots. No significant correlations were found between subjective LE and alpha power (p>0.1). There were no correlations between WER and alpha power either (p>0.1).

5.4 Discussion

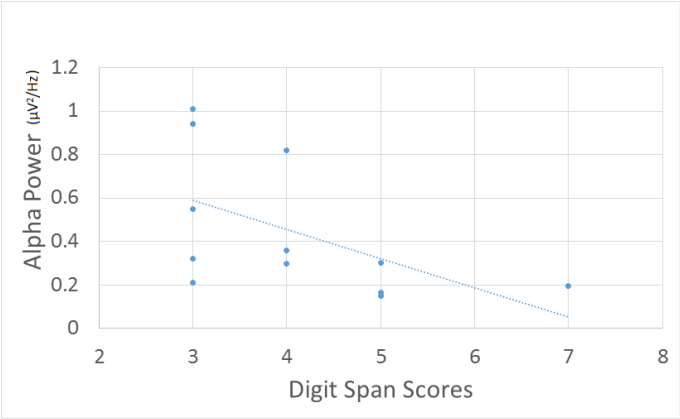
The aim of the study was to look at differences in neural correlates (alpha power) of LE when listening to HS and OS. In addition, we were interested in knowing which factors (amongst SI, subjective LE and cognitive abilities) influence these neural correlates at an individual level.

The OS speaker we chose was amongst the most intelligible OS speakers in the database. This choice of a highly intelligible OS speaker was intentional as we wanted to look at differences in LE while reducing the impact of differences in SI. In the full sentence transcription SI task, we observed that although OS was less intelligible compared to HS, the intelligibility of both OS and HS was over 80 percent.

Our observations from the current study are that OS was associated with higher alpha power in the centro-parietal region compared to HS. Our results corroborate with some previous research on degraded speech perception. Listening to degraded speech required more cognitive effort [114] and neurophysiological resources[142] compared to non-degraded speech. McMahon et. al [82] observed increased alpha activity in the parietal region, lower speech recognition scores and higher perceived effort scores for more degraded speech (decreasing Signal-to-Noise



(a) Alpha Power vs Digit Span Scores for HS



(b) Alpha Power vs Digit Span Scores for OS

Figure 5.4: Scatter plots for alpha power and digit span scores for HS and OS

ratio). Similarly, higher recognition accuracy, shorter reaction times and decreased alpha power was reported for stimuli with more acoustic detail in two other studies ([163] and [7]).

When looking at the individual level independent of speech type, there was no correlation between subjective LE and alpha power. This is unexpected and different from the results of [163], but this divergent behaviour of subjective LE and objective LE is a commonly reported behaviour in LE measurement tests [70]. As explained by Lemke et. al [70], it could be because objective LE (e.g. (neuro-)physiological data) measures "processing effort" or the "amount of processing resources allocated to a task" while subjective LE measures "perceived effort" or a "subjective estimation of how taxing a listening task was", thereby measuring different constructs.

Backward digit span scores was found to be negatively correlated with alpha power (lower digit span scores associated with more alpha power). In other words, for participants with a smaller working memory capacity, more alpha power was observed during the LE task. This suggests that possibly working memory capacity offers a buffer for LE.

The two observations in our experiment (higher alpha for participants with smaller working memory and for OS) are in line with the explanations by Lemke et. al [70] that processing effort arises from the interaction of listener-internal factors (cognitive capacities) and listener-external factors (adverse listening conditions, task difficulty).

A limitation of this study is its small sample size. Also, in order to get a clean measure of LE we wanted to control for differences in SI. However, given the nature of the stimuli it was not possible to achieve equal SI rates. Therefore, differences in LE are likely at least partially influenced by SI which is evident from the negative correlation between the SI scores and LE ratings (also observed by [94] for electrolaryngeal speech). However, it is important to point out that SI rates were still very high and that despite these high rates of SI, significant differences in LE emerged. A similar effect was reported in a previous study on tracheoesophageal speech [95]. This means that even if an OS speaker achieves high intelligibility, the inherent degradation in OS makes it more effortful to understand compared to HS.

5.5 Conclusions

We found significantly higher alpha power and subjective LE and significantly lower SI scores for OS compared to HS, suggesting a higher LE for OS. There was no correlation between alpha power and subjective LE, possibly because subjective and objective LE measure separate constructs. Working memory capacity negatively correlated with alpha power. In other words, more effortful processing (indicated by higher alpha power) was observed in participants with poorer working memory capacities (lower digit span scores).

Listening to OS (even if highly intelligible) is a mentally taxing task. This can cause fatigue, especially in people who interact with OS speakers for long periods of time, such as family members, clinicians and therapists. This makes it difficult to hold long conversations with an OS speaker. Measuring LE (subjective as well as physiological) gives us a more realistic idea of the challenges of OS, even when intelligibility is high.

This study revealed the differences in intelligibility and LE between OS and HS and which factors influenced them. These insights were useful in developing and evaluating enrichment systems (Chapter 7) for OS with which we aim to reduce these differences. Following a laryngectomy, OS speakers undergo regular sessions with a speech therapist which helps in improving their intelligibility over time. However, due to physical constraints, the intelligibility does not reach to the level of healthy speakers. Our future work (Chapter 6) involves developing software-based interventions for OS by increasing its intelligibility even further and reducing its LE for

those interacting with these patients. Reducing LE in addition to improving intelligibility will help OS speakers hold longer conversations without causing fatigue in the listener. This will improve the communication experience and quality of life of OS speakers.

The findings of this study are useful for researchers looking at perception and restoration of OS as well as other types of degraded speech such as dysarthric speech, cleft lip and palate speech, whispered speech, stuttering, deaf speech and cochlear implant speech. All these communication disorders face similar challenges of intelligibility and LE.

The findings of this chapter are in preparation to be published as a paper titled 'Oesophageal Speech and Effortful Listening: an EEG Study'.

Chapter 6

Enrichment Systems

"Mend your speech a little, lest it mar your fortunes."

—William Shakespeare

In the previous chapters, I described characteristics and limitations of OS and the gaps in intelligibility and LE between OS and HS. In this chapter I present some experiments we performed with the aim of enriching OS. The contents of this chapter have featured in previously presented research titled: A multifaceted enrichment of oesophageal speech [110] and a previously published paper titled 'Enrichment of Oesophageal Speech: Voice Conversion with Duration-matched Synthetic Speech as Target' [111].

6.1 Introduction

OS is less intelligible and more effortful to process compared to HS (See Chapter 4 and 5). Poor intelligibility and increased LE hinders verbal communication possibilities for OS speakers, even in non-noisy environments. Lack of intelligibility means that the OS speakers have difficulty in meeting some important needs such as telephonic conversations, calling for a medical appointment, asking for directions and ordering food. Apart from these basic needs, OS speakers find it difficult to engage in family gatherings and public speaking, and to use voice-activated digital devices. All these challenges have amplified in the COVID era with additional barriers such as wearing masks and the reduction of face-to-face interactions. Therefore, enriching OS with software interventions is a very promising aid for OS speakers to facilitate easier communication.

In this chapter, I present some of my own contributions towards enrichment of OS. The methodologies ranged from simple modifications (Section 6.3) on the OS signal to elaborate DNN-based VC methods (Section 6.2).

6.2 Experiment 1: DNN-based OS Enrichment

As stated in Section 2.3, one of the possible approaches to enrich OS is to use a 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. In the OS enrichment context, utterances of an OS speaker can be mapped to a healthy speaker's utterances, thereby having the OS acquire characteristics of HS. Like our previous approaches [126, 124], this proposed method is also based on VC.

VC systems may be parallel (requires temporally aligned source target utterance pairs) or non-parallel (requires 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.

The authors of [45] 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 longer durations of a phoneme, 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 [126], 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 SS with explicit phone durations. This resulted in SS that matches in duration with the source OS utterances, and thus, 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.

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 verbal communication 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 an objective measure of intelligibility and ASR performance.

To sum up, in this section, we present a novel, DTW-free, parallel VC system for OS enrichment which includes an SS target. There is a single speaker or speaker dependent system as well as a multi-speaker or speaker independent system. We evaluate its outputs for ASR performance, STOI and a preference test (for single speaker system only) in comparison with unprocessed OS.

6.2.1 Materials and Methods

Data

We chose four OS speakers with a wide range of intelligibility from the original corpus (See Section 3.5). In the original database, the four speakers were identified as '02M3', '04M3', '16M3', '25F3' and we continue to use these IDs. Average stimulus duration, speaking rates and intelligibility of the four speakers are presented in Table 6.1.

For each speaker, we used a parallel dataset of all the 100 phonetically-balanced Spanish sentences, where the source was OS and the target was SS. The procedure followed to generate the parallel SS will be explained in the next section. As stated before, 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).

	Average duration per stimulus (seconds)	Average speaking rate (syllables per second)	ASR scores (WER in %)
02M3	7.48±1.67	4.32±1.80	56.25
04M3	9.27±2.36	3.84±1.71	74.34
16M3	12.52±3.61	2.59±1.19	90.39
25F3	7.85±2.02	4.24±1.86	43.38

Table 6.1: Average stimulus duration, speaking rates and intelligibility of the four OS speakers

Proposed VC System

The proposed VC system, BLSTM with SS as target (BLSTMSS), is a Neural Network based system with OS as source and SS with matching durations as target (see Figure 6.1). The procedure is described in detail in the following steps.

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 [107] was unsuitable for OS. Therefore, using the Montreal Forced Alignment tool [72], new models were created by using

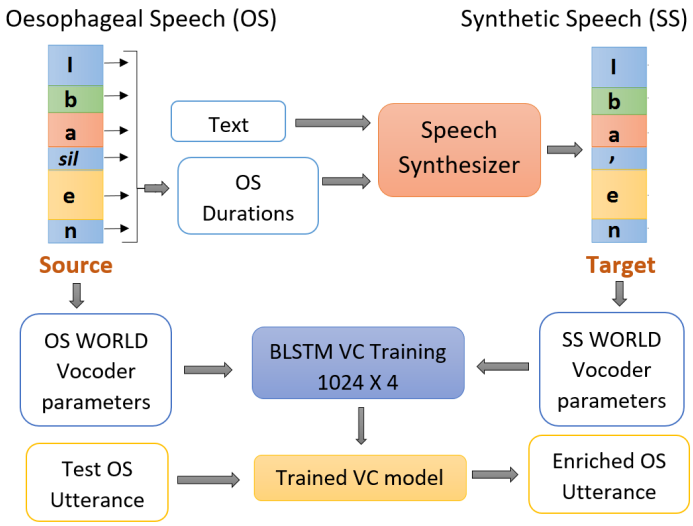


Figure 6.1: The proposed OS-HS VC system: BLSTMSS

OS as the training material (See Section 3.4). Automatic alignment with this forced aligner gave us the phone labels and their durations for the source OS utterances.

Generating Target Synthetic Speech

Using the labels, their durations and the utterance text, SS was generated by explicitly assigning these durations to the phones. The text-to-speech system used was a Hidden Markov Models (HMM) based synthesis system [34] which was originally developed for the Basque language. The Spanish version is described in [119]. This 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.

Voice Conversion Neural Network

Voice conversion was performed with the VC recipe of the Merlin toolkit [165]. Parametrisation and resynthesis was done using the WORLD Vocoder [91]. The extracted parameters included 60 Mel Cepstral Coefficients (MCC), 1 excitation parameter (log F0), 1 Band Aperiodicity Parameter (BAP), the deltas of of the MCC, log F0 and BAP, the delta deltas of the MCC, 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 X (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, the alignment step in

the training process was skipped. The train parameters were normalised to 0 mean and unit variance and then fed into a 4 layered BLSTM (4 X 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.

Multi-speaker system

A multi-speaker version of the BLSTMSS method was implemented using the same 100 sentences from 11 high intelligibility OS speakers. They are the most intelligible OS speakers in the database (less than 60% WER) based on the ASR system trained with HS. These 11 speakers are rightmost 11 speakers [1] in Figure 3.5 (blue bars), not including the TOS speaker 09MT.

For the multi-speaker system, we combined the data from all the 11 speakers instead of performing VC for each speaker separately. Each speaker's utterance of a certain sentence had its own different durations and hence the corresponding target signal was also different. Therefore, each utterance had a paired SS target, a total of 1100 utterances (100 sentences from 11 speakers). Ninety utterances from each speaker (the same 90 sentences from all speakers) and the corresponding target SS were put in the source and the target training set respectively. The training and conversion process was the same as that for the single speaker system described in 'Voice Conversion Neural Network' from Section 6.2.1.

6.2.2 Evaluations and Results

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

The ASR evaluation for the multispeaker system, evaluation was performed using one ASR system (ASR 3) and with STOI scores. A comparison of the speaker dependent system and the speaker independent multispeaker system is also made.

ASR Evaluation for the Speaker Dependent Systems

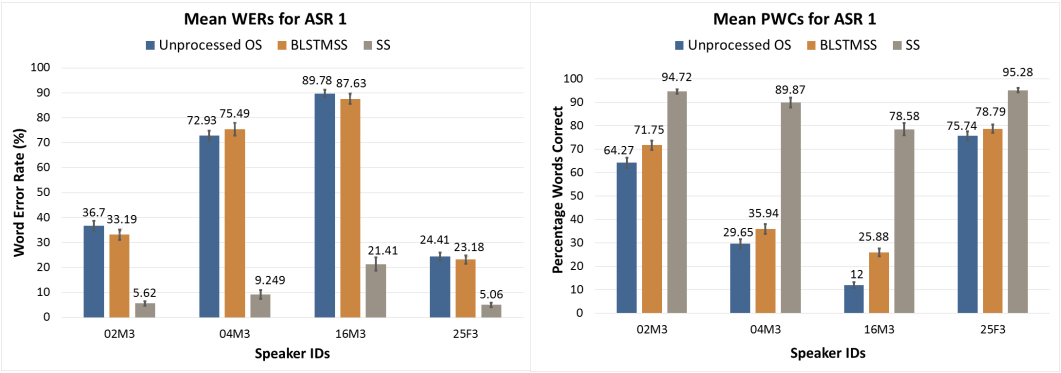
We evaluated the outputs of our proposed enrichment system using three ASR systems: the speech-to-text system from Microsoft Azure using the python azure-cognitive services-speech

[1] Speaker IDs: 01M3, 02M3, 03M3, 08M3, 12M3, 19M3, 22M3, 24M3, 25F3, 28F3, 29M3

library (ASR 1) [85], the Elhuyar speech recognition system (ASR 2) [30] and a Kaldi based system (ASR 3) [107, 126, 124]. The input files to these ASR systems were the 100 single channel speech signals sampled at 16000 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 customisable ASR with full control of all the components such as the language model, dictionary etc. ASR 3, which uses a limited lexicon and unigram language model was used in our previous studies [126, 124]. The advantage of this ASR 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: WER and PWC. WER and PWC were calculated using Equation 2.1 and Equation 2.2 respectively. Both the concepts are explained in Section 2.2.1.

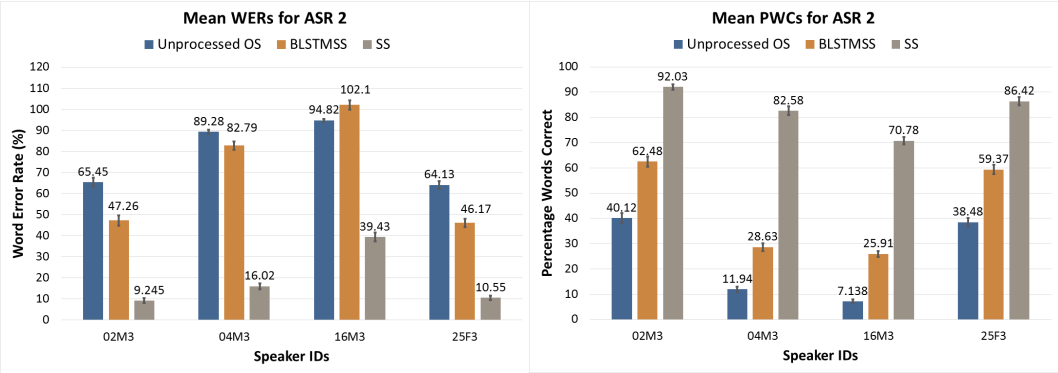


(a) Word Error Rates

(b) Percentage Words Correct

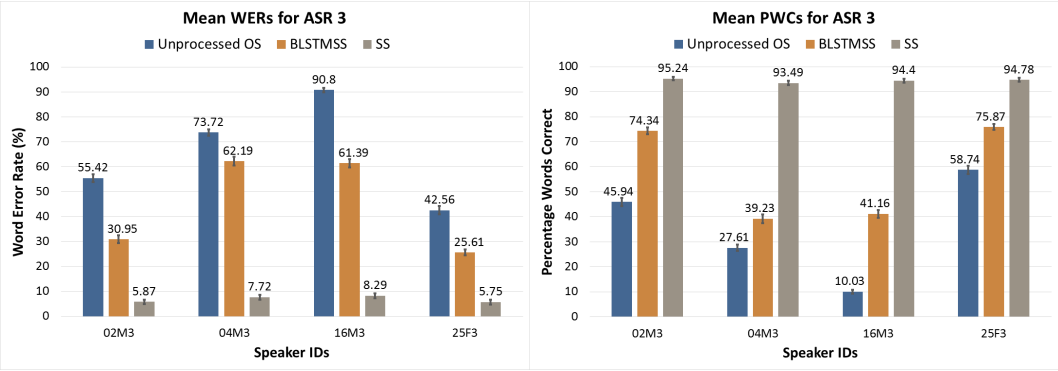
Figure 6.2: ASR 1 WER and PWC scores for unprocessed OS (source), the BLSTMSS converted outputs and target SS (target). Error bars show standard errors.

Figure 6.2, Figure 6.3 and Figure 6.4 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.



(a) Word Error Rates (b) Percentage Words Correct

Figure 6.3: 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 6.4: ASR 3 WER and PWC scores for unprocessed OS (source), the BLSTMSS converted outputs and target SS (target). Error bars show standard errors.

ASR Evaluation for the Multi-speaker Systems

Figure 6.5 shows the ASR scores from ASR 3 for the multi-speaker system. There was ASR improvement for four speakers (speaker 02M3, 03M3, 19M3, 22M3). For the other 7 speakers, the WER increased.

Figure 6.6 shows the comparison of the ASR scores for the single speaker system and the multi-speaker system for the two OS speakers (02M3 and 25M3) that were part of both the systems. As can be observed, the multi-speaker version did not did not have better ASR scores compared to the speaker dependent system.

STOI Scores for Speaker Dependent Systems

We calculated STOI (See Section 2.2.1 for details) 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.

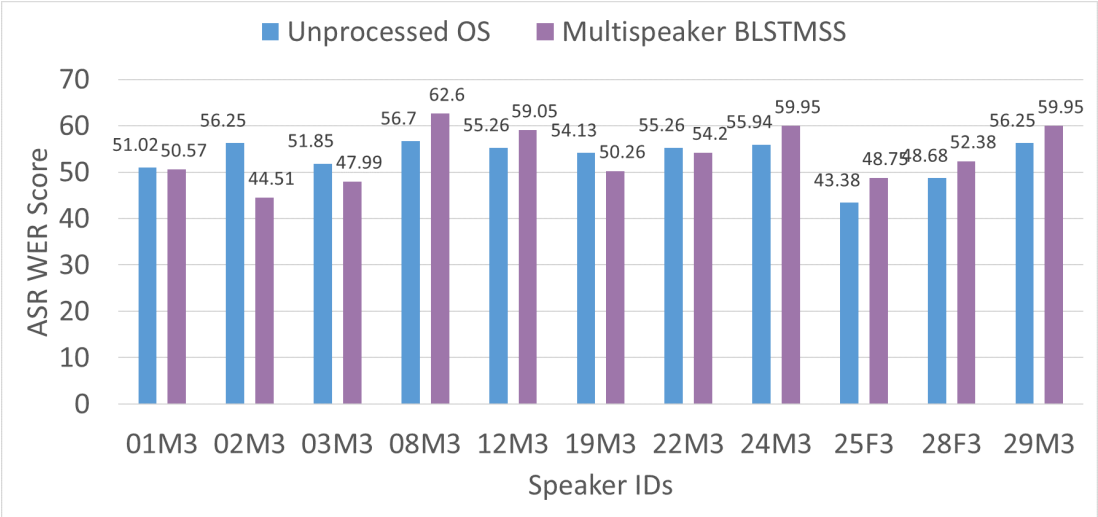


Figure 6.5: ASR scores for the multi-speaker system containing 11 OS speakers.

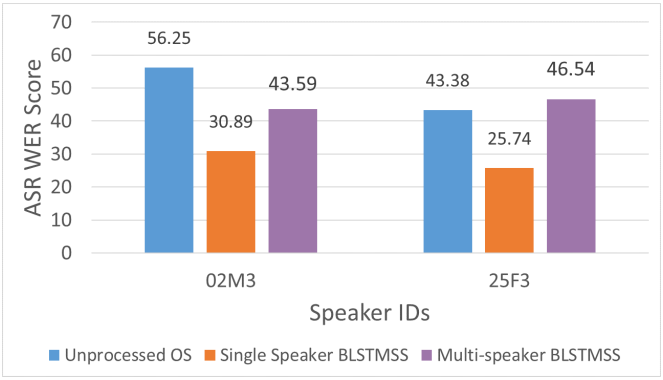


Figure 6.6: ASR scores comparison for the single speaker and the multi-speaker BLSTMSS system.

The STOI results for the single speaker BLSTMSS method are shown in Figure 6.7. 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 62 percent was observed for all the BLSTMSS samples.

STOI Scores for Multi-speaker Systems

Figure 6.8 shows the STOI scores for the multi-speaker system. We can observe that the STOI scores improved for all the 11 speakers.

Figure 6.9 shows the comparison of the STOI scores for the single speaker system and the multi-speaker system for the two OS speakers (02M3 and 25M3) that were part of both the systems. Both the enrichment systems have higher STOI scores, but there were no significant differences between the two systems.

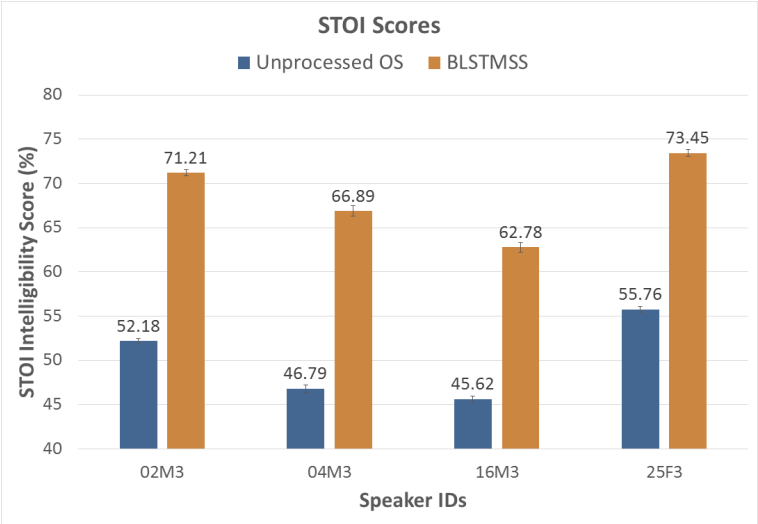


Figure 6.7: STOI scores for the four OS speakers and the enriched versions. Reference signal for STOI is duration-matched SS. Error bars show standard errors.

Subjective Tests

Subjective tests were performed only for the speaker dependent systems and not for the multi-speaker system.

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.

Participants listened to pairs of samples, one unprocessed OS sentence and the corresponding BLSTMSS enriched output of the same sentence. There were 10 pairs for each speaker, a total of 40 pairs. The chosen 10 pairs were the shortest sentences in the set, as that allowed us to have maximum number of evaluations while keeping the test under 20 minutes. The presentation of all the pairs, as well as the order of BLSTMSS and OS 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 stimulus they preferred amongst the two. The options they were given were '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).

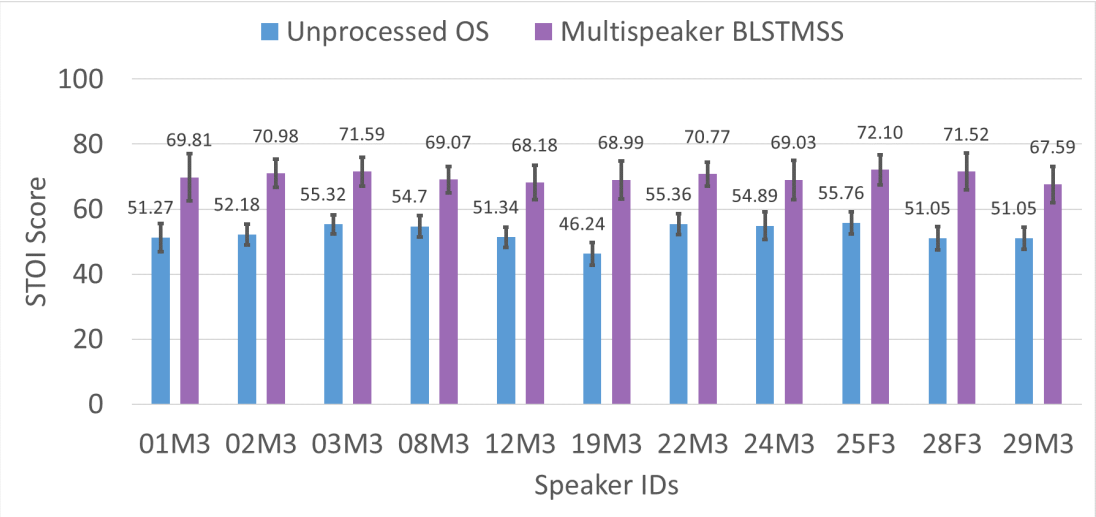


Figure 6.8: STOI scores for the multi-speaker system containing 11 OS speakers. Reference signal for STOI is duration-matched SS. Error bars show standard deviations.

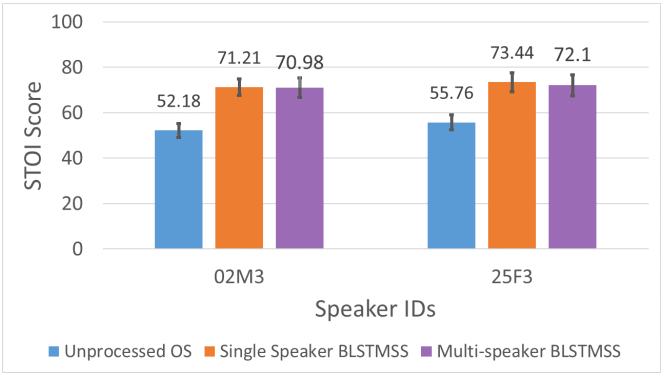


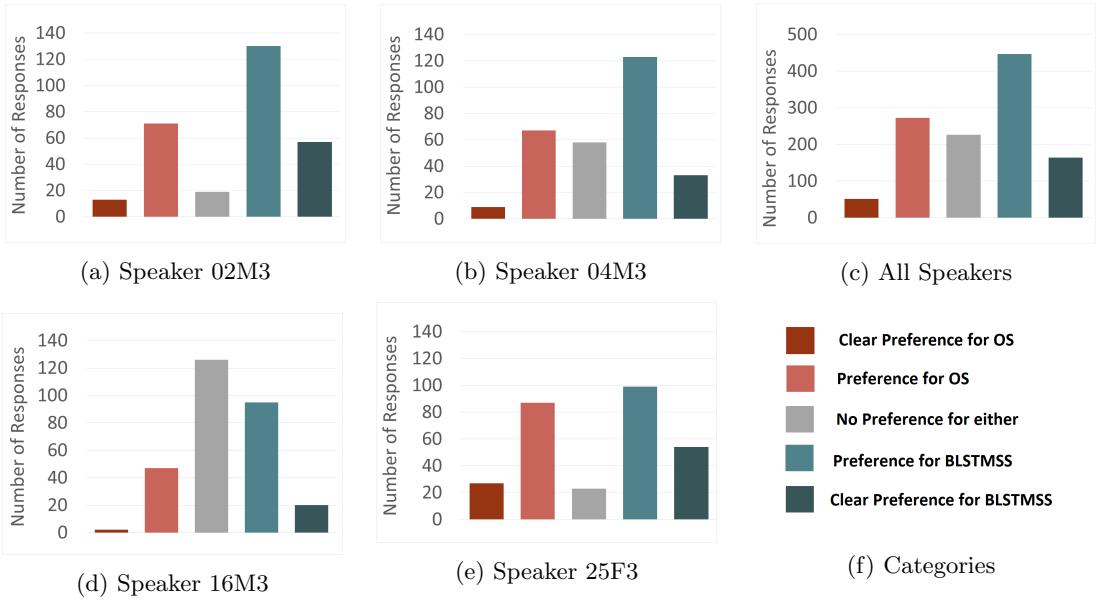
Figure 6.9: STOI scores comparison for the single speaker and the multi-speaker BLSTMSS system. Reference signal for STOI is duration-matched SS. Error bars show standard errors.

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 participants marking the options of at least 3 of these 4 pairs correctly were included in the analysis. This ensured reliability of the participants' responses.

We asked the participants to describe the audio equipment they used to listen to the samples. 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.

A total of 32 native Spanish participants performed the listening test. 2 of them were rejected

because they failed the control test. One other participant used bad listening equipment and was excluded too. 16 out of the chosen 29 listeners had no experience with using speech technologies. 5 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.



(a) Speaker 02M3 (b) Speaker 04M3 (c) All Speakers

(d) Speaker 16M3 (e) Speaker 25F3 (f) Categories

Figure 6.10: Histogram plots for the preference scores of the four speakers separately and All together

Overall, the most chosen option was 'Preference for BLSTMSS' as can be observed in Figure 6.10c. When looking at speakers separately we observed that speaker 16M3 (Figure 6.10d) has a different trend compared to other speakers. For speakers 02M3 (Figure 6.10a), 04M3 (Figure 6.10a) and 25F3 (Figure 6.10e), 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 undecided category. The next more preferred option was 'Preference for BLSTMSS'.

Paired samples Student t-tests was performed to compare the overall preference for OS or BLSTMSS. The overall preference score for BLSTMSS was obtained by adding the number of responses in the 'preference for BLSTMSS' and 'clear preference for BLSTMSS' categories. Similarly the number of responses in 'preference for OS' and 'clear preference for OS' categories were added to obtained an overall preference for OS. This overall preference score was calculated for all speakers together as well as each speaker separately.

For all speakers taken together, there was a significantly higher overall preference for BLSTMSS compared to OS (t(28)=2.953, p=0.006). When looked at speakers separately, the overall preference was significantly higher for BLSTMSS for speaker 02M3 ($t(28) = 3.465, p =$

0.002), speaker 04M3 (t(28)=3.048, p=0.005) and speaker 16M3 (t(28)=2.635, p=0.014). The preference for BLSTMSS was higher for speaker 25F3 as well but did not reach statistical significance (t(28)=1.314, p=0.200).

6.2.3 Discussion

In this study, we have employed and evaluated a DNN-based voice conversion system aimed at enriching OS. The evaluations involved subjective (preference test) as well as objective (intelligibility, ASR) aspects. The evaluations were performed on unprocessed-enriched pairs of samples from four OS speakers.

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.

STOI as an objective intelligibility measure is usually applied in cases where already available clean signals are degraded with noise. In 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. Therefore, the absolute STOI values must be interpreted keeping this in mind. Nonetheless, for all the four speakers, a higher STOI value for the proposed system (over 62 to 73 percent) compared to unprocessed OS (45 to 55 percent) indicates that there was an improvement in intelligibility.

The preference test revealed a preference for the proposed method for three of the four speakers. For the remaining fourth speaker, there was not preference for the proposed system, but no preference for the 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 long duration of stimuli, phones and silences, and the resulting slow and unnatural rhythm were present in the converted version too. With further assistance from a speech therapist, this poorly intelligible speaker can improve their inherent intelligibility and speech rhythm. With a higher intelligibility, the enrichment would be beneficial as observed from the results of other speakers. Nonetheless, there was an overall significant preference for BLSTMSS.

In a previous experiment of enriching OS using a DNN-based VC system [126], 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 [124], we did not achieve intelligibility improvement

or a preference for the proposed method. Both the above experiments involved only one OS speaker who also was one of the most intelligible speakers of the dataset (speaker 02M3). 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.

Between the single speaker and the multi-speaker systems, the single speaker system was the better system as far as objective measurements such as ASR and STOI are concerned. As subjective evaluations were not performed, it is not known whether they are more or less preferred by listeners.

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.

6.3 Experiment 2: Light weight OS enrichments

In the previous section, it was demonstrated that VC can be used to enrich OS by improving its intelligibility. However, VC is a time consuming process as it requires extensive data preparation, right choice of conversion methods and a long training process. In this section I describe some experiments which are simpler to implement.

Laryngectomy involves the removal of the source or fundamental frequency of the speech. Therefore, one way of enrichment involves the addition or reconstruction of the fundamental frequency [129]. Another way of enrichment is by removing the additional artefacts introduced while speaking, such as swallowing air to speak. These pauses of swallowing that are necessary for OS speech production, disturb the durations and hence, the rhythm of the speech. Smearing of rhythm of speech is known to affect recognition [27]. Moreover, the swallowing sounds are unpleasant to listen to, affecting the process of listening.

The objective of this study is to implement some light-weight strategies directly on the speech, such as improving the rhythmic structure of the speech by eliminating undesired silences and improving spectral characteristics by resynthesising with a high quality natural sounding vocoder. We have used one high intelligibility OS speaker (ID: 02M3, WER: 56.25%) and one low intelligibility OS speaker (ID: 16M3, WER: 90.39%) in this study.

6.3.1 Materials and Methods

Removal of Undesired Silences

In Section 6.2.1, the process of generating phonetic labels was explained. The output of that process was more accurate labels for OS, with additional labels that corresponded to pauses. We removed portions in the OS signal that corresponded to these pauses or silences. Most of these pauses were between words. However, there were several instances of pauses within words, especially for 16M3. Figure 6.11 shows an OS signal with and without the pauses.

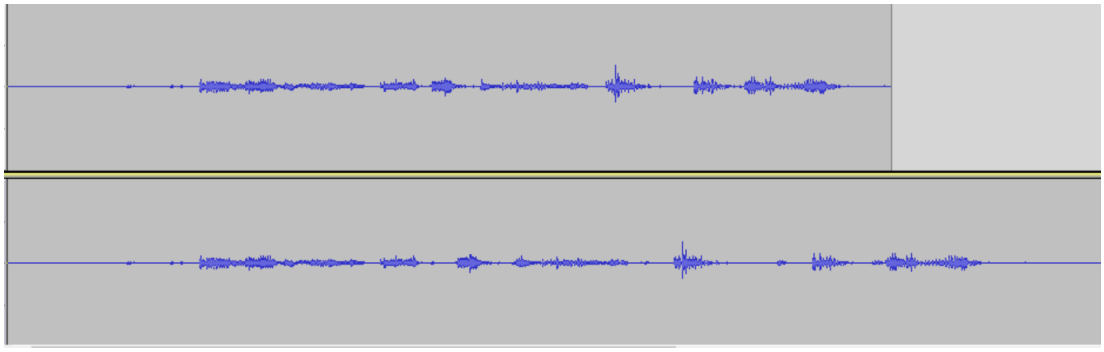


Figure 6.11: Unprocessed OS signal (bottom) and OS signal with pauses removed (top)

Wavenet Synthesis

Wavenet [148] is a speech synthesis system that has been shown to work as a vocoder generating high quality natural sounding speech [140]. In the wavenet speech synthesis configuration, the phonetic labels of the speech signal are used for "local conditioning". This means that the phonetic labels guide the process of training and synthesis of speech. One idea would be to take the labels of OS and generate speech from wavenet. However, in the real world, it would not be possible to generate accurate labels and that too in real time. Additionally, this method would require a very large amount of OS data, which is difficult. Therefore, this idea would not lead to a practical usage.

On the other hand, in the vocoder configuration, the local conditioning is done using mel filter bank parameters. This means that wavenet is trained by having the mel filter bank parameters as a guide. Once trained, new mel filter bank parameters are fed to the trained wavenet model and the output speech is synthesised. Unlike phonetic labels, it is possible to generate mel filter bank parameters in real time or with short delay. Therefore, this method has possible practical usage.

We trained a wavenet vocoder using HS speech samples and their mel filter bank parameters. The training data was taken from the AhoSyn database [120]. This included 3995 sentences

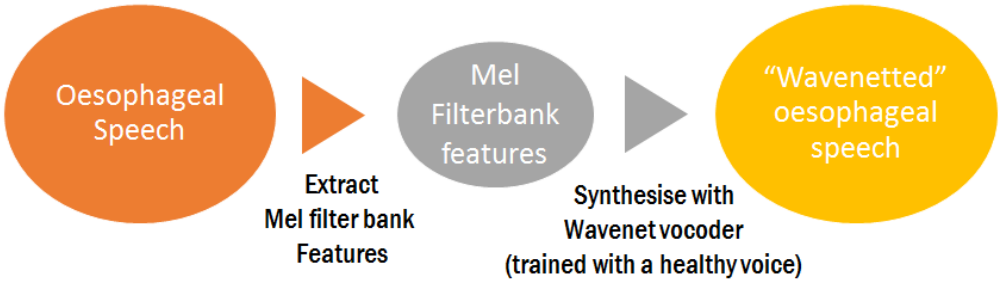


Figure 6.12: Wavenet Synthesis

spoken by a male native Spanish speaker. We then used this trained wavenet by extracting features (Mel filter bank parameters) from OS and synthesising samples using the mel features as local conditioning. The output was OS with a more pronounced fundamental frequency, but some additional artefacts. This final synthesis process is described in Figure 6.12.

6.3.2 Results and Discussion

The proposed systems were evaluated using the ASR 3 system described in the Experiment 1. Here, instead of calculating the WER, the inbuilt 'best WER' score from Kaldi was used. These numbers are quite close to the scores obtained using the WER Matlab toolkit in the previous experiment.

Removing silences showed a small improvement for speaker 16M3 (1.44%, see Table 6.2) but not for 02M3. This is possibly because 02M3 had fewer pauses. The mean number of pauses per utterance and mean duration of pauses for 16M3 (13.99±4.18, 280±53 ms) was higher than those for the 02M3 (7.56±2.29, 169±60 ms).

Even though the speech generated from Wavenet had a more pronounced presence of a normal fundamental frequency, it was not successful in improving ASR. This is possibly because of addition of some artefacts and glitches introduced during the speech generation process. The training data used for this wavenet vocoder was HS and therefore, it may have struggled to correctly translate the mel features of OS to speech. During informal listening tests, the wavenet outputs were reported to be preferred over other enrichment methods (silence removal, BLSTMHS, BLSTMSS and GMM). However it cannot be stated conclusively without proper evaluation.

A comparison of these light weight methods with VC methods is shown in Table 6.2. As can be observed, VC methods have far more success in improving intelligibility compared to these light-weight methods.

Speech type	ASR Scores (WER in %)	
	02M3	16M3
Unprocessed OS	56.32	90.39
GMM VC	37.93*	NA
BLSTMHS VC	40.35*	NA
BLSTMSS VC	30.89*	61.32*
Silence removal	57.99	88.95*
Wavenet	56.09	91.37

Table 6.2: ASR scores for unprocessed and enriched OS for high (02M3) and low intelligibility (16M3) OS. Text marked with * shows numbers where improvement was observed

6.4 Enrichments Demonstration

I created a web-based interactive demonstration simulating an OS speaker speaking with an enriched voice. For this demonstration, we recorded an OS speaker speaking a passage (See Section 3.3.2). Audio was recorded simultaneously with the same recording equipment of the original database to obtain a better quality recording. This audio was then passed through three different enrichment systems: A GMM based system (Enrichment 1), BLSTMSS (Enrichment 2) and a a pitch modified version of enrichment 2 to better suit the speaker (Enrichment 3).

In the interactive demo, the user can play the video and choose amongst four options for the audio: the original speech and the 3 enrichment systems. Whenever the user chooses an audio type (by clicking the corresponding button), the video plays with the chosen audio in a synchronised manner. In this way, it is possible to visualise the OS speaker speaking in the original as well as enriched version of the speech. The demo can be found in the following link: https://aholab.ehu.eus/users/sneha/london_demo/test.html.

This demo was presented in the Royal Institution, London [2] and as a part of a show and tell session in a virtual conference [3].

6.5 Conclusions

The BLSTMSS system had better ASR scores and objective intelligibility measures compared to unprocessed OS. In recent times, communication with digital assistants and other devices is on the rise. Therefore, an improvement is this direction is desirable for efficient communication with digital devices and dialogue systems.

The BLSTMSS system was preferred by listeners compared to unprocessed OS. A slight exception was observed in case of the least intelligible speaker of the set. For this speaker,

[2] https://www.rigb.org/whats-on/events-2020/march/public-easy-speaking-effortless-listening
[3] https://cmsworkshops.com/ICASSP2020/Papers/ViewPaper.asp?PaperNum=6191

there was not a decisive preference for the BLSTMSS method. We presume that when this less intelligible speaker gains more experience in OS with the aid of a speech therapist, their intelligibility will improve and they will gain more benefit from the enrichment.

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

Some other methods to enhance OS that may benefit from further exploring are the eigenvoices method [23] and a GMM-based open source VC system called sprocket [60]. The results from the DIFFGMM algorithm in this VC system seemed promising, but a formal evaluation could not be performed. Therefore it would be useful to explore this method more and perform formal evaluations.

In addition to ASR scores, STOI and preference tests, we are interested in investigating subjective and physiological LE 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. LE provides useful additional information about whether enriched OS is easier to perceive and process compared to unprocessed OS. Therefore, our future studies (Chapter 7) will focus on LE in addition to intelligibility and listener preferences.

The demonstration in Section 6.4 has helped the general public, researchers and the OS speakers themselves to visualise the expected outputs of an enrichment system.

The DNN-based system was published as a journal paper [111] and the light weight enrichment systems were presented in a conference [110].

Chapter 7

Final Enrichment Evaluations

"The most basic of all human needs is the need to understand and be understood. The best way to understand people is to listen to them."
—Ralph G. Nichols

The final step in the OS enrichment problem is the evaluation of outputs of the enrichment processes. The aim of OS enrichment was to close the gaps in intelligibility and LE between OS and HS. In other words, we aimed to have the metrics of the enriched outputs to be as close as possible to HS. Some evaluations were presented in the previous chapter but these evaluations only compared some machine intelligibility related scores of the particular novel algorithm compared to OS. In this chapter, I describe some objective, subjective and physiological evaluations of the intelligibility and LE of all the OS enrichment tasks we have performed so far, including work from other researchers in the lab. These metrics are compared with those of unprocessed OS and clear speech (HS and SS). We shall see through the evaluations how the algorithms we adopted contributed to the enrichment of OS.

7.1 Introduction

When speech is disordered, the deficiency in producing clear speech makes listening to the speech difficult. We have seen in Chapter 4 and Chapter 5 that listening to OS poses such challenges to listeners compared to HS. In Chapter 6, we described some algorithms that were developed to transform OS with the objective that listening to an OS speaker would be easier. Our findings suggest that we had success in enriching speech in some areas (ASR scores, STOI, preference scores). These results are encouraging and in many software based enrichment studies [102, 166, 12], the success of the enrichment methods are evaluated based on these objective measures or simple Likert scale based MOS scores. But these evaluations are unidimensional

and do not tell us the whole story.

In Chapter 5, we described an in depth evaluation of OS and HS by conducting a listening experiment in a laboratory setting and measuring SI, subjective LE and EEG activity. This helped us attain a deeper understanding of the differences in HS and OS. In this chapter, I describe a similar in-depth experiment with all the major enrichment methods developed by us so far in addition to the OS and HS samples. The goal is to know which is the winning enrichment method. Ideally it would be the one that beats OS and the other enrichment methods to be more intelligible and less effortful to process.

All the methods used in this experiment and the motivations to use them were explored and explained in previous chapters. This chapter is just a coming together of all the enrichments and all the types of evaluations to determine which is (if we do have one) the best enrichment system that we made.

7.2 Stimuli

Five systems were evaluated in this experiment: OS, HS and three versions of enriched OS. The three versions of enriched OS were outputs of three different DNN based enrichment systems (BLSTMHS [126], PPG [124], and the new single speaker BLSTMSS method described in Chapter 6). A subjective word recognition task also included the multi-speaker version of the BLSTMSS system.

In this evaluation, we have chosen samples only from speaker 02M3. Using samples from all speakers was not feasible as that would increase the conditions in the listening tests and would make them too long, making it difficult for the listeners to sustain attention. Speaker 02M3 had a high intelligibility (ASR WER: 56.25%). He was the only speaker who had performed additional recordings of words and passages which were useful for more evaluations.

For all the conditions, we used all the 100 phonetically-balanced Spanish sentences described in Section 3.1. The sentences were syntactically and semantically predictable but had some difficult low frequency words. They were sufficiently difficult for a SI and LE rating task.

In addition to the sentences, we used 150 low frequency difficult words for a word recognition task. The details of these words are described in Section 3.3.1 and the list of words in Appendix D.

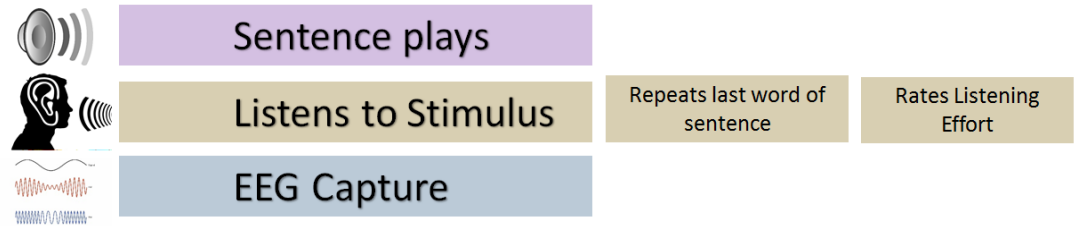


Figure 7.1: LE and SI Task Schematic Representation

7.3 Experimental Procedure

32 native Spanish speakers (9 male, 23 female, Age: 25.63 ± 5.11) participated in a listening test, presented with a psychopy [103] interface. Each participant listened to 100 sentences, 20 sentences in each of the 5 conditions: OS, PPG, BLSTMHS, BLSTMSS and HS. After listening to each sentence, they performed an SI task and an LE rating task. No sentences were heard more than once. All the stimuli underwent loudness normalisation in accordance with EBU R 128 Standard [29].

The 100 sentences were presented in 5 blocks of 20 sentences where each block contained stimuli from all the 5 conditions in a randomised order. The stimuli were counterbalanced across participants and conditions, such that each sentence in a particular condition was listened to an equal number of times across all participants.

7.3.1 Behavioural Tasks

The behavioural tasks consisted of an SI task and an LE task with sentences and an SI task with word stimuli.

In the sentences SI task, the participants repeated aloud the last word of the sentence they just heard. Their response was recorded and later checked for correctness. The SI score per stimulus was whether or not the listener got the last word right: 1 for correct, 0 for wrong. The 13-point ACALES scale (See Section 2.2.2) was used to rate LE. Unlike the previous experiments, here every stimulus had an SI as well as LE rating. Figure 7.1 shows a schematic representation of this task.

An additional SI task was performed with the 150 isolated words. This was a simple test with the participant listening to the word and then repeating it. The experimenter scored the responses for correctness later: 1 for correct, 0 for wrong. Figure 7.2 shows a schematic representation of this task.

In all, the test lasted 1.5-2 hours which included 30 minutes of EEG setup, 30 minutes of the combined SI-LE task, 15 minutes of the words SI task and 15-30 minutes of post test question-

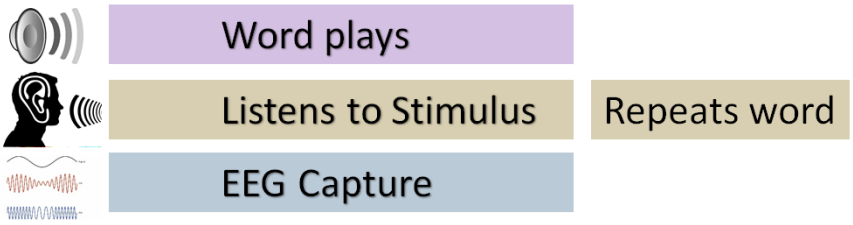


Figure 7.2: Word Recognition Task Schematic Representation

naires and dismounting of the EEG cap. The participants were given monetary compensation. The experiment was approved by the ethics committee of the Basque Centre on Cognition, Brain and Language.

7.3.2 EEG Acquisition and Analysis

A continuous EEG was recorded while participants were performing the LE rating task. The brain activity was recorded from 32 electrode sites mounted into an elastic EEG cap (EasyCap, Herrsching, Germany) and arranged according to the International 10-20 system [57] (See Appendix B for more details). BrainVision Recorder were used to record EEG data. The EEG was recorded at a sampling rate of 1000 Hz. EEG data offline processing and analysis was conducted using EEGLab v.14 [18].

Amongst the 32 electrodes, the scalp electrodes were electrodes 1 to 27. Electrode 28 was in the right mastoid position and electrodes 29, 30, 31 and 32 were placed on the forehead and eye area to record ocular activity. An impedance of $<$ 5kOhms was ensured on each electrode during the EEG cap mounting.

EEG recordings were re-referenced off-line to an average of the right mastoid electrode. The EEG data was then filtered with a 0.1Hz to 45Hz bandpass filter. Excessive ocular artefacts, such as eye blinks, and other EEG artefacts were identified and corrected, using an independent component analysis as implemented in EEGLab.

Epoch extraction procedure was the same as in Section 5.2.1. Variable length epochs were extracted with durations matching the speech signals with an additional 500ms pre-stimulus as baseline. Similarly, the frequency analysis was performed as per the process described in Section 5.2.1. It involved a PSD calculation with a 99 percent overlap. Here, as the sample rate was 1000, 2000 points (corresponding to 2 seconds) were used to calculate the Fourier transform. As before, the alpha power was the mean of the power values between 8 and 12 Hz. Although electrode layouts and placements were slightly different in this experiment compared to the experiment in Chapter 5, the analysis here was also focused on the centro-parietal region

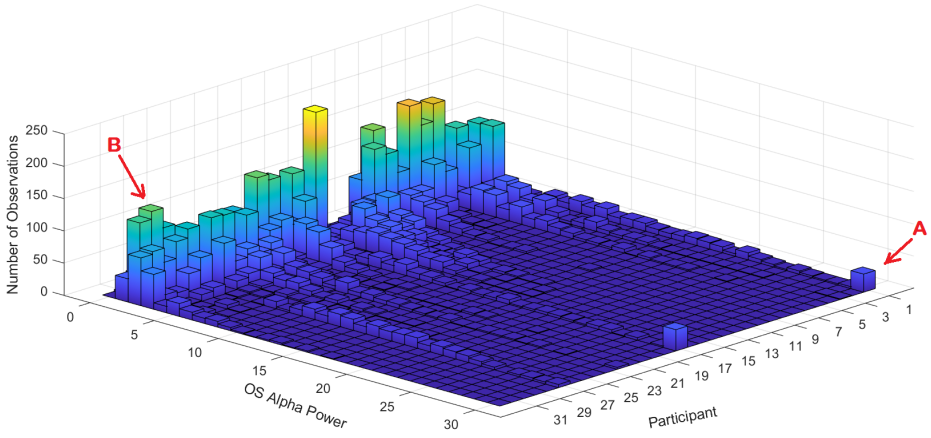


Figure 7.3: Distribution of Alpha Power Value for all the 32 participants.

(CP1, CP2, CP5, CP6, C3, C4, CZ, P1, P2, P3, P4, P7, P8, PZ).

With 32 participants, 5 conditions, 20 stimuli per condition and 12 electrodes, we had 32*5*20*12 i.e. 38400 data series. Out of these, some trials were rejected due to noisy data and we were left with 37944 data points.

Alpha power is affected by individual brain activity, environmental factors, the mental and physiological state of the participant at the time of the experiment and other such factors [96]. Figure 7.3 shows the distribution of alpha power for the 32 participants for OS. Here, the 'Participant' axis represents each of the 32 participants and the 'OS Alpha Power' axis represents the 30 bins of alpha power values. The 'Number of Observations' axis is the number of alpha power values observed in a particular alpha power bin. Here are two example points in the graph to help you understand the graph better. At point A, participant number 1 has 0-50 observations where alpha was in the range of 30-31 $\mu V^2/Hz$ (an outlier). At point B, participant number 30 has close to 100 observations where alpha was in the range of 1-2 $\mu V^2/Hz$. As a general observation, we can see that each participant has a different alpha profile or distribution. A subset of 3 participants provides a simplified representation of the issue. Figure 7.4 shows the varying power distribution of 3 of the 32 participants.

To remove all these between subjects differences and to focus only on the alpha power differences between conditions, a power normalisation was performed. There are several ways to perform normalisations for data like this such as z-scores, dividing by the max value and dividing by resting state alpha power. Each of these processes are explained in the following paragraphs.

A z-scored alpha power was generated by using the mean and standard deviation of all the

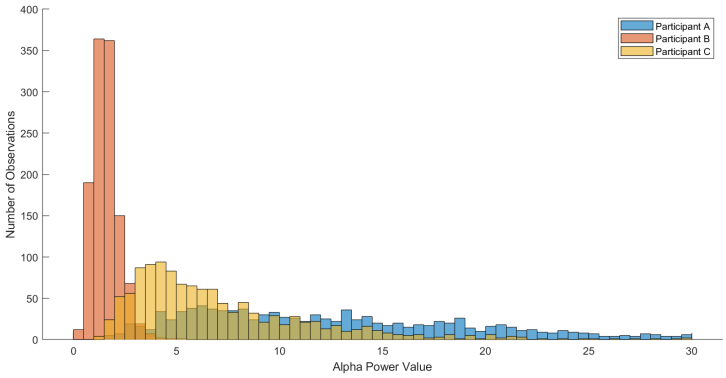


Figure 7.4: Distribution of Alpha Power Value for a subset of 3 participants. Each participant has a separate range of alpha power values.

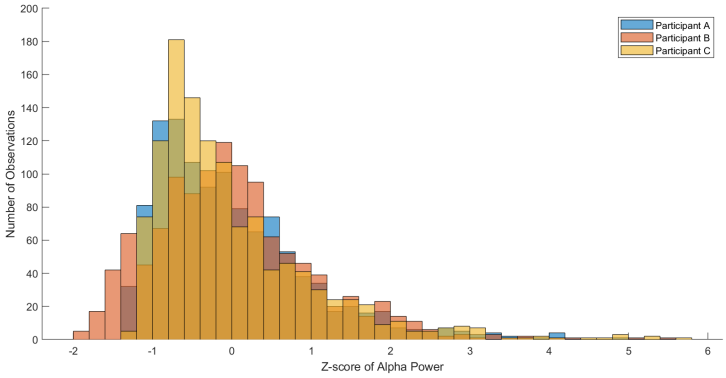


Figure 7.5: Distribution of z-scored alpha power for a subset of 3 participants.

data points from each participant. The distribution post applying the z-score is shown in Figure 7.5 for the same three participants considered before. As it can be observed, the distribution of z-scored alpha power for the three participants is overlapping. This eliminated inter subject variability.

Although the z-score approach of normalising helps us get rid of the problems of inter-subject variability, it assumes that the data follows a normal distribution. We can see that the alpha power data does not follow a normal distribution (See Figure 7.4). It has a slight skewness and it looks more like a log normal distribution.

The second way to normalise the data is by dividing the entire dataset from a participant by the maximum alpha power of that participant. As alpha power only has positive values, this process fits all the datapoints from a participant between 0 and 1. See Figure 7.6 for the normalised alpha power distributions for the 3 participants.

Another typical approach for EEG data normalisation is to divide the alpha power from one participant by the baseline resting state alpha power. The resting state corresponds to the

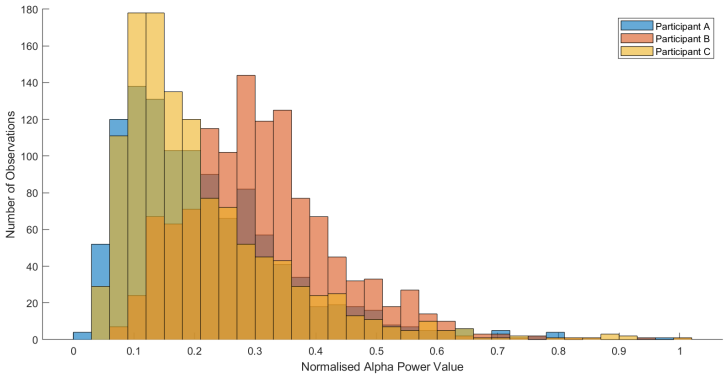


Figure 7.6: Distribution of Normalised alpha power (normalised by max value) for a subset of 3 participants.

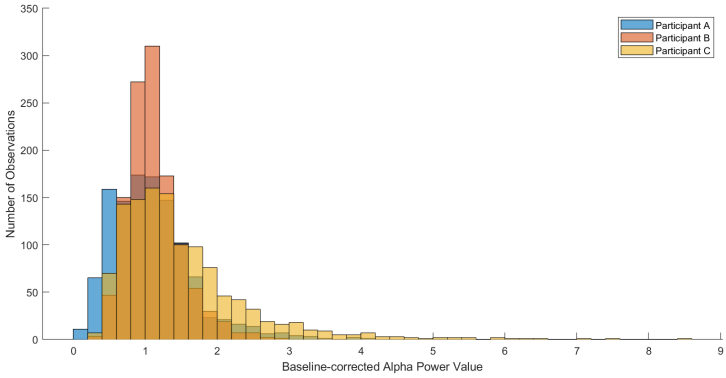


Figure 7.7: Distribution of baseline corrected alpha power for a subset of 3 participants.

time when the participant is not engaged in any task. In this case, the participant was asked to focus on a cross on the screen for 60 seconds. The idea here is that this normalisation will take away the alpha power components that pertain to the participants inherent brain activity without the effect of any task performance. We divided the alpha power for each stimulus by the average alpha power during the 60 seconds of resting state EEG data. This was performed separately for each of the 12 centro-parietal electrodes. See Figure 7.7 for the normalised alpha power distributions for the 3 participants.

7.4 Results

7.4.1 Speech Intelligibility

Figure 7.8 shows the averaged SI scores (% words correct or PWC scores) from the 32 participants. HS had the highest SI score ($p<0.001$). SI score for OS was higher than that of all the enrichments ($p<0.001$). Within the enrichments, the proposed system had significantly higher

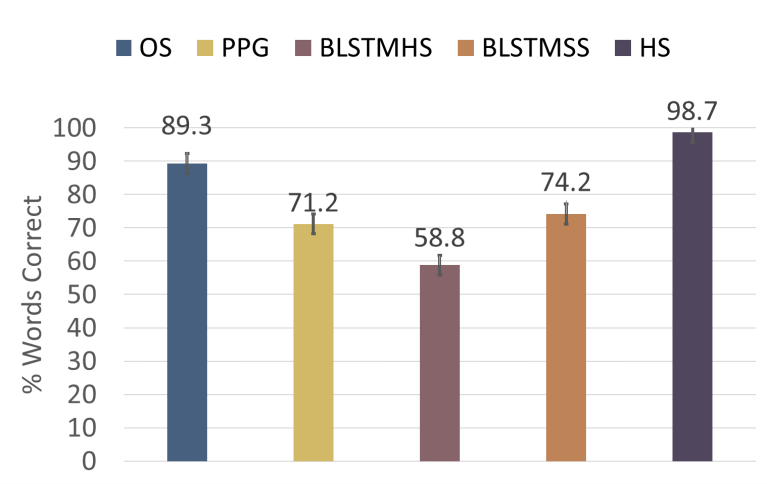


Figure 7.8: Speech Intelligibility (SI) task performance for the three enrichment systems (BLSTMHS, PPG and BLSTMSS), HS and OS. Error bars show 95% confidence intervals.

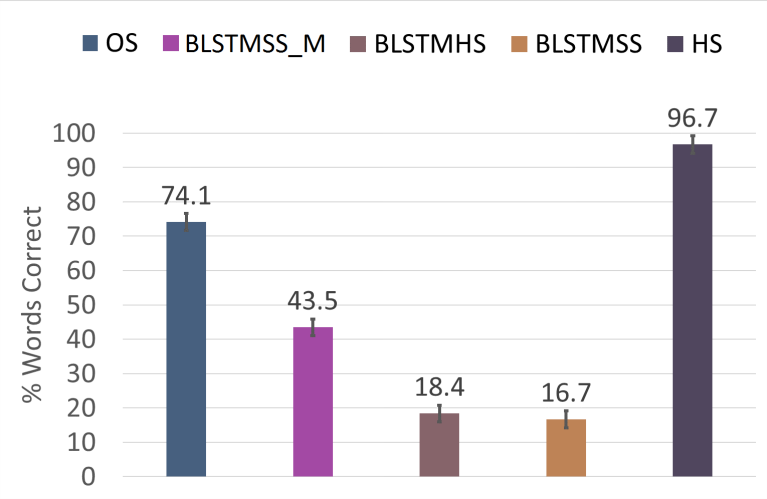


Figure 7.9: SI scores for isolated words.

scores compared to BLSTMHS ($t(99) = 4.59$, $p<0.001$) and was comparable to PPG ($p=0.186$, null hypothesis for equal means accepted at alpha=0.05).

The isolated words SI scores are shown in Figure 7.9. We can see that multi-speaker version has higher SI scores compared to the single speaker BLSTMSS as well as the BLSTMHS method.

7.4.2 Listening Effort

Only 30 out of the 32 participants were considered as the LE data for 2 participants failed to save.

Figure 7.10 shows the LE results from 30 participants. We can observe from the figure that HS has far less LE compared to OS ($t(99)=-33.59$, $p<0.001$), as expected. Secondly, comparing

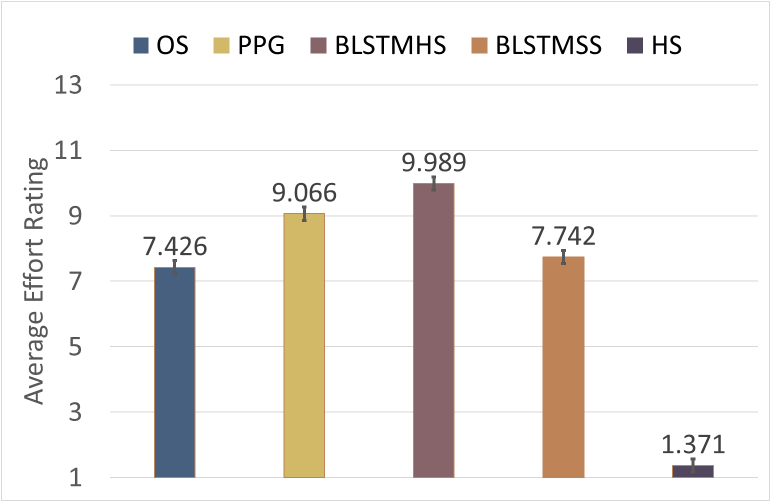


Figure 7.10: Average Listening Effort (LE) for the three enrichment systems, HS and OS. Error bars show 95% confidence intervals.

Table 7.1: Median LE Ratings for OS, HS and the three enrichments

Condition	Median LE	Corresponding LE Rating
OS	8	Moderate to considerable effort
PPG	9	Considerable Effort
BLSTMHS	11	A lot of effort
BLSTMSS	8	Moderate to considerable effort
HS	1	No effort

the enrichments to OS, only the BLSTMSS method was comparable to OS. ($t(99)=-1.198$, $p=0.233$, null hypothesis of equal means accepted at alpha=0.05). LE for PPG ($t(99)=-8.77$, $p<0.001$) and BLSTMHS ($t(99)=-11.65$, $p<0.001$) was higher than that for OS. This means that while BLSTMSS has managed to significantly reduce LE compared to our previous systems, it has still not managed to go lower than OS. This is also evident from the median LE ratings in Table 7.1.

LE rating had a significant negative correlation (Spearman's $rho=-0.459$, $p<0.001$) with SI scores, suggesting that less intelligible stimuli were more effortful, as expected. Order of presentation had no effect on subjective LE.

7.4.3 Response Times

In addition to LE and SI measurements, we saved the Response Times (RT) of both the tasks. This was done using the in-built reaction time option in psychopy. SI RT was the time between the end of the sentence playing and the end of the verbal response from the participant. LE RT was from the start of the appearance of the rating scale to the marking of the response.

Figure 7.11 shows the SI and LE RTs for the five conditions. The lowest SI RT was for

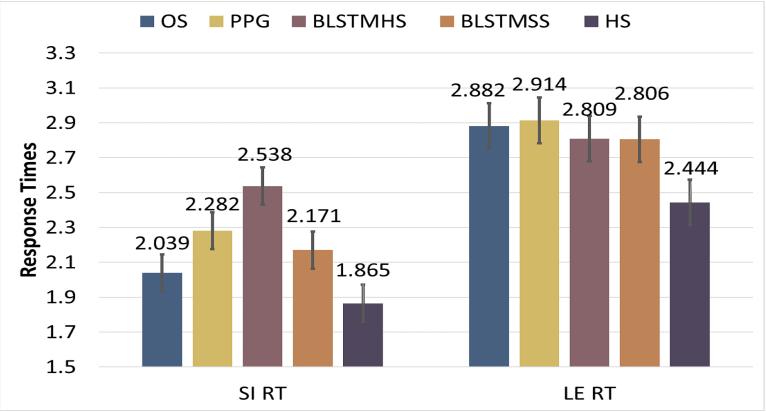


Figure 7.11: Response times (RT) of SI and LE tasks for the three enrichment systems, HS and OS. Error bars show 95% confidence intervals.

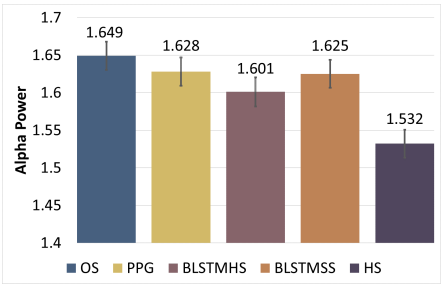
HS ($p<0.001$). OS and HS had faster SI responses compared to all the enrichment systems ($p<0.009$). Amongst the enrichment systems, the BLSTMSS system had significantly faster RTs than BLSTMHS ($t(99)=-3.66$, $p<0.001$) and faster than PPG ($t(99)=-1.53$, marginally significant at $p=0.065$). It can be seen that SI RT (left graph) follows a trend like that of LE ratings (Figure 7.10). There was a significant correlation (Spearman's $rho=0.26$, $p<0.001$) between them. This suggests that where the SI task took longer to respond, a higher LE was reported. There was a significant negative correlation (Spearman's $rho=-0.34$, $p<0.001$) between SI score and SI RT suggesting that for a more intelligible stimulus, the response was faster.

In the LE RT graph, there was a short RT for HS. For all the other conditions, the RT was higher than the RT for HS ($p<0.001$) but not significantly different amongst themselves ($p>0.28$, null hypothesis for equal means accepted at $alpha=0.05$). A significant weak positive correlation (Spearman's $rho=0.12$, $p<0.001$) was found between LE rating and LE RT.

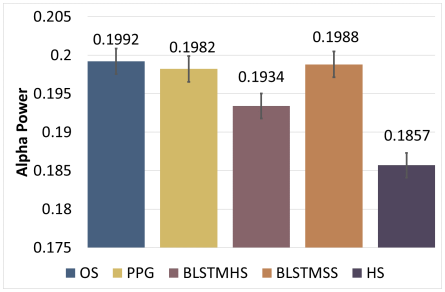
7.4.4 EEG Activity

A one way ANOVA with alpha power as dependent variable and condition as independent variable revealed a significant effect ($p < 0.001$) of condition. Post hoc tests revealed a significantly higher ($p < 0.001$) alpha for OS compared to HS. There were no significant differences between the enrichments. There was no significantly reduced alpha power for any of the enrichment methods compared to OS. Amongst the three enrichments, the BLSTMHS method had the lowest mean alpha power, however this difference was not significant.

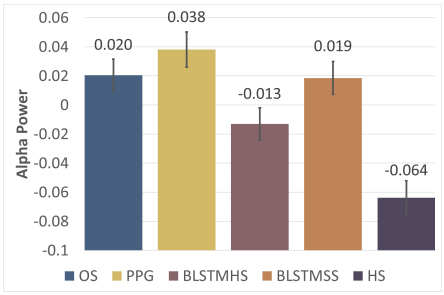
Figures 7.12a, 7.12b and 7.12c show the alpha power scores obtained by baseline normalisation, max normalisation and z score normalisation respectively. Irrespective of the style of normalisation, the results of alpha power remain the same: Lower alpha for HS compared to



(a) Baseline-normalised



(b) Max-normalised



(c) Z scores

Figure 7.12: Alpha Power for the five conditions

all other conditions and no significant winner amongst the enrichments.

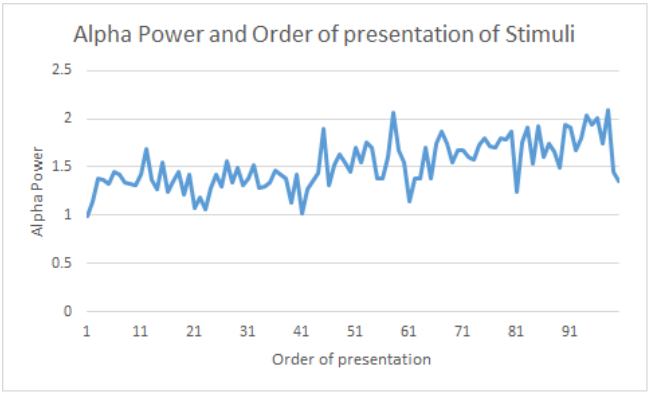


Figure 7.13: Average Alpha power for all participants as the experiment progresses. The x axis represents the presentation order including all conditions.

A one way ANOVA also revealed a significant effect ($p < 0.001$) of the 'order of presentation of the stimuli'. Stimuli that were presented earlier in the trial had a lower alpha compared to those presented later. Figure 7.13 shows that the participant's alpha power increased as the experiment progressed. Figure 7.14 represents a condensed version of the same data (5 blocks of 20 stimuli each) separated by condition. In this figure we can see that this upward alpha trend was observed for all conditions except for the abrupt drop in alpha activity that was found for unprocessed OS from block 4 to block 5.

Alpha activity was not found to be correlated with Subjective LE, SI scores, LE RT or SI

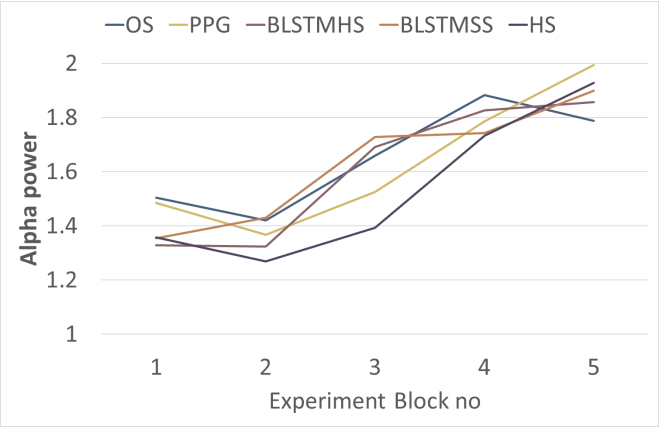


Figure 7.14: Alpha power progression across blocks (time) for OS, PPG, BLSTMHS, BLSTMSS and HS

RT. Spearman's rho values were close to zero.

7.4.5 ASR

The ASR evaluation was done with the ASR 3, i.e. the Kaldi system described in the previous chapter.

Figure 7.15 shows the WER scores of BLSTMSS in comparison to our previous methods, PPG [124] and BLSTMHS [126]. It can be observed that BLSTMSS was able to significantly reduce ASR errors in comparison to previous methods.

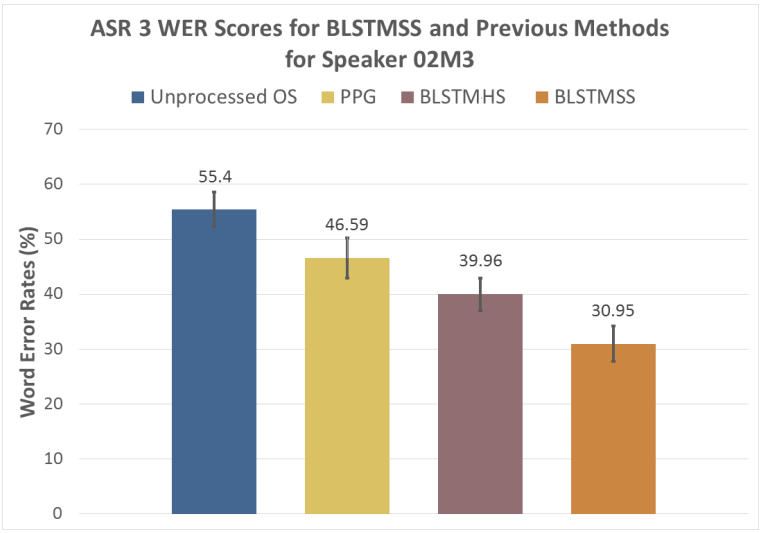


Figure 7.15: WER scores for Unprocessed OS, previous systems (PPG and BLSTMHS) and BLSTMSS as calculated by ASR 3 for speaker 02M3. Error bars show standard errors.

7.4.6 STOI

We compared the STOI scores of BLSTMSS with those of our previous methods (see Figure 7.16). The duration matched SS signals that were used as the VC target were also used as references to calculate STOI. BLSTMSS has higher STOI scores (about 5 percent) compared to previous systems and unprocessed OS.

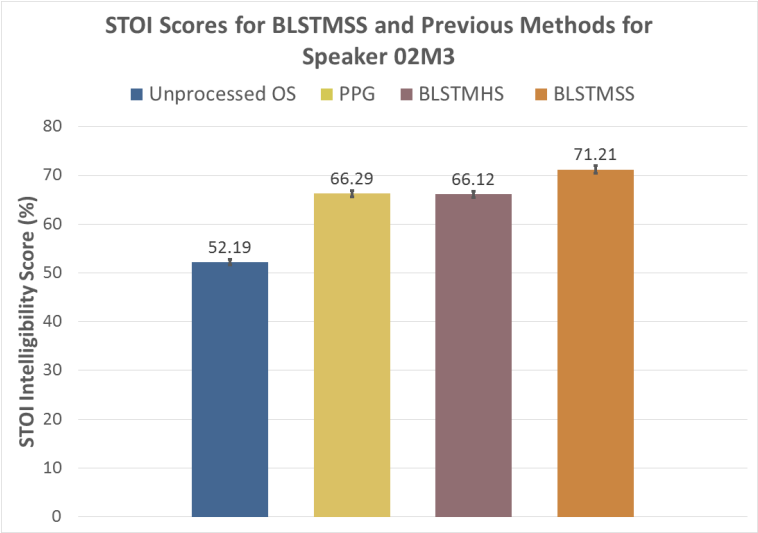


Figure 7.16: STOI scores for Unprocessed OS, previous systems (PPG and BLSTMHS) and BLSTMSS for speaker 02M3. Reference signal for STOI is duration-matched SS. Error bars show standard errors.

7.5 Discussion

The listening test measured how listeners responded to OS and its enrichments. In this area, our latest enrichment method (BLSTMSS) had better performance (higher or comparable SI scores, lower LE ratings) compared to our previous methods, but not better than unprocessed OS. The multi-speaker version of the BLSTMSS method had better SI scores for the isolated words SI task showing that the multi-speaker version may hold better possibilities compared to the single speaker version for newly recorded stimuli. However, in this case too, the OS samples outperformed the transformed OS samples.

Possibly, the distortion introduced from the VC process and the loss of some phonemic cues of unprocessed OS made the VC outputs less intelligible to listeners than OS. This may also explain the inability of VC to improve LE. A preference towards unprocessed OS was also evident in the preference tests in [126] and [124].

A similar listening test was performed for the speech of children with dysarthria [15], where RT was a significant predictor of LE ratings, and high accuracy in word recognition was related

to shorter RTs and lower LE. Our results of LE, SI and their RTs are along the same lines.

The EEG activity data reiterates our findings from Chapter 5: Listening to OS entails higher alpha power compared to HS. There were no significant differences in alpha power between the OS condition and the enriched OS conditions. Although the BLSTMHS method had lower alpha compared to other enrichments, this was not significant and hence we cannot conclude if this enrichment provided any improvement in cognitive load.

We observed that the alpha power increased as the experiment progressed. This may indicate increase in fatigue or decrease in alertness as was observed by Antons et. al [3] too. While shorter experiments would reduce these tendencies, it reduces the data points necessary to reach statistical significance. Therefore, such listening tests must be designed by keeping this trade off in mind.

In the case of unprocessed OS, there was a drop in alpha from the fourth block to the final block. A possible explanation for this is that when the LE demands are too high, the listener often "gives up" at the task of listening and starts exerting lesser LE. For example when you are in extreme fatigue, say, a jetlag, you cannot meet the required attentional demands and may exert lesser effort [136]. This seems to have happened with unprocessed OS in the last block, where the listener was so fatigued that they exerted a lower amount of effort. This pattern of a drop in physiological LE for stimuli above a certain level of difficulty has been often observed in many LE studies [99, 100, 164]. However, this same trend was not observed for the other conditions. A longer experiment with more blocks would be helpful in understanding this disengagement effect better.

The lack of correlation between LE and alpha power, and the connection of alpha power and fatigue could mean that alpha power is an indicator of fatigue and not LE in the experiment that we performed. In general, the participant was fatigued as the experiment progressed and the fatigue was less evident for HS compared to OS and the synthetic enrichment outputs. The fatigue associated with OS reached to the extent that the participants may have disengaged from the task of listening to OS.

An evaluation of systems via listening tests is time consuming and cumbersome. However, it helps us know how our improvements are received by human listeners. In the end, the aim is that the laryngectomees and the people interacting with them benefit from the enrichments. Therefore, people's opinions in this case are crucial and valuable.

In the ASR and the STOI evaluations, BLSTMSS had better scores compared to our previous systems, as well as unprocessed OS. Therefore, the enriched outputs would be useful in improving human-machine interactions. As for human-human interactions, more research

would be needed to develop an enrichment system that appeals more to human listeners than unprocessed OS. Nonetheless, an improvement in SI and LE compared to previous enrichment systems suggests that we have been able to move some steps forward in the OS enrichment research.

7.6 Conclusions

The BLSTMSS system (developed as part of this thesis) outperformed our previous systems in STOI and ASR, as well as SI scores and subjective LE scores. When compared with OS, there was an improvement in ASR scores and STOI, but not in SI scores and subjective LE. EEG activity revealed more alpha power when listening to OS compared to HS indicating an effect of task difficulty. Additionally, alpha power increased as the experiment progressed indicating fatigue. No conclusions could be made based on the alpha power of enriched OS. Therefore, although there is an improvement in some objective measures, further improvements are needed to make the transformed OS more preferable and understandable to human listeners.

The findings of this chapter are in preparation to be published as a paper titled 'Evaluation of Enriched Oesophageal Speech: an EEG Study'.

Chapter 8

Conclusions

> "*Begin at the beginning, the King said gravely, "and go on till you come to the end: then stop."*
>
> —Lewis Carroll, *Alice in Wonderland*

I set foot on this journey of enrichments and evaluation of OS with some very clear and straightforward aims. It is when I got into the problem that the depths and complications of this task became apparent. All the obstacles and successes in this processes have taught me some important lessons, which will be helpful for me in research as well as any problem solving activity. I cannot say that I have totally solved the problem that I aimed to solve, but surely I have progressed a bit and gained a lot of knowledge. As progress is limitless, my journey with this problem could keep going on and on. But as the King says, the journey has come to an end and I must stop. Here I summarise all the findings of my thesis and prepare my baton to be passed to future researchers who wish to take this journey forward.

8.1 A Recap of the Problem Statement and Aims

OS lacks the intelligibility and quality of HS due to physical abnormalities: absence of larynx and vocal folds and separation of trachea and oesophagus. The lack of intelligibility and quality also increases the cognitive load when listening to OS. These issues impede OS speakers from social and familial engagements, expressing their needs, using voice-controlled digital devices and in general, navigating this world verbally.

The main aim of this thesis was to enrich OS signals by increasing intelligibility and quality and reducing LE. The other aim was to decide and implement appropriate metrics to evaluate OS and enriched OS.

8.2 Methods Used and Takeaways

A wide range of methodology was used to do the research in this thesis. All these methods have expanded my knowledge base and enabled me in making better research decisions.

A thorough understanding of OS was gained by studying the OS database. The variety of data and some acoustical measures of that data present in the database helped me a great deal in understanding the characteristics of OS.

For enrichment of OS, the main methods used were machine learning and other statistical learning tools. Machine learning is being used increasingly in a wide variety of industries such as healthcare, business, entertainment, social media, transport and other important industries. The success of machine learning in these fields is encouraging and intriguing. Through this project, I was able to gain some understanding of the inner workings of machine learning and neural networks such as learning rates, training data, normalisation, optimisation algorithms and so on. In the quest of finding a DNN system that could enrich OS, I went through several of these DNN terminologies as well as new and upcoming architectures that have shown promising results. In the case of this thesis, a DNN based method was found to be the most effective in achieving the goals of enrichment. Apart from DNNs and machine learning, a great deal of signal processing knowledge was applied too. This was essential for the understanding of speech characteristics such as fundamental frequency and formants as well as during EEG signal processing for an effective treatment of the signal. My main takeaway was that to solve any problem it is important to thoroughly understand the data in hand and have updated information about the tools available to work on that data.

A great deal of expertise on software languages such as matlab and python was needed and acquired during the processes of this project. Knowing how to program in these languages as well as knowing how to use programs written by other collaborators in these languages was crucial in achieving the goals of this project.

Statistical analysis was a key part of interpreting data and drawing conclusions. Providing statistical analyses made the results scientifically sound and credible. Over the course of the project I gained a lot of knowledge about various analysis tools and methods such as parametric and non-parametric tests, ANOVAs and t-tests and open source statistical tools such as JASP. I learned the importance of applying the right kind of statistical analysis on any given data set. Although I have not come out of this project as a statistics sage, I can say that I have made a few dips into the pool of statistical wisdom. And surely they have been great tools to make sense of the data that I have collected.

As this is a study of speech and LE, listening tests were another major indispensable part of

the research. The only true and correct way to know listeners' opinions and their response to the modifications we made was by conducting listening tests. I came across several factors that are crucial while conducting listening tests such as conducting pilot tests before the actual test. Some of these are choosing your participant demographic carefully, preparing for a no-show of participants, meticulous storage of listener data, deciding the mode of the listening test (lab or web-based), avoiding glitches in the listening test software (especially the ones that do not save data properly), maintaining anonymity of participants and ethical requirements and so on. Experiment design i.e the number and type of conditions and stimuli to be used, the manner of presentation (randomised, avoiding memory effect and priming), usage of right software tool and audio setup, length of the experiment and experiment breaks and other such factors are important in a successful listening experiment. To put it simply, conducting a listening test is like a grand orchestra where each part (small or big) plays an important role in the final result.

An unexpected course in this project was the use of a biomedical tool such as EEG to measure LE. This decision came in midway through the thesis and as an engineer, I was completely unprepared to conduct an EEG experiment. However, the experience and insights I have gained from this method were mind blowing and the biggest learning points. The part objective part subjective nature of this method leaves a lot of room for continually thinking about the outcomes. Nonetheless, the fact that this data came from an actual human being makes it very interesting and valuable. Fitting a participant with an EEG cap is a time consuming and meticulous process. After testing around 50 participants, I feel I am prepared to conduct another EEG experiment.

EEG signals interpretation and analysis is a vast world with a lot of possibilities. Most of the EEG research I have come across use a standard ensemble of tools to interpret and analyse EEG data. With the aid of my supervisors who are adept at signal processing, I was able to look at the EEG data with more depth. For example, I tried different window lengths and windowing techniques while extracting spectral characteristics of EEG. I found a way to analyse the EEG data corresponding to the entire length of the signal as opposed to just a fixed length of the signal, which is the usual practice. I tried to understand the inner workings of the independent component analysis process which is a sub-process of EEG data analysis. There was so much more that I could explore in the EEG experiments, but I was constrained by time. In sum, I appreciate the chance to study EEG signals in the context of disordered speech perception and gain several insights from it. I hope this tool is used in the future by researchers in this field.

8.3 Research Outcomes

Chapters 4 to 7 described all the experiments performed as part of this thesis. Each experiment revealed some interesting insights that were useful in shaping the next study and hopefully also to the readers of this thesis for their future research ideas. Here I will summarise the outcomes of all the experiments performed in this thesis.

In Chapter 4, we performed two experiments that collect intelligibility and self-reported LE metrics for OS and HS speakers. These experiments revealed the gaps in intelligibility and LE between HS and OS. OS was more effortful to process and less intelligible compared to HS. Even when the intelligibility of OS was high, there was significantly more LE associated with OS compared to HS. One interesting effect was that of listeners' familiarity to OS. OS was not more intelligible to familiar listeners, but it was reportedly less effortful to process for familiar listeners compared to non-familiar listeners. Machine intelligibility was also poorer for OS compared to HS.

In Chapter 5 we explored EEG as a way to measure LE. The findings suggested that alpha power, a neural measure known to measure LE, was higher for OS compared to HS. However, this measure was not correlated with subjective LE. The reasons for this non-correlation is an area that can be explored further. An interesting observation was that alpha power was affected by individual cognitive capacities of the participants - the participants who had poorer working memory had higher alpha power.

In Chapter 6, we implemented a novel DNN-based voice conversion system which mapped OS to duration-matched SS. This system outperformed our previous systems in STOI scores and ASR scores as well as SI scores and LE obtained with a listening test. When compared with unprocessed OS, there was an improvement in ASR scores, STOI and subjective preference scores, but not in SI scores and LE. Therefore, although there is an improvement in objective measures, further improvements are needed to make the transformed OS more preferable and understandable to human listeners. Some other quick and light weight exploratory strategies provided some benefit for a low intelligibility OS speaker, but not for high intelligibility OS. As there are infinite possibilities for OS enrichment, several experiments remained either not properly explored, or not evaluated completely. Therefore, there is definitely a lot of space in this area to improve results further.

In Chapter 7, three chosen enrichments were evaluated using a wide range of evaluation metrics. This included subjective metrics such as LE ratings, behavioral measurements such as SI scores and reaction times, objective measurements such as ASR scores and STOI scores and alpha power. The evaluations reveal that the proposed novel enrichment method was successful

in improving machine intelligibility metrics compared to previous methods. However, there was not the same level of success in improving subjective outcomes. Like the experiment in Chapter 5, a higher alpha power was observed for OS compared to HS, but nothing conclusive could be said about the enrichment methods. Overall, there was a steady increase in alpha power as the experiment progressed indicating the participants' fatigue or reduced attention.

8.4 Future Directions

The main limitation of the work described in this thesis is that the enriched outputs did not outperform unprocessed OS in subjective evaluations. Although there was some success in the preference tests described in Chapter 6, the HSR and subjective LE scores suggest a win for OS compared to the enrichments. Therefore, investigating the reasons behind this and developing enrichments that would appeal to human listeners as well would be a possible and necessary future direction.

It would be interesting to explore listeners' EEG signals further in the context of disordered speech. Some questions that can be asked are: What do the other bands of the EEG signal (other than alpha band) tell about disordered speech? Can listeners' EEG signals reveal insights about other unexplored aspects of disordered speech such as prosody, rhythm and emotion recognition? A machine learning based analysis of EEG signals instead of the traditional ERP and frequency band analysis would also be an interesting area to explore.

In the future, it would be interesting to install this enrichment system as a face-to-face communication aid in a stand-alone device or a smartphone where 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 can 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.

8.5 Contributions

This thesis contributed to the research and development of systems that will aid laryngetomees speaking in OS and possibly speakers of other pathologies too. Some important contributions

are listed below.

- Database and manual labelling: While automatic phonetic labelling software exist to label large speech databases, they did not work well for OS. Therefore a set of manual labelling was done for one OS speaker. When the manual labels were used to evaluate the accuracy of a customised automatic labelling process. These improved automatic labels were useful in developing an OS-friendly ASR system as well as novel methods for enrichment.
- New enrichment techniques: A voice conversion based approach was used to enrich OS with a novel idea of using SS with matching durations as target. This eliminated the need for the alignment of the source OS signal to the target, which can be the cause of erroneous results.
- Exploration of non VC techniques for enrichment: Apart from VC, some non-VC approaches were explored such as removing unwanted artefacts and improving the spectral characteristics of the signal.
- Wavenet vocoder for Spanish: As part of one of the enrichment processes, a new high quality vocoder (Wavenet) was developed for Spanish. This vocoder can be used in future studies to generate speech from acoustic parameters of speech.
- Parallel SS data for each speaker: Also as part of the enrichment process, a set of parallel SS was created using the phone labels. This synthetic speech dataset which matches OS in duration can be useful for future OS enrichment studies.
- Behavioural data for OS, HS and enrichments for one speaker: An extensive set of evaluations were performed for one OS speaker and a control HS speaker. This included objective and subjective intelligibility measurements, LE measurements and preference tests.
- EEG data from 12 participants when listening to an OS speaker and an HS speaker.
- EEG data from 32 participants when listening to an OS speaker, an HS speaker and three OS enrichments.
- ASR, STOI and preference test data for OS and enrichments of 4 speakers.
- Subjective and objective LE (physiological) as an additional evaluation tool.
- Code for listening tests: An interface to conduct listening tests with sentence transcription and subjective measures was created. A separate code for extraction of results and calculate intelligibility (transcription errors) is available too.

8.6 Publications

8.6.1 Peer-reviewed Journal Papers

- Serrano, L., **Raman, S.**, Hernaez, I., Navas, E., Sanchez, J., Saratxaga, I. (2020). A Spanish Multispeaker Database of Esophageal Speech. Computer Speech and Language, Volume 66, March 2021, 101168. DOI: 10.1016/j.csl.2020.101168

- **Raman, S.**; Serrano, L.; Winneke, A.; Navas, E.; Hernaez, I. Intelligibility and Listening Effort of Spanish OS. Appl. Sci. 2019, 9, 3233. DOI: 10.3390/app9163233

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

8.6.2 Papers in Preparation

- A paper titled 'Oesophageal Speech and Effortful Listening: an EEG Study' based on the findings of Chapter 5.

- A paper titled 'Evaluation of Enriched Oesophageal Speech: an EEG Study' based on the findings of Chapter 7.

8.6.3 Peer-reviewed Conference Papers

- Serrano, L., **Raman, S.**, Tavarez, D., Navas, E., Hernaez, I. (2019) Parallel vs. Non-Parallel Voice Conversion for Esophageal Speech. Proc. Interspeech 2019, 4549-4553, DOI: 10.21437/Interspeech.2019-2194.

- Serrano, L., Tavarez, D., Sarasola, X., **Raman, S.**, Saratxaga, I., Navas, E., Hernaez, I. (2018) LSTM based voice conversion for laryngectomees. Proc. IberSPEECH 2018, 122-126, DOI: 10.21437/IberSPEECH.2018-26

- **Raman, S.**, Hernaez, I., Navas, E., Serrano, L. (2018) Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech. Proc. IberSPEECH 2018, 107-111, DOI: 10.21437/IberSPEECH.2018-23

8.6.4 Poster Presentations

- **Raman, S.**, Hernaez, I., Navas, E., Serrano, L. A Multifaceted Enrichment of Oesophageal Speech. ICA 2019 Conference Proceedings, pp. 5739-5741. DOI: 10.18154/RWTH-

CONV-239415

- **Raman, S.**, Winneke, A., Hernaez, I., Navas, E. (2020) Listening effort and oesophageal speech: An EEG study. SpiN 2020 9-10 January 2020, Tolouse, France. Abstract p.60-61.

- **Raman, S.**; Serrano, L.; Winneke, A.; Navas, E.; Hernaez, I. A Study of Intelligibility and Self Reported Listening Effort of Spanish Oesophageal Speech, Poster session at 2018 Speech Processing Courses in Crete Summer School.

8.7 Other Activities and Achievements

8.7.1 Awards

- 2nd Best Student Poster Award at 2018 Speech Processing Courses in Crete Summer School.

- Selected paper "Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech" in Iberspeech 2018 Conference to be part of "IberSPEECH 2018: Speech and Language Technologies for Iberian Languages" Special Issue in the Multidisciplinary Digital Publishing Institute Journal.

8.7.2 Workshop and Conference Attendances

- INTERSPEECH 2017, August, 20-24,2017, Stockholm, Sweden

- Challenges in Hearing Assistive Technology (CHAT-2017) Collocated with Interspeech 2017

- Workshop in Amsterdam (topic: The Application of Pupillometry in Hearing Science to assess Listening Effort, 7-8 September 2017)

- Summer School in Crete E4 (topic: Towards Intelligible and Conversational Speech Synthesis Engines, 24-28 July 2017)

- INTERSPEECH 2018, September, 2-6,2018, Hyderabad, India

- Speech Production Summer School collocated with INTERSPPECH 2018, 9-11 September, DAIICT Gandhinagar, India. Participated in 5 min PhD contest. Number of people reached:50

- Oral presentation 'Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech' at IberSpeech 2018 conference. November 2018, Barcelona, Spain. Number of people reached:200
- ICA 2019, Aachen, Sept 2019, attendance and poster presentation ""A Multifaceted Enrichment of Oesophageal Speech"" - Number of people reached: 80
- SpiN 2020, Tolouse, Jan 2020, attendance and poster presentation ""Listening Effort and Oesophageal Speech: An EEG Study""" Number of people reached:50
- ICASSP 2020, Barcelona, 4-8 May 2020, virtual attendance and demo presentation ""Enriched Speech for Effortless Listening"" CONFERENCE, Number of people reached: (virtual attendance, - 100 people approx"
- Course on Python Programming, Vitoria-Gasteiz, Spain, 23-24 October 2017
- 'Cognitive Effort and Speech Styles' workshop, Nijmegen, The Netherlands, 7-9 March, 2018
- 'Speech Modifications' workshop, The Hague, The Netherlands, 29-30 September 2018
- 'Hearing Impairment' Workshop, Oldenburg, Germany, 26 Feb-1 March 2019

8.7.3 Research Visits

- One month visit to the University of Crete, Greece to learn about new speech synthesis techniques (02/07/2018 to 27/07/2018).
- Three months visit to Fraunhofer Institute of Digital Media Technology, Oldenburg, Germany to learn techniques of recording EEG and to perform a pilot speech perception experiment (26/02/2019 to 17/05/2019).
- Two months visit to Basque Centre on Cognition, Brain and Language, San Sebastian, Spain to conduct a full-fledged speech perception experiment using EEG (16/12/2019 to 07/02/2020).

8.7.4 Public Engagement

- Demo presentation to the general public at The Royal Institution, London on the topic 'Easy speaking and effortless listening' (March 3 2020)
- Participation in project explainer video (`https://youtu.be/_2W52Y3IE_Y`).

- Participated in a video interview that aimed to motivate women to take up engineering and other STEM courses (`https://youtu.be/d8BDR6Bh4KU` and `https://youtu.be/XYrHjF1takw`).

Bibliography

[1] Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen. A non-intrusive short-time objective intelligibility measure. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5085–5089, 2017.

[2] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara Van Gog. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438, 2010.

[3] Jan-Niklas Antons, Robert Schleicher, Sebastian Arndt, Sebastian Möller, and Gabriel Curio. Too tired for calling? a physiological measure of fatigue caused by bandwidth limitations. In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 63–67. IEEE, 2012.

[4] Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[5] Martin J Ball. *An Introduction to Speech Disorders*. Routledge, 2021.

[6] Jon Barker, Emmanuel Vincent, Ning Ma, Heidi Christensen, and Phil Green. The pascal chime speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633, 2013.

[7] Robert Becker, Maria Pefkou, Christoph M Michel, and Alexis Georges Hervais-Adelman. Left temporal alpha-band activity reflects single word intelligibility. *Frontiers in Systems Neuroscience*, 7:121, 2013.

[8] Suzanne Bennett and Bernd Weinberg. Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech, Language, and Hearing Research*, 16(4):608–615, 1973.

[9] Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347, 2001.

[10] Giulia Borghini and Valerie Hazan. Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, 12:152, 2018.

[11] Teresa Cervera, José L Miralles, and Julio González-Àlvarez. Acoustical analysis of spanish vowels produced by laryngectomized subjects. *Journal of Speech, Language, and Hearing Research*, 44(5):988–996, 2001.

[12] Chen-Yu Chen, Wei-Zhong Zheng, Syu-Siang Wang, Yu Tsao, Pei-Chun Li, and Ying-Hui Lai. Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system. In *INTERSPEECH*, pages 4686–4690, 2020.

[13] John M Christensen and Bernd Weinberg. Vowel duration characteristics of esophageal speech. *Journal of Speech and Hearing Research*, 19(4):678–689, 1976.

[14] Sarah Colby and Bob McMurray. Cognitive and physiological measures of listening effort during degraded speech perception: Relating dual-task and pupillometry paradigms. *Journal of Speech, Language, and Hearing Research*, pages 1–26, 2021.

[15] Kimberley J Cote-Reschny and Megan M Hodge. Listener effort and response time when transcribing words spoken by children with dysarthria. *Journal of Medical Speech-Language Pathology*, 18(4):24–35, 2010.

[16] Walter L Cullinan, Catherine S Brown, and P David Blalock. Ratings of intelligibility of esophageal and tracheoesophageal speech. *Journal of Communication Disorders*, 19(3):185–195, 1986.

[17] Delphine Dahan and James S Magnuson. Spoken word recognition. In *Handbook of psycholinguistics*, pages 249–283. Elsevier, 2006.

[18] Arnaud Delorme and Scott Makeig. Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[19] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3893–3896. IEEE, 2009.

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

[21] Tuan Dinh, Alexander Kain, Robin Samlan, Beiming Cao, and Jun Wang. Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency. In *Proceedings of Interspeech 2020*, pages 4781–4785, 2020.

[22] Hironori Doi, Keigo Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Enhancement of esophageal speech using statistical voice conversion. 2009.

[23] Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models. *IEICE TRANSACTIONS on Information and Systems*, 93(9):2472–2482, 2010.

[24] Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Statistical approach to enhancing esophageal speech based on gaussian mixture models. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4250–4253. IEEE, 2010.

[25] Hironori Doi, Tomoki Toda, Keigo Nakamura, Hiroshi Saruwatari, and Kiyohiro Shikano. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):172–183, 2014.

[26] Thomas Drugman, Myriam Rijckaert, Claire Janssens, and Marc Remacle. Tracheoesophageal speech: A dedicated objective acoustic assessment. *Computer Speech & Language*, 30(1):16–31, 2015.

[27] Rob Drullman, Joost M Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064, 1994.

[28] Michael D'Zmura, Siyi Deng, Tom Lappas, Samuel Thorpe, and Ramesh Srinivasan. Toward eeg sensing of imagined speech. In *International Conference on Human-Computer Interaction*, pages 40–48. Springer, 2009.

[29] R EBU-Recommendation. Loudness normalisation and permitted maximum level of audio signals. *Citeseer*, 2011.

[30] Elhuyar. Aditu - el reconocedor del habla de elhuyar basado en inteligencia artificial y redes neuronales. `https://aditu.eus/`, Accessed October 2020.

[31] Charles W Eriksen. The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, 2(2-3):101–118, 1995.

[32] Daniel Erro, Inma Hernaez, Agustin Alonso, D García-Lorenzo, Eva Navas, Jianpei Ye, H Arzelus, Igor Jauk, Nguyen Quy Hy, Carmen Magariños, et al. Personalized synthetic voices for speaking impaired: Website and app. *16th Annual Conference of the International Speech Communication Association*, 2015.

[33] Daniel Erro, Inma Hernáez, Eva Navas, Agustın Alonso, Haritz Arzelus, Igor Jauk, Nguyen Quy Hy, Carmen Magarinos, Rubén Pérez-Ramón, M Sulír, et al. Zuretts: Online platform for obtaining personalized synthetic voices. *Proceedings of eNTERFACE*, pages 1178–1193, 2014.

[34] Daniel Erro, Iñaki Sainz, Iker Luengo, Igor Odriozola, Jon Sánchez, Ibon Saratxaga, Eva Navas, and Inma Hernáez. Hmm-based speech synthesis in basque language using hts. *Proceedings of FALA*, pages 67–70, 2010.

[35] Tiago H Falk, Chenxi Zheng, and Wai-Yip Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010.

[36] R Holly Fitch, Steve Miller, and Paula Tallal. Neurobiology of speech perception. *Annual Review of Neuroscience*, 20(1):331–353, 1997.

[37] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, Julien Tardieu, Cynthia Magnen, Pascal Gaillard, Xavier Aumont, and Christian Füllgrabe. Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research*, 60(9):2394–2405, 2017.

[38] B Garcia, Ibon Ruiz, and Amaia Méndez. Oesophageal speech enhancement using poles stabilization and kalman filtering. In *2008 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1597–1600. IEEE, 2008.

[39] Luis Serrano García, Sneha Raman, Inma Hernáez Rioja, Eva Navas Cordón, Jon Sanchez, and Ibon Saratxaga. A Spanish multispeaker database of esophageal speech. *Computer Speech & Language*, page 101168, 2020.

[40] N Gómez-Merino, F Gheller, G Spicciarelli, and P Trevisi. Pupillometry as a measure for listening effort in children: a review. *Hearing, Balance and Communication*, 18(3):152–158, 2020.

[41] Amy Jane Hall, Axel Winneke, and Jan Rennies-Hochmuth. *EEG alpha power as a measure of listening effort reduction in adverse conditions.* Universitätsbibliothek der RWTH Aachen, 2019.

[42] R Hannemann, Jonas Obleser, and Carsten Eulitz. Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, 1153:134–143, 2007.

[43] Anne Hauswald, Anne Keitel, Ya-ping Chen, Sebastian Rösch, and Nathan Weisz. Degradation levels of continuous speech affect neural speech tracking and alpha power differently. *European Journal of Neuroscience*, 2020.

[44] Mark S Hawley, Phil Green, Pam Enderby, Stuart Cunningham, and Roger K Moore. Speech technology for e-inclusion of people with physical disabilities and disordered speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[45] Elina Helander, Jan Schwarz, Jani Nurminen, Hanna Silen, and Moncef Gabbouj. On the impact of alignment on voice conversion performance. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[46] Candace Bourland Hicks and Anne Marie Tharpe. Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, 2002.

[47] Sven Hilbert, Tristan T Nakagawa, Patricia Puci, Alexandra Zech, and Markus Bühner. The digit span backwards task. *European Journal of Psychological Assessment*, 2014.

[48] Jens Hjortkjær, Jonatan Märcher-Rørsted, Søren A Fuglsang, and Torsten Dau. Cortical oscillations and entrainment in speech processing during working memory load. *European Journal of Neuroscience*, 51(5):1279–1289, 2020.

[49] Norman D Hogikyan and Girish Sethuraman. Validation of an instrument to measure voice-related quality of life (v-rqol). *Journal of Voice*, 13(4):557–569, 1999.

[50] Sandra Cavanaugh Holley, Jay Lerman, and Kenneth Randolph. A comparison of the intelligibility of esophageal, electrolaryngeal, and normal speech in quiet and in noise. *Journal of Communication Disorders*, 16(2):143–155, 1983.

[51] Valtteri Hongisto. A model predicting the effect of speech of varying intelligibility on work performance. *Indoor Air*, 15(6):458–468, 2005.

[52] Dee J Hubbard and Deanie Kushner. A comparison of speech intelligibility between esophageal and normal speakers via three modes of presentation. *Journal of Speech, Language, and Hearing Research*, 23(4):909–916, 1980.

[53] Adam Jacks, Katarina L Haley, Gary Bishop, and Tyson G Harmon. Automated speech recognition in adult stroke survivors: Comparing human and computer transcriptions. *Folia Phoniatrica et Logopaedica*, 71(5-6):286–296, 2019.

[54] Barbara H Jacobson, Alex Johnson, Cynthia Grywalski, Alice Silbergleit, Gary Jacobson, Michael S Benninger, and Craig W Newman. The voice handicap index (vhi) development and validation. *American Journal of Speech-Language Pathology*, 6(3):66–70, 1997.

[55] Parvaneh Janbakhshi, Ina Kodrasi, and Hervé Bourlard. Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6405–6409. IEEE, 2019.

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

[57] Herbert H Jasper. The international "10–20" system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–375, 1958.

[58] J. Jensen and C. H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

[59] Joshua Y Kim, Chunfeng Liu, Rafael A Calvo, Kathryn McCabe, Silas CR Taylor, Björn W Schuller, and Kaihang Wu. A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*, 2019.

[60] K. Kobayashi and T. Toda. sprocket: Open-source voice conversion software. 2018.

[61] Minako Koike, Noriko Kobayashi, Hajime Hirose, and Yuki Hara. Speech rehabilitation after total laryngectomy. *Acta Oto-Laryngologica*, 122(4):107–112, 2002.

[62] Birger Kollmeier and Matthias Wesselkamp. Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421, 1997.

[63] John Kominek, Tanja Schultz, and Alan W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68, 2008.

[64] Christian Kothe, David Medine, and Matthew Grivich. Lab streaming layer (2014). *URL: https://github. com/sccn/labstreaminglayer (visited on 02/01/2019)*, 2018.

[65] Melanie Krueger, Michael Schulte, Thomas Brand, and Inga Holube. Development of an adaptive scaling method for subjective listening effort. *The Journal of the Acoustical Society of America*, 141(6):4680–4693, 2017.

[66] Karl D Kryter. Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697, 1962.

[67] Evelyne Lagrou, Robert J Hartsuiker, and Wouter Duyck. The influence of sentence context and accented speech on lexical access in second-language auditory word recognition. *Bilingualism: Language and Cognition*, 16(3):508–517, 2013.

[68] Sophie Landa, Lindsay Pennington, Nick Miller, Sheila Robson, Vicki Thompson, and Nick Steen. Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology*, 16(4):408–416, 2014.

[69] Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 2020.

[70] Ulrike Lemke and Jana Besser. Cognitive load and listening effort: Concepts and age-related considerations. *Ear and Hearing*, 37:77S–84S, 2016.

[71] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10, No. 8:707–710, 1966.

[72] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32(3):35–52, 2015.

[73] Richard P Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[74] Hanjun Liu, Mingxi Wan, Supin Wang, Xiaodong Wang, and Chunmei Lu. Acoustic characteristics of mandarin esophageal speech. *The Journal of the Acoustical Society of America*, 118(2):1016–1025, 2005.

[75] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*, 2018.

[76] Andreas Maier, Tino Haderlein, Ulrich Eysholdt, Frank Rosanowski, Anton Batliner, Maria Schuster, and Elmar Nöth. Peaks–a system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437, 2009.

[77] Alfredo Mantilla, Héctor Pérez-Meana, Daniel Mata, Carlos Angeles, Jorge Alvarado, and Laura Cabrera. Recognition of vowel segments in spanish esophageal speech using hidden markov models. *15th International Conference on Computing*, pages 115–120, 2006.

[78] Kenji Matsui, Noriyo Hara, Noriko Kobayashi, and Hajime Hirose. Enhancement of esophageal speech using formant synthesis. *Acoustical Science and Technology*, 23(2):69–76, 2002.

[79] Megan J McAuliffe, Phillipa J Wilding, Natalie A Rickard, and Greg A O'Beirne. Effect of speaker age on speech recognition and perceived listening effort in older adults with hearing loss. *Journal of Speech, Language, and Hearing Research*, 55(3):838–47, 2012.

[80] Ronan McGarrigle, Kevin J Munro, Piers Dawes, Andrew J Stewart, David R Moore, Johanna G Barry, and Sygal Amitay. Listening effort and fatigue: What exactly are we measuring? a british society of audiology cognition in hearing special interest group 'white paper'. *International Journal of Audiology*, 53(7):433–440, 2014.

[81] Sharynne McLeod and Sadanand Singh. *Speech sounds: A pictorial guide to typical and atypical speech*. Plural Publishing, 2009.

[82] Catherine M McMahon, Isabelle Boisvert, Peter de Lissa, Louise Granger, Ronny Ibrahim, Chi Yhun Lo, Kelly Miles, and Petra L Graham. Monitoring alpha oscillations and pupil dilation across a performance-intensity function. *Frontiers in Psychology*, 7:745, 2016.

[83] Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. Silent speech recognition as an alternative communication device for

persons with laryngectomy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2386–2398, 2017.

[84] John E Meyers, Kurt Volkert, and Anh Diep. Sentence repetition test: Updated norms and clinical utility. *Applied Neuropsychology*, 7(3):154–159, 2000.

[85] Microsoft. Microsoft azure cognitive services speech-to-text. `https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/get-started-speech-to-text`, Accessed October 2020.

[86] Catherine Middag, Tobias Bocklet, Jean-Pierre Martens, and Elmar Nöth. Combining phonological and acoustic asr-free features for pathological speech intelligibility assessment. *12th Annual Conference of the International Speech Communication Association*, 2011.

[87] Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt. Dia: A tool for objective intelligibility assessment of pathological speech. *6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 165–167, 2009.

[88] Jose L Miralles and Teresa Cervera. Voice intelligibility in patients who have undergone laryngectomies. *Journal of Speech, Language, and Hearing Research*, 38(3):564–571, 1995.

[89] Seyed Hamidreza Mohammadi and Alexander Kain. An overview of voice conversion systems. *Speech Communication*, 88:65–82, 2017.

[90] E Ann Mohide, Stuart D Archibald, Michelle Tew, J Edward Young, and Trish Haines. Postlaryngectomy quality-of-life dimensions identified by patients and health care professionals. *The American Journal of Surgery*, 164(6):619–622, 1992.

[91] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[92] Tova Most, Yishai Tobin, and Ravit Cohen Mimran. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders*, 33(2):165–181, 2000.

[93] Murray J Munro. The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, 20(2):139–154, 1998.

[94] Kathleen F Nagle and Tanya L Eadie. Perceived listener effort as an outcome measure for disordered speech. *Journal of Communication Disorders*, 73:34–49, 2018.

[95] Kathy F Nagle and Tanya L Eadie. Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders*, 45(3):235–245, 2012.

[96] Aljoscha C Neubauer, Andreas Fink, and Roland H Grabner. Sensitivity of alpha band erd to individual differences in cognition. *Progress in Brain Research*, 159:167–178, 2006.

[97] Paul L. Nunez and Ramesh Srinivasan. Electroencephalogram. *Scholarpedia*, 2(2):1348, 2007. revision #91219.

[98] Jonas Obleser, Malte Wöstmann, Nele Hellbernd, Anna Wilsch, and Burkhard Maess. Adverse listening conditions and memory load drive a common alpha oscillatory network. *Journal of Neuroscience*, 32(36):12376–12383, 2012.

[99] Barbara Ohlenforst, Dorothea Wendt, Sophia E Kramer, Graham Naylor, Adriana A Zekveld, and Thomas Lunner. Impact of snr, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hearing Research*, 365:90–99, 2018.

[100] Barbara Ohlenforst, Adriana A Zekveld, Thomas Lunner, Dorothea Wendt, Graham Naylor, Yang Wang, Niek J Versfeld, and Sophia E Kramer. Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, 351:68–79, 2017.

[101] Ibon Oleagordia-Ruiz and Begonya Garcia-Zapirain. Harmonic to noise ratio improvement in oesophageal speech. *Technology and Health Care*, 23(3):359–368, 2015.

[102] Imen Ben Othmane, Joseph Di Martino, and Kaïs Ouni. Enhancement of esophageal speech obtained by a voice conversion technique using time dilated fourier cepstra. *International Journal of Speech Technology*, 22(1):99–110, 2019.

[103] Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, 162(1-2):8–13, 2007.

[104] M Kathleen Pichora-Fuller, Sophia E Kramer, Mark A Eckert, Brent Edwards, Benjamin WY Hornsby, Larry E Humes, Ulrike Lemke, Thomas Lunner, Mohan Matthen, Carol L Mackersie, et al. Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel). *Ear and Hearing*, 37:5S–27S, 2016.

[105] Eduard Polityko. Word error rate. *https://www.mathworks.com/examples/matlab/community/19873-word-error-rate, access date: 20th February 2018*, 2018.

[106] Gerasimos Potamianos and Chalapathy Neti. Automatic speechreading of impaired speech. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*, 2001.

[107] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[108] DA Preece. Latin squares, latin cubes, latin rectangles. *Wiley StatsRef: Statistics Reference Online*, 2014.

[109] Sneha Raman, Inma Hernaez, Eva Navas, and Luis Serrano. Listening to laryngectomees: A study of intelligibility and self-reported listening effort of spanish oesophageal speech. *Proceedings of IberSPEECH 2018*, pages 107–111, 2018.

[110] Sneha Raman, Inma Hernaez, Eva Navas, and Luis Serrano. *A multifaceted enrichment of oesophageal speech.* Universitätsbibliothek der RWTH Aachen, 2019.

[111] Sneha Raman, Xabier Sarasola, Eva Navas, and Inma Hernaez. Enrichment of oesophageal speech: Voice conversion with duration–matched synthetic speech as target. *Applied Sciences*, 11(13):5940, 2021.

[112] Sneha Raman, Luis Serrano, Axel Winneke, Eva Navas, and Inma Hernaez. Intelligibility and listening effort of spanish oesophageal speech. *Applied Sciences*, 9(16):3233, 2019.

[113] Shakti P Rath, Daniel Povey, Karel Veselỳ, and Jan Cernockỳ. Improved feature processing for deep neural networks. *14th Annual Conference of the International Speech Communication Association*, pages 109–113, 2013.

[114] Gerard B Remijn, Mitsuru Kikuchi, Yuko Yoshimura, Kiyomi Shitamichi, Sanae Ueno, Tsunehisa Tsubokawa, Haruyuki Kojima, Haruhiro Higashida, and Yoshio Minabe. A near-infrared spectroscopy study on cortical hemodynamic responses to normal and whispered speech in 3-to 7-year-old children. *Journal of Speech, Language, and Hearing Research*, 60(2):465–470, 2017.

[115] Jan Rennies, Henning Schepker, Inga Holube, and Birger Kollmeier. Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *The Journal of the Acoustical Society of America*, 136(5):2642–2653, 2014.

[116] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

[117] Jana Roßbach, Saskia Röttges, Christopher F Hauth, Thomas Brand, and Bernd T Meyer. Non-intrusive binaural prediction of speech intelligibility based on phoneme classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 396–400. IEEE, 2021.

[118] Marinela Rosso, Ljiljana Širić, Robert Tićac, Radan Starčević, Igor Šegec, and Nikola Kraljik. Perceptual evaluation of alaryngeal speech. *Collegium Antropologicum*, 36(2):115–118, 2012.

[119] I Sainz, D Erro, E Navas, I Hernáez, J Sánchez, I Saratxaga, I Odriozola, I Luengo, et al. Aholab speech synthesizers for albayzin2010. *Proceedings of FALA*, 2010:343–348, 2010.

[120] Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sanchez, Ibon Saratxaga, and Igor Odriozola. Versatile speech databases for high quality synthesis for basque. In *LREC*, pages 3308–3312. Citeseer, 2012.

[121] Iñaki Sainz, Daniel Erro, Eva Navas, Inmaculada Hernáez, J Sanchez, I Saratxaga, and Igor Odriozola. Versatile Speech Databases for High Quality Synthesis for Basque. *8th international Conference on Language Resources and Evaluation (LREC)*, pages 3308–3312, 2012.

[122] Odette Scharenborg. Reaching over the gap: A review of efforts to link human and automatic speech recognition r'esearch. *Speech Communication*, 49(5):336–347, 2007.

[123] Luis Serrano. *Técnicas para la mejora de la inteligibilidad en voces patológicas*. PhD thesis, University of the Basque Country (UPV/EHU), 2019.

[124] Luis Serrano, Sneha Raman, David Tavarez, Eva Navas, and Inma Hernaez. Parallel vs. non-parallel voice conversion for esophageal speech. *Proceedings of Interspeech 2019*, pages 4549–4553, 2019.

[125] Luis Serrano, David Tavarez, Igor Odriozola, Inma Hernaez, and Ibon Saratxaga. Aholab system for albayzin 2016 search-on-speech evaluation. *Proceedings of IberSPEECH*, pages 33–42, 2016.

[126] Luis Serrano, David Tavarez, Xabier Sarasola, Sneha Raman, Ibon Saratxaga, Eva Navas, and Inma Hernaez. LSTM based voice conversion for laryngectomees. In *Proceedings of IberSPEECH 2018*, pages 122–126, 2018.

[127] Luis Serrano, David Tavarez, Xabier Sarasola, Sneha Raman, Ibon Saratxaga, Eva Navas, and Inma Hernaez. Lstm based voice conversion for laryngectomees. *Proceedings of IberSPEECH*, pages 122–126, 2018.

[128] A. Sesma and Asunción Moreno. Corpuscrt 1.0: Diseno de corpus orales equilibrados. *Computer Program]: http://gps-tsc. upc. es/veu/personal/sesma/CorpusCrt. ph p3*, 2000.

[129] Hamid Reza Sharifzadeh, Ian V McLoughlin, and Farzaneh Ahmadi. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *IEEE Transactions on Biomedical Engineering*, 57(10):2448–2458, 2010.

[130] Dushyant Sharma, Yu Wang, Patrick A Naylor, and Mike Brookes. A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication*, 80:84–94, 2016.

[131] Kåre Sjölander and Jonas Beskow. Wavesurfer-an open source speech tool. In *Sixth International Conference on Spoken Language Processing*. Citeseer, 2000.

[132] Constantin Spille, Birger Kollmeier, and Bernd T Meyer. Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language*, 52:123–140, 2018.

[133] Smiljka Štajner-katušić, Damir Horga, Maja Mušura, and Dubravka Globlek. Voice and speech after laryngectomy. *Clinical Linguistics & Phonetics*, 20(2-3):195–203, 2006.

[134] Herman JM Steeneken. The measurement of speech intelligibility. *Proceedings of Institute of Acoustics*, 23(8):69–76, 2001.

[135] Antje Strauß, Malte Wöstmann, and Jonas Obleser. Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience*, 8:350, 2014.

[136] Daniel J Strauss and Alexander L Francis. Toward a taxonomic model of attention in effortful listening. *Cognitive, Affective, & Behavioral Neuroscience*, 17(4):809–825, 2017.

[137] Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016.

[138] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[139] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.

[140] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. Speaker-dependent wavenet vocoder. In *INTERSPEECH*, pages 1118–1122, 2017.

[141] JASP Team. Jasp (version 0.14)[computer software] `https://jasp-stats.org/`, 2020.

[142] C. Theys and M. McAuliffe. Listening to disordered speech results in early modulations of auditory event-related potentials. In *Groningen, The Netherlands: 7th International Conference on Speech Motor Control. 5/7/2017-8/7/2017. Stem-, Spraak- en Taalpathologie.*, 2017.

[143] Cristina Tiple, Silviu Matu, Florina Veronica Dinescu, Rodica Muresan, Radu Soflau, Tudor Drugan, Mircea Giurgiu, Adriana Stan, Daniel David, and Magdalena Chirila. Voice-related quality of life results in laryngectomies with today's speech options and expectations from the next generation of vocal assistive technologies. In *2015 E-Health and Bioengineering Conference (EHB)*, pages 1–4. IEEE, 2015.

[144] Noe Tits. Exploring the parameters describing the quality and intelligibility of alaryngeal voices. *Master Thesis, University of Mons*, 2017.

[145] Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1249. IEEE, 2007.

[146] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[147] Rene L Utianski, John N Caviness, and Julie M Liss. Cortical characterization of the perception of intelligible and unintelligible speech measured via high-density electroencephalography. *Brain and Language*, 140:49–54, 2015.

[148] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. In *SSW*, volume 125, 2016.

[149] Kristin J Van Engen and Jonathan E Peelle. Listening effort and accented speech. *Frontiers in Human Neuroscience*, 8:577, 2014.

[150] Steven Van Kuyk, W Bastiaan Kleijn, and Richard Christian Hendriks. An evaluation of intrusive instrumental intelligibility metrics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2153–2166, 2018.

[151] Dianne J Van Tasell and Jeilry L Yanz. Speech recognition threshold in noise: effects of hearing loss, frequency response, and speech materials. *Journal of Speech, Language, and Hearing Research*, 30(3):377–386, 1987.

[152] Mahesh Viswanathan and Madhubalan Viswanathan. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer Speech & Language*, 19(1):55–83, 2005.

[153] Vibha Viswanathan, Hari M Bharadwaj, and Barbara G Shinn-Cunningham. Electroencephalographic signatures of the neural representation of speech during selective attention. *Eneuro*, 6(5), 2019.

[154] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[155] Bernd Weinberg. Acoustical properties of esophageal and tracheoesophageal speech. *Laryngectomee Rehabilitation*, pages 113–127, 1986.

[156] Bernd Weinberg, Yoshiyuki Horii, and Bonnie E Smith. Long-time spectral and intensity characteristics of esophageal speech. *The Journal of the Acoustical Society of America*, 67(5):1781–1784, 1980.

[157] Tara L Whitehill and Christy C-Y Wong. Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology*, 14(4):335–342, 2006.

[158] Conor J Wild, Afiqah Yusuf, Daryl E Wilson, Jonathan E Peelle, Matthew H Davis, and Ingrid S Johnsrude. Effortful listening: The processing of degraded speech depends critically on attention. *Journal of Neuroscience*, 32(40):14010–14021, 2012.

[159] Richard H Wilson and MS Jaffe. Interactions of age, ear, and stimulus complexity on dichotic digit recognition. *Journal of the American Academy of Audiology*, 7(5):358–364, 1996.

[160] Matthew B Winn, Jan R Edwards, and Ruth Y Litovsky. The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, 36(4):e153, 2015.

[161] Axel H Winneke, Michael Schulte, Matthias Vormann, and Matthias Latzel. Effect of directional microphone technology in hearing aids on neural correlates of listening and memory effort: An electroencephalographic study. *Trends in Hearing*, 24:2331216520948410, 2020.

[162] Matthew G Wisniewski, Alexandria C Zakrzewski, Destiny R Bell, and Michelle Wheeler. Eeg power spectral dynamics associated with listening in adverse conditions. *Psychophysiology*, 58(9):e13877, 2021.

[163] Malte Wöstmann, Björn Herrmann, Anna Wilsch, and Jonas Obleser. Neural alpha dynamics in younger and older listeners reflect acoustic challenges and predictive benefits. *Journal of Neuroscience*, 35(4):1458–1467, 2015.

[164] Yu-Hsiang Wu, Elizabeth Stangl, Xuyang Zhang, Joanna Perkins, and Emily Eilers. Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing*, 37(6):660, 2016.

[165] Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An open source neural network speech synthesis system. In *SSW*, pages 202–207, 2016.

[166] Seung Hee Yang and Minhwa Chung. Improving dysarthric speech intelligibility using cycle-consistent adversarial training. *arXiv preprint arXiv:2001.04260*, 2020.

[167] Kathryn M Yorkston and David R Beukelman. A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders*, 11(6):499–512, 1978.

[168] Deokgyu Yun, Hannah Lee, and Seung Ho Choi. A deep learning-based approach to non-intrusive objective speech intelligibility estimation. *IEICE TRANSACTIONS on Information and Systems*, 101(4):1207–1208, 2018.

Acronyms

ERP Event Related Potentials. 48, 95, 117, 132

ACALES Adaptive Categorical Listening Effort Scaling. 17, 41, 77, 117

AI Articulation Index. 11, 12, 117

ALS Amyotrophic Lateral Sclerosis. 14, 117

ANOVA Analysis of Mean and Variance. 36, 38, 43, 44, 84, 85, 117

ASR Automatic Speech Recognition. xi, xv–xvii, xix, 9, 11–14, 22, 23, 27–29, 32, 33, 35, 36, 38, 39, 45, 46, 58, 59, 61–64, 68, 69, 71–73, 75, 76, 86, 88, 89, 94, 96, 117, 140–143

BAP Band Aperiodicity Parameter. 60, 117

BLSTM Bidirectional Long Short-Term Memory. 117

BLSTMHS BLSTM with Healthy Speech as target. xvii, 71, 72, 76, 77, 82–84, 86–88, 117

BLSTMSS BLSTM with Synthetic Speech as target. xvi, xvii, 59–68, 71–73, 76, 77, 82–84, 86–89, 117

CMOS Comparative Mean Opinion Score. 14, 22, 117

CMVN Cepstral Mean and Variance Normalisation. 35, 117

DNN Deep Neural Networks. x, 12, 21, 22, 57, 58, 68, 69, 76, 92, 94, 117

DTW Dynamic Time Warping. 58, 59, 117

EEG Electroencephalogram. x–xii, xvii, 18, 19, 40–43, 47–52, 76–78, 80, 81, 84, 88, 89, 92–96, 98, 99, 117, 125–132

ELS Electrolaryngeal Speech. 8, 20, 22, 117

ESTOI Extended Short Term Objective Intelligibility. 11, 12, 117

fMRI functional Magentic Resonance Imaging. 18, 117

GAN Generative Adversarial Networks. 22, 117

GMM Gaussian Mixture Models. 21, 22, 71–73, 117

HMM Hidden Markov Models. 60, 117

HNR Harmonics-to-Noise Ratio. 11, 22, 117

HS Healthy (Laryngeal) Speech. x, xv–xvii, xix, 3, 4, 8–12, 14, 20–23, 26, 27, 29, 31–41, 44–54, 57, 58, 60, 61, 70, 71, 73, 75–77, 81–84, 86, 88, 89, 91, 94–96, 117

HSR Human Speech Recognition. xv, 9, 11, 14, 23, 32, 36–39, 45, 46, 95, 117

KS Kolmogorov–Smirnov. 37, 41, 117

LE Listening Effort. xv–xvii, xix, 15–20, 23, 26, 31–33, 35–55, 57, 73, 75–78, 82–85, 87–89, 91–96, 117

LSTM Long Short-Term Memory. 22, 97, 117, 141

MCC Mel Cepstral Coefficients. 60, 117

MCD Mel Cepstral Distortion. 13, 21–23, 117

MFCC Mel-Frequency Cepstral Coefficients. 35, 117

MOS Mean Opinion Score. 13, 21–23, 75, 117

NIRS Near-infrared Spectroscopy. 48, 117

OOV Out of Vocabulary. 35, 117

OS Oesophageal Speech. ix–xi, xv–xvii, xix, 3–5, 7–11, 14, 15, 17, 19–23, 25–29, 31–41, 43–55, 57–73, 75–77, 79, 81–89, 91, 92, 94–97, 117

PESQ Perceptual Evaluation of Speech Quality. 13, 117

PET Positron Emission Tomography. 18, 117

PPG Phonetic Posteriorgrams. xvii, 22, 76, 77, 82–84, 86, 87, 117

PSD Power Spectral Density. 50, 78, 117

PSM Perceptual Similarity Measure. 117

PWC Percentage Words Correct. xvi, 13, 14, 62, 63, 68, 81, 117

RT Response Time. 117

SD Standard Deviation. 41, 59, 117

SI Speech Intelligibility. xv–xvii, 11, 14, 35, 40, 42, 49, 52, 54, 76, 77, 81–85, 87–89, 94, 117

SNR Signal-to-Noise Ratio. 11, 117

SPL Sound Pressure Level. 41, 117

SRT Speech Reception Threshold. 13, 14, 117

SS Synthetic Speech. xvi, xvii, 28, 58–63, 65, 66, 68, 69, 75, 87, 94, 96, 97, 117

STEM Science, Technology, Engineering and Mathematics. 117

STI Speech Transmission Index. 11, 12, 117

STOI Short Term Objective Intelligibility. xi, xvi, xvii, 11, 12, 14, 59, 61, 63–66, 68, 69, 73, 75, 87–89, 94, 96, 117

TOS Tracheoesophageal Speech. 8, 14, 20, 22, 25, 61, 117

V-RQOL Voice-related Quality of Life. 20, 117

VC Voice Conversion. xvi, 13, 20–23, 57–61, 68, 69, 71–73, 87, 96, 117

VHI Voice Handicap Index. 20, 117

WER Word Error Rate. xv–xvii, xix, 12–14, 29, 33, 36–40, 43–46, 50–52, 59, 61–63, 68, 69, 71, 72, 76, 86, 117

WM Working Memory. 48, 117

Appendix A

30 sentences Used in the Experiment in Chapter 4

1. Una fiesta en Florida Park con glamur
2. De Filadelfia vino el grupo Judíos por una Paz Justa
3. Ello h acía intuir un duelo en toda regla
4. Deja mucha buena obra hecha pero me rehuye el balance
5. Hoy jueves dieciocho de julio de dos mil trece
6. Ha podido reunirse con Chávez tras su elección
7. Quizá ustedes no lo advirtieron por eso lo refiero ahora
8. Abdulá Abdulá ministro de Exteriores va más allá
9. Da igual no importa de dónde extraiga uno la emoción
10. Sólo el chileno Mark González ponía una pizca de orgullo
11. Pero usted ya conoce por dentro el mundo del cine
12. Hay voces que ya hablan de indulto sería factible
13. Lo que no cree nadie aquí es que cacen a Sadam con vida
14. Unos días de euforia y meses de atonía
15. Blasco Ibáñez hizo alguna vez la misma cosa

16. Tenía la voz alborotada y la amistad ruidosa

17. Apliqué el oído a esta rayita y percibí un murmullo

18. Gastó todo el agua incluyendo el agua de las lluvias

19. Si el club no hubiera cambiado se hubiera ido

20. Goliat estuvo a punto de engullir a David

21. Tal vez fue hace siglos o acaso hace tan sólo unas décadas

22. Aún no sabemos qué fue a hacer a Taiwan

23. El pueblo noruego rechazó vía referéndum la adhesión

24. Con este álbum llegará seguro al número uno de ventas

25. Mi aldea estaba a la orilla de un riachuelo como éste

26. Fui yo por consejo del señor Regueiro Souza

27. Hoy no juega al golf y el traje es azul cielo y oro

28. Occidente y el islam son dos miedos que se acechan

29. Núñez ya tiene a su hijo predilecto en casa

30. Qué diferencia hay entre el caucho y la hevea

Appendix B

EEG Data Terminology and Procedures

B.1 EEG Recording Equipment

B.1.1 Cap and Electrodes

An EEG cap is a stretchable head covering with holes (for electrodes) that fit the subject's head snugly. As the head sizes of subjects vary, there are different sizes of caps available, typically: 54cm, 56 cm, 58cm and 60cm. What size cap fits a subject is decided by measuring the head circumference with a measuring tape passing through the forehead. It is important to provide some tolerance as the cap should not be too tight. For example, a person with head circumference measuring 57cm would be fitted with a 58 cm cap. In addition, it must be ensured that the cap is placed in the right position. For this, the centre-most point of the cap is matched to the centre of the head, which is midway between the centre forehead (nasion) and the natural bump at the centre back (inion), as well as midway between the 2 ears. The caps may have 32, 64 or 128 holes depending on the type and requirement.

The electrodes are arranged on an EEG cap as per the International 10-20 system for measuring EEG. It is known as the 10-20 system because the distance between any two adjacent electrodes is either 10% or 20% of the front-back or right left distance.

The position and the naming convention of the electrodes follows a [brian area][direction code] system. The brain areas are Parietal (P), Frontal(F), Central(C), Temporal(T) and Occipetal(O). Intermediate or overlapping areas are named as Centro-parietal (CP), Front-central (FC), Tempero-parietal (TP), Front-Temporal(FT), Parietal-occipetal (PO) and so on.

The directions codes are z for the front-back midline area, odd numbers (1,3,5) for the left of the midline and even numbers (2,4,6) for the right of the midline. Numbers are assigned incrementally from front to back. Based on these conventions, the electrodes are named as F1 (frontal left), P6 (Parietal Right), CPz (centro-parietal midline) etc.

Apart from the above-mentioned electrodes, there are reference and ground electrodes. The EEG signal is a potential difference between the scalp potential and a reference potential. This reference potential is obtained by placing an electrode away from the scalp, on a relatively neutral area such as behind the ears (mastoids) or on the nose. As these areas are not impacted by brain activity, a difference of the scalp electrode potential to this reference potential results in the potential difference associated with the brain activity in the said scalp electrode. A ground electrode is usually placed on the cap and its function is mainly to remove power line noise.

In addition, some electrodes may be placed on the temples and forehead to record ocular activity. These electrodes can then later be used to remove ocular activity artefacts.

B.1.2 Gel

An electrolyte gel is applied (with needle-less syringes) where the electrodes come in contact with the scalp. The function of this gel is to increase the electrical connectivity of the scalp to the electrode. The gel from one electrode site must not seep through to another electrode site as it would create a short circuit which is undesirable. Sometimes the gel may contain a slightly abrasive or exfoliating component that helps in scraping off the dead skin cells in the top layer of the scalp. This also improves the connectivity.

B.1.3 Amplifier

The EEG amplifier amplifies the EEG signal and also converts it from digital to analogue. This is so that the recording software on the computer can receive and store the EEG signals digitally and amplified. The sampling rate of the signals are typically 250 to 2000 Hz.

B.1.4 Recording software

While placing electrodes and recording EEG signals, it is helpful to see the status in real time. This helps in trouble shooting and ensuring good quality of data. In a recording software (such as BrainVision), we can visualise the electrodes and their impedance while mounting the cap and electrodes (See Figure B.1). The aim in this stage is for all the electrodes to have a low impedance (preferably $< 5k\Omega$). Additionally, in the recording software interface, it is possible

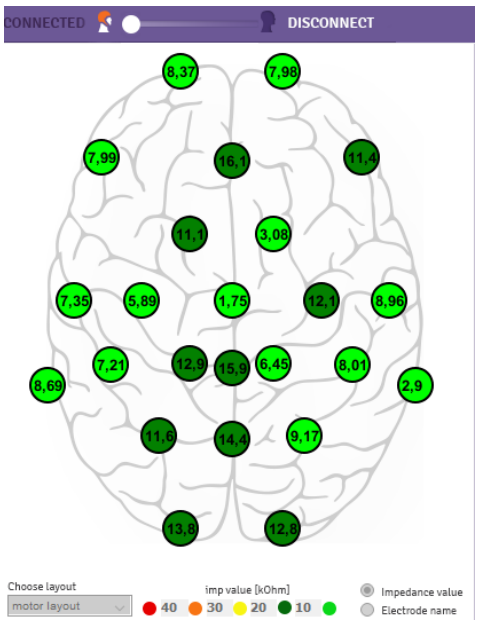


Figure B.1: EEG recording software showing impedance values of electrodes.

to see the EEG signals in real time and that can help the experimenter identify noisy channels, restless or tired participants etc.

B.2 Synchronisation

While conducting an EEG experiment, the experiment software sends out a trigger to the EEG recording software when the audio starts playing. However it is possible that there are a few milliseconds of delay between the experiment software sending the trigger and the EEG recording software receiving those triggers. It is necessary to ensure that the EEG markers (or triggers) and the audio stimuli start points are synchronised. This synchronisation may be achieved using hardware or manually. In the hardware method, a clock device synchronises the audio input and the EEG markers. Alternatively, it may be done manually by identifying the delay between the playing of the audio signal and the start of EEG recording and then adding the appropriate amount of delay later to synchronise the EEG and audio channels.

B.3 EEG Recording Process

Once the ground, reference and all the data electrodes are placed correctly with the right impedance, ocular and muscular activity are checked to unsure that the EEG data are being recorded correctly. This can be observed by spikes in the data when the participant blinks or dense activity when the participant clenches their jaw or bites.

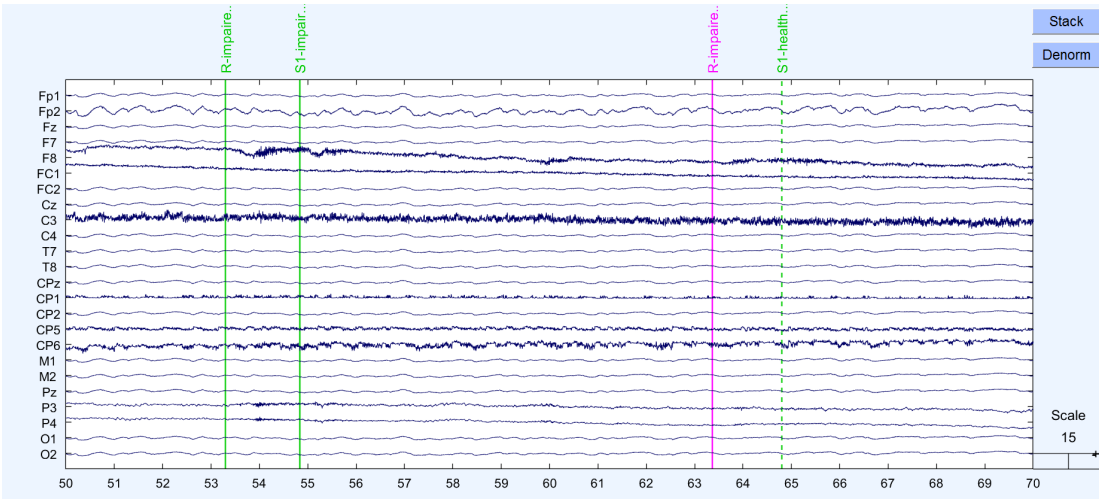


Figure B.2: Raw EEG data

In case there are noisy electrodes, they must be fixed by either putting more gel or correcting other connectivity issues. The signals must be constantly monitored to avoid noisy data. To avoid noisy data from the participant. they are instructed to be as still as possible and to look straight on to the screen.

When the participant is ready and all the electrodes look fine, we can start recording the EEG data. Then the experiment software which displays the controls for the behavioural data and plays the audio files is run. Once the experiment is over, the EEG data recording is stopped and the EEG file is appropriately saved. There are several different formats of EEG data such as .eeg, .xdf etc.

B.4 Raw EEG Data

Raw EEG data is difficult to process as it is and needs to undergo several steps of preprocessing to extract meaningful data from it. To view and process these raw EEG data files, an EEG processing software is used. One such software is the EEGLab software.

When the EEG file is loaded in EEGLab, we can first observe the raw EEG data. This is a time series recorded by each one of the electrodes. Figure B.2 shows the raw EEG data from a participant. The vertical axis is the electrode name and the horizontal axis is time. The vertical colourful lines are the triggers or the event markings where the participant started hearing a stimulus or performed some action. It can be observed that the data is quite noisy.

Each data series is assigned a channel location which is the two-dimensional or three-dimensional location of that electrode on the head. The channel locations are obtained from the manual of the EEG cap set. Because of this step, it is possible to view the electrode data

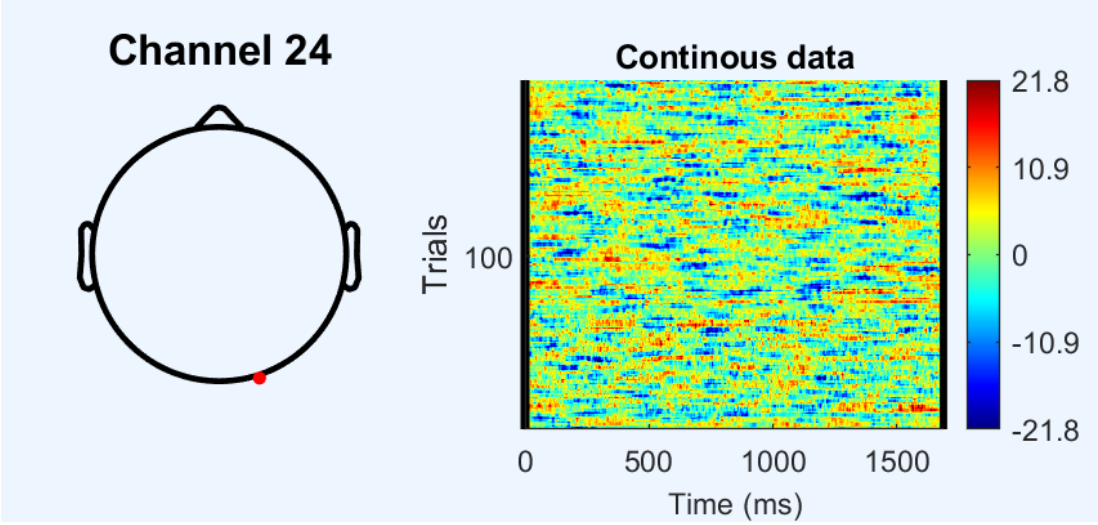


Figure B.3: Raw EEG data

in a 2-D or 3-D topographical plot. For example, in Figure B.3, we can see the topographical representation of channel 24, which is the O2 (occipetal right) electrode. The red dot on the topographical plot shows the position of the O2 electrode on the head.

Some possible artefacts and noise we can notice in the raw EEG data are noisy channels (such as channel C3 in Figure B.2). The occasional dips in the Fp1 and Fp2 channels correspond to the participant blinking. This will be clearer in the data post the filtering process.

Rereferencing is the process of getting the potential difference between the electrodes and the reference electrode. Typically, the reference electrodes are the mastoid electrodes (the electrodes behind the ears). It may be one mastoid electrode or an average potential of the two mastoid electrodes. Alternatively, it is also possible to have an average reference. In this method, an average of potentials from all the electrodes is calculated and then each this average potential is subtracted from each of the electrodes.

B.5 Cleaning Up Raw EEG Data

B.5.1 Filtering

Raw EEG data has a lot of high frequency noise and power line noise which is removed by band pass filtering. Typically the filtering is performed by band passing between 1 and 45 Hz as EEG signals of interest lie between these frequencies. Sometimes an additional 50Hz or 60Hz notch filter is added to remove power line noise. The filtered EEG signal is shown in Figure B.4. Here the ocular activity is visible in the form of abrupt dips around the 12th, 13th and 14th seconds.

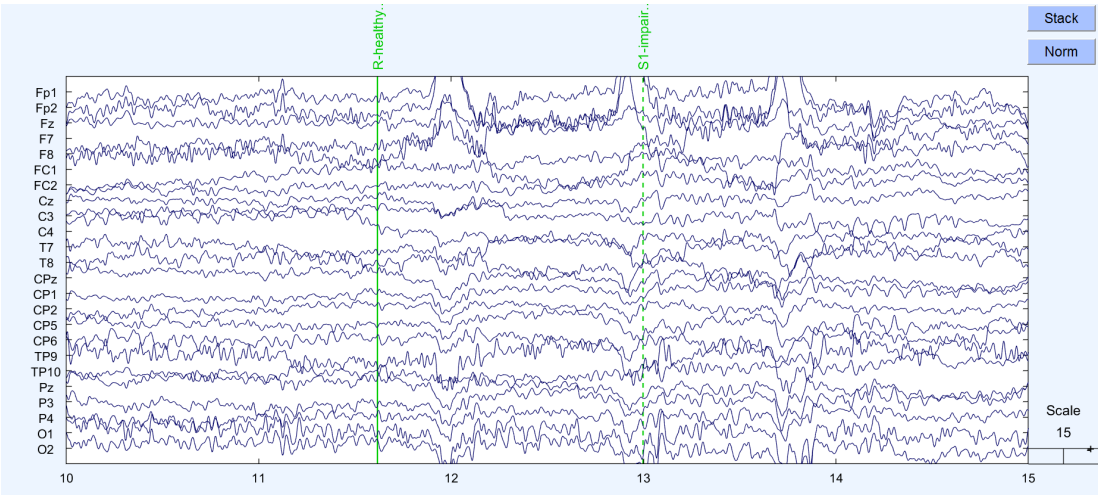


Figure B.4: Band pass filtered EEG data (1-45 Hz)

B.5.2 Independent Component Analysis

Independent Component Analysis (ICA), is the process of splitting your EEG data into several components. This is done by defining a fixed number of components (same as the number of electrodes, say N) and representing each channel as a weighted sum of each of these components. This results in N weights, one each for each component per channel. For N channels, there will be N such lists resulting in a N X N matrix of weights. The idea of performing ICA is to extract components that span across different channels and to identify channels with strange components.

Removal of components can be a subjective or objective process. They may be removed based on some fixed thresholds objectively, or by looking at the visual representation of the components. We have followed the subjective approach. Eye blinks, eye movements, and strange behaviours in some channels were removed with visual inspection. Some components present only in a single trial or two were also removed. At least 2 and at most 5 components were removed for each participant.

One way of removing components by visual inspection involves looking at the components time series and identifying the components with noise or the components that indicate ocular activity such as the blink dips. The other way is to look at the topographical plots such as the one shown in Figure B.5. In this figure, component number 1 indicates the eye blink movements. The eye muscular activity are picked up most by the frontal electrodes. They are also the components with the most amplitude and therefore usually appears as the first component. Component number 2 (red and blue crescents in the front two electrodes) also indicates eye muscular activity. In this case it is the participant moving their eyes from side to

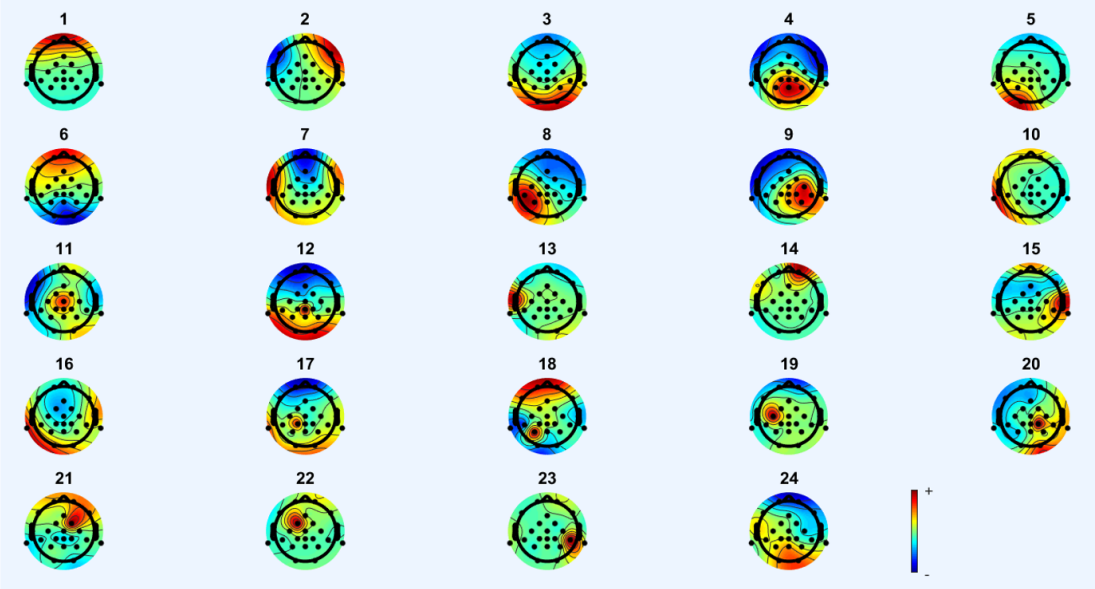


Figure B.5: ICA components for EEG data from one participant

side. Some components such as component 22, 23, 13 etc. are limited to one single electrode and it is a sign of some abnormality in that electrode, such as noise. They may be removed, but not necessarily. If these noisy electrode components form one of the major components (from 1 to 10, say) then it would be more important to remove them.

B.5.3 Epoching

Once the EEG data has undergone filtering and artefacts removal by ICA, it is ready to be studied. The first step is to look for the time stamps of interest within the entire EEG data series. These time stamps are usually when the participant began and ended listening to the auditory stimulus. These time stamps or event markers were already marked into the EEG data that was being recorded and saved. Epochs are the time series of interest in the EEG data which correspond to the participants listening activity. For example, if the participant started listening to the first sentence of the experiment at second 13, an event marker will have been placed at second 13 indicating the details of the stimulus (condition, stimulus ID etc.).

An epoch will be extracted a few milliseconds before this marker (known as baseline) and the respective stimulus duration after the marker. An epoch may be fixed or variable. If all the presented stimuli are of a certain length (say 3-4 seconds), then a fixed 3 second epoch may be extracted. This is the typical process. However, the stimulus duration may have high variation, as is the case in the experiments of this thesis, and a variable epoch which corresponds to the length of that particular stimulus is extracted. A baseline is extracted because it represents the brain activity when the participant is not engaged in the task yet. By normalising the task

related data with the baseline activity, we can remove the effect of non-task-related background activity.

Once all the epochs are obtained, a visual inspection may be performed to remove noisy epochs. Alternately, an objective trial rejection may be performed by setting thresholds for intensity and amplitude.

B.6 Data of Interest in Clean EEG Data

After obtaining the data that correspond to the actual listening activity, we can proceed to investigate the components or frequencies of the EEG signal. EEG signals may be analysed in several ways such as in the time domain, the frequency domain, time-frequency domain and recently even with neural networks. Here, two main types of analysis used frequently in speech perception research are illustrated: Event Related Potentials (ERP) components and frequency components.

After a certain stimulus is presented to the listener, their brain undergoes a series of activities as a response to that stimulus. These activities or ERPs are indicated by positive and negative polarities in the milliseconds of time following the stimulus. ERP components are named in the format [polarity][time in milliseconds]. For example, a P300 indicates the positive potential observed at approximately 300 ms post the event being investigated. Similarly, an N400 refers to the negative potential observed post 400 ms after a stimulus is presented. Some ERP components are known to be associated with certain typical events. For example, the P300 is observed as a response to any novel stimulus and an N400 response is observed when presented with any meaningful stimuli such as words, pictures etc.

An EEG signal is a time series which is a combination of several frequency components. The signal is a sum total of these different frequency components that are emitted by the brain in response to various layered and complex activities it performs. Extensive research in the field of EEG frequencies has resulted in the definition of several frequency bands that have been found to correlate with several cognitive functions. Some of these bands are the alpha (8-12Hz), beta (16-31Hz), gamma (>32Hz), theta (4-7Hz) and delta (0.5-4Hz) bands.

Appendix C

Contents of the Passage Corpus Described in Section 3.3.2

C.1 Passage 1

El día 9 de mayo de l453, el Imperio turco se hizo dueño de la vieja ciudad de Constantinopla; para los países europeos, el comercio con Asia ya no era posible. Fue entonces cuando Portugal, abierto al Atlántico, empezó a buscar un nuevo camino por mar. El plan era sencillo, pero lento: seguir la costa de África, encontrar el paso al Océano Índico, y desde allí ir hacia la India. En 1487, Bartolomé Días dio la vuelta al cabo de Buena Esperanza: Portugal había encontrado su camino.

Por aquellas mismas fechas, un hombre llamado Cristóbal Colón, intentaba conseguir la ayuda de los reyes españoles, doña Isabel y don Fernando, para probar un camino distinto: él quería ir siempre hacia el Oeste, cruzando el Atlántico.

Durante mucho tiempo Colón no pudo convencer a nadie. Todos pensaban que era un viaje imposible: los que no creían todavía que la Tierra era redonda, por supuesto, pero también los que sí lo creían; para éstos, la distancia entre Asia y Europa era demasiado grande. Sin embargo, por fin, en l492, los Reyes Católicos decidieron ayudar a Colón. Este cambio de opinión fue importantísimo porque el 12 de octubre de ese mismo año, tres barcos españoles encontraban tierra al otro lado del Atlántico. De esta manera, Europa había llegado a América.

C.2 Passage 2

Una batalla de tomates en medio de una plaza, parece una película de los hermanos Marx; sin embargo, una fiesta así existe. Se celebra cada verano, el último miércoles de agosto, en Buñol, un pueblo de Valencia. La tomatina es una de las fiestas más insólitas y divertidas de España.

Esta fiesta empezó en 1944, cuando los vecinos del pueblo, enfadados con los concejales, les lanzaron tomates durante las fiestas locales. Se lo pasaron tan bien que decidieron repetirlo cada año.Y con el tiempo se ha convertido en una verdadera batalla campal en la que participan miles de personas y en la que las armas siguen siendo los tomates. Durante los años de la dictadura del general Franco, el gobierno prohibió esta fiesta porque no era religiosa. Pero a la muerte del dictador, los vecinos empezaron a celebrarla de nuevo, en los años setenta.

Aunque la fiesta empezó en contra del Ayuntamiento, hoy en día, es este quien la paga. Para que los vecinos de Buñol, los veraneantes y los forasteros que se unen a la fiesta se diviertan, el Ayuntamiento compra unos cincuenta mil kilos de tomates, que llegan cargados en varios camiones.

El día de la tomatina, sobre las once de la mañana, la multitud está congregada en la plaza Mayor, que está en el centro del pueblo, y en las calles de alrededor. La gente no acude vestida con sus mejores galas sino con la ropa más vieja que tiene, porque después de la batalla hay que tirarla a la basura.

C.3 Passage 3

He llegado a La Mancha para asistir al tradicional concurso de la Monda de la Rosa del Azafrán, que se celebra, el último domingo de octubre, en la gran plaza de Consuegra.

La plaza está tan llena de gente que apenas se puede pasar. Los jóvenes van vestidos con trajes típicos; un grupo de danzas interpreta los bailes regionales y en unos puestos improvisados se pueden degustar los sabrosos quesos manchegos.

Las mesas para el concurso están preparadas. Sobre los manteles blancos hay montones de flores malva, las rosas del azafrán, que guardan en su interior unos valiosos estigmas.

Estos diminutos estigmas son los verdaderos protagonistas de la fiesta, porque con ellos se hace el azafrán, la especia más cara del mundo, que se usa tradicionalmente en la cocina española para dar sabor y el color amarillo a platos típicos como la paella. El uso de esta especia es muy antiguo. Se han encontrado restos de azafrán en las momias egipcias; Homero lo menciona en sus escritos y los romanos crearon con él un afrodisíaco.

El mismo día del concurso, la familia de José Moya, que me ha invitado a asistir a las fiestas,

se levanta antes de salir el sol. Están cansados después de varios días de duro trabajo, pero entre ellos reina un ambiente festivo. Para ellos, como para tantas otras familias de la zona, hoy es el último día de la cosecha del azafrán y sólo les quedan por recoger las flores de un campo.

C.4 Passage 4

La multitud te arrastra. Es una auténtica locura. En esas circunstancias es casi imposible sacar una foto. Pero si no la saco nadie va a creerme. Mi jefa pensará que estoy loca o que he bebido demasiado y he imaginado la historia. Varias veces intento incorporarme para enfocar la cámara y todas ellas recibo un tomatazo en la cara. La multitud es implacable. El ácido del tomate se me mete por los ojos y por la boca y me pica.

El delirio dura dos horas. Hacia la una, el cuarto camión se aleja despacio, vacío. Suena otro cohete. Significa que la batalla ha terminado. Nadie puede lanzar ni un solo tomate, si alguien lo hace tendrá que pagar una multa. Es mi oportunidad. Mi cámara está cubierta de tomate, pero todavía funciona. El rojo es el único color que aparecerá en las fotos.

Cansada, sucia y muerta de risa, bajo con la multitud hacia el río, donde el Ayuntamiento ha instalado unas duchas públicas en una explanada. Todos estamos cubiertos de arriba abajo de salsa de tomate. Después de una ducha ligera, sin desnudarse, la gente sube hacia el pueblo, con la ropa mojada pegada al cuerpo y con las pepitas, las semillas del tomate, en el pelo. Todos tienen un aspecto deplorable, el mismo aspecto que debo de tener yo. Están exhaustos pero contentos, después de unas horas de diversión y desahogo. Ahora la verdadera ducha espera en casa.

C.5 Passage 5

Veréis, mi nombre ahora no tiene importancia. Nací en un pequeño pueblo de pescadores y era muy joven cuando empecé a trabajar como marinero en Portugal. Yo fui de los primeros en viajar por África, hacia Guinea, y en conocer todas las rutas de estos mares. Así llegué a ser piloto de la Santa Susana. Hace dos años, a la vuelta de un viaje a Guinea, quisimos encontrar una nueva ruta hacia Portugal, más rápida que las otras. Pero, ya cerca de Cabo Verde, los vientos cambiaron de una manera muy extraña. El barco tomó de repente la dirección del Oeste y nosotros no pudimos hacer nada para cambiarla. El barco iba más rápido que nunca, siempre hacia el Oeste.

Después de quince días así, mis hombres empezaron a asustarse. Sabían que nos encontrábamos en algún lugar del mundo que no conocía nadie, y que ya no era posible volver

atrás. Cada mañana me preguntaban lo mismo: ¿qué nos esperaba al final de aquel viaje? Yo no lo sabía, no podía contestarles. La comida ya nos faltaba, estábamos preparados para morir, cuando entonces... Entonces, después de veinte o más días en aquel barco que no podíamos conducir... ¡vimos tierra!

La voz del marinero se hizo cada vez mas débil; hablaba de las extrañas gentes y lugares de un mundo distinto, de sus islas, de sus costumbres. Repitió muchas veces más aquel nombre, Cibao, un lugar de esas tierras donde había visto montañas de oro. Y ya no dijo otra cosa en toda la noche. Cuando el sol iba a salir, el marinero volvió a llamar a Colón con la mano. Cristóbal se acercó para oírlo mejor.

Appendix D

Contents of the Words Corpus Described in Section 3.3.1

- abrigadero
- achicadora
- aclamadoras
- acopladuras
- afiladura
- aguardenteros
- alacranera
- alcohómetro
- almohadazo
- alpargateros
- amansadora
- amoladeras
- anfibiótico
- animalescos
- antiyanquismo
- apresadora
- apretaduras
- arrastraderas
- aserraduras
- atracadoras
- autobiógrafo
- autoconvenzan
- avellanate
- avergüéncese
- ayudadoras
- bachillerada
- besalamano
- blasfemadores
- cabeceada
- cachiporrazo
- calabaceras
- caricatures
- cascabeleros
- censuradoras
- chacinerías
- codiciadores
- comediógrafa
- congenérico
- conserjerías
- contrabalanza
- contraescarpas
- contraintuitiva
- contrasubversión
- convidadora
- cosmetólogos
- cubreobjetos
- demandadero
- depopulador
- derrocadero
- desarticulo
- desatándonos
- descargaderos
- desenvuelto
- desgranadora
- deshidratantes
- deslindadora
- despobladores
- desquiciadora
- destripadora
- desvirtuadoras
- disciplinante
- dimisionarias
- discutidora
- disparatador
- divagadores
- electrizador
- embaladoras
- embarrotado
- embotadora
- empedernecen
- emplomadura
- encaballado
- enceradores
- encomiadora
- engastadura
- engrapadora
- enlazadora
- entronizamos
- epicicloides
- equilibrismos

- escafandrista
- escarbadura
- escribidora
- escrupulillos
- escurribanda
- esgrimidores
- espeluznados
- establecedor
- estatalismos
- estenógrafa
- estrujadora
- exhortadoras
- exhumadores
- explicaderas
- exsacerdote
- farisaísmo
- filosofador
- fulminadora
- fustigadora
- geométridos
- graneadores
- herborizador
- hidrobátidos
- hiperopía
- horribilidad
- humilladora
- ilustradoras
- imprimadoras
- incircunscrita
- inconfidencia
- indefensibles
- inelegancia
- inmovilizo
- insinuadoras
- limpiapiscinas
- malgastadora
- manifacturas
- manzanilleras
- meteorismos
- minoridades
- mitigatorio
- ocultadores
- ofrecedora
- ojaladura
- opalescencias
- ostentadores
- otorgadoras
- pantalonera
- paseadoras
- pateaduras
- perjudicador
- picosegundo
- polifarmacia
- porteadoras
- preacordado
- premiosidades
- primeridades
- profetizador
- publicitantes
- quirología
- rebotadora
- recapturamos
- reconfortable
- recriminador
- refortalecer
- regaladores
- relajadoras
- reprobadores
- requeridores
- retardadoras
- ribeteador
- rompehogares
- sacadineros
- saltacaballos
- saqueadora
- semiacabada
- semidespertar
- sobornadora
- sobreveedor

Appendix E

Resumen

E.1 Descripción del problema

Un gran número de personas tiene dificultades para producir mensajes hablados. Esta dificultad se convierte en una barrera para la comunicación en las interacciones entre humanos y máquinas. Como resultado, las personas con trastornos del habla tienen dificultades para expresar sus pensamientos y necesidades y en consecuencia tienen una vida social insatisfactoria.

La producción del habla es un proceso que supone un gran esfuerzo para las personas con trastornos del habla. Adicionalmente, también es necesario un esfuerzo importante por parte del oyente para comprender con precisión y eficacia el habla con determinadas patologías y para acostumbrarse a los patrones de habla patológica. Por otro lado, existe una tendencia creciente a utilizar dispositivos digitales para la comunicación, ya sea para pedir comida, programar una alarma o comunicarse con el banco. Así, los dispositivos digitales y los asistentes virtuales también deberían poder manejar con eficiencia y precisión el habla patológica si quieren beneficiar a las personas con trastornos del habla.

En resumen, el habla patológica es un reto tanto para el hablante como para el oyente. El objetivo de la tesis es explorar formas de superar estos retos, primero comprendiendo el alcance de los mismos, luego enriqueciendo el habla patológica para hacerla menos difícil de procesar y, finalmente, explorando un conjunto diverso de medidas para evaluar adecuadamente el habla patológica y sus versiones enriquecidas.

El habla patológica en la que nos centramos en esta tesis es el habla esofágica. Es uno de los mecanismos de producción del habla adoptados tras una laringectomía (extirpación de la laringe). El habla esofágica, generada a partir de las vibraciones del esófago, carece de frecuencia fundamental y contiene ruidos y efectos no deseados, como los sonidos de deglución,

que afectan al ritmo y la prosodia del habla. Esto hace que el habla esofágica sea difícil de procesar.

E.2 Recogida y preparación de datos

Ya se disponía de una base de datos de 32 hablantes esofágicos. Ésta incluía 100 frases fonéticamente equilibradas, 14 palabras y 5 grabaciones de vocales sostenidas. Se grabaron otras 150 palabras y 5 pasajes. Uno de los pasajes también se grabó en forma de vídeo.

Las 100 frases ya estaban etiquetadas fonéticamente. Estas etiquetas se utilizaron para los procesos de conversión de voz y de Reconocimiento Automático del Habla ASR. Un sistema ASR adaptado utilizando el habla esofágica como datos de entrenamiento tuvo un mejor rendimiento para la voz esofágica en comparación con el entrenado con el habla sana.

Se creó una base de datos de habla sintética paralela con la duración de los fonemas forzada a las duraciones de los fonemas del habla esofágica.

Se realizó un estudio en profundidad de las características lingüísticas y de señal de la base de datos de habla esofágica disponible para tomar decisiones informadas sobre qué procesos de enriquecimiento y evaluaciones pueden utilizarse en el futuro.

E.3 Evaluación preliminar de los datos del habla esofágica

Realizamos un experimento a través de una interfaz web para recopilar métricas de inteligibilidad y de esfuerzo auditivo autodeclarado de participantes con audición normal para la voz de 4 hablantes esofágicos y 2 hablantes sanos de control. También se comparó el rendimiento del (ASR) entre las dos condiciones.

Este experimento reveló que el tipo de hablante (habla sana o habla esofágica) tenía un efecto tanto en la inteligibilidad como en el esfuerzo de escucha. Se observó una correlación negativa significativa entre la inteligibilidad y el esfuerzo de escucha. Los oyentes familiarizados con el habla esofágica obtuvieron los mismos resultados de inteligibilidad que los que no lo estaban. Sin embargo, declararon menos esfuerzo al escuchar el habla esofágica que los oyentes no familiarizados. El rendimiento de la ASR fue peor para el habla esofágica en comparación con el habla sana.

El segundo experimento preliminar se realizó en un entorno controlado de laboratorio. Se recogieron respuestas de esfuerzo de escucha conductual, inteligibilidad del habla y datos de EEG continuo mientras los participantes escuchaban un habla sana y un habla esofágica seleccionada para tener un nivel de inteligibilidad comparable a la del habla sana.

Este experimento demostró que, aunque la inteligibilidad del habla esofágica era cercana a la del habla sana, había una diferencia considerable en el esfuerzo de escucha subjetivo. Se sabe que la potencia de la señal EEG en la banda alfa indica el esfuerzo de escucha del habla degradada. Encontramos una mayor potencia alfa para el habla esofágica en comparación con el habla sana, lo que indica un mayor esfuerzo de escucha para el habla esofágica. También encontramos una relación entre las capacidades cognitivas de los participantes y la potencia alfa. Se observó una menor potencia alfa en los participantes con mejores capacidades cognitivas (es decir, capacidad de memoria de trabajo).

E.4 Enriquecimiento

El enriquecimiento del habla esofágica se llevó a cabo mediante un novedoso método de conversión de voz. La conversión de voz paralela es un método para convertir un discurso de origen en un discurso de destino en el que el material de entrenamiento es paralelo, es decir, se utilizan las mismas frases del hablante de origen y del hablante de destino. Este proceso implica la alineación temporal del habla de origen y del habla de destino, lo que resulta problemático para las muestras de habla esofágica, debido a sus pobres características espectrales y temporales. Por lo tanto, el nuevo sistema de enriquecimiento consistió en eliminar este proceso de alineación mediante el uso de un habla sintética que coincidiera en contenido y duración con el habla esofágica de origen. La conversión de la voz se llevó a cabo mediante una red neuronal profunda conocida como red de memoria a corto plazo bidireccional LSTM. Se implementaron sistemas dependientes e independientes del hablante.

Además, se utilizaron algunos enfoques menos intensivos para el enriquecimiento, como la mejora de la frecuencia fundamental del habla esofágica utilizando un vocoder de última generación conocido como wavenet, y la eliminación de ruidos no deseados de la señal del habla esofágica.

E.5 Evaluación del habla esofágica enriquecida

El habla esofágica original, tres versiones del habla esofágica enriquecida y el habla sana se evaluaron utilizando un conjunto diverso de medidas, como la inteligibilidad objetiva, el esfuerzo de escucha autodeclarado y basado en el EEG, los tiempos de reacción y la inteligibilidad del habla.

La transformación basada en wavenet no consiguió mejorar las puntuaciones de ASR, probablemente debido a los artefactos generados por el proceso de síntesis de wavenet. Se observó

cierta mejora en el enfoque de eliminación de silencios para el hablante esofágico con baja inteligibilidad, pero no para el hablante esofágico de alta inteligibilidad.

Una prueba de preferencia reveló que las muestras de la nueva red neuronal profunda eran preferidas a las del habla esofágica sin procesar. Además, este sistema de enriquecimiento superó a nuestros sistemas anteriores en las puntuaciones de inteligibilidad objetiva y ASR, así como en las puntuaciones de inteligibilidad del habla y de esfuerzo auditivo obtenidas con una prueba de escucha. Sin embargo, en comparación con el habla esofágica sin procesar, se produjo una mejora en las puntuaciones de ASR y en las medidas de inteligibilidad objetiva, pero no en las puntuaciones de inteligibilidad del habla ni en el esfuerzo de escucha.

En resumen, aunque hay una mejora en las medidas objetivas, se necesitan más mejoras para que el habla esofágica transformada sea más preferible y comprensible para los oyentes humanos. Además, esto también revela un desacuerdo entre las evaluaciones subjetivas y objetivas y la importancia de evaluar los resultados de la mejora del habla mediante métodos subjetivos elaborados.

Se preparó una interfaz web de demostración para visualizar a un hablante de habla esofágica hablando con una voz enriquecida [1].

E.6 Relevancia científica de los resultados

- Disponibilidad de una base de datos paralela de habla esofágica y habla sintética de la misma duración que ha sido etiquetada fonéticamente y evaluada para varias características acústicas y medidas de inteligibilidad. Esta base de datos es útil para realizar experimentos de restauración de la voz en el habla esofágica, así como para estudiar las características del habla esofágica

- Disponibilidad de métricas de esfuerzo de escucha e inteligibilidad para hablantes esofágicos.

- Un nuevo método basado en DNN para la transformación de la voz del habla esofágica que puede ser utilizado y mejorado por la comunidad investigadora.

- Los experimentos de esfuerzo de escucha basados en EEG son populares para el habla degradada por ruido, el habla con acento extranjero, etc., pero se han realizado muy pocos experimentos para el habla patológica. La evaluación del esfuerzo de escucha basado en EEG que hemos realizado para el habla esofágica y sus versiones enriquecidas puede ayudar a mejorar la investigación basada en EEG para el esfuerzo de escucha del habla deteriorada.

[1] https://aholab.ehu.eus/users/sneha/london_demo/test.html

E.7 Relevancia de los resultados para la sociedad

- El enriquecimiento del habla esofágica tiene como objetivo una mayor inteligibilidad, un menor esfuerzo de escucha y menos errores de ASR que el habla esofágica sin procesar. Una mayor inteligibilidad y un menor esfuerzo de escucha del habla esofágica enriquecida suponen una menor fatiga en la escucha. La menor cantidad de errores se traduce en una mayor precisión en la interacción con los dispositivos digitales. Todo esto sería beneficioso para mejorar la experiencia de comunicación tanto para los hablantes como para los oyentes del habla esofágica.

- La evaluación del habla esofágica en las áreas de inteligibilidad, esfuerzo de escucha, ASR y otras características acústicas pueden ser herramientas útiles para los logopedas en la formación de los hablantes de habla esofágica y su seguimiento.