Efficient Speech Intelligibility Evaluation Using Keyword Spotting

Khanh Linh Tran

Supervisor: Professor Jon Barker

A report submitted in fulfilment of the requirements for the degree of BSc in Computer Science

in the

Department of Computer Science

May 8, 2024

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Khanh Linh Tran

Signature: Khanh Linh Tran

Date: May 8, 2024

Abstract

We reproduce an approach for human speech intelligibility evaluation recently proposed by the University of Edinburgh based on keyword spotting. The original study indicates that the method serves as an effective alternative for speech intelligibility assessment employing audio-visual stimuli. Similar to the original approach, participants listen to a stimulus and select the word they hear from a set of alternatives. To identify the target word and suitable candidates, we also analysed stimuli using a phonetic dictionary and language model. One notable feature of the novel approach is its independence from specially designed sentences, so this project investigates if the approach produces similar outcomes using a different materials, audio-only recordings. Upon comparing our reproduced method to a naive baseline where candidate words are randomly generated, we found limited data to evidently distinguish the two methods. Nevertheless, our reproduced method demonstrates a similar trend and effectiveness as observed in the original study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims and Objectives

Within human interaction, communication plays a crucial role, especially in information exchanges, and it can be conveyed through different methods including speech, written text or technological interfaces. The clarity in communication is thus increasingly significant in today's interconnected society, which lies in the concept of intelligibility. In this realm of digital communication, the ability to deliver clear and understandable messages is continuously more demanding from daily conversations to advanced technical instructions. Meanwhile, intelligibility evaluation becomes a useful tool to assess the message clarity and communication effectiveness, and typically is utilised in speech enhancement or synthesis technologies such as those found in hearing aids, public announcement systems or assistive communication devices.

The Centre for Speech Technology Research of the University of Edinburgh has recently brought out a promisingly efficient method for assessing intelligibility in speech enhancement by using keyword spotting[29]. The test format is also promised to be user-friendly with the multiple choice format. Test materials will consist of a set of short sentences and their audios which are mixed with an interferer at different signal-to-noise-ratios (SNRs) to vary the difficulty level. The audio is played only once during the test, so the listeners are encouraged to carefully attend to each stimulus.

According to University of Edinburgh's paper, keyword spotting method is compared against transcription-type test with the same set of stimuli. Both methods demonstrated a similar trend of decreasing accuracy in lower SNRs. However, the method developed by Edinburgh's Research Center took only about half the time required by the transcription method.

This project will implement that novel approach introduced by University of Edinburgh for subjective human speech intelligibility assessment, and evaluate its effectiveness. The approach still relies on keyword spotting combined with multiple-choice question quizzes. It will employ the materials with varying phoneme complexity, while the background noise will adapt according to the performance of the participants. Nevertheless, a limitation lies in the necessity to select the meaningful yet sufficiently complex alternatives, which can be addressed with the use of a statistical language model trained on domain data. In this test, a human participant listens to a sentence and is presented with a list of candidate words that it may have included.

The mission of the project is to design candidate word lists in a manner that ensures the correct answer is not obvious when mixed with other candidates unless the sentence has good intelligibility. The listening tests will compare this novel method against a naive baseline, i.e., selecting random common words as alternatives. The approach will then be evaluated by measuring how well it can predict the intelligibility of sentences which have already been scored for intelligibility in a previous study.

1.2 Overview of the Report

So as to fulfill the objectives presented in section 1.1, the investigation and implementation are divided into chapters as follow:

- **Chapter 2** provides the basic understanding of ”speech intelligibility” and explores the potential impacts on communication clarity, the cause to misunderstanding in speech delivery, as well as existing methods for assessing speech intelligibility.

- **Chapter 3** presents a comparative analysis between the original approach and the reproduced one, regarding both differences and similarities. It also delves into the factors discussed in Chapter 2 that need to be considered to attain the project goals.

- **Chapter 4** describes the technical implementation of the elements analysed in Chapter 3, alongside improvements discerned during the implementation phase. It also brings up the criteria and components taken into account, and the choice of test delivery interface for the assessment design.

- **Chapter 5** discusses the results of the delivered test which utilises the compiled design materials guided by essential elements and criteria outlined in previous chapters. The evaluation also includes a comparison between the novel approach and a naive baseline method.

- **Chapter 6** provides a summary of the project’s aims, the investigations conducted, the implemented measures, the analysis of outcomes, along with recommendations for potential improvements.

Chapter 2

Literature Survey

To start with, this chapter covers the definition of speech intelligibility. There are various factors that contribute to making a speech less understandable and affecting the intelligibility. The common elements embraces similar pronunciations, language comprehension, and memory lapses. With the growing demands in speech enhancement, there has been ongoing exploration of techniques to evaluate the intelligibility. Despite existing methods, the pursuit of new approaches still continues, driven by the need for improvements, commonly in the matter of efficiency and accuracy, which solidifies its importance among different applications in this era. All these aspects will be further explored and discussed in this chapter.

2.1 Speech Intelligibility

In linguistic context, intelligibility means how well comprehensive a speech is when delivered from a speaker to a listener. Therefore, speech intelligibility means how much you understand a speech, and that depends on the context, depends on the ongoing conversation, depends on if the speaker speaks clearly, etc. For example, even though the listener does not know about the full context, or did not necessarily hear all the words spoken by the other, but he still roughly understands what the speaker is trying to say to him, then in that case, the speech is considered intelligible. However, even though the listener can transcribe all the spoken words, but he either has not understood or misunderstands the saying, then it is said the speech is not intelligible to him. Being able to write down the words is not necessary and not sufficient, but understanding the words is important; and sometimes a speech is just half-way delivered to be comprehended by the listener. As a result, speech intelligibility is considered if the the communication is successful, where communication means a person A is trying to communicate a concept to person B via speech.

2.2 Factors that can Reduce Speech Intelligibility

In the situation of decreasing clarity and comprehensibility in communication, there are several factors contributing to reduced speech intelligibility. One of the common reasons results from the speech quality. Speech quality means how well an utterance is produced, how precise is the articulation with the variation in rhythm and intonation. Person A produces a poor speech which involves frequent hesitations, repetitions, or lengthening the sounds; thus making it more challenging to listen and follow the messages. This then reduces intelligibility to the listener, person B. This is not because person B's hearing is bad, but due to the communication quality. Cognition is another component, which would be required to understand a context. For example, a person B have heard all the words, his hearing is working perfectly well, but his brain's function is impaired, having difficulties in paying attention and processing language, alternatively, that person is not familiar with the language being spoken or the accent being used. Thereby, he struggles to understand the communication. Or in another case, the person B has poor hearing, thus he misses important words or phrases as the conversation goes on. Therefore, it is likely that he misinterprets the speaker's points and finds it hard to keep following the flow of the conversation. In addition, a noisy environment can cause reduction in the speech intelligibility when communicating through speech.

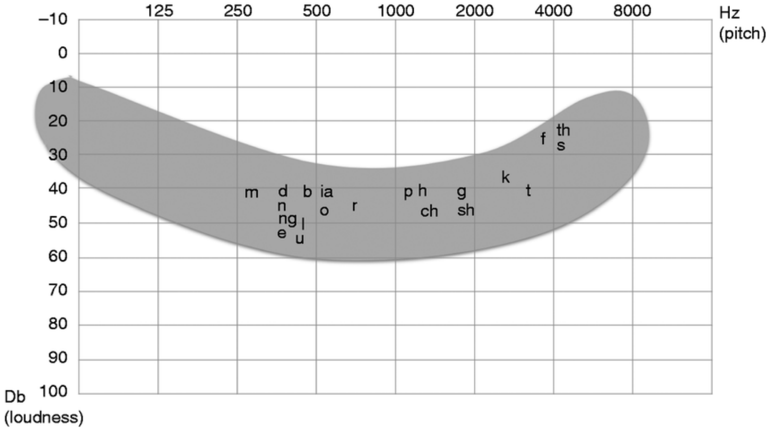


Figure 2.1: The Speech Banana[8]

The Speech Banana is a visual demonstration of the different sounds produced when talking. Each sound is plotted at various ranges of loudness that they can be heard. Therefore, with external factors of noise, i.e., in a loud environment, the human speech will be delivered at an unusual level of decibels(dBs); thus, some of the phonemes will be harder to be heard, which can cause the misunderstanding in words used and context. For example, person A says ”Do you have a fan?”, but with loudness of the environment being more than 40dB, the person B cannot hear the word ”fan” properly, but heard something like ”Do you have a van” or cannot hear the word at all, just something ending with */ae n/*. This means that the communication did not succeed, and the hearing evidently has an impact on intelligibility.

2.3 Types of Speech Misunderstanding

When it comes to understanding a speech, there will be misunderstanding a speech, which can be due to a reason as phonetic perception. Each word has its own sounds when uttering, but some words share the same pronunciation despite their different spellings. One of the most commonly confusing homophones include ”two”, ”too” and ”to”. The misunderstanding of the intended homophone can lead to miscommunication and confusion. For example, given a scenario where a group of friends plan to buy concert tickets. One person asks if anyone has already bought the tickets and Person One says ”I got three”. Then Person Two interrupts and says ”Me two” as he bought 2 tickets, which is however mistaken by his friends for ”Me too”, which means he also bought three tickets, potentially resulting in the wrong number of tickets need to buy more. Beside homophones, confusion can stem from the fact that some words share closely similar pronunciation sounds i.e. ”dog” and ”dock”. In certain accents, the minor distinctions are even less noticeable, thus altering the whole intended meaning of a given context.

Another type of misunderstanding is about language understanding which relates to the cognitive effects of language comprehension and memory. There is a study on the close relationship between misunderstanding and language comprehension by Inaad Mutlib Sayer[26]. In this journal, there explains what misunderstand is, as quoted from Humphreys-Jone, ”misunderstanding occurs when a hearer fails to understand correctly the proposition which a speaker expresses in an utterance”. Language perception relates to word interpretation which contribute to information retrieval. For instance, in a group meeting, the leader is giving suggestions on a project’s scope. However, another team member misconceive the concept as a ”goal” due to their perception of the two terms as similar. Consequently, in upcoming team meetings, the member’s frequent interchange of these two terms can bring about confusion among the team and other stakeholders. In Inaad’s study, there mentions different characteristics of linguistic inputs, which include syntactic processing which is about the listener’s ability to construct syntactic structures to derive meaning of sentences, or inferential processing which is to use contexts and background knowledge to perceive the information, or comprehension and emotion which indicates how emotions impact understanding and responding processes, i.e. one’s saying with the intention of offering help but considered a criticism by another. Memorising also plays a part in determining either successful or ineffective communication. A simple example is in terms of memorising word order, i.e. a person was reading out a series of number like ”354325”, but the other person forgot the order and noted that down as ”353425” which is clearly a different series.

The types of speech misunderstanding as discussed above arise from various elements, comprising phonetic similarities, linguistic perception, and memory recall.

2.4 Measurement and Assessment

There have been various approaches to deliver intelligibility tests introduced, broadly categorised into: by type of the material that the tests used; by the way in which noise levels are varied; and by the way that the results are reported based on the number of words of different noise thresholds, each of which will be discussed in this chapter.

Regarding assessing by the types of material used in the assessments, the speech intelligibility tests typically make use of digit sequence (Digit Triplet Test[30]), random word sequences, or keyword embedded in carrier sentences[3], or spoken sentences (transcription task[31]), conversation material. Despite some overall benefits, these materials still have certain limitations, i.e. digit sequence, and random word sequences have limits in capturing the complexity of speech in real-life situations, due to the lack of phonetic challenge or the context of complete sentences; the other approaches have challenges in terms of sentence designs which require the standardised testing topics, or time-consumption.

Another material type is how the noise levels are modified. Commonly in speech intelligibility tests, either the speech volume or background noise undergoes the alteration with an aim to increase the complexity of the tests. This approach does help the assessments one step closer to practicality. For example, in Digit Triplet Test, although it uses simple and straightforward phonetic words, altering the sound levels of speaker's voice makes the test a bit more challenging.

The final approach mentioned above is regarding how the results are reported. Generally in simplified Matrix test approach, there sets different noise thresholds and record the number of correct responses per threshold, with an aim to assess listeners' ability to recognise speech sounds under different conditions. This is considered an extremely accurate measurement tool that almost resembles everyday situations[4].

Different approaches are in use to determine the speech intelligibility for corresponding different purposes. One primary aim is to rate speech quality. Speech quality assessment is more vital with the increasingly demanding usage of speech processing algorithm in widely applicable fields in the current era, such as telecommunications or listening tests [19] which are typically utilised in communication channel by i.e. mobile phone designers or utilised by speech therapists.

Another purpose is to assess the listeners, where audiologists conduct cognitive assessments to test listeners' hearing abilities, or language comprehension or even cognitive processing skills.

2.5 Summary

Literature Survey has covered potential effects as well as previous research on speech intelligibility evaluation. There also covers the factors influencing speech intelligibility including speech quality, individual's impairment, or external environment, together with the causes of misunderstanding. By analysing different methodologies, we have gained insights into the strengths and further required improvements of existing approaches used to evaluate speech comprehension. With the procedure outlined by the University of Edinburgh's Technology Center, this study aims to follow the novel method to address the gaps by integrating those elements discussed above into the design of the delivered tests. Thereby, the study attempts to fulfill its objective of assessing the hearing people while guaranteeing the options in the quiz are not readily noticeable, thus minimising the possibility of guessing but emphasising intelligibility. With the target on hearing individuals and true intelligibility, the disrupting components such as impairments, inferential processing requirement or emotions would be avoided or mitigated in the design materials.

Chapter 3

Analysis

This chapter looks into the motivation of the project. It then explores the relevant requirements and design elements necessary to execute the project's goals. The study aims to apply a new, more efficient method to assess human speech intelligibility subjectively. The application is based on keyword spotting with the delivery method in a multiple-choice question (MCQ) format introduced by the Speech Research Center of the University of Edinburgh[29], which consists of multiple noisy sentences and a list of alternative words. The prime aspect of the project is on the design of the candidate word lists, such that it is guaranteed the answer is not obvious unless the sentence demonstrates good intelligibility. The novel approach will be evaluated on how predictable the sentences are, in comparison with a fundamental method, i.e. candidate word lists are generated randomly.

3.1 Evaluation Dataset and Strategy

The motivation behind this project derives from designing an robust and desirable intelligibility evaluation drawn from the University of Edinburgh's recently proposal on speech enhancement. The idea in this project also involves the ability to be performed by the elderly to evaluate hearing aid algorithms for speech in noise processing. Thereby, the design elements require time efficiency, low cognitively demand, ease of performance and typically the accessibility for people with poor keyboard skills. All those requirements serve as the reasons behind the MCQ format of the delivery method, where the human participants will play the audio to listen to the sentences and select their answers.

According to procedures of the Edinburgh's Center, the stimuli are short video clips from TED and TEDx, and sentences extracted from the talks' manual transcriptions. The recruited participants are native British speakers with self-reported normal hearing and normal vision. Their participants are presented with a question "Which of the following words did the speaker in the video say?" and a list of options consisting of four candidate words with a special option "none of the above". Each question has only one correct answer, either a candidate was actually spoken, or there is the chance that none were so. The test is conducted on personal devices like desktop or laptop with headphones in a quiet environment.

Participants are presented with audios mixed with interfere types of distinct noise sources (i.e. microwave, washing-machine, hairdryer and soundscape) and competing speaker (half of each type) at different SNRs in an increasing difficulty order. The videos are played only once, thus requiring the listeners to pay close attention to each stimulus.Edinburgh's test is compared against transcription method using the same set of sentences in the same order. Ultimately both methods showed a similar trend of decreasing scores in increasing noise levels, yet keyword-spotting method achieves a (20%) higher word accuracy. Moreover, the novel method is reported to consume approximately half of the time compared to the transcription task on average.

My project shares some similarities with the approach above, yet also introduces some variants. The test format remains consistent, presenting identical questions and five choices, including one "none of the above", with only one correct answer. In addition, the participants recruited in this study share the same characteristics, environment and device requirements to carry out the test also remain unchanged. Instead of using video clips, this project utilises non-visual clean audios with the sentences extracted from that source. The audios are further enhanced by mixing with pink noise, an interferer, at five different SNR thresholds in a mixed order of difficulty, ensuring that the participants do not undergo fatigue either early on or later towards the end of the test. It is similarly instructed to play the audios only once throughout the test. There is also a comparison of methodology in this project. Instead of explicitly comparing time consumption, the comparison concentrates on highlighting the importance of thoughtful choice selection in test design. In half of the test, word choices are carefully selected with the purpose of confusing the listeners, while the remaining half have the list of choices randomly generated.

In short, the task is to turn the prompts into MCQ responses. The goal is then to make the system generate those MCQs based on the given sentences. Given that part of this solution is to develop the response; that is requirement for the implementation to go from sentences to the suitable MCQ responses.

3.2 Project Design

Regarding the intended design for this project, we want to focus on the listeners, focus on assessing the hearing. In order to ensure the cognitive problems are not measured, the first step is to make sure all of the participants did not have cognitive impairments or dysfunctions. The targeted participants for this project have a normal hearing and preferably to be native English speakers, since the stimuli and tests is designed in English. So as to avoid cognitive issues in testing, the sentences are chosen such that they would not tax the cognitive ability, for example, making the sentences very short, i.e., having seven to ten words per sentence; otherwise the listeners have to remember. Nevertheless, even with short sentences, if the phrase is like "I went shopping and bought apples, bananas, chocolates and pencils", how likely will the listeners remember "apples, bananas, chocolates and pencils" in the right order? Even though they get all those words right but if in the wrong order, when it comes to scoring, it can be marked as wrong because the order is not matched despite the fact that they heard all the words. The order does not actually matter the meaning of the context in this case; yet in some cases, the order does matter. For example, if one were to say ”I went shopping and bought an apple and then a banana”, interchanging ”an apple” and ”a banana” in this sentence definitely changes the context meaning. Hence, it is preferable to avoid those cases that might cause the confusion to the concept and purpose of the test. Specifically, sentences that list out objects or have some kind of schematic structure in it should be avoided. Besides, it is not completely true that, if the listeners heard the words right then it means they understand the concept right; or if they get the words wrong, they get the concept wrong. Sometimes even just halfway through a conversation, the listeners can immediately grasp the idea being conveyed, leading to successful communication. On the other hand, the listeners can hear all individual words correctly, but with all the words assembled into a sentence, they fail to communicate back appropriately, commonly due to misinterpreting a word’s homophone.

Homophones are also avoided in usage in this project. If the choices include homophones of the original word, it will require the listeners to understand the contexts so as to make the right choice. Nevertheless, each question in the listening tests records a short sentence without any preceding context. Therefore, homophones can mislead the listeners, potentially resulting in incorrect evaluations although they clearly heard the correct information. For example, one is saying ”Please right your name on the paper.”, but the options are ”yours”, ”write”, ”right”, and ”None of the above”. The participants will hesitate between ”write” and ”right” because they both sound /r ay t/ and both fit linguistically in the sentence. To conclude, homophones are excluded from test options due to their potential for misleading effects.

Accordingly, the test design becomes important, aiming to elicit the suitable responses for the given sentences. The criteria of suitability here means that the responses are the alternative but fit sensibly into the context, or in another word, the responses are good-phonetic and good-linguistic matches. In Chapter 2, we discussed about distinctive types of misunderstanding, arising from factors such as phonetic ambiguity and linguistic understanding. The candidate words will be those having high phonetic match scores and linguistic match scores to fit in phonetically and linguistically, or else, it is obvious for the participants to guess the answer based on evidently different pronunciation or by ruling out contextually inappropriate words. For example, the original statement is ”A cat sits on the matt”, it can baffle the listener if the candidate of ”cat” is ”cap”, but if it is ”sad”, then they would not mistake for this option since ”sad” patently does not fit in the context.

3.3 Phoneme Distance

In designing the test, with the purpose of confusing the listeners, it needs to ensure that the pronunciation of the original word and the candidates would cause certain challenges to listen to. As such, there comes an approach to deduce the appropriate substitutes which is to gather distance-n phonemes of each word occurring in the prompts. This practise is implemented using edit distance method, specifically Levenshtein distance[20][10]. The original idea is applied to words, however, it is modified to apply to phonemes for this test design, which is also called Phonetic Edit Distance (PED)[28].

PED is a metric that quantifies how dissimilar two sequences of the phonemes are to one another, which is by calculating the minimum operations (inserting, deletions, or substitutions) required to transform one sequence of phonemes into another. Commonly, PED takes International Phonetic Alphabet (IPA) strings as input and returns the phonetic distance between them. In this project, the phonemes are in ARPAbet format[6] instead of IPA because while IPA can represent the sounds of multiple languages, this test materials are specifically in English and ARPAbet are typically tailored for English sounds, using ASCII characters to represent phonemes, which is more straightforward to use on English words and easier for English speakers to understand.

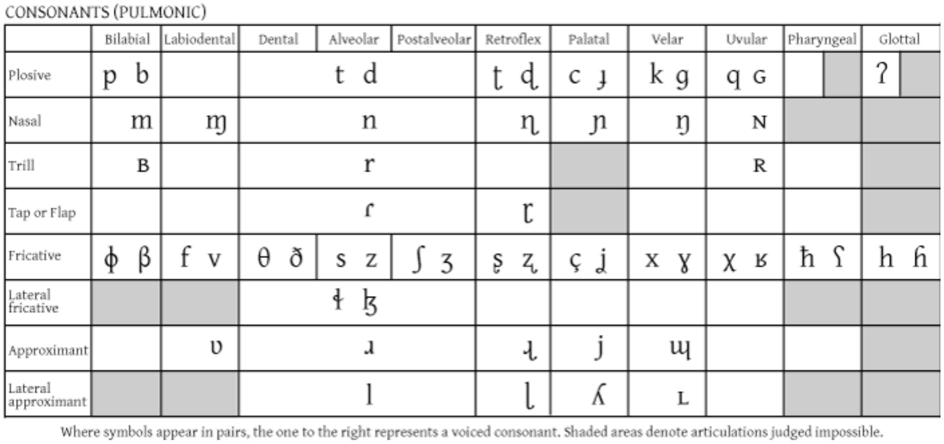
WORDS	**PHONEME**	**DISTANCE** $\phi(word_1, word_2)$
cat	/k ae t/	original word
cats	/k ae t s/	1.0
scats	/s k ae t s/	2.0

Table 3.1: Phonetic edit distance table
Note: ϕ denotes function for standard PED

Explanation: According to Table 3.1, the distance between the phonemes of ”cat” and ”cats” is 1 because their phoneme elements share the same /k/, /ae/ and /t/, but there is an addition of /s/ at the end of ”cats”, costing 1. Meanwhile, when comparing ”cat” and ”scats”, there is an addition of /s/ at the beginning of the pronunciation of ”scats”, costing 1; there is another addition of /s/ at the end, also costing another 1; so the sum of all operational costs equal to 2 between ”cat” and ”scats”.

Going beyond just counting the number of phonetic edits, we can apply a similar approach to recognise that some phoneme pairs are not only different, but the difference is more recognisable than the others.

Explanation: According to Figure 3.1, ”t” and ”d” share more articulatory similarity compared to ”t” and ”p”. Consequently, in Table 3.2, the distance between two words ”CAT” and ”CAD” is estimated to be less than the distance between two words ”CAT” and ”CAP”. In the provided subjective example within Table 3.2, the distance between ”CAT” and ”CAD” is estimated as 0.5, while the distance between ”CAT” and ”CAP” is 0.8.



CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 3.1: IPA Chart Consonants table[28]

WORDS	PHONEME	DISTANCE $\Psi(word_1, word_2)$
CAT	/k ae t/	original word
CAD	/k ae d/	0.5
CAP	/k ae p/	0.8

Table 3.2: Phonetic distance table

Note: Ψ denotes the PED, with the operation cost of replacement is not fixed as 1.0

3.4 Language Model Scoring

In addition to phonetic compatibility, there requires linguistic compatibility within the sentence as well, so as to minimise the likelihood of making informed guesses. To check on this compatibility feature, an approach called Language Model Scoring is utilised.

In Natural Language Processing (NLP), a language model[14] is a computational model that uses machine learning techniques and pretrained language data to generate a probability distribution over words, and then predicts the likelihood of a sequence of words occurring in a given context based on the known contents, as well as based on the syntactic and semantic coherence criteria. By analysing the probability of a sequence's occurrence, one of its utilisation is to determine the coherence of a sequence of words arranged into a sentence.

Generative Pretrained Trasformer 2 (GPT-2) is a widely used and excellent pretrained model in NLP these days [12]. GPT-2 comes in different variants trained on different scale of data for different purposes, such as gpt2, gpt2-medium, gpt2-large, gpt2-xl or DistilGPT-2. gpt2 is the base and the smallest version of the GPT-2 model whose fundamental objective is to produce coherent and contextually relevant texts[23], while gpt2-xl is the largest and most complex version of the model which is more suitable for advanced language processing tasks. DistilGPT-2 [7] however is a smaller, size-reduced variant compared to the gpt2 version and is developed by Hugging Face, a company specialising in NLP. The reduction in size makes DistilGPT-2 faster than GPT-2 while still preserving much of its performance and capabilities. As a matter of fact, the larger the model, the more data was used to pretrain the model. As a result, it can generate more accurate results, but it also requires more compiling time, and it is more suitable for applications involving multi-languages. Yet, this project is limited to British English and has time constraints, so medium-size language model, gpt2-medium is the most suitable choice in use to achieve the project's goals.

Thereby, this project makes use of a pretrained language model, GPT-2, to evaluate sentences that are differed by only one word which is a phonetically compatible candidate of the initial word. The intention is to estimate the sensibility of each candidate sentence and select the most phonetically and linguistically coherent options only.

3.5 Summary

This project adopts the framework from University of Edinburgh's recent intelligibility evaluation on audio-visual speech enhancement. While it shares certain similarities with the proposed method, this study introduces some adjustments and differences, one of which lies in the stimuli that is in audio format only. Next chapter will delve more into the implementation process along with the discussed elements.

Chapter 4

Project Implementation

The intention of the test design is to generate the top phonetically and linguistically ideal alternatives to the given prompts. So as to find the most suitable candidates of each word, phoneme distance combined with language model scoring are then in use.

In this chapter, we will be considering the implementation of the system that generates audibly confusing alternatives to the content words in sentence prompts. This shall involve utilising Python code along with a British English pronunciation dictionary, BEEP dictionary, and sentence prompts sourced from a published dataset of sentences used as stimuli in the project.

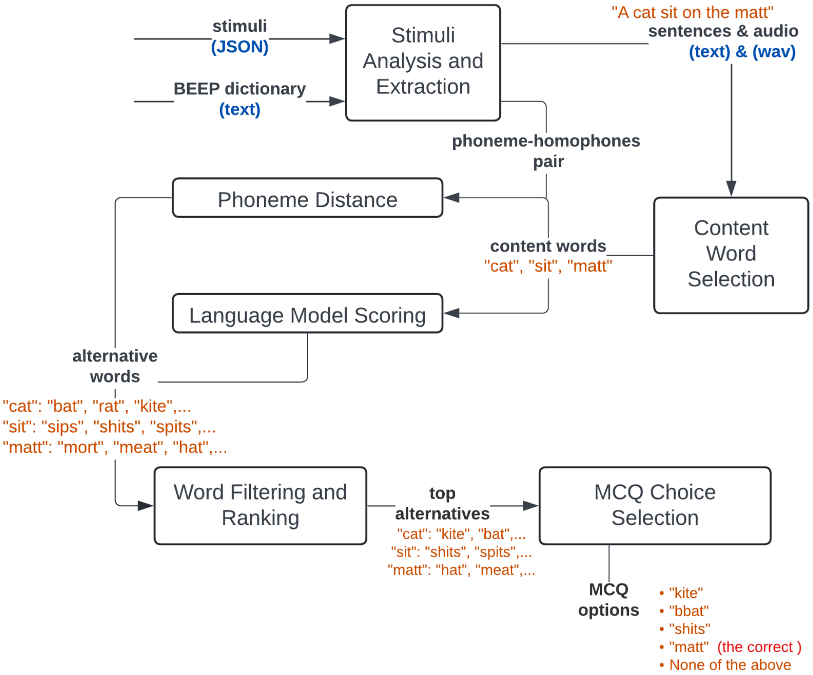


Figure 4.1: Illustration of the Development Workflow: Inputs including Stimuli in JSON Format and a Dictionary Resulting in MCQ Test Options.

Figure 4.1 represents the overall implementation workflow to achieve the final design materials for generating multiple-choice options. The implementation process shall be divided into different stages:
1. Stimuli Analysis and Extraction: Sentence prompts and phoneme-word pairs are derived from the stimuli and pronunciation dictionary respectively for the later application in the implementation process.
2. Content Word Selection: Only content words from each sentence are selected to find their substitutes.
3. Phoneme Distance and Language Model Scoring: Implementation details are discussed based on the aforementioned theories explained in Chapter 3.3 and 3.4.
4. Word Filtering and Ranking: Alternatives are analysed further to identify the top suitable options (i.e the top ten most suitable candidates for each content word in every sentence).
5. MCQ Choice Selection: Some techniques and criteria are applied to choose four options from the pool of alternatives for each question in the MCQ test.

4.1 Stimuli Analysis and Extraction

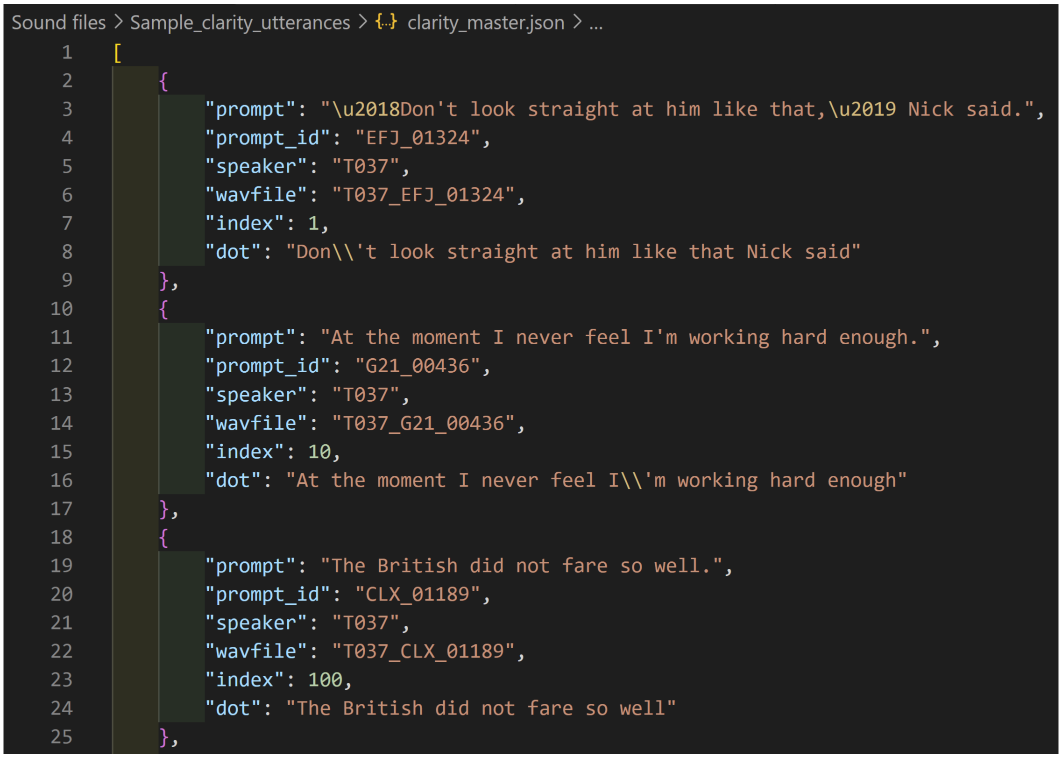
4.1.1 Stimuli Analysis

In order to design the test, appropriate sentences and their audio versions are required as a starting point. The project makes use of a set of sentences and corresponding audio recordings sourced from a published dataset of sentences that have been designed for intelligibility testing [15]. This project also incorporates BEEP dictionary[25] which is an English dictionary consisting of over 250,000 English words associated with British English pronunciations in ARPAbet. The whole project is implemented in Python. The original audio files are clean audios while the objective is to evaluate how intelligible the sentences are across varying levels of noise environments. Therefore, these audio files are upgraded with pink noise[2] at five equally different SNR levels ranges from -3dB to 3dB, representing the spectrum from distinctly noisy to less noisy conditions.

4.1.2 Stimuli Extraction

The set of sentences is stored in dictionary types in a JSON file called *clarity_master.json* (Figure 4.2) that contains the sentences used for the test quizzes and their corresponding audio file name. Initially, these sentences are extracted based on the keyword *dot*, so they can be called as **dot sentences**. Following this extraction, unique words across them are gathered with the aim to find out the specific vocabularies are used in the materials.

The list of distinctive words in the stimuli is cross-referenced with the word set in the BEEP dictionary so as to identify any vocabularies present in the materials but not included in the dictionary. Subsequently, the BEEP dictionary is updated to include those missing words along with their corresponding phonemes.

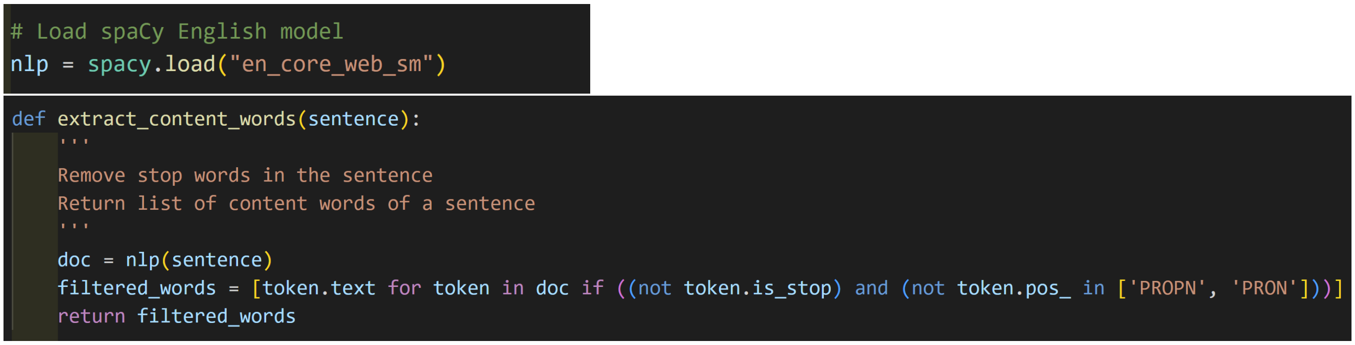

Sound files > Sample_clarity_utterances > {..} clarity_master.json > ...

```
[
    {
        "prompt": "\u2018Don't look straight at him like that,\u2019 Nick said.",
        "prompt_id": "EFJ_01324",
        "speaker": "T037",
        "wavfile": "T037_EFJ_01324",
        "index": 1,
        "dot": "Don\\'t look straight at him like that Nick said"
    },
    {
        "prompt": "At the moment I never feel I'm working hard enough.",
        "prompt_id": "G21_00436",
        "speaker": "T037",
        "wavfile": "T037_G21_00436",
        "index": 10,
        "dot": "At the moment I never feel I\\'m working hard enough"
    },
    {
        "prompt": "The British did not fare so well.",
        "prompt_id": "CLX_01189",
        "speaker": "T037",
        "wavfile": "T037_CLX_01189",
        "index": 100,
        "dot": "The British did not fare so well"
    },
```

Figure 4.2: clarity_master.json file

4.2 Content Word Selection

In each sentences, there is always a combination of stop words and content words. Stop words are common words in a language that contribute little to the overall context of a sentence, while content words carry more meaning to the text. Unlike stop words, the changes in content words typically lead to more significant alterations in the overall meaning of a conveyed sentence or speech. With the example of the sentence ”A cat sits on the matt”, ”cat”, ”sits” and ”matt” are considered content words; on the contrary, the rest are stop words (”a”, ”the”, ”on”), with or without which, the sentence's content still remains understandable. By analysing the impacts of the word substitutions, prioritising content words for substitution of each sentence is thus preferred. This approach also saves the compiling time by avoiding unnecessary candidates word searches and focuses on more notable semantic modifications that occur with each substitution. As a result, for each unique dot sentence of the stimuli, only content words are retrieved to search for their most suitable replacements. This objective is then achieved using spaCy[18], a Python library for NLP[13] with the task to eliminate stop words from the sentence. Apart from stop words, proper nouns which are names of a specific person or place[5] are considered integral to the whole sentence, so their substitution would lead to unnecessary confusion or loss of specificity. Additionally, it is unlikely that proper nouns have both phonetically and semantically suitable alternatives. Therefore, in the substituting task of this project, it is decided to exclude proper nouns from the list of words considered for substitutions so as to preserve the accuracy and specificity of the sentences.The function was tested with the example sentence ”A cat sits on the matt”, which returned ”cat”, ”sits”, and ”matt” as content words, as expected. Additionally, a small subset of stimuli sentences, some containing proper nouns and some not, were selected for the preliminary test, producing the anticipated results.



```
# Load spaCy English model
nlp = spacy.load("en_core_web_sm")
```

```
def extract_content_words(sentence):
    '''
    Remove stop words in the sentence
    Return list of content words of a sentence
    '''
    doc = nlp(sentence)
    filtered_words = [token.text for token in doc if ((not token.is_stop) and (not token.pos_ in ['PROPN', 'PRON']))]
    return filtered_words
```

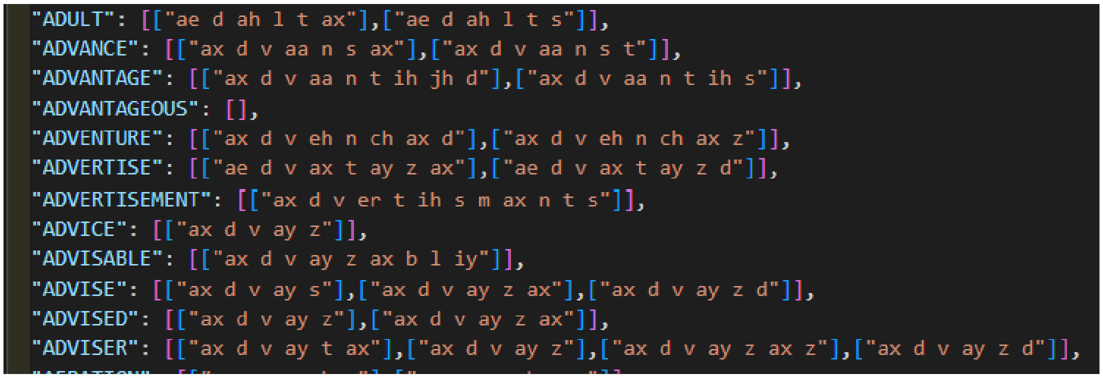
Figure 4.3: Retrieve content words using spaCy

4.3 Phoneme Distance

As outlined in the Project Plan, the substitutes need to be phonetically confusing; thus resulting in the application of PED. Essentially, the closer two words are in terms of phoneme distance, the more similarly they sound, together with the emphasis on hearing, distance-1 phonemes are deliberately chosen for the test set, and are implemented based on Levenshtein Distance method in Python[9].

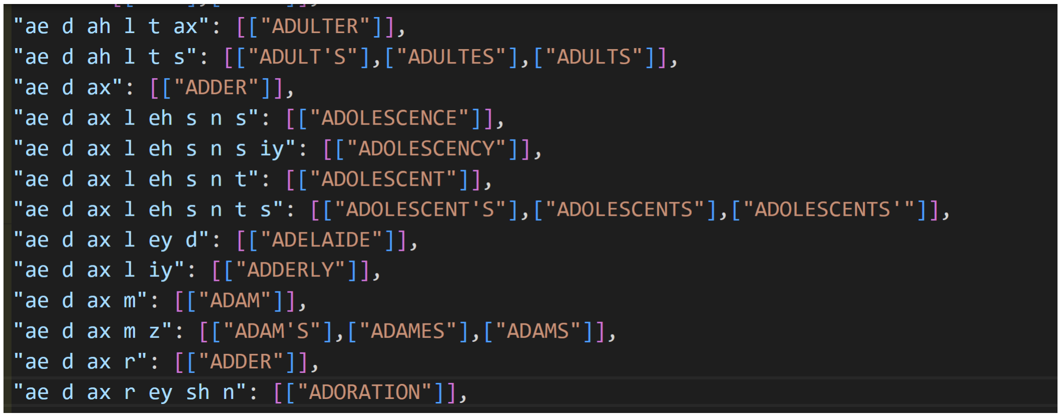
The unique words from dot sentences are paired with corresponding distance-1 phonemes based on the original phonetic transcription from the BEEP dictionary. These pairs are stored in a JSON file named *words_to_alternative_phonemes.json* (Figure 4.4). After obtaining suitable candidate phonemes for the words used in the test set, the subsequent task is to map each unique phoneme to its corresponding words or a set of homophones also based on the BEEP dictionary. The pairs are then stored in another JSON file called *phonemes_to_corr_words.json* (Figure 4.5).

Beside Phonetic Edit Distance, the phonetic distance between words is also taken into consideration. In another word, since some phonemes are more likely to be confused than the others, the differences between individual phonemes are then considered. According to Consonant Distance Table (Figure 4.6) and Vowel Distance Table (Figure 4.7), the distance between each consonant/vowel and itself is always 0. However, the distance between each of them and other consonants/vowels is varied. For example, the distance between /b/ and /d/ is 0.25, while the distance between /b/ and /k/ is 0.40. This indicates that the sound of /b/ is more similar to and more likely to be confused with /d/ rather than with /k/. In implementation, the method of phonetic proximity calculation is slightly different depending on whether a candidate word is formed through substitution or insertion/deletion. In terms of



```
"ADULT": [["ae d ah l t ax"],["ae d ah l t s"]],
"ADVANCE": [["ax d v aa n s ax"],["ax d v aa n s t"]],
"ADVANTAGE": [["ax d v aa n t ih jh d"],["ax d v aa n t ih s"]],
"ADVANTAGEOUS": [],
"ADVENTURE": [["ax d v eh n ch ax d"],["ax d v eh n ch ax z"]],
"ADVERTISE": [["ae d v ax t ay z ax"],["ae d v ax t ay z d"]],
"ADVERTISEMENT": [["ax d v er t ih s m ax n t s"]],
"ADVICE": [["ax d v ay z"]],
"ADVISABLE": [["ax d v ay z ax b l iy"]],
"ADVISE": [["ax d v ay s"],["ax d v ay z ax"],["ax d v ay z d"]],
"ADVISED": [["ax d v ay z"],["ax d v ay z ax"]],
"ADVISER": [["ax d v ay t ax"],["ax d v ay z"],["ax d v ay z ax z"],["ax d v ay z d"]],
```

Figure 4.4: words_to_alternative_phonemes.json

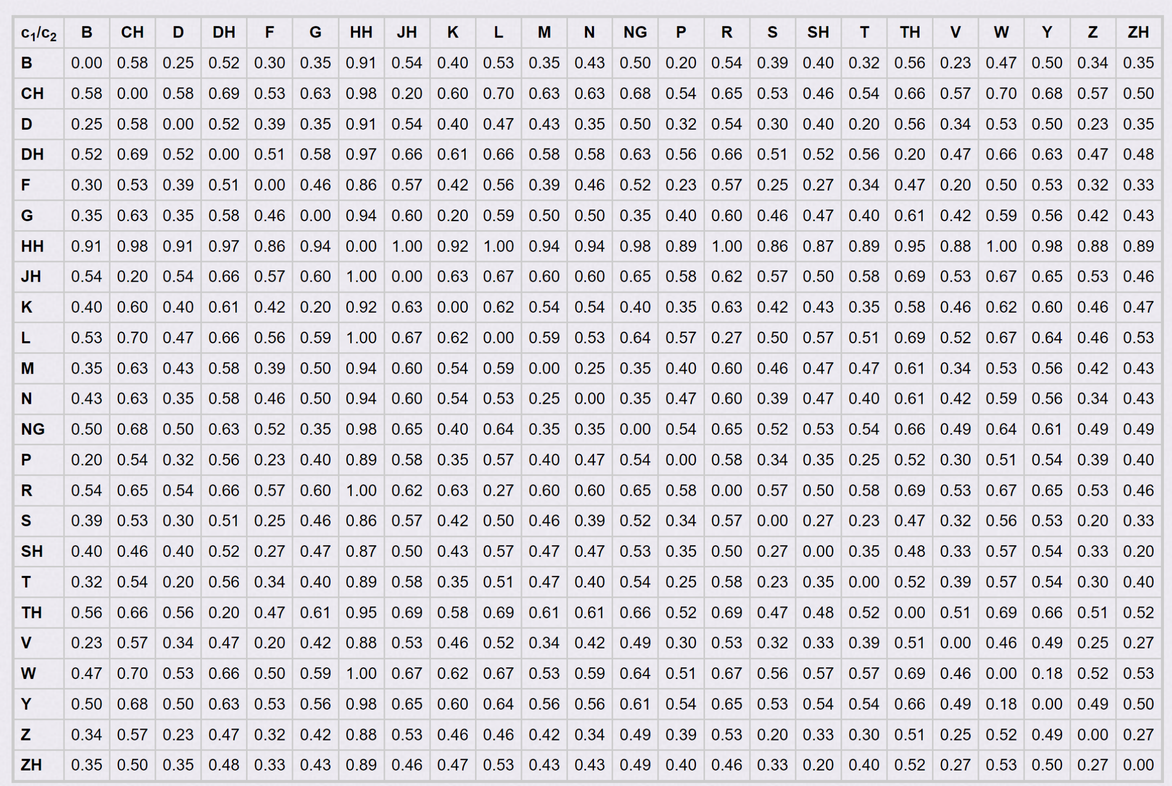


```
"ae d ah l t ax": [["ADULTER"]],
"ae d ah l t s": [["ADULT'S"],["ADULTES"],["ADULTS"]],
"ae d ax": [["ADDER"]],
"ae d ax l eh s n s": [["ADOLESCENCE"]],
"ae d ax l eh s n s iy": [["ADOLESCENCY"]],
"ae d ax l eh s n t": [["ADOLESCENT"]],
"ae d ax l eh s n t s": [["ADOLESCENT'S"],["ADOLESCENTS"],["ADOLESCENTS'"]],
"ae d ax l ey d": [["ADELAIDE"]],
"ae d ax l iy": [["ADDERLY"]],
"ae d ax m": [["ADAM"]],
"ae d ax m z": [["ADAM'S"],["ADAMES"],["ADAMS"]],
"ae d ax r": [["ADDER"]],
"ae d ax r ey sh n": [["ADORATION"]],
```

Figure 4.5: phonemes_to_corr_words.json

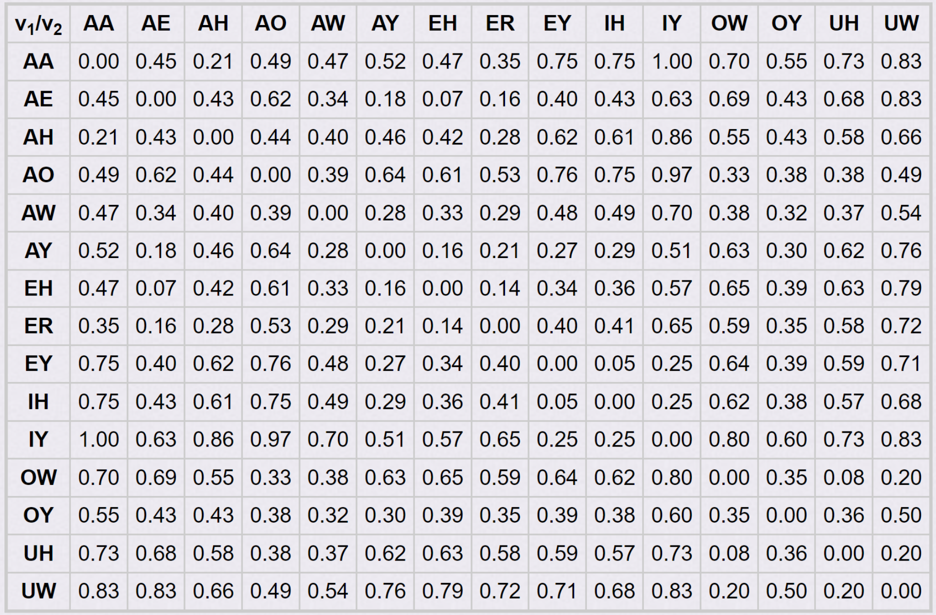
phonetic substitution, the score is determined by calculating the difference among individual phonetic units of candidates and the original word. Regarding insertion or deletion, it is determined by the phonetic units that are not present in one another. There involves a subjective decision that the difference score of each extra unit is relatively 1 by default since one additional unit could alter the sound noticeably. The detailed implementation is described in Figure 4.8 and has been validated with about five examples, covering all three edit operations and varying word lengths. This process produced the expected values, verifying the proper functioning of the code implementation.

In short, Levenshtein distance assist with identifying those words with distance-1 phonemes based on ARPAbet pronunciation from BEEP dictionary, while consonant and vowel distances help estimate the articulatory similarity between each substitutes and the original word. These phoneme scores are then combined with linguistic scores for subsequent ranking.



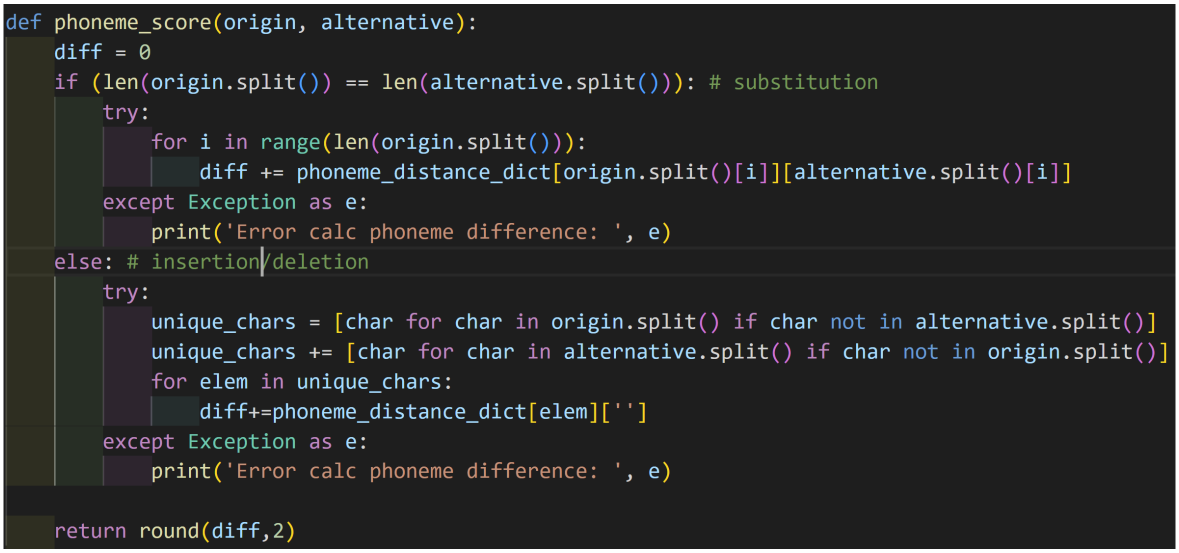
c_1/c_2	B	CH	D	DH	F	G	HH	JH	K	L	M	N	NG	P	R	S	SH	T	TH	V	W	Y	Z	ZH
B	0.00	0.58	0.25	0.52	0.30	0.35	0.91	0.54	0.40	0.53	0.35	0.43	0.50	0.20	0.54	0.39	0.40	0.32	0.56	0.23	0.47	0.50	0.34	0.35
CH	0.58	0.00	0.58	0.69	0.53	0.63	0.98	0.20	0.60	0.70	0.63	0.63	0.68	0.54	0.65	0.53	0.46	0.54	0.66	0.57	0.70	0.68	0.57	0.50
D	0.25	0.58	0.00	0.52	0.39	0.35	0.91	0.54	0.40	0.47	0.43	0.35	0.50	0.32	0.54	0.30	0.40	0.20	0.56	0.34	0.53	0.50	0.23	0.35
DH	0.52	0.69	0.52	0.00	0.51	0.58	0.97	0.66	0.61	0.66	0.58	0.58	0.63	0.56	0.66	0.51	0.52	0.56	0.20	0.47	0.66	0.63	0.47	0.48
F	0.30	0.53	0.39	0.51	0.00	0.46	0.86	0.57	0.42	0.56	0.39	0.46	0.52	0.23	0.57	0.25	0.27	0.34	0.47	0.20	0.50	0.53	0.32	0.33
G	0.35	0.63	0.35	0.58	0.46	0.00	0.94	0.60	0.20	0.59	0.50	0.50	0.35	0.40	0.60	0.46	0.47	0.40	0.61	0.42	0.59	0.56	0.42	0.43
HH	0.91	0.98	0.91	0.97	0.86	0.94	0.00	1.00	0.92	1.00	0.94	0.94	0.98	0.89	1.00	0.86	0.87	0.89	0.95	0.88	1.00	0.98	0.88	0.89
JH	0.54	0.20	0.54	0.66	0.57	0.60	1.00	0.00	0.63	0.67	0.60	0.60	0.65	0.58	0.62	0.57	0.50	0.58	0.69	0.53	0.67	0.65	0.53	0.46
K	0.40	0.60	0.40	0.61	0.42	0.20	0.92	0.63	0.00	0.62	0.54	0.54	0.40	0.35	0.63	0.42	0.43	0.35	0.58	0.46	0.62	0.60	0.46	0.47
L	0.53	0.70	0.47	0.66	0.56	0.59	1.00	0.67	0.62	0.00	0.59	0.53	0.64	0.57	0.27	0.50	0.57	0.51	0.69	0.52	0.67	0.64	0.46	0.53
M	0.35	0.63	0.43	0.58	0.39	0.50	0.94	0.60	0.54	0.59	0.00	0.25	0.35	0.40	0.60	0.46	0.47	0.47	0.61	0.34	0.53	0.56	0.42	0.43
N	0.43	0.63	0.35	0.58	0.46	0.50	0.94	0.60	0.54	0.53	0.25	0.00	0.35	0.47	0.60	0.39	0.47	0.40	0.61	0.42	0.59	0.56	0.34	0.43
NG	0.50	0.68	0.50	0.63	0.52	0.35	0.98	0.65	0.40	0.64	0.35	0.35	0.00	0.54	0.65	0.52	0.53	0.54	0.66	0.49	0.64	0.61	0.49	0.49
P	0.20	0.54	0.32	0.56	0.23	0.40	0.89	0.58	0.35	0.57	0.40	0.47	0.54	0.00	0.58	0.34	0.35	0.25	0.52	0.30	0.51	0.54	0.39	0.40
R	0.54	0.65	0.54	0.66	0.57	0.60	1.00	0.62	0.63	0.27	0.60	0.60	0.65	0.58	0.00	0.57	0.50	0.58	0.69	0.53	0.67	0.65	0.53	0.46
S	0.39	0.53	0.30	0.51	0.25	0.46	0.86	0.57	0.42	0.50	0.46	0.39	0.52	0.34	0.57	0.00	0.27	0.23	0.47	0.32	0.56	0.53	0.20	0.33
SH	0.40	0.46	0.40	0.52	0.27	0.47	0.87	0.50	0.43	0.57	0.47	0.47	0.53	0.35	0.50	0.27	0.00	0.35	0.48	0.33	0.57	0.54	0.33	0.20
T	0.32	0.54	0.20	0.56	0.34	0.40	0.89	0.58	0.35	0.51	0.47	0.40	0.54	0.25	0.58	0.23	0.35	0.00	0.52	0.39	0.57	0.54	0.30	0.40
TH	0.56	0.66	0.56	0.20	0.47	0.61	0.95	0.69	0.58	0.69	0.61	0.61	0.66	0.52	0.69	0.47	0.48	0.52	0.00	0.51	0.69	0.66	0.51	0.52
V	0.23	0.57	0.34	0.47	0.20	0.42	0.88	0.53	0.46	0.52	0.34	0.42	0.49	0.30	0.53	0.32	0.33	0.39	0.51	0.00	0.46	0.49	0.25	0.27
W	0.47	0.70	0.53	0.66	0.50	0.59	1.00	0.67	0.62	0.67	0.53	0.59	0.64	0.51	0.67	0.56	0.57	0.57	0.69	0.46	0.00	0.18	0.52	0.53
Y	0.50	0.68	0.50	0.63	0.53	0.56	0.98	0.65	0.60	0.64	0.56	0.56	0.61	0.54	0.65	0.53	0.54	0.54	0.66	0.49	0.18	0.00	0.49	0.50
Z	0.34	0.57	0.23	0.47	0.32	0.42	0.88	0.53	0.46	0.46	0.42	0.34	0.49	0.39	0.53	0.20	0.33	0.30	0.51	0.25	0.52	0.49	0.00	0.27
ZH	0.35	0.50	0.35	0.48	0.33	0.43	0.89	0.46	0.47	0.53	0.43	0.43	0.49	0.40	0.46	0.33	0.20	0.40	0.52	0.27	0.53	0.50	0.27	0.00

Figure 4.6: Consonant Distance Matrix Table: Phonetic Distance between Consonants [22]



v_1/v_2	AA	AE	AH	AO	AW	AY	EH	ER	EY	IH	IY	OW	OY	UH	UW
AA	0.00	0.45	0.21	0.49	0.47	0.52	0.47	0.35	0.75	0.75	1.00	0.70	0.55	0.73	0.83
AE	0.45	0.00	0.43	0.62	0.34	0.18	0.07	0.16	0.40	0.43	0.63	0.69	0.43	0.68	0.83
AH	0.21	0.43	0.00	0.44	0.40	0.46	0.42	0.28	0.62	0.61	0.86	0.55	0.43	0.58	0.66
AO	0.49	0.62	0.44	0.00	0.39	0.64	0.61	0.53	0.76	0.75	0.97	0.33	0.38	0.38	0.49
AW	0.47	0.34	0.40	0.39	0.00	0.28	0.33	0.29	0.48	0.49	0.70	0.38	0.32	0.37	0.54
AY	0.52	0.18	0.46	0.64	0.28	0.00	0.16	0.21	0.27	0.29	0.51	0.63	0.30	0.62	0.76
EH	0.47	0.07	0.42	0.61	0.33	0.16	0.00	0.14	0.34	0.36	0.57	0.65	0.39	0.63	0.79
ER	0.35	0.16	0.28	0.53	0.29	0.21	0.14	0.00	0.40	0.41	0.65	0.59	0.35	0.58	0.72
EY	0.75	0.40	0.62	0.76	0.48	0.27	0.34	0.40	0.00	0.05	0.25	0.64	0.39	0.59	0.71
IH	0.75	0.43	0.61	0.75	0.49	0.29	0.36	0.41	0.05	0.00	0.25	0.62	0.38	0.57	0.68
IY	1.00	0.63	0.86	0.97	0.70	0.51	0.57	0.65	0.25	0.25	0.00	0.80	0.60	0.73	0.83
OW	0.70	0.69	0.55	0.33	0.38	0.63	0.65	0.59	0.64	0.62	0.80	0.00	0.35	0.08	0.20
OY	0.55	0.43	0.43	0.38	0.32	0.30	0.39	0.35	0.39	0.38	0.60	0.35	0.00	0.36	0.50
UH	0.73	0.68	0.58	0.38	0.37	0.62	0.63	0.58	0.59	0.57	0.73	0.08	0.36	0.00	0.20
UW	0.83	0.83	0.66	0.49	0.54	0.76	0.79	0.72	0.71	0.68	0.83	0.20	0.50	0.20	0.00

Figure 4.7: Vowel Distance Matrix Table: Phonetic Distance between Vowels [22]



```
def phoneme_score(origin, alternative):
    diff = 0
    if (len(origin.split()) == len(alternative.split())): # substitution
        try:
            for i in range(len(origin.split())):
                diff += phoneme_distance_dict[origin.split()[i]][alternative.split()[i]]
        except Exception as e:
            print('Error calc phoneme difference: ', e)
    else: # insertion/deletion
        try:
            unique_chars = [char for char in origin.split() if char not in alternative.split()]
            unique_chars += [char for char in alternative.split() if char not in origin.split()]
            for elem in unique_chars:
                diff+=phoneme_distance_dict[elem]['']
        except Exception as e:
            print('Error calc phoneme difference: ', e)

    return round(diff,2)
```

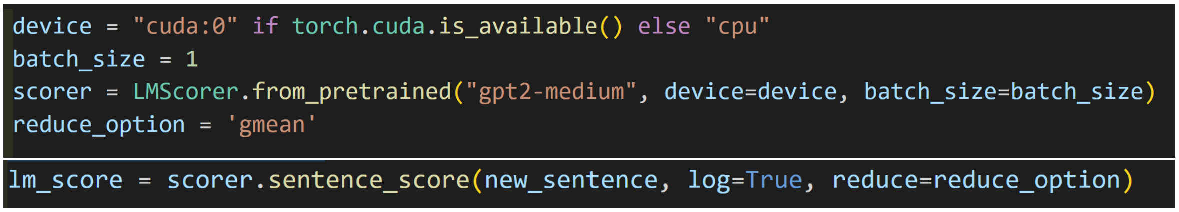
Figure 4.8: Phoneme Score Calculation

4.4 Language Model Scoring

In terms of scoring the sensibility of sentences with the aid of language models, the project implementation uses lm-scorer[27], a Python package that offers a simple interface to score sentences using various machine learning language model variants. As outlined in Chapter 3.4, the variant gpt2-medium is chosen for the optimal balance of performance and runtime efficiency.

In Figure 4.9, the implementation is based on the lm-scorer library source code, combining with logarithm to calculate the word probability distribution or the sentence score. This is because without it, the overall scores are too small (i.e. 0.000897199 or even less), requiring additional effort when it comes to tracking and analysing the comparison manually. With the application of logarithm, the multiplication of the probability turns into addition, resulting in more comparable scores (i.e. -7.016232).

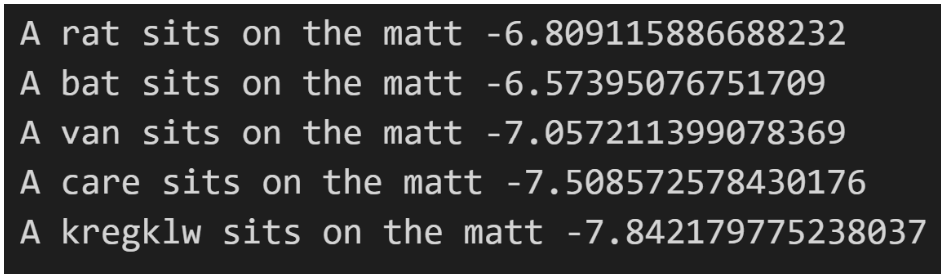
The implementation is tested with alternatives for ”cat” in the sample sentence ”A cat sits on the matt”. The substitutes are from comparatively sensible to nonsensical to identify if a higher score corresponds to more sensible candidates, along with examining if the scores indicate the sensibility of those new sentences replacing ”cat” with alternatives as expected.



```
device = "cuda:0" if torch.cuda.is_available() else "cpu"
batch_size = 1
scorer = LMScorer.from_pretrained("gpt2-medium", device=device, batch_size=batch_size)
reduce_option = 'gmean'
```

```
lm_score = scorer.sentence_score(new_sentence, log=True, reduce=reduce_option)
```

Figure 4.9: Sentence score by lm-scorer



```
A rat sits on the matt -6.809115886688232
A bat sits on the matt -6.57395076751709
A van sits on the matt -7.057211399078369
A care sits on the matt -7.508572578430176
A kregklw sits on the matt -7.842179775238037
```

Figure 4.10: Linguistic score test cases

In Figure 4.10, the left hand side represents alternative sentences and their respective language modelling scores compiled from the code presented in Figure 4.9 on the right hand side. The first two sentences are anticipated to be relatively sensible candidates, while the subsequent three are expected to be progressively less sensible, with the last one considered non-sensible. The scores are shown to follow the trend where the higher the score is, the more comprehensible its sentence is, and the scores decreasing are specifying sentences becoming less coherent, thus as anticipated.

Hence language modelling score is computed using lm-scorer, a Python package, with gpt2-medium variant. The technical execution is examined with various example sentences to confirm the score trend and ensure that the execution functions properly.

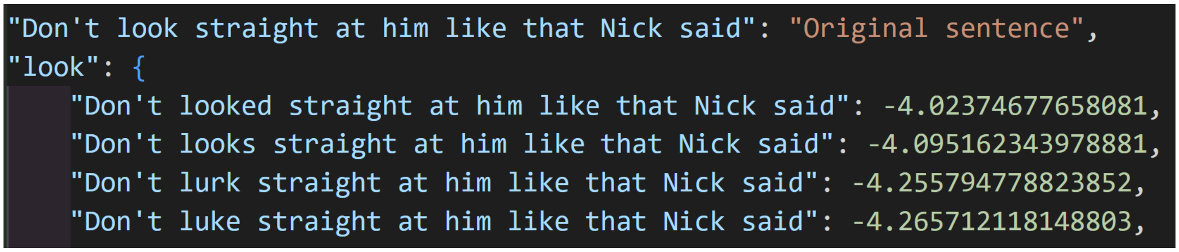
4.5 Word Filtering and Ranking

4.5.1 Word Filtering

Based on the observation of the performance and results during the initial implementation, there has been a re-evaluation regarding the pool of alternative options for individual selected words. Specifically, there is a refinement to exclude the lemma of original words from consideration in alternative options.

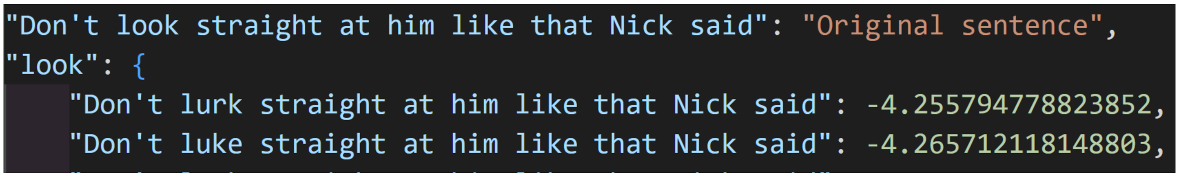
After compiling the list of top candidates for each prompt sentence, it becomes evident that the top alternative words tend to be the inflection[21] of a root word. This observation stems from the fact that, despite a slight variation to form the inflections, the overall structure and meaning of the sentence are still considered rather coherent, leading to high scores from the LM.

In the sentence above, the preferred alternatives of ”look” include ”looked” and ”looks”. Similarly, in another example like ”A cat sits on the matt”, the top substitutes would include ”cats” and ”catted”. Nevertheless, if the candidates throughout the test primarily consist of such grammatical variations of the original words, only differing slightly in their ending sounds such as /s/, /t/, /ih d/, this potentially shift the focus of the listeners towards spotting these subtle distinctions. This, however, is not the intended purpose of the project which is about hearing evaluation rather than merely pointing out the phonetic difference. Accordingly, along with excluding homophones of the original, lemmas of original words are



```
"Don't look straight at him like that Nick said": "Original sentence",
"look": {
    "Don't looked straight at him like that Nick said": -4.02374677658081,
    "Don't looks straight at him like that Nick said": -4.095162343978881,
    "Don't lurk straight at him like that Nick said": -4.255794778823852,
    "Don't luke straight at him like that Nick said": -4.265712118148803,
```

Figure 4.11: Before removing lemma



```
"Don't look straight at him like that Nick said": "Original sentence",
"look": {
    "Don't lurk straight at him like that Nick said": -4.255794778823852,
    "Don't luke straight at him like that Nick said": -4.265712118148803,
```

Figure 4.12: After removing lemma

also excluded to maintain the naturalness and accuracy of the test design. For instance, the lemma of all the words ”go”, ”going”, ”goes”, or ”went” is ”go”. In order to obtain lemmas of a word, lemmatisation is then applied, which is a fundamental technique commonly utilised in NLP and machine learning[11] to simplify words to their base form (lemma), reducing the variants while ensuring the consistency in text analysis. Test cases were executed using a target word along with a word set containing its lemma subset and other random words, which successfully returned the expected results with solely non-lemma words.

4.5.2 Ranking The Alternatives

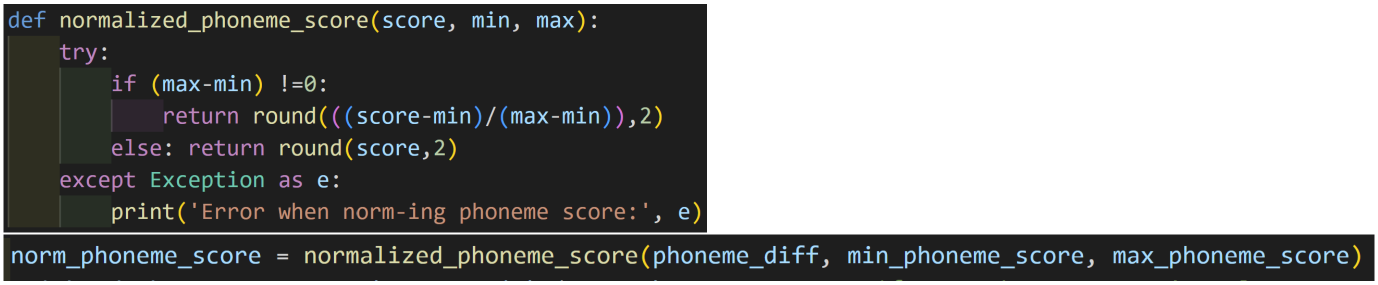
Combined scores

The score of each candidate sentence is determined based on both phoneme score and language model (LM) score, with a greater weight on the LM score. This is because ensuring the sensibility of the sentence is crucial to guaranteeing the intelligibility of the conveyed speech. In addition, since the phonemes are of distance-1, there is a higher chance that they sound close to each other, the extending task is to categorise which differences among the selected phoneme sets are less obvious than the others.

Upon computing the phoneme scores as outlined in section 4.3, these are normalised based on the minimum and maximum scores of each score comparison set (Figure 4.13). The higher the linguistic scores are, the more sensible sentences are; while greater phoneme scores reflect more distant sounds between two words. Therefore, the formula to calculate the final sentence score (Figure 4.14) is

$$sentence\_score = language\_score * higher\_weight - phoneme\_score * lower\_weight$$

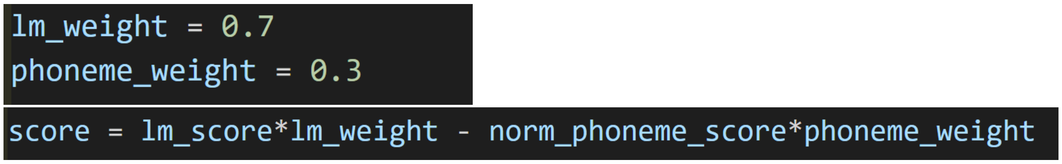
.



```
def normalized_phoneme_score(score, min, max):
    try:
        if (max-min) !=0:
            return round(((score-min)/(max-min)),2)
        else: return round(score,2)
    except Exception as e:
        print('Error when norm-ing phoneme score:', e)
```

```
norm_phoneme_score = normalized_phoneme_score(phoneme_diff, min_phoneme_score, max_phoneme_score)
```

Figure 4.13: Normalise phoneme scores



```
lm_weight = 0.7
phoneme_weight = 0.3
```

```
score = lm_score*lm_weight - norm_phoneme_score*phoneme_weight
```

Figure 4.14: Sentence score calculation

Ultimately, a final score deduced by combining phoneme distance and LM scores using lm-scorer library for each alternative sentences. The scores are then ranked in descending order for each unique dot sentence, ensuring the most ideal candidates are placed at the top of the each list.

Content word position-based ranking

Initially, the candidate sentences are ranked in accordance to final score for each provided sentence. However, the performance at the stage reveals that the replacement occurring towards the end of the sentence tends to yield higher LM scores. This is due to the nature of probability distribution over a set of words[24]. Across a sentence, the language model analyses each word and calculates the likelihood of the next word based on the ones before it. Consequently, if there is an unusual word near the start of a sentence, the model interprets that as an unfamiliar pattern, resulting to low scores for the remainder of the sentence due to its lack of coherence. In contrast, if the lack of coherence occurs towards the end of the sentence, it may not be identified till later, allowing the preceding words to maintain a higher probability. Therefore, it would be biased to group then rank all the alternative sentences together. A more impartial approach is to assess them based on the position of each substituted word. This thus allows the most suitable options to be placed at the forefront for each replacement position, ensuring a more precise evaluation.

4.6 MCQ Choice Selection

The implementation above produces a selection of top alternatives for each content word within prompt sentences. These aim to lead to confusion for the listeners when engaging with sentence audios. However, if the options are not thoroughly designed, there is a risk to inadvertently guide the listeners towards the answer i.e. options tend to point towards a specific word. Typically, if the option list consist of ”different, difference, differenced, sometimes” besides ”None of the above”, the participants are inclined to discard ”sometimes” and focus on identify the rest three related words. Thereby, it is crucial to design substitutes carefully to prevent participants from excluding options quickly or making speculation at ease. The quiz involves two approaches for later comparison: the new approach which utilises options derived from implementation process as outlined above, and a naive baseline where the options are generated randomly. The process of choosing choices for these two approaches share certain similarities yet comprising particular differences.

In both approaches, a correct answer is initially identified for each prompt, either a word in the sentence or ”none of the above”. In the new approach, for the former scenario, other options are the candidates of other content words in that prompt rather than alternative to the correct answer. The ”None of the above” option is typically applied when a sentence contains either too few content words (such as only one) or too few alternative options. In the baseline approach, the random word options are generated by an online generator[1] limited to a word size of 3, with a view to avoiding unnecessarily complex or lengthy words. The criteria for choosing random words includes ensuring that alternatives are audibly distinguishable from the words in the sentences. This involves selecting random words that are not present in the prompt, and are phonetically distinct from the given prompt’s words, making sure they are not lemmas of any words in the sentence. All considerations discussed are intent on minimising the likelihood of participants making speculation quickly and easily.

In addition, the chosen test delivery platform is Google Form. Due to time constrains to implement and inspect the approach, we opted for using an existing interface to deliver the MCQ listening test. Nevertheless, few platforms support audios embedding in the sentences. Besides, we aim to deliver an approximately ten-minute test, thus with short sentences lasting a maximum of six seconds derived from the dataset, a total of 100 sentences are included. With such a substantial number of sentences together with an advanced feature like audio embedding, existing platforms typically require payment to proceed. Google Form, on the other hand, offers to have audio files embedded and accommodates a large-scale test at no cost. As a result, Google Form was selected as the platform to deliver the listening test in this project.

4.7 Summary

This chapter has been delving deeply into technical aspects of attaining test materials in preparation for designing the MCQ quiz from given stimuli with those elements considered significant for a thorough design. This includes the implementation of phonetic and linguistic matches, as well as the consideration and implementation of further improvements based on observations of the implementation performance. Moreover, some particular components are taken into account when choosing the options from the compiled test materials to incorporate into quiz's choices. Consequently, the outcomes of the test, which will illustrate the comparison between the approach with carefully designed options and another with a randomly generated option list, are discussed in the following chapter.

Chapter 5

Evaluation

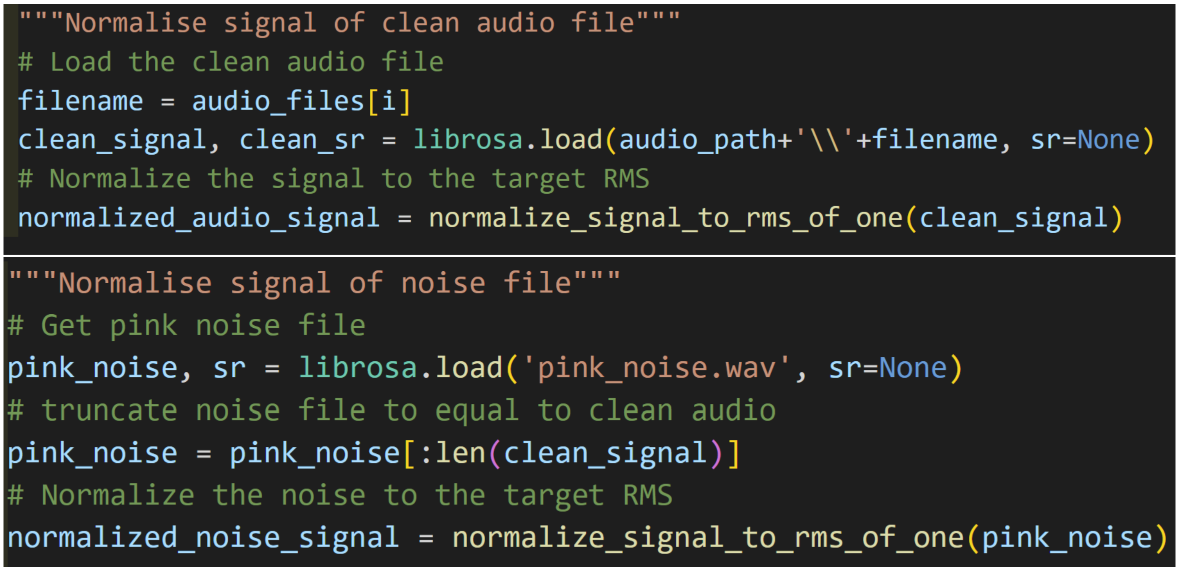
The project aims to reproduce a speech intelligibility evaluation method originally introduced by the University of Edinburgh with some modifications. The previous chapters detail the analysis and implementation necessary to meet the requirements and objectives of this study. After completing all the executions, it shifts to evaluating the effectiveness of the novel approach which is conveyed in an MCQ listening test format with meticulously selected word options. Furthermore, a naive baseline method where the word options are generated randomly while adhering to specific criteria discussed in Chapter 4.6, for comparison purpose in the evaluation. This chapter will provide more insights into the outcome analysis based on the provided materials along with the research compiled.

5.1 Methodology

5.1.1 Materials

The evaluation is delivered as an MCQ test consisting of 100 sentences which are derived from a published British dataset designed for speech intelligibility assessment[15]. Half of the sentences utilise compiled alternative word materials while the remaining half employ the baseline method. In the stimuli, each sentence is paired a corresponding clean speech audio recording. However, with the focus on assessing the hearing ability, these clean audios are then mixed with pink noise taken from a public online source[2], composed at five different SNRs evenly spaced between -3.0dB and 3.0dB. Thereby, each approach has 10 sentences at each SNR threshold to assess the intelligibility under different noise conditions and different word option strategies.

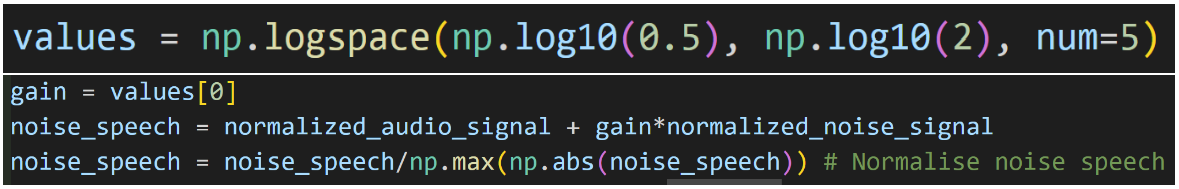
The initial step is to extract signal data from speech recordings and the pink noise file. Subsequently, the speech and noise signals are normalised to achieve a consistent amplitude level of 1 before mixing signals together. The normalisation process uses root mean square (RMS) computation. However, prior to normalisation, the pink noise data is clipped to match the length of each speech audio file.



```python
"""Normalise signal of clean audio file"""
# Load the clean audio file
filename = audio_files[i]
clean_signal, clean_sr = librosa.load(audio_path+'\\'+filename, sr=None)
# Normalize the signal to the target RMS
normalized_audio_signal = normalize_signal_to_rms_of_one(clean_signal)
```

```python
"""Normalise signal of noise file"""
# Get pink noise file
pink_noise, sr = librosa.load('pink_noise.wav', sr=None)
# truncate noise file to equal to clean audio
pink_noise = pink_noise[:len(clean_signal)]
# Normalize the noise to the target RMS
normalized_noise_signal = normalize_signal_to_rms_of_one(pink_noise)
```

Figure 5.1: Normalised a speech audio and pink noise



```python
values = np.logspace(np.log10(0.5), np.log10(2), num=5)
```

```python
gain = values[0]
noise_speech = normalized_audio_signal + gain*normalized_noise_signal
noise_speech = noise_speech/np.max(np.abs(noise_speech)) # Normalise noise speech
```

Figure 5.2: Gain values and noisy signal data

Following normalisation, the gain values are computed for five SNRs, each is subsequently applied to the normalised noise signal, adding the normalised speech signal, then resulting in the final noisy audio files of the corresponding noise ratio (refer to code in Figure 5.2). Ultimately, the noisy signal data is also normalised to prevent clipping, making sure that all signals remain in the range of -1 and +1.

After finishing design materials setup, the listening tests then begin where subjects listen to audio prompts and make a multiple-choice selection for each sentence. Half of the trials have choice alternatives generated by the novel method, while the remaining half have alternatives generated randomly. Questions are presented at different SNR levels in a mixed order throughout the test, with the ultimate aim of ordering them by intelligibility. Afterwards, the test scores using data from the two techniques will be analysed to determine whether the new approach better predicts this trend or not.

5.1.2 Subjects

The MCQ quiz is performed by 8 British participants aged between 20 and 30 years old, all reporting no impairments. All subjects are recruited outside the University of Sheffield and were sourced from my personal network. To start with, the invitations were sent via email containing the Google Form link where the test is delivered. The test did not ask for any personal information aside from their email addresses. Before proceeding with the test, all subjects must select option boxes as making agreement to certain conditions including the permission to use their data anonymously for future research and learning.

To ensure ethical compliance, the project underwent an ethical review and obtained approval, as documented in Appendix A. In addition, a participation information sheet and consent form for evaluation were prepared and also included in Appendix A.

5.1.3 Test Procedure

Upon all agreements are made and participants are deemed eligible to proceed with the test, they are initially provided with a sample audio mixed with the highest degree of noise interference (-3dB). Preconditions and instructions are also provided to guarantee participants' safety and the reliability of the outcomes. The participants are required to conduct the test in a quiet room using a desktop or laptop. They can listen either via a speaker or with a headphone, with headphone being more recommended. The test combines with potentially loud noise so participants are instructed to start with the volume at the minimum level, gradually increasing it to their comfortable level while listening to the sample audio. The instructions also emphasise that the audio must be played only once so the listeners need to pay close attention to each presented questions. Given the design of the Google Form where audio files are likely opened in a new tab, the listeners are advised to review the options before listening to the audios. Once all preconditions are met and instructions are followed carefully, participants can begin the test.

5.2 Results

The test employs two approaches, each covering half of the test, and the audio recordings are conducted at five levels of SNR. Accordingly, there are ten sample data points at each SNR threshold for each method plotted.

According to Figure 5.3, the baseline approach demonstrates a fluctuating trend while the novel approach reflects similar findings with the University of Edinburgh's Speech Center, where scores decline in noisier environments. The test is reported to take a maximum of approximately 20 minutes, denoting its efficiency and thus meeting the requirements. The error bars on the graph represent standard errors. The wide error bars indicate that the true means of the two approaches at each noise ratio are potentially overlapped. In this figure, a notable fall is observed in Baseline Approach at the SNR of 3.0dB. 3dB SNR means being asociated with the minimal noise, resulting in highly intelligible sentences, consequently high

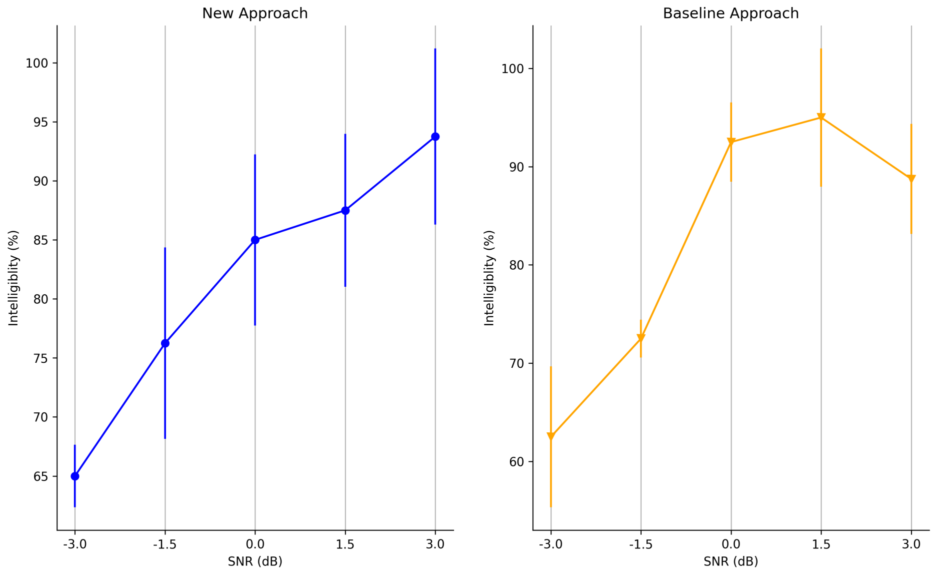


Figure 5.3: New Approach vs Baseline

scores are expected. However, the sudden drop observed here is due to a specific sentence where the correct word is almost fully masked by the noise. As a consequence, listeners tend to interpret it as another word which is not among the provided options. Thereby, a majority of participants opted for ”None of the above”, leading to the observed decline in scores.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
New Approach (/50)	39	39	40	42	45	42	39	40
Baseline Approach (/50)	37	40	41	42	47	40	43	39
Total(/100)	76	79	81	84	92	82	82	79

Table 5.1: Individual performance table
Note: P1 denotes Participant 1

The outcomes are also analysed using *t-test*, a statistical test used to compare the means of two groups. Specifically, the Paired Samples T-Test is utilised in this study to investigate the results obtained from the MCQ listening test, with the aim of determining whether there is a difference between two groups and evaluating the significance of that statistic difference. T-Test assesses whether the observed difference between two groups reflects a true gap or just occurs by chance[17].

According to Table 5.2, the p-value is 0.3241 exceeding the common significance level (typically set at 0.05). This also suggests that there is about 32.41% chance of observing those score differences between two approaches purely by chance, assuming that the two methods are not truly distinct, and this is a relatively high probability [a1]. Moreover,

Group	New Approach	Baseline Approach
Mean	39.63	41.13
Standard deviation (SD)	4.37	3.00
Standard error of the mean (SEM)	1.55	1.06
Sample size (N)	8	8
Two-tailed P value	0.3241	
95% confidence interval of the difference	from -4.84 to 1.84	
t	1.0607	
Degree of freedom	7	
Standard error of difference	1.414	

Table 5.2: Results of Paired T-Test analysis (using online T-Test calculator[16])

the 95% confidence interval of difference ranges from -4.84 to 1.84, which includes 0. This indicates that there is no statistically significant difference between the means of two subjects [a2]. The t-value, which represents a ratio of the difference between the means of the two sample sets and the variation within the sample sets[17], is just approximately 1.0 [a3].

In summary, these three findings [a1], [a2], and [a3] propose that there is insufficient evidence to confidently confirm a significant difference in performance of two approaches.

5.3 Result Conclusions

According to the potential overlap of true means between two approaches as depicted in Figure 5.3 and the statistic analysis based on Table 5.2, it is not evident that the two systems perform differently. However, the novel approach demonstrates a consistent trend and the effectiveness, which align with the findings in the original research.

Chapter 6

Conclusions

6.1 Project Summary

This project's task is to reproduce an efficient method for assessing speech intelligibility in speech enhancement, drawing inspiration from a technique recently introduced by The Center for Speech Technology Research of the University of Edinburgh (June, 2023). While sticking to specific procedures of the original study, this project introduces some variations. Different from the University of Edinburgh which uses audio-visual stimuli, the project utilises published audio recordings, previously designed for intelligibility testing, to gather the sentence prompts. Ultimately it demonstrates that the technique is reproducible and achieves similar results. In addition to replicating the novel approach, the project also seeks to discover the differences between this approach with carefully designed word options and a naive baseline where alternatives are generated randomly.

The design materials are implemented by making use of the fact that there are various factors having impacts on speech intelligibility, which include cognition, hearing ability and environmental conditions. Additionally, the potential elements contributing to confusion during listening or misunderstanding encompass phonetic similarities, linguistic perception, and memory recall. By putting the emphasis on assessing hearing and contemplating these components, the test is designed to require minimal cognitive effort, taking advantage of phonetic and linguistic matches along with varied noise levels to create an effective assessment platform.

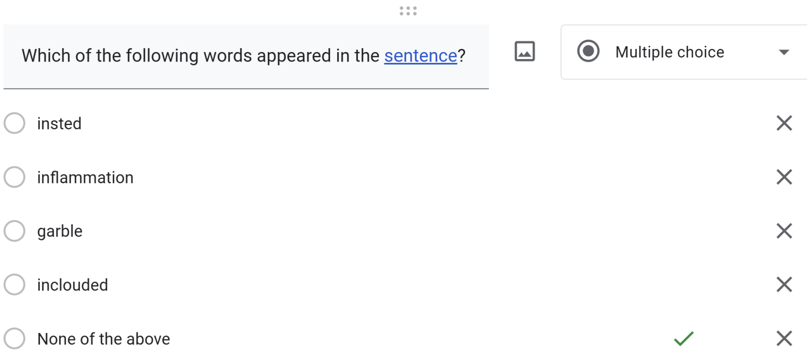
The entire project is implemented using Python and Python libraries. Typically, a Python library, lm-scorer, is utilised to score the linguistic match, assessing the coherence of candidate sentences. Word pronunciations are represented in ARPAbet format. Phoneme matching is then evaluated using Levenshtein Distance metric, equipped with more detailed analysis based on consonant and vowel distance. Through observation of the performance of the execution of the hypotheses, further improvements have been identified and implemented accordingly. The enhancements involve modifying the scoring and ranking system to rank changes according to each content word of new sentences, refining the selection of content words per prompt to exclude proper nouns alongside stop words, together with avoiding using lemmas as alternative options. On top of diligent considerations required to bring out optimal design materials for multiple-choice options, there are also standard criteria to take into account to pick substitutes either from the compiled set or at random.

The MCQ listening test comprises 100 sentences, evenly divided between the novel approach and the naive baseline method. The survey is completed by 8 native British participants without reported impairments. The survey is reported to take approximately 20 minutes to complete, thus proving to meet efficiency criteria. Recruited participants' performance revealed ambiguous distinctions between the deliberately chosen word method and the baseline. Moreover, the reproduced method demonstrates a similar trend to the results observed in the original study.

6.2 Potential Improvements

The stimuli actually contains a larger set of prompts, with candidates for all sentences compiled and stored in JSON files. In this project, the test is delivered via Google Form, gathering the first 100 sentences and their corresponding audios from the stimuli with alternative options manually chosen from the compiled JSON files. However, an interface could be developed with configuration settings to automatically select the sentences from the stimuli and alternative options from the compiled materials. This customisable interface potentially introduces more versatile features, such as a flexible test scale based on individual time limits or purposes, differing noise levels based on participant performance during the test, and the incorporation of diverse types of noises mixed with the original clean recordings rather than solely pink noise in use.

Furthermore, this project quiz has been conducted with a limited number of participants, less than ten in total. So far, it has not provided conclusive evidence to support the hypothesis of noticeable distinction in intelligibility measurement between the two techniques. Expanding the scale of participants with consistent recruitment criteria probably produces more accurate analysis and insights into the contrast between these two systems.



Which of the following words appeared in the sentence?

● Multiple choice

○ insted ×

○ inflammation ×

○ garble ×

○ inclouded ×

○ None of the above ✓ ×

Figure 6.1: Strange words in the option potentially help participants narrow down correct choice and affect the test quality

Last but not least, the phoneme dictionary utilised in this project is BEEP dictionary which is last updated in 1996. Due to a prolonged period without revision, the dictionary allegedly contains outdated or unusual words that could be mistaken for misspellings. For example, as in Figure 6.1, words like ”insted” or ”inclouded” are included the dictionary with their own distinct phonemes (/ih n s t ih d/ and /ih n k l aw d ih d/ respectively), thus appearing in the option list. However, individuals might assume these words are misspelled versions of ”instead” and ”included” (even though they have different pronunciations in the dictionary). Subsequently, they are prone to dismissing these options automatically, thereby affecting the quality of the test. As a result, using a more updated phoneme dictionary if available could enhance the test quality further.

In short, a customisable interface enables greater flexibility in test scale and caters to a range of testing purposes via different configurations. Besides, by exploiting a more up-to-date dictionary and expanding the number of participants, the test can yield more conclusive and reliable result.

Bibliography

[1] Random word generator [online], n.d. Available at https://randomwordgenerator.com/.

[2] ARC, A. Pink noise test track, February 2021. Available at https://www.arcaudio.com/post/pink-noise-test-track. Accessed on 23-04-2024.

[3] Bonino, I. A., Leibold, L., and Buss, E. Release from perceptual masking for children and adults: Benefit of a carrier phrase. *Ear and Hearing 34* (January 2013), Pages 3–14. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3529824/.

[4] Buschermöhle, M. International matrix tests — reliable speech audiometry in noise, n.d. Available at https://www.mack-team.de/pdf/ht-internationalermatrixtest.pdf. Accessed on 29-11-2023.

[5] Caulfield, J. What is a proper noun?, December 2022. Available at https://www.scribbr.co.uk/nouns/proper-noun/#:~:text=A%20proper%20noun%20is%20a,%2C%20songs%2C%20and%20other%20media. Accessed on 26-04-2024.

[6] Dan Jurafsky, J. H. M. Chapter h: Phonetics. *Speech and Language Processing (3rd ed. draft)* (February 2024), Figure H.1 Page 2. Available at https://web.stanford.edu/~jurafsky/slp3/H.pdf.

[7] Face, H. distilbert/distilgpt2, April 2023. Available at https://huggingface.co/distilbert/distilgpt2#distilgpt2. Accessed on 22-02-2024.

[8] Fordington, S., and Holland Brown, T. Speech banana image. *An evaluation of the Hear Glue Ear mobile application for children aged 2–8 years old with otitis media with effusion* (October 2020), Page 2. Available at https://www.researchgate.net/figure/Speech-banana-showing-the-range-of-frequencies-used-in-everyday-speech-The-letters_fig1_346409806. Accessed on 14-03-2024.

[9] Gad, A. F. Implementing the levenshtein distance in python, April 2021. Available at https://blog.paperspace.com/

implementing-levenshtein-distance-word-autocomplete-autocorrect. Accessed on 20-02-2024.

[10] Gad, A. F. Measuring text similarity using the levenshtein distance, April 2021. Available at https://blog.paperspace.com/measuring-text-similarity-using-levenshtein-distance. Accessed on 15-2-2024.

[11] GeeksforGeeks. Python—lemmatization with nltk, January 2024. Available at https://www.geeksforgeeks.org/python-lemmatization-with-nltk/. Accessed on 27-04-2024.

[12] GeeksforGeeks. Top 5 pretrained models in natural language processing (nlp), January 2024. Available at https://www.geeksforgeeks.org/top-5-pre-trained-models-in-natural-language-processing-nlp/. Accessed on 19-02-2024.

[13] GeeksforGeeks. Removing stop words with spacy, n.d. Available at https://www.geeksforgeeks.org/removing-stop-words-nltk-python/amp/#removing-stop-words-with-nltk. Accessed on 11-04-2024.

[14] Google. Introduction to large language models, n.d. Available https://developers.google.com/machine-learning/resources/intro-llms#what_is_a_language_model.

[15] Graetzer, S., Akeroyd, M., Barker, J., Cox, T., Culling, J., Naylor, G., Porter, E., and Viveros-Muñoz, R. Sample_clarity_uttereances. *Dataset of British English speech recordings for psychoacoustics and speech processing research: The clarity speech corpus* (April 2022). Available at https://salford.figshare.com/articles/dataset/Dataset_of_British_English_speech_recordings_for_psychoacoustics_and_speech_processing_research/16918180?file=33988673.

[16] GraphPad. T test calculator, n.d. Available at https://www.graphpad.com/quickcalcs/ttest1/. Accessed on 30-04-2024.

[17] Hayes, A. T-test: What it is with multiple formulas and when to use them, December 2023. Available at https://www.investopedia.com/terms/t/t-test.asp#:~:text=The%20t%2Dtest%20produces%20two,of%20the%20two%20sample%20sets. Accessed on 29-04-2024.

[18] Honnibal, M., and Montani, I. spacy 3.7.4, February 2024. Available at https://pypi.org/project/spacy. Accessed on 03-04-2024.

[19] Loizou, P. C. Speech quality assessment. *Studies in Computational Intelligence* (January 2011), Pages 623–654. Avaialble at https://ecs.utdallas.edu/loizou/cimplants/quality_assessment_chapter.pdf.

[20] Nam, E. Understanding the levenshtein distance equation for beginners, February 2019. Available at `https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0`. Accessed on 15-02-2024.

[21] Nordquist, R. Efficient intelligibility evaluation using keyword spotting: A study on audio-visual speech enhancement, July 2019. Available at `https://www.thoughtco.com/inflection-grammar-term-1691168`. Accessed on 27-04-2024.

[22] Occen. Phonetic distance between words with application to the international spelling alphabet, October 2018. Available at `https://www.occasionalenthusiast.com/phonetic-distance-between-words-with-application-to-the-international-spelling-alphabet` Accessed on 14-04-2024.

[23] OpenAI. Better language models and their implications, February 2019. Available at `https://openai.com/research/better-language-models`. Accessed on 22-02-2024.

[24] Priyanka. Perplexity of language models, February 2023. Available at `https://medium.com/@priyankads/perplexity-of-language-models-41160427ed72`. Accessed on 26-04-2024.

[25] Robinson, T. BEEP dictionary, August 1996. Available at `https://www.openslr.org/14/`. Accessed on 05-12-2023.

[26] Sayer, I. M. Misunderstanding and language comprehension. *Procedia: social behavioral sciences 70* (January 2013), Pages 738–748. Available at `https://www.sciencedirect.com/science/article/pii/S1877042813001195`.

[27] simonepri. Language model based sentences scoring library, 2020. Available at `https://github.com/simonepri/lm-scorer`. Accessed on 07-11-2023.

[28] Tafseer Ahmed, Muhammad Suffian Nizami, M. Y. K. Discovering lexical similarity using articulatory feature-based phonetic edit distance. Pages 2–4. Available at `https://arxiv.org/ftp/arxiv/papers/2008/2008.06865.pdf`.

[29] Valentini-Botinhao, C., Aldana Blanco, A. L., Klejch, O., and Bell, P. *Efficient Intelligibility Evaluation Using Keyword Spotting: A Study on Audio-Visual Speech Enhancement* (June 2023). Available at `https://salford.figshare.com/articles/dataset/Dataset_of_British_English_speech_recordings_for_psychoacoustics_and_speech_processing_research/16918180?file=33988673`.

[30] Van den Borre, E., Denys, S., van Wieringen, A., and Wouters, J. The digit triplet test: a scoping review. *International journal of audiology 60*, 12 (April 2021), Pages 946–963. Available at `https://pubmed.ncbi.nlm.nih.gov/33840339/`.

[31] XUE, W., VAN HOUT, R., CUCCHIARINI, C., AND STRIK, H. Assessing speech intelligibility of pathological speech: test types, ratings and transcription measures. *Clinical linguistics phoneticsa 37* (December 2021), Pages 52–76. Available at `https://www.tandfonline.com/doi/full/10.1080/02699206.2021.2009918#:~:text=According%20to%20this%20definition%2C%20an,to%20use%20only%20existing%20words`.

Appendices

Appendix A

Ethical Application and Approval

The
University
Of
Sheffield.

Downloaded: 07/05/2024
Approved: 12/04/2024

Khanh Tran
Registration number: 200188605
Computer Science
Programme: Computer Science

Dear Khanh

PROJECT TITLE: Efficient Speech Intelligibility Evaluation Using Keyword Spotting
APPLICATION: Reference Number 059710

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 12/04/2024 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 059710 (form submission date: 09/04/2024); (expected project end date: 08/06/2024).
- Participant information sheet 1135656 version 1 (09/04/2024).
- Participant consent form 1135657 version 1 (09/04/2024).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Luke Whitham
Ethics Administrator
Computer Science

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/research-services/ethics-integrity/policy
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

Efficient Speech Intelligibility Evaluation Using Keyword Spotting

Information Sheet: version April 15th 2024

You are being invited to take part in a research project. Before you decide whether or not to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

What is the project's purpose?

Your participation means to conduct listening tests for evaluating an undergraduate project. The project has concerned implementing a novel approach for subjective human speech intelligibility assessment, and evaluating its effectiveness. The approach is based on keyword spotting combined with multiple-choice question quizzes. A human participant listens to a sentence and is presented with a list of candidate words that it may have included. The main focus of the project is to automate the design of the candidate word lists, i.e., such that the correct answer is not obvious unless the sentence has good intelligibility. The listening tests will therefore be comparing this novel method against a naive baseline, i.e., selecting random common words as alternatives. I will then evaluate the approach by measuring how well it can predict the intelligibility of sentences which have already been scored for intelligibility in a previous study.

The results will be made publicly available so that all researchers can learn what works well and what does not.

Why have I been chosen?

You have been selected because you responded to our call for volunteers and you matched our inclusion criteria, i.e., you believe that your hearing is normal and you are a fluent speaker of English.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and you will be asked to sign a consent form. You can still withdraw at any time up until May 8th 2024. After this date, your data will have been fully anonymised and it will no longer be possible to remove it from the dataset). If withdrawing, you do not have to give a reason. If you wish to withdraw from the research, please contact the Principle Investigator, using the contact details at the end of this Information Sheet. Withdrawing will carry no negative consequences.

What will happen to me if I take part? What do I have to do?

The study will consist of listening tests delivered online via a web interface which you will perform from your own computer in a quiet room wearing your own headphones. Clear instructions will be provided and you will have the opportunity to choose a date that is convenient for you.

The experiment will involve listening to a sequence of noisy sentences (about 1-6 seconds each). For each sentence, you will be asked a multiple choice question (MCQ) of the form “Which of the following words appeared in the sentence”, and provided with four options, including “None of the above”. You will record your responses using a simple interface.

The experiment will be delivered over a web interface using https://quizizz.com/join. Prior to the experiment, you will provide consent by answering a series of Yes/No questions. You will not be permitted to proceed to the experiment unless you have agreed your consent. Before the experiment begins, you will need to adjust your volume to a comfortable listening level. The experiment will last no more than 15 minutes including time for reading the instructions and completing. The interface will record your MCQ responses and no other information.

In addition to your ratings, we will collect your age, which is important information for a listening test as age is known to influence hearing ability.

What are the possible disadvantages and risks of taking part?

The experiment requires listening to sounds over headphones. This could have the potential to cause hearing damage if care was not taken. To avoid this we advise you to make sure that the volume controls are initially set to low levels before being adjusted upwards to a comfortable listening level, and not to exceed safe levels.

If you do experience any unexpected discomfort during the experiment, you are asked to stop or take a break.

What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this work will lead to advances in future speech intelligibility testing approaches. These advances will improve the performance of everyday devices such as noise-canceling headphones, augmented reality headsets and commercial hearing aids.

Will my taking part in this project be kept confidential?

Yes. Any personally identifying information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. You will not be able to be identified in any reports or publications. The only data that we will publish is your anonymised rating scores and your anonymised age.

What is the legal basis for processing my personal data?

According to data protection legislation, we are required to inform you that the legal basis we are applying in order to process your personal data is that 'processing is necessary for the performance of a task carried out in the public interest' (Article 6(1)(e)). Further information can be found in the University's Privacy Notice https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

What will happen to the data collected, and the results of the research project?

We will collect your age and your rating scores for the sound samples heard during the different listening sessions of the experiment. The collected data will be associated with an anonymised participant identifier. Up until August 15th 2024, we will keep a file that links your anonymised identifier with your name and contact information. This file will be kept securely at the University of Sheffield and used to allow us to remove your data if you wish to withdraw from the experiment. These personal details will not be revealed to people outside the project. After August 15th 2024 the file containing your name and contact information will be completely destroyed, the collected data will therefore become entirely anonymised.

The anonymised collected data will also be used to disseminate the results of the proposed approach of the project in scientific publications. In an open science approach, the anonymised opinion scores that you provided during the listening test will be made publicly available so they can be used for future research and learning.

Who is organising and funding the research?

This project is being organised by the Department of Computer Science of University of Sheffield. There is no funding for this project.

Who is the Data Controller?

The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly.

Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by the Department of Computer Science.

What if something goes wrong and I wish to complain about the research?

Please contact the Principal Supervisor, Prof. Jon Barker (jp.p.barker@sheffield.ac.uk), if you wish to raise a complaint. If you feel your complaint has not been handled to your satisfaction, you can also contact the Head of Department, Prof. Heidi Christensen (heidi.christensen@sheffield.ac.uk), who will then escalate the complaint through the appropriate channels. If the complaint relates to how the participants' personal data has been handled, information about how to raise a complaint can be found in the University's Privacy Notice:

https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

Who can I contact for further information?

If you require additional information please contact the primary researcher, Khanh Linh Tran via Email: kltran1@sheffield.ac.uk

You will be given a copy of this information sheet and a signed consent form to keep.

Thank you for your participation in this study.

University of Sheffield

“Efficient Speech Intelligibility Evaluation Using Keyword Spotting” Consent Form

Please tick the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information sheet dated 15th April 2024 and the project has been fully explained to me. (If you will answer No to this question please do not proceed with this consent form until you are fully aware of what your participation in the project will mean.)		
I have been given the opportunity to ask questions about the project.		
I agree to take part in the project.		
I understand that taking part in the project will include listening to noisy sentences over headphones and using a software interface to carry out the test.		
I understand that choosing to participate as a volunteer in this research does not create a legally binding agreement nor is it intended to create an employment relationship with the University of Sheffield.		
I understand that my taking part is voluntary and that I can withdraw from the study at any time before May 8th 2024; I do not have to give any reasons for why I no longer want to take part and there will be no adverse consequences if I choose to withdraw.		
How my information will be used during and after the project		
I understand my personal details such as name, phone number, address and email address etc. will not be revealed to people outside the project.		
I understand and agree that other authorised researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.		
I give permission for the anonymised data that I provide (i.e, listening tests’ scores and my age in years) to be made publicly available so they can be used for future research and learning.		
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.		

Name of participant [printed] Signature Date

Name of Researcher [printed] Signature Date

Project contact details for further information:

- If you require additional information, have questions, or wish to withdraw please contact the primary researcher :
  Khanh Linh Tran, Email: kltran1@sheffield.ac.uk
- If you want to report an incident, wish to raise a complaint, or if you have some concerns about the project, you can contact the Principal Supervsior:
  Professor Jon Barker, Email: j.p.barker@sheffield.ac.uk

If you feel your complaint has not been handled to your satisfaction, you can also contact the Head of Department, Prof. Heidi Christensen (heidi.christensen@sheffield.ac.uk), who will then escalate the complaint through the appropriate channels.