University of Sheffield

Computer Says 'Know': ASR Confidence and Transcription

Adam Spencer

Supervisor: Professor Jon Barker

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

in the

Department of Computer Science

May 23, 2023

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: Adam Spencer

Signature:

Date: May 23, 2023

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims and Objectives

1.2 Overview of the Report

Chapter 2

Literature Survey

The purpose of this chapter is to develop an understanding of the existing literature on the topics of automatic speech recognition, neural network confidence, speech corpora, and computer-aided transcription. Each topic is discussed in its own section, with subsections to explore parts of each topic in greater depth.

2.1 Automatic Speech Recognition

2.1.1 What is ASR?

Automatic Speech Recognition (ASR) is a term used to describe technology which allows computers to recognise and produce a text transcription of spoken language. The research and development of technology involving speech has been a part of computer science since the late 1930s[1, 2], with rudimentary ASR systems being constructed as early as the 1950s[3]. These early attempts at recognising human speech treated it as a 'pattern matching' problem, the theory being that words could be constructed by matching the pattern created in a speech signal to corresponding spoken phonemes[1]. Despite speech recognition fundamentally being a problem of matching speech patterns to text, these early attempts were not particularly robust; requiring re-tuning by a human operator in order to match the specificities of each new speaker's speech patterns[3].

The 1970s saw the application of statistical techniques to improve the robustness of ASR systems[1], the most widely adopted method being the application of 'hidden Markov models' (HMMs)[4]. Use of HMMs continued through the '90s[5] and is still in use today[6].

2.1.2 Hidden Markov Models

In order to better understand the way in which modern ASR systems have developed, why they function the way they do, and what limitations have yet to be solved, it's important to explore the approach which historically saw the widest adoption; hidden Markov models.

First, what is a 'Markov model' (also known as a Markov process)? In his 1960 work[7], Dynkin describes a Markov process using the example of a randomly-moving particle in space;

“ If the position of the particle is known at the instant t, supplementary information regarding the phenomena observed up till the instant t (and in particular, regarding the nature of the motion until t) has no effect on prognosis of the motion after the instant t (for a known "present", the "future" and the "past" are independent of each other). ”

From his description, we can draw the following assumptions for modelling a system as a Markov process;

- The system consists of states.
- The system is *in motion*, i.e. moving between states.
- This motion is random.
- The motion observed prior to t (e.g. $t-1$) does not influence the motion following $t+k$ where $k \geq 1$.
- Because the particle is constantly moving between states, the state at time t depends only on the state immediately prior, $t-1$.
- There is some probability, p, that the system moves from one state to another.

In a *Hidden* Markov Model (HMM), the states and transition probabilities between them are known, but for some output sequence the order and selection of states used to produce the output is not known. Knowing both the states and transition probabilities, it is therefore possible to calculate the most probable set of inputs used to produce the output.

To apply this model to speech, treat speech as a continuous sequence of discrete states, where each state is a feature vector representing an acoustic signal (either whole words, phonemes or even sub-phonetic features[5]). Assuming that each state is generated from a probabilistic distribution correlated with other states in the model[8] (i.e. probability that one state follows another) and having trained these distributions on known data, the output signal (i.e. the speech signal) can be used to determine the most probable sequence of tokens spoken. These tokens may then be decoded by a language model to construct a transcription[5].

Despite making up much of the research foundational to modern ASR, Markov models have a crucial flaw when applied to speech; parts of speech are dependent on more than just the part immediately before (i.e. $t-1$). For instance, in a presentation discussing *hats* it is unlikely that the word *cat* would be used, despite the phonetics of the word being largely the same.

2.1.3 The Transformer

Skipping ahead from the mid-1980s to the current day 'state-of-the-art', ASR has moved towards what is known as the 'encoder-decoder' model. At a high level, this model consists of two key parts; an encoder and a decoder. The encoder processes (*encodes*) input audio into features, these features are aligned with language and then processed by the decoder (*decoded*)

to produce an output transcript[9]. The key difference between modern approaches and the classical HMM-based approach is the use of widely researched 'machine learning' techniques, including various forms of neural network[10, 11, 12, 13].

Recent research has proposed a new network architecture called the *Transformer*[14], aiming to reduce the computational complexity of encoder-decoder models by forgoing convolutional or recurrent neural networks (CNNs and RNNs) and instead relying on 'attention'. The motivation for the *Transformer* can be understood as follows;

- CNNs (e.g. [15]) and RNNs (e.g. [16]), while popular, have greater per-layer computational complexity than self-attention[14].
- Recurrent neural networks must perform $O(n)$ sequential operations for a sequence length n, whereas self-attention has a constant (i.e. $O(1)$) maximum number of sequential operations, enabling parallel computation[14].
- By allowing each layer in the encoder and decoder to attend to the whole output of the previous layer,

In a multilayer network, attention layers are used to build relations between separate parts of an input sequence by allowing each node (or 'attention head'[17]) to attend to all outputs from the previous layer. A technique refered to as 'multi-head attention'[14] enables attention to be calculated in parallel for all inputs in a sequence.

The calculation for attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Where

- Q is known as the *Query* vector,
- K is known as the *Key* vector, and
- V is known as the *Value* vector.

The output of the attention function is described as "a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key"[14]. To understand what these vectors denote in the context of speech recognition, we must understand what the different types of attention *are* and how their use *differs* in the encoder and decoder;

- self-attention layers (in both the encoder and decoder) take input from the previous layer within the same block (i.e. encoder or decoder). In this layer, Q, K, and V are all calculated by multiplying an input embedding (a vector representation of the raw input sequence) by three different trained weight vectors.

- cross-attention layers (present only in the decoder) enable mixing of the encoder's output with the output of the decoder's previous self-attention block, allowing the decoder to attend over the whole encoded input. In this layer, only K and V are taken from the encoder; Q is dervied from the previous decoder block.

For the sake of clarity, a simplified version of the operation of the transformer (as explained in [14]) is as follows;

1. An input sequence is transformed into an embedding, which is then input into the *encoder stack* (the layers which make up the encoder) after being modified using *positional embeddings*, which act to preserve information about the position of each part of the input sequence.

2. The *attention head* of each self-attention layer in the encoder first multiplies the input by three trained weight matrices to acquire the values of Q, K, and V, then calculates Attention using these values as input. Each attention head is trained with its own weight matrix, so there are an equal number of Attention results and attention heads. The resulting attention matrices are concatenated and then multiplied by another weight matrix to produce the layer's output to be fed forward to the next layer.

3. The output of the final encoder layer is transformed into a pair of K and V vectors and used as input (via *cross-attention* for the decoder, which works almost identically to the encoder (i.e. a stack of self-attention blocks). The key difference between the encoder and decoder is that the decoder uses previous decoded output as its Q vector and it may not self-attend to all of the output at once, only the output earlier to its current position (unlike the encoder which attends to the whole sequence simultaneously)

4. Finally, the output of the decoder goes through some normalisation steps to find and output the most probable *output embedding*.

An *output embedding* may be thought of as a way to link together an expected output (e.g. a word) and a vector for use in one-hot encoding space. In the Transformer's decoder this allows a score represented in one-hot encoding space to be immediately matched to an output.

Though this is all rather complex, the key points to understand are that a Transformer uses both an encoder and decoder, the decoder can attend to only previously decoded output, and these two 'blocks' work together to produce an output.

2.1.4 Evaluating ASR Systems

As with any computational model, the performance of an ASR system is evaluated in terms of both speed and accuracy. Calculating the speed of a model is simple enough; just time it! Accuracy, however, is not so simple because there are multiple metrics upon which a reference and ASR-generated transcript (known as a 'hypothesis') may differ.

There are three different levels at which to calculate the similarity between a reference and hypothesis; the word-level, phoneme-level, and character-level[18] Comparing words and characters is relatively self-explanatory, the transcripts are either split up into individual words or characters for comparison. A phoneme-level comparison involves transforming textual words into the parts of speech which constitute them, known as phonemes.

The accuracy calculation to derive an 'error-rate' (also known as the edit-distance[19]) is as follows;

$$WER = \frac{S + D + I}{S + D + C}$$

where:

- S is the number of *substituted* words (words which appear in the place of another correct word)
- D is the number of *deleted* words (words which aren't present in the hypothesis but appear in the reference)
- I is the number of *inserted* words (words which do not appear in the reference but appear in the hypothesis)
- C is the number of *correct* words (words which appear in both the reference and hypothesis)

Word error-rate (WER) is the most commonly used metric in the literature[20], though arguments can be made that it is not entirely representative of the degree to which a system has an understanding of speech. Take, for example, a reference which contains a compound word like 'soundproof' or 'eggshell' – if an ASR system were to output 'sound proof' or 'egg shell' it would have incurred one substitution and one deletion equating to a WER of 200% despite having produced an output which is representative of the input.

Judging by calls to work to produce a metric to replace WER still being made today[21, 22] and the relative lack of published ASR papers discussing results using metrics other than WER, there is not yet a viable alternative.

2.1.5 Problems in ASR

Despite their ubiquity, modern ASR systems aren't without fault.

Cutting edge systems like *(*wav2vec) are touted as being capable of achieving 'greater-than-human' scores on specific datasets[23, 24, 25] such as *LibriSpeech*[26], achieving as low as 1.4% error[27]. LibriSpeech consists entirely of English audiobook recordings which have been selected in-part based on their quality[26]; not particularly representative of everyday speech[22], lacking features such as speaker overlap which are common in conversation[28]. An evaluation of modern proclaimed 'state-of-the-art' ASR systems found WER scores averaging approximately

17% when faced with real-world conversational data[22] despite reporting results below 4% on LibriSpeech.

Racial disparities in the accuracy of ASR systems has also been reported, with average WER scores for black speakers almost double that of white speakers[29]. The authors of this article attribute this performance difference to the acoustic models used in the ASR systems rather than the language models, thus concluding that there is a lack of training data from black speakers.

These two problems serve indicate that, in order to improve accuracy, ASR systems should be trained on more diverse data in terms of the context of the speech and the demographic of speakers.

2.2 Whisper

In late September 2022, the OpenAI research laboratory (known for such projects as GPT-3/4 and ChatGPT) released a new open-source ASR system called 'Whisper' [30] which uses the encoder-decoder Transformer architecture discussed in section 2.1.3. It is described as a 'zero-shot' model, meaning it is expected to produce good results *without* any dataset-specific fine tuning. Whisper is unique in being very large (trained on 680,000 hours of speech data), open-source, and fully supervised; all the training data used to create the model has been accurately labeled and quality-checked by humans, unlike much larger unsupervised (or semi-supervised) models such as 'BigSSL' (1,000,000+ hours of data) [24].

Unsupervised training is appealing for training speech recognisers because there is a wealth of unlabeled recordings, and labeled recordings are uncommon for less widely-spoken languages[31]. Unsupervised systems have a clear disadvantage, however, when compared to supervised; they lack clear decoder mappings[30], meaning that even for a successfully encoded input there may not be a clear mapping from that input into a speech token. To solve this, fine-tuning to map encodings to decoded text tokens is done on the part of the model's developers, though this is a precarious route to overfitting; if the model is too fine-tuned to its training data, performance will suffer when faced with data which isn't well represented in the training set. For example, if an unsupervised model were trained using the voices of young people, it may perform with considerably poorer accuracy when used to transcribe elderly speakers due to differences inherent to their speech[32].

The way Whisper functions is discussed in detail in section 4.3.

2.3 Confidence

Neural networks will always produce an output for an accepted input, no matter how likely or unlikely the output is of being correct. The network may even assign a very large probability to its given output; consider, for example, a classifier trained to predict the city in the UK which someone was born in given the co-ordinates of their current address. The classifier simply assigns the closest city to their address and often makes a correct prediction. However,

if the input co-ordinates were somewhere in Iceland it may assign a high probability to their place of birth being in Inverness because it is much closer than any of the other UK cities, even though the true probability that this hypothetical Icelander was born in Inverness is quite low.

The point of this example is to illustrate that the probability assigned by a neural network is not always a good indication of correctness. An estimation of a model's expected correctness is refered to as *confidence*, i.e. how *confident* is the model that its prediction is correct?

2.4 Speech Corpora

2.5 Transcription

2.5.1 Manual Transcription

2.5.2 Fully-Automatic Transcription

2.5.3 Semi-Automatic Transcription

2.6 Summary

Chapter 3

Requirements and Analysis

The objective of this work is not to produce a fully-working, infallible system which aims to receive actual use by transcribers, rather, the aim of this work is to explore the current state of the field of ASR and to understand the extent that current ASR technology could provide aid to a human transcriber. With the purpose of facilitating an extensive evaluation, this chapter shall list the requirements for this work to meet its objective and provide a detailed analysis of each requirement.

3.1 Understand the motivations of computer-aided transcription

Before exploring how a computer system may aid a human transcriber, it is important to understand;

- *why* a human transcriber may require aid;
- to *whom* a computer-aided transcription system would provide benefit; and
- what the *extent* of such a benefit would be.

If this report does not properly motivate computer-aided transcription to a reader, it will have partially failed in its purpose

3.2 Generate transcripts using ASR

Evaluating the quality of ASR transcription requires a key set of data; ASR-generated transcripts. Rather than comparing different ASR systems, Whisper[30] has been chosen as the only system to use for generating transcripts because;

- it is new (made available in September 2022);
- it is entirely free and open-source, meaning it is easily modifiable and available to be used without licence; and

- it reportedly achieves very good results across different speech corpora.

As mentioned in the literature review, Whisper is implemented using *PyTorch*, meaning this work would benefit greatly from access to high-performance GPUs. This would enable fast turnaround times when transcribing large speech corpora and thus enable rapid evaluation and tweaking of settings to minimise erroneous results.

The key to generating useful transcripts is some high-quality speech recordings from a speech corpus. While preliminary testing of Whisper may use data from any available corpora, it would be very useful to obtain some data which is;

- not present in Whisper's training data, to prevent the model from regurgitating labels for data it has already seen;
- is well-suited to Whisper's particularities, aiming to maximise the usefulness of results; and
- is representative of real data which would benefit from computer-aided transcription, as to enable more practical evaluation.

Two preliminary aims, therefore, are to understand what kind of data is suited to Whisper and then what kind of data would benefit from computer-aided transcription. Once these aims are understood, a suitable dataset may be gathered and used for evaluation, however it is also useful to understand the caveats related to using well-suited data! The naïve assumption that all data seen by *any* computer system is 'perfect' would misrepresent the usefulness of the system in question. To combat this, this work must properly acknowledge the limited extent to which a computer-aided transcription system using Whisper is viable, and evaluate how the viability could be increased to be more applicable to real-world tasks.

3.3 Implement various confidence measures

Neural network confidence is widely discussed in the literature. Considering the aim of this work, it shall focus on re-creating and applying existing measures of confidence to Whisper, whether through modification to the model itself or through inference of the model's output.

If some such modifications fall out of scope of the project, this work would still benefit from a discussion of how those approaches may be applied to a future system and what advantages they may bring.

3.4 Understand the effectiveness of selected confidence measures

Evaluation of the extent to which Whisper may aid human transcription may be done by comparing the accuracy of Whisper's predictions against a reference and the reported model confidence.

This type of evaluation requires taking the following steps to complete;

1. Selection of suitable measure(s) of system accuracy
2. Extensive normalisation of text to facilitate accurate measurements
3. Visualisation of results

The standard across surveyed literature is *word error rate* (WER), despite potential limitations. For the sake of evaluation, other measurements such as *phone error rate* (PER) and *character error rate* (CER) should be calculated alongside WER.

3.5 Explore designs for computer-aided transcription

It would be of great utility to understand how system confidence may aid a human transcriber as this would further refine the 'lens' through which the system may be evaluated, and as such is vital to the completion of this work. There's limited use in a purely theoretical exploration of a computer system such as this, which is designed to be interfaced with by a human. Instead, demonstrating the benefits of a computer-aided transcription system would be easily facilitated using a graphical program.

A number of considerations are required for the design of this system specifically due to its intended nature to serve as an example rather than a final implementation, including;

- several design iterations should be produced to demonstrate different extents of computer aid;
- each design decision should be discussed thoroughly to demonstrate the intended effect and any notable caveats; and
- a suitable number of screenshots are required in the appendix to demonstrate use of the system.

3.6 Evaluation

3.7 Ethical, Professional and Legal Issues

Chapter 4

Design

4.1 Speech Data

According to its authors, Whisper's robustness is due likely in part to its use of a language model in its decoder[30]. Though likely beneficial for keeping track of sentence context, this poses a potential threat to the models accuracy in a number of circumstances, including;

- Misspoken words or sentences with improper syntax (e.g. 'then' instead of 'than'), for these errors may be corrected by the model, despite being inaccurate to the original recording.
- Disjoint terms (e.g. 'book purple dish soap'), as such terms are highly unlikely to occur in sequence and thus the language model will not consider them a probable output.

To combat these drawbacks, this work will use a conversational speech corpus rather than one made of spoken disjoint terms. While not representative of all speech, an argument can be made that the majority of speech which must be transcribed (e.g. conference recordings, courtroom hearings, lectures, etc.) has a maintained context throughout and is not significantly formed of disjoint terms, though the potential for disjoint terms to be present in a recording should be acknowledged as a potential area of weakness for Whisper.

While introducing the corpus selected for this work that the original objective was to explore the impact of age-related changes to speech on ASR performance, and for this goal the *LifeLUCID* corpus[33] was determined to be the best match. Despite the scope of the project having since changed, the data gathered fits the new objective very well, as it is formed of 52 recordings of conversations between 104 discrete speakers aged between 8 and 85 years old. They are solving a 'spot-the-difference' task, and the data selected for this work was recorded in normal conditions (that is, they can hear and communicate with each other normally).

The corpus' authors mention that the reference transcripts were generated by an ASR system and only one channel's audio was human-corrected. The ability to compare the quality of ASR output to a reference transcript is required to evaluate this work, thus, only

the human-corrected transcript and corresponding audio channel were used to ensure the references are reliable. This leaves 52 10-minute recordings of individual speakers with gaps where the other participant is speaking.

4.1.1 TextGrid Format

The reference transcripts are supplied in *Praat TextGrid* format which is produced by the Praat software suite[34]. The TextGrid format consists of each individual part of speech (words, hesitations, mid-sentence silences, silences when the other participant is speaking etc.) being present in consecutive entries with the time in the recording which they start and finish.

```
intervals [14]:
  xmin = 21.05
  xmax = 21.47
  text = "BUSH"
```

Figure 4.1: Example of an entry in TextGrid format

Figure 4.1 is an example of a single 'interval' in the TextGrid format. A larger example is available in Appendix 7 You may observe that `xmin` and `xmax` denote the points in the recording at which the section starts and ends, and `text` denotes the content of the section. Considering that each entry appears consecutively and that there are over 1,000 in each file, the format is not easily human-readable.

4.1.2 Data Preparation

There are a number of issues with using the data in its original format, including;

1. Whisper struggles to maintain alignment when transcribing long form data[30], so the approximate 10-minute length of each recording requires shortening to maintain system performance;
2. TextGrids are not human-readable; and
3. The output of the ASR system should be stored alongside the reference transcripts to ease evaluation, which is not possible using TextGrids.

The solution to the first problem would best be solved by splitting the conversation recordings into individual utterances. Luckily, the human-evaluated references include metadata which shows the start- and stop-times of each word and non-word part of speech, meaning it can easily be split into individual utterances without using voice activation detection or other automatic techniques. Using the documentation for LifeLUCID it was possible to determine that there are two types of non-speech token;

1. 'Break' tokens – these are tokens which denote the speaker is not mid-utterance; either listening to the other participant or engaged in irrelevant discussion (these latter parts are silenced in the recordings).

2. 'Junk' tokens – these denote either:

   - The speaker has paused (but the other participant is not speaking).
   - A bell or dog bark is being played as part of their task (these are silent in the recording).
   - Hesitations (e.g., 'umm', 'uhhh', etc.)
   - Other non-speech, non-breaking tokens (not specified in their documentation but present in the transcripts).

For the purpose of this work, an utterance is defined as an uninterrupted piece of speech without any long pauses. By providing threshold values for the minimum length of a 'break' token and maximum length of a 'junk' token, the boundaries at which utterances start and end can be easily computed from the reference TextGrids. The utterance boundaries can then be used to extract individual utterances from the full-length recordings, resulting in a series of numbered audio files.

The second and third problems can be solved together by changing from the TextGrid format to JSON (JavaScript Object Notation). This format is human-readable[35] and able to hold all the data and metadata required for this work, including a way to reference the original piece of audio it represents.

4.2 Running Whisper

Whisper comes packaged with models of various sizes, requiring between approximately 1 and 10GB of VRAM and an increasing amount of time to produce transcripts. Considering the objective to create an automatic transcription system which uses entirely free and open-source software, it is worth assuming that the individuals or institutions who would benefit the most from this system are those without the resources to rely on professional manual transcription. It follows, then, that these 'target users' would not have access to high-powered computers and instead rely on consumer-grade hardware to generate transcriptions. Thus, for the purposes of this work, the medium English-language model was selected as it requires only 5GB of VRAM and takes approximately half as much time to produce transcripts as the larger models[30].

4.2.1 High-Powered Computing

4.3 Confidence Scoring

Whisper does not have a clear confidence scoring system in its unaltered state. Therefore, for it to be viably used to aid a human transcriber, some method to estimate system

confidence must be implemented. At a high level, there are two sources from which to estimate confidence: Whisper's standard output, and its internal processes. Table 4.1 gives a brief comparison of the benefits and drawbacks associated with each of these sources of confidence;

Score Source	Benefits	Drawbacks
Standard Model Output	Does not require any modification to Whisper.	Only shows an average probability score per utterance.
Model Internal Scoring	Allows access to per-word scoring and the steps the model takes to converge on an output.	Requires modification to Whisper.

Table 4.1: Comparison between sources of model confidence

The following subsections provide a detailed explanation of how each approach works and should help illustrate the benefits and drawbacks of each.

4.3.1 Confidence From Model Output

Though it does not yield a clear confidence score, Whisper does output various data relating to its processing of the input. These include;

- `avg_logprob` – The average of the log token-probability for a segment of speech (discussed further below)
- `compression_ratio` – The ratio of the length of the UTF-8-encoded text to its gzip-compressed representation. Due to the way gzip operates, this ratio indicates the 'repetetive-ness' of the decoded text; a higher ratio means that the result is more repetetive, suggesting that there may have been a decoding error.
- `temperature` – Before producing an output, if the `avg_logprob` is below a certain threshold or the `compression_ratio` is above some threshold, the model will treat the decoding as failed and compute a new output with an increased temperature parameter. Temperature is used to introduce some randomness while computing predictions, therefore the higher the final output `temperature` value is, the more randomness had to be introduced in order to determine the given output.
- `no_speech_prob` – Whisper is trained to complete many tasks, one of which being the detection of non-speaking moments in a recording. This value indicates the model's predicted probability that there is no speech in the input audio file.

Of these dataThis work will focus on the `avg_logprob` metric, which is the average of the log probabilities for each token in a segment of the input, which are computed as the

log softmax of the *logits*, which can be thought of as unnormalised 'scores' that have been assigned to each token while decoding. The softmax function is used to normalise these scores such that their sum is equal to 1 (allowing them to be used as probabilities).

The logits themselves are calculated during each forward pass through the decoder, derived from both learned positional and token embeddings. In the context of the Transformer's decoder, learned positional embeddings can be thought of as a way to learn information about token positions in the transcript[36]. Token embeddings act like an abstraction on a one-hot encoding layer, matching a token to a weight vector. By using token- and learned position-level weights with cross-attention from the encoder[14], Whisper acts like an "audio-conditional language model"[30]; the scores (logits) assigned to tokens are based on both the encoded audio input and a model of language.

4.3.2 Confidence From Model Internals

4.4 A System For Transcription

Chapter 5

Implementation and Testing

5.1 Preparing the Data

5.1.1 Generating Utterances

The contents, beginning, and end of every utterance where computed using the data available in the *TextGrid* files using a Python script named `get_utterances`. This script operates over a directory containing *TextGrid* files, writing out the utterances as files in *JSON* format.

The script also takes as args; a minimum time between tokens required to end the utterance and a maximum pause time allowed within one utterance. These thresholds allow utterances to be fine-tuned by a user, leading to fewer drawn-out or unreasonably short utterances.

JSON was selected due to its ability to be easily read and understood by a human, unlike *TextGrids*. This allowed for simple verification of the data without the need for more specific software to view the files.

5.1.2 Audio Segmentation

Another Python script named `segment_audio` was created to generate audio files for each utterance. Given two directories as input; one containing `.json` files (as output by the `get_utterances` script) and the other containing `.wav` files representing each audio recording, the audio is split along the beginning and end times of each utterance and output to a new directory.

This script uses the *python-soundfile* module[37] to load audio files into *NumPy*[38] arrays. By multiplying the sampling rate of the audio by the start- and end-times of each utterance, the array indices at the start and end of each utterance are computed. Array slices between these indices represent each utterance, which can then be saved to new audio files using the *python-soundfile* module.

5.2 ASR With Whisper

Whisper is available as a Python module named `whisper`[39]. The module features a `transcribe()` function to transcribe audio files given as a parameter to the function and return an object containing the output of Whisper.

Chapter 6

Results and Discussion

6.1 Simulating Computer-Aided Transcription

An effective computer-aided transcription system must reduce the human cost required to lower the amount of errors a transcription to an acceptable level. By ordering transcribed utterances using some confidence metric and letting a human transcriber make corrections, an effective system should see the error rate fall more rapidly than if the results were corrected in a random order.

A simple simulation was devised in order to test the effectiveness of each potential measure of system confidence by using the formula for WER as follows;

1. For a set of M utterances $U = \{u_0, u_1, \dots u_M\}$ (either a single conversation or entire corpus), the subsitutions (S_i), deletions (D_i), insertions (I_i), and number of words in the reference (N_i) are calculated for each utterance, u_i.

2. U is ordered using a confidence metric.

3. The WER, w_i, is calculated for a slice of U containing all items including and following some utterance, u_i;
$$w_i = \frac{\sum_{j=i}^{M}(S_j + D_j + I_j)}{\sum_{k=0}^{M} N_k}$$

4. Increment i by 1 then repeat the previous step for all $0 \leq i \leq M$. Notice that the denominator is the same for all slices but the numerator changes to simulate each utterance being corrected in order.

5. Output is a set $W = \{w_0, w_1, \dots w_M\}$, where some item w_i is equal to the WER of U after correcting all utterances prior to u_i.

These results shall be analysed by plotting graphs of WER against the percentage of utterance which have been human-corrected, henceforth refered to as 'Cost'.

6.2 Simulation Results

This section details the results acquired from running this simple simulation using results ordered with each of the following metrics;

- `avg_logprob` taken directly from Whisper's standard output;
- utterance-average confidence score; and
- utterance-minimum and -maximum word-confidence scores.

Where "utterance-minimum and -maximum word-confidence" refers to an ordering of utterances based on each utterances minimum and maximum word-level confidence score.

Results which show a per-conversation average have a shaded section to show the standard deviation from the mean, where the mean is the coloured line on the graph.

Dashed 'random order' lines show the WER/Cost trend expected if the results were manually corrected in a random order. A performant metric for ordering results should have a plot showing a line which dips below the 'random order' line. A metric which follows the trend of the 'random order' line is thus performing the same as if the results were randomly ordered and therefore ineffective for computer-aided transcription.

6.2.1 Utterance-average confidence versus `avg_logprob`

Figure 6.1: Per-conversation average WER when evaluating in non-descending order of utterance `avg_logprob`

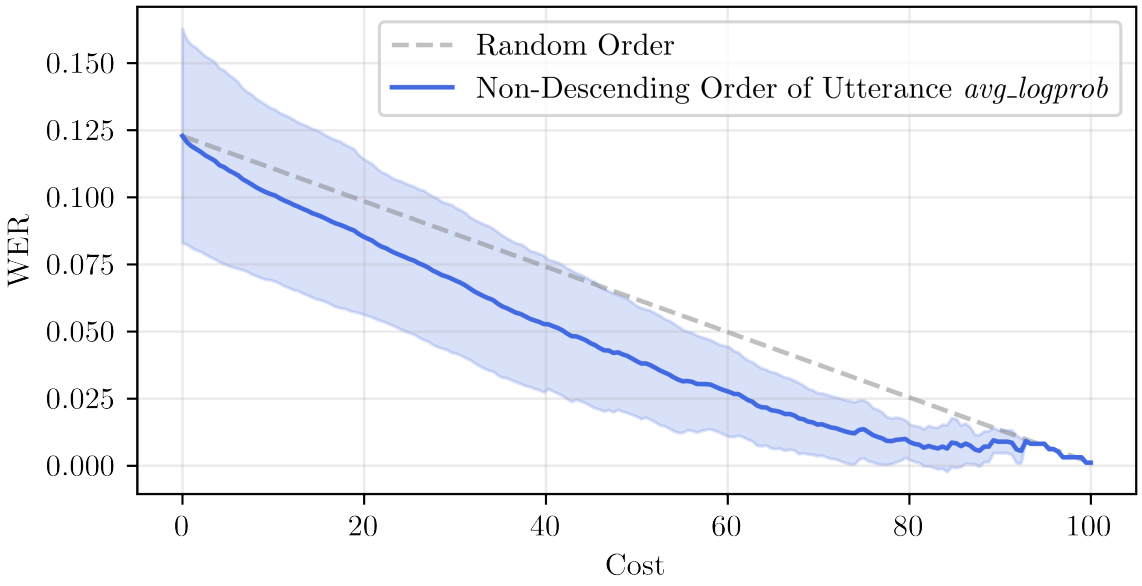


Figure 6.2: Per-conversation average WER when evaluating in non-descending order of utterance-average confidence

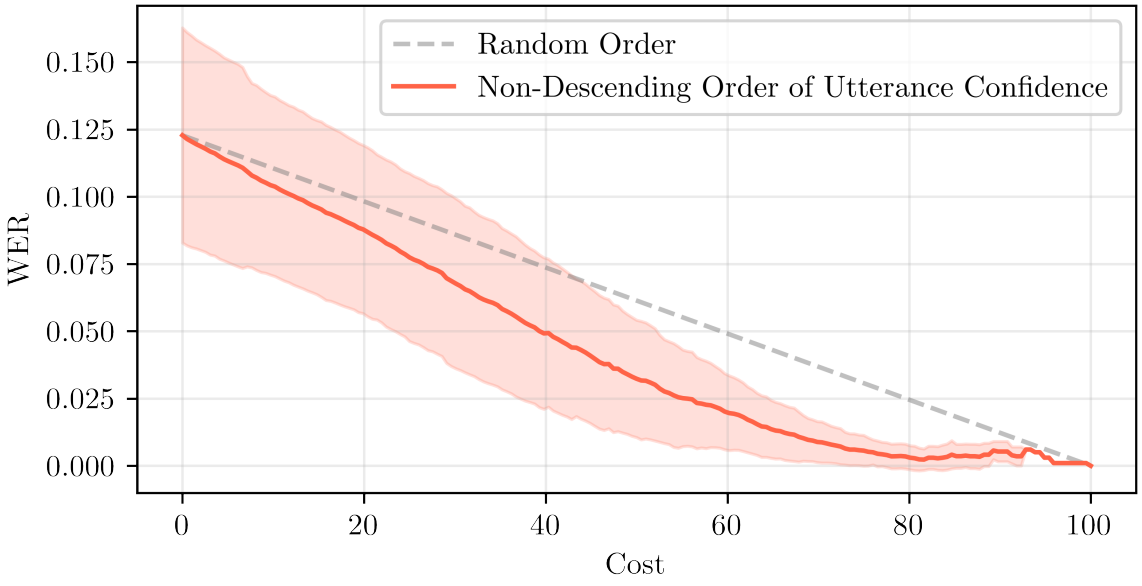


Figure 6.3: Comparing per-conversation average performance of confidence and `avg_logprob`

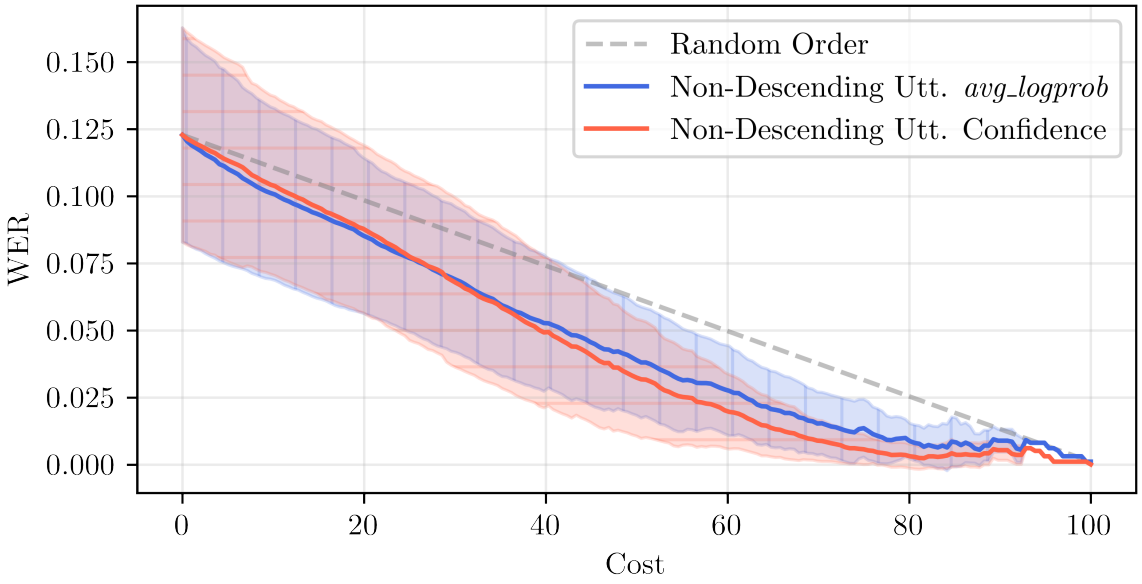
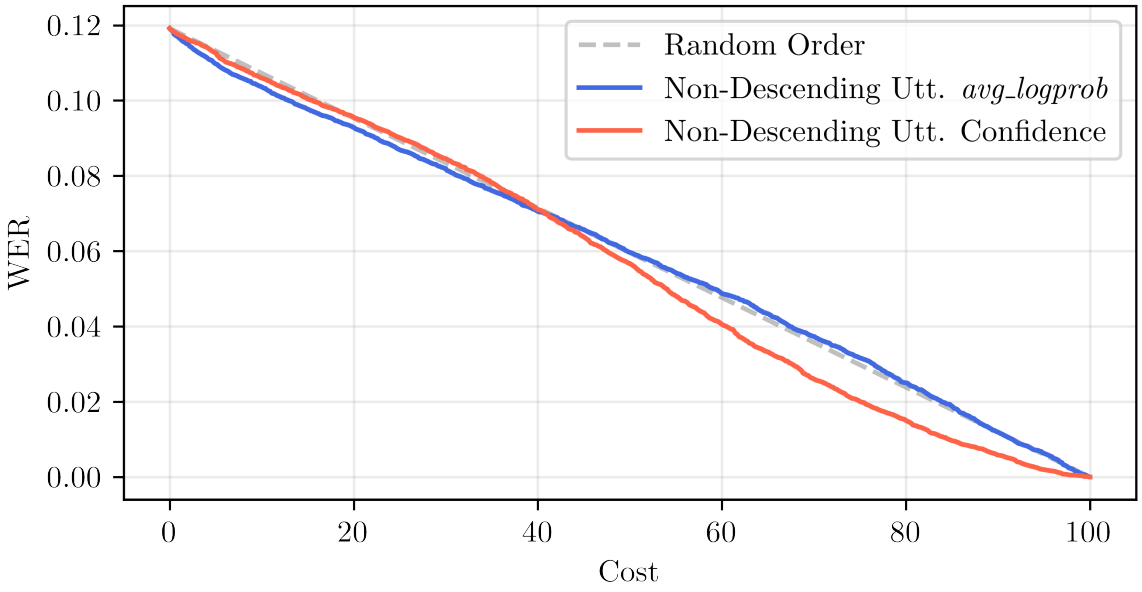


Figure 6.4: Comparing whole-corpus performance of confidence and `avg_logprob`



6.2.2 Per-conversation average results using word-confidence

Figure 6.5: Ordered by non-descending utterance-minimum word-confidence

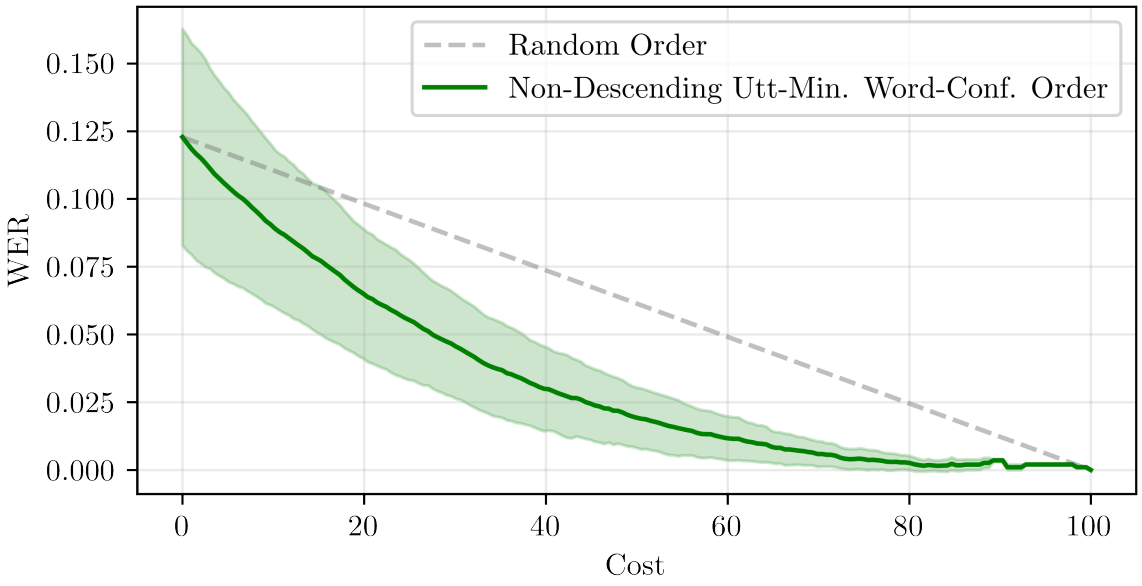


Figure 6.6: Ordered by non-ascending utterance-minimum word-confidence

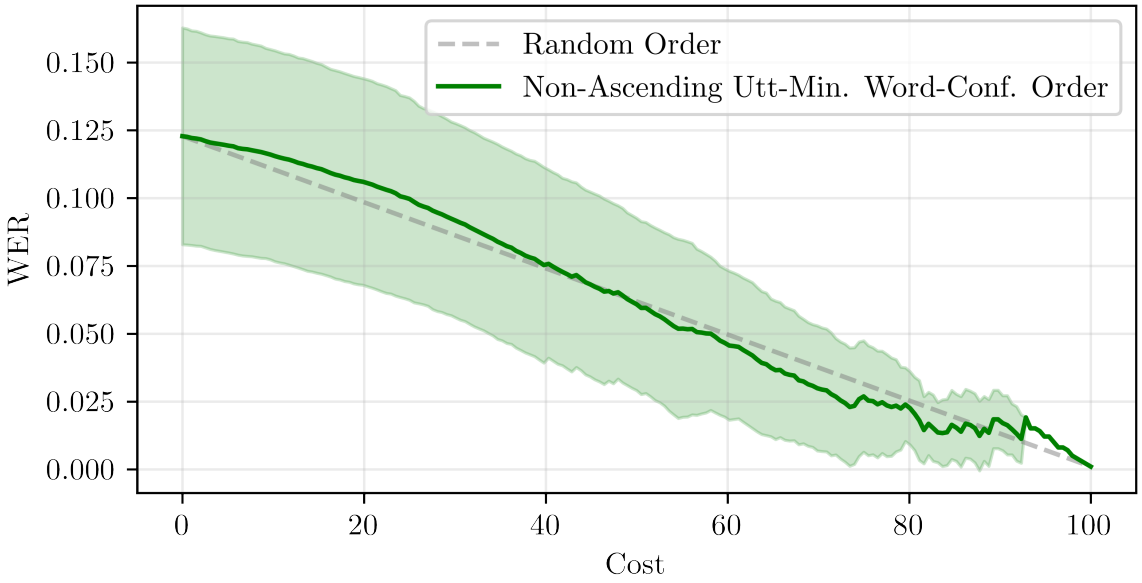


Figure 6.7: Ordered by non-descending utterance-maximum word-confidence

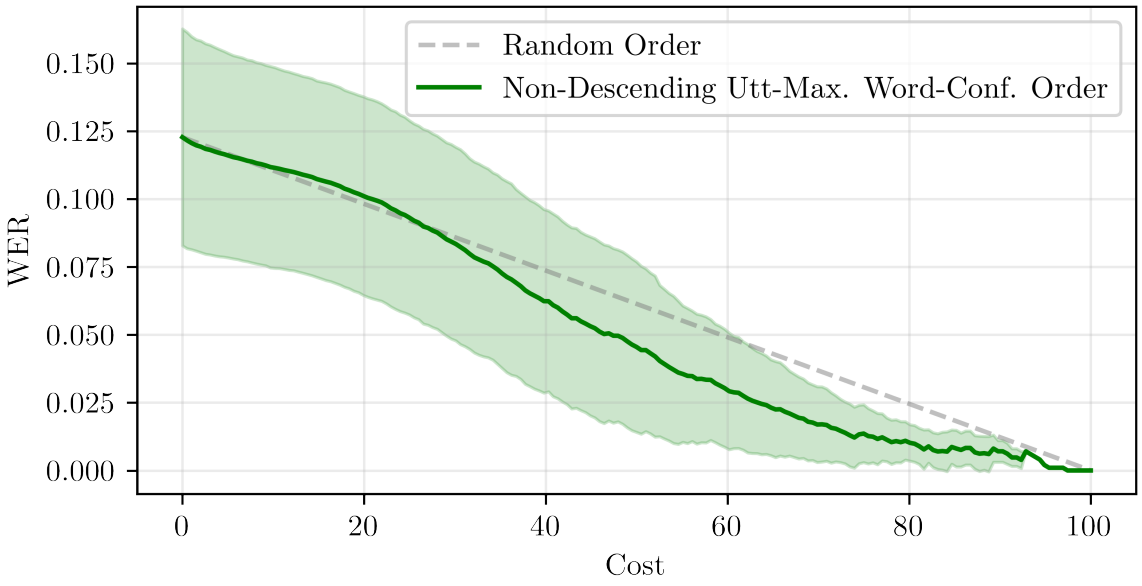
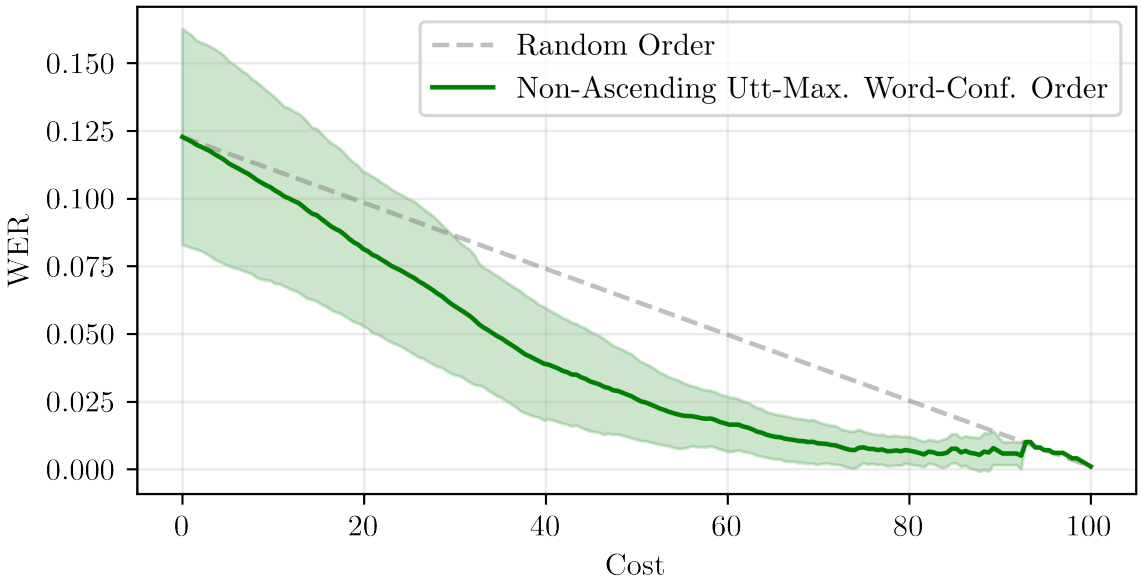


Figure 6.8: Ordered by non-ascending utterance-maximum word-confidence



6.2.3 Corpus-wide comparisons with word-level confidence ordering

Figure 6.9: Comparing whole-corpus evaluation performance with each word-confidence ordering

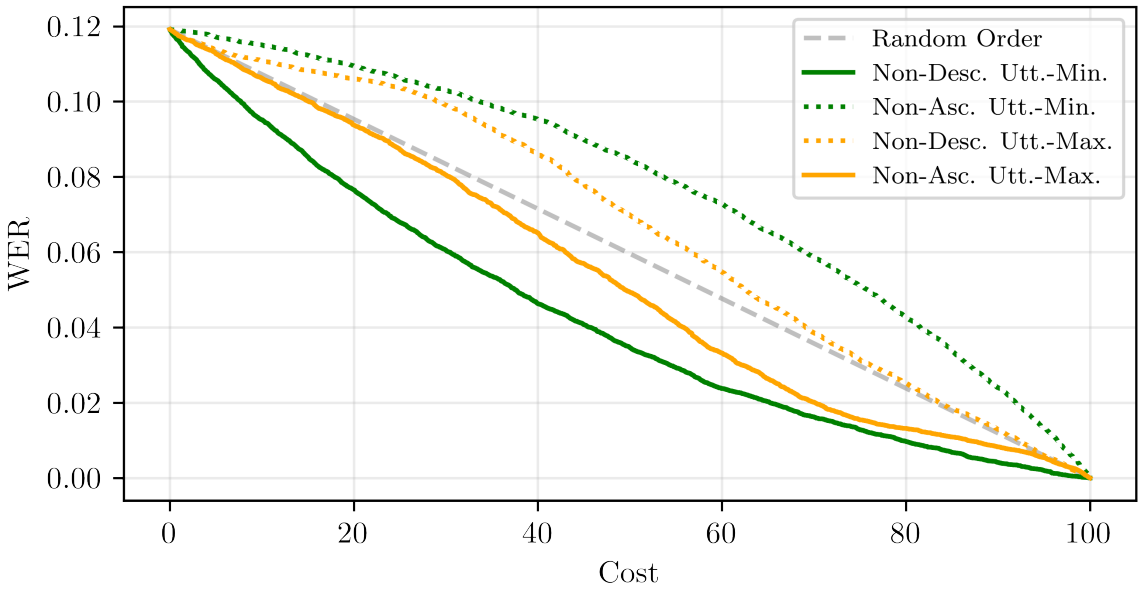
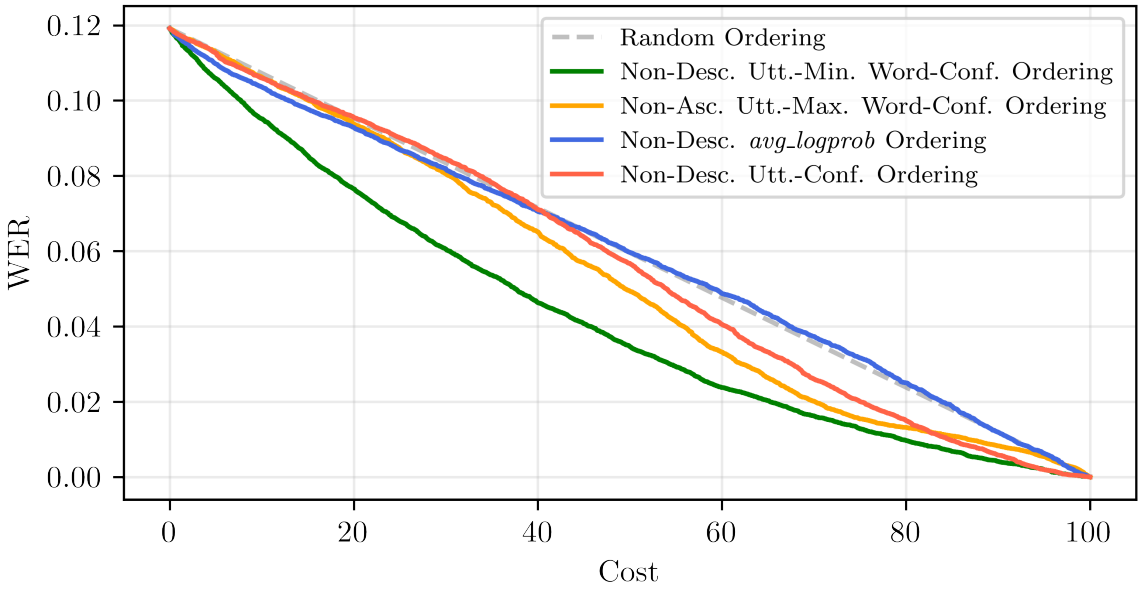


Figure 6.10: Comparing whole-corpus evaluation performance with each ordering metric



6.2.4 Different word-level confidence scoring techniques

Figure 6.11: Comparing whole-corpus evaluation performance with different word-level confidence metrics

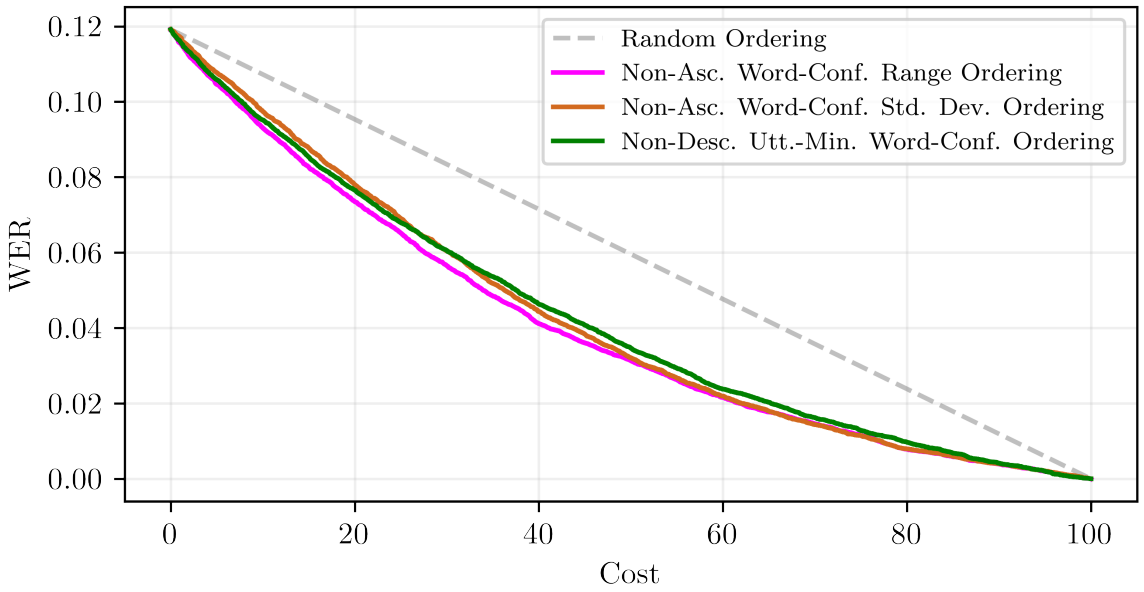
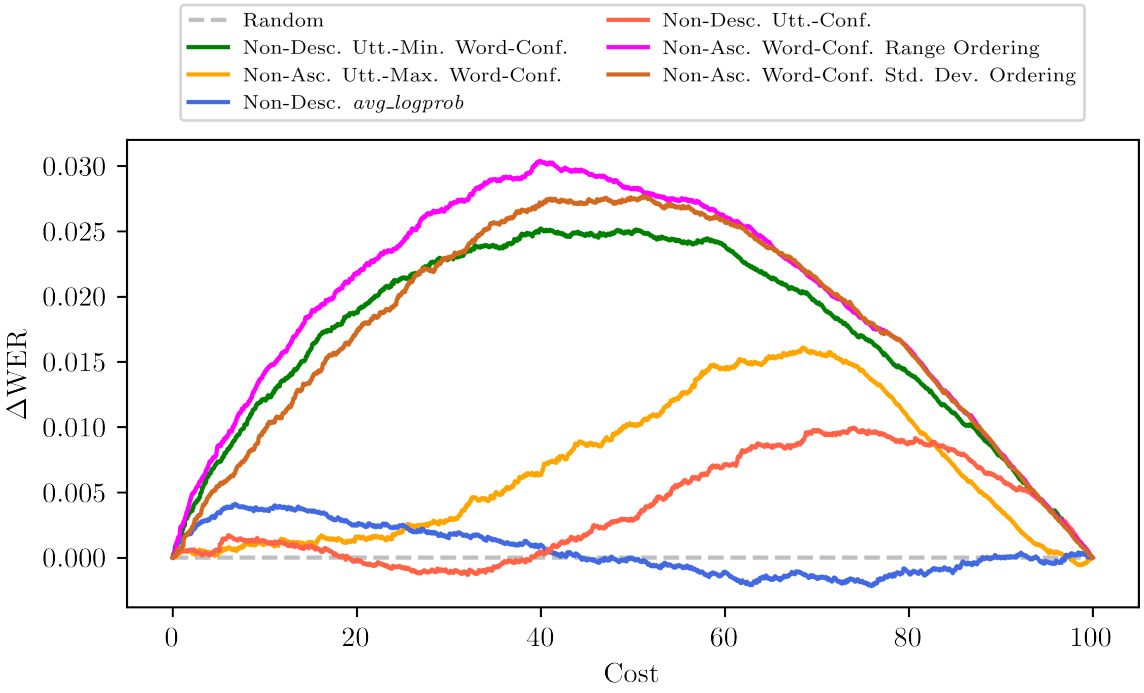


Figure 6.12: WER difference between metrics and random ordering



6.3 Discussion of Results

This section shall discuss findings which can be drawn from these results.

6.3.1 Utterance-average confidence and `avg_logprob`

The results show that `avg_logprob` and utterance-average confidence perform relatively similarly on average for a given conversation (fig. 6.3), with similarly poor performance when used to order the entire corpus (fig. 6.4). Utterance-average confidence appears to have marginally superior performance in both comparisons shown in section 6.2.2.

The similarity between the plots of the utterance-average word confidence and Whisper's `avg_logprob` output is likely due to their similarity in calculation. Confidence is calculated by *whisper-timestamped*[40] from the average of the word-level confidence scores in an utterance, which are calculated from the average of the sub-word log-probabilities for each word. Whisper[30] calculates `avg_logprob` as simply the average of the log-probabilities of all tokens in an utterance. Perhaps the slight performance increase seen when using utterance-average confidence scores could be explained due to it being calculated from word-level scores rather than token-level and the accuracy metric operating at the word-level rather than token-level too.

6.3.2 Word-level confidence metrics

The results presented in section 6.2.2 show the per-conversation average relationship between WER and Cost for the non-ascending and non-descending utterance-minimum and -maximum word-confidence metrics.

It is apparent that for the utterance-minimum word-confidence metric it is best to use non-descending (fig. 6.5) rather than non-ascending order (fig. 6.6), as non-ascending order closely follows the random order (i.e. is ineffective). This makes intuitive sense because ordering results on non-ascending minimum confidence would mean correcting the most confident results first, thus the opposite of an effective ordering method.

Notice that non-descending utterance-minimum word-confidence ordering (fig. 6.5) has the narrowest shaded area of all the plots presented, meaning it has the lowest standard deviation from the mean and therefore the most reliable representation of performance on a given conversation.

As for using utterance-maximum word-confidence ordering, non-ascending order (fig. 6.8) shows superior performance to non-descending (fig. 6.7). The reason for this is less intuitive; it would make more sense that ordering from lowest maximum confidence to highest (i.e. non-descending) would put utterances with lower confidence scores first. This could be due to word-confidence being derived from word-level probability scores, though this is an area which needs more experimentation to better understand this result.

When comparing whole-corpus evaluation results using word-confidence measures (fig. 6.9), non-descending utterance-minimum word-confidence ordering has the best (largest difference from random) and most consistent (smoothest line) performance of all the metrics, followed by non-ascending utterance-maximum word-confidence. The other two metrics are shown to

be a hinderance, achieving worse performance than if the results were evaluated in random order and should thus be ignored as metrics to use in a computer-aided transcription system.

6.3.3 Comparing word-level and utterance-level metrics

Figure 6.10 presents a comparison between the performance of each metric when evaluating the whole corpus. The metric with the best performance is clearly non-descending utterance-minimum word-confidence; it is much more consistent than the others and remains much further from random ordering, meaning it has the highest WER reduction for the same cost as all other metrics.

Utterance Ordering Metric	**Cost to reduce WER by 50%**
Random	50.0%
Non-descending utterance-minimum word-confidence	30.7%
Non-ascending word-confidence range	**27.9%**
Non-ascending word-confidence standard deviation	30.6%
Non-descending `avg_logprob`	50.1%
Non-descending utterance-confidence	47.7%
Non-descending utterance-maximum word-confidence	43.2%

Table 6.1: Cost required to halve WER.

Figure 6.11 shows the performance of non-ascending ordering using the both range scores and standard deviation of word-confidence for each utterance. Notice the slight performance gains over non-descending utterance-minimum word-confidence.

These metrics show slight performance gains over non-descending utterance-minimum word-confidence, as illustrated in figure 6.12 which plots the difference in WER from using random ordering (ΔWER) against Cost.

Table 6.3.3 presents the cost required to halve WER, with non-ascending word-confidence range being the best performer. According to these results, a computer-aided transcription system using this metric would require only 28% of the results to be checked in order to halve the WER of the ASR output, in this case falling to approximately 6% WER. Considering that the corpus has been segmented into 7312 utterances, fewer than 2050 of them would require manually correcting to reduce the WER by half when ordered using this metric. For reference, to achieve this result by correcting the results in random order would require checking over 3600 pieces of audio, or 44% more.

6.4 Future Work

This section shall highlight some of the areas which this project has failed to adequately address and present potential avenues for further work.

6.4.1 Increase test corpus

This work has focused on only one speech corpus, *LifeLUCID*[33]. Though it presents a diverse range of speaker ages, it has made for a relatively small test sample (less than 9 hours total). Thorough testing on a variety of speech corpora would provide validatation to the results presented earlier in this chapter. Other languages could be used to increase the diversity of the test corpus because Whisper is capable of operating on various different languages (though is trained on 65% English data[30]).

Despite the limited test corpus used in this work, the results of ordering utterances using metrics based on word-level confidence scores show a very clear benefit to the efficiency of a semi-automatic transcription system. This benefit may differ in magnitude across corpora, though it seems highly unlikely that it would not provide a benefit to other similar corpora (i.e. consisting of English speakers).

6.4.2 Acquire results from human transcribers

The rudimentary simulation presented in section 6.1 is built on the assumption that a result is free from errors once human-corrected. In reality, this is not always the case; humans often make mistakes, and therefore would produce different results from what was presented in this chapter.

An experiment using human participants would require a working computer-aided transcription software; the software demonstration presented for this project is not fully-featured and would require a small amount of modification before it could be used for transcription. Specifically, it is not capable of taking text input from the user, though it can order, highlight and blank-out utterances, as well as play accompanying audio files.

6.4.3 Experiment with other derivations of confidence

This work has used confidence scoring derived from a single unmodified output from Whisper (`avg_logprob`) and internal probability scoring. While ordering utterances based on minimum word-level confidence scores derived from internal probability scoring has shown a degree of effectiveness, there is expansive literature discussing methods of neural network confidence measures which have not been implemented in this work.

For example, use of a neural network (e.g. a multilayer perceptron) which takes Whisper's output as input and outputs a confidence metric could be shown to provide a reliable metric for confidence. Similar methods are proposed in the literature, such as using recurrent neural networks[41, 42] or a naïve Bayes classifier[43]. The benefit of this method would be that Whisper doesn't require any (or minimal) modification to apply them.

To avoid spending time experimenting with modifications to Whisper's architecture, a mildly modified version of Whisper called *whisper-timestamped*[40] was used in this project. An 'entropy-based' approach for calculating word-confidence is proposed in the literature[44, 45], though implementation would require considerable modification to the Transformer which Whisper uses.

A solution to this problem may be to use a different model than Whisper, though at the time of writing there doesn't appear to be another free and open-source ASR model with similar performance.

Chapter 7

Conclusions

Bibliography

[1] L. R. Rabiner, "Automatic Speech Recognition - A Brief History of the Technology Development," *Scinapse*, Jan. 2004. [Online]. Available: https://www.scinapse.io/papers/187290754

[2] D. W. H., "The vocoder," *Bell. Labs. Rec.*, vol. 18, p. 122, 1939. [Online]. Available: https://cir.nii.ac.jp/crid/1572261551231523968

[3] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952. [Online]. Available: https://doi.org/10.1121/1.1906946

[4] J. K. Baker, *Stochastic modeling as a means of automatic speech recognition.* Carnegie Mellon University, 1975.

[5] Y. Bengio *et al.*, "Markovian models for sequential data," *Neural computing surveys*, vol. 2, no. 199, pp. 129–162, 1999.

[6] X. Dong, W. Cao, H. Cheng, and T. Zhang, "Hidden markov model-driven speech recognition for power dispatch," in *Tenth International Conference on Applications and Techniques in Cyber Intelligence (ICATCI 2022)*, J. H. Abawajy, Z. Xu, M. Atiquzzaman, and X. Zhang, Eds. Cham: Springer International Publishing, 2023, pp. 760–768.

[7] E. B. Dynkin, *Theory of Markov Processes*, T. Köváry, Ed. Pergamon Press, 1960.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[9] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[10] M. K. Mustafa, T. Allen, and K. Appiah, "A comparative review of dynamic neural networks and hidden markov model methods for mobile on-device speech recognition," *Neural Computing and Applications*, vol. 31, pp. 891–899, 2019.

[11] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech

recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[12] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[13] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 4835–4839.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, and R. Collobert, "Fully convolutional speech recognition," *arXiv preprint arXiv:1812.06864*, 2018.

[16] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[17] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[18] A. Fang, S. Filice, N. Limsopatham, and O. Rokhlenko, "Using phoneme representations to build predictive models robust to asr errors," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 699–708. [Online]. Available: https://doi.org/10.1145/3397271.3401050

[19] S. Nießen, F. J. Och, G. Leusch, H. Ney *et al.*, "An evaluation tool for machine translation: Fast evaluation for mt research." in *LREC*, 2000.

[20] Y. Park, S. Patwardhan, K. Visweswariah, and S. C. Gates, "An empirical analysis of word error rate and keyword error rate." in *INTERSPEECH*, vol. 2008, 2008, pp. 2070–2073.

[21] H. B. Pasandi and H. B. Pasandi, "Evaluation of automated speech recognition systems for conversational speech: A linguistic perspective," *arXiv preprint arXiv:2211.02812*, 2022.

[22] P. Szymański, P. Żelasko, M. Morzy, A. Szymczak, M. Żyła-Hoppe, J. Banaszczak, L. Augustyniak, J. Mizgajski, and Y. Carmiel, "Wer we are and wer we think we are," *arXiv preprint arXiv:2010.03432*, 2020.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[24] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wu, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, oct 2022. [Online]. Available: https://doi.org/10.1109%2Fjstsp.2022.3182537

[25] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," 2021. [Online]. Available: https://arxiv.org/abs/2108.06209

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[27] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020.

[28] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation." in *Interspeech*. Citeseer, 2001, pp. 1359–1362.

[29] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv:2212.04356, 2022.

[31] A. Baevski, W.-N. Hsu, A. CONNEAU, and M. Auli, "Unsupervised speech recognition," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 27 826–27 839. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/ea159dc9788ffac311592613b7f71fbb-Paper.pdf

[32] W. S. Horton, D. H. Spieler, and E. Shriberg, "A corpus analysis of patterns of age-related change in conversational speech." *Psychol. Aging*, vol. 25, no. 3, p. 708, 2010.

[33] O. Tuomainen, L. Taschenberger, and V. Hazan, "LifeLUCID Corpus: Recordings of Speakers Aged 8 to 85 Years Engaged in Interactive Task in the Presence of Energetic and Informational Masking, 2017-2020," *UK Data Service*, May 2021. [Online]. Available: https://reshare.ukdataservice.ac.uk/854350

[34] "Praat: doing Phonetics by Computer," Mar. 2023, [Online; accessed 8. Apr. 2023]. [Online]. Available: https://www.fon.hum.uva.nl/praat

[35] N. Nurseitov, M. Paulson, R. Reynolds, and C. Izurieta, "Comparison of json and xml data interchange formats: a case study." *Caine*, vol. 9, pp. 157–162, 2009.

[36] Y.-A. Wang and Y.-N. Chen, "What do position embeddings learn? an empirical study of pre-trained language model positional encoding," *arXiv preprint arXiv:2010.04903*, 2020.

[37] B. Bechtold, "python-soundfile," 2013, [Online; accessed 8. Apr. 2023]. [Online]. Available: https://github.com/bastibe/python-soundfile

[38] C. R. Harris, K. J. Millman, S. f. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.

[39] "openai-whisper," Apr. 2023, [Online; accessed 10. Apr. 2023]. [Online]. Available: https://pypi.org/project/openai-whisper

[40] J. Louradour, "whisper-timestamped," https://github.com/linto-ai/whisper-timestamped, 2023.

[41] P.-S. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng, "Predicting speech recognition confidence using deep learning with word identity and score features," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7413–7417.

[42] K. Kalgaonkar, C. Liu, Y. Gong, and K. Yao, "Estimating confidence scores on asr results using recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4999–5003.

[43] A. Sanchis, A. Juan, and E. Vidal, "A word-based naïve bayes classifier for confidence estimation in speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 565–574, 2011.

[44] A. Laptev and B. Ginsburg, "Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 152–159.

[45] D. Qiu, Q. Li, Y. He, Y. Zhang, B. Li, L. Cao, R. Prabhavalkar, D. Bhatia, W. Li, K. Hu *et al.*, "Learning word-level confidence for subword end-to-end asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6393–6397.

Appendices

Appendix A

Example of the TextGrid Format

This example serves to illustrate the lack of readability of a TextGrid file. To aid formatting, it is displayed in two columns, though the original file is a long series of *intervals*.

It represents a piece of audio which is under three seconds in length (from a 10 minute-long recording) and consists of the text "a bush with a yellow duck on top".

```
intervals [12]:
  xmin = 20.899
  xmax = 20.971783458461772
  text = "SIL"
intervals [13]:
  xmin = 20.971783458461772
  xmax = 21.05
  text = "a"
intervals [14]:
  xmin = 21.05
  xmax = 21.47
  text = "BUSH"
intervals [15]:
  xmin = 21.47
  xmax = 21.66
  text = "with"
intervals [16]:
  xmin = 21.66
  xmax = 21.720024609817834
  text = "A"
intervals [17]:
  xmin = 21.720024609817834
  xmax = 22.1
  text = "SIL"
intervals [18]:
  xmin = 22.1
  xmax = 22.49
  text = "yellow"
intervals [19]:
  xmin = 22.49
  xmax = 22.84
  text = "duck"
intervals [20]:
  xmin = 22.84
  xmax = 23.06
  text = "ON"
intervals [21]:
  xmin = 23.06
  xmax = 23.769
  text = "top"
```

Appendix B

Another Appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.