university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Enhancing Speech Recognition of Welsh for Older Adults Using Data Augmentation Techniques

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Phat Do (Voice Technology, University of Groningen)
with the external supervisor being
Dr. Defne Abur (Faculty of Arts, University of Groningen)

Wansu Zhu (s5551390)

June 11, 2024

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported me throughout the course of this research and the writing of this thesis.

First and foremost, I am profoundly grateful to my supervisor, Dr. Phat Do, for his unwavering guidance, invaluable insights, and constant encouragement. I would also like to express my appreciation to my external supervisor, Dr. Defne Abur, for her kind support in providing possible speech data and for her efforts in facilitating my integration into her lab. Their expertise and dedication were instrumental in shaping this work, and I am truly fortunate to have had the opportunity to learn from them."

I am also indebted to my family for their endless support and belief in me. To my parents, thank you for your unconditional love, sacrifices, and for always being my greatest champions. Your faith in my abilities has been a constant source of motivation. To my elder sister, your advice and encouragement have been indispensable. Your support has helped me navigate through the most challenging times during this programme.

To my friends, your camaraderie and understanding have made this journey not only bearable but enjoyable. Your support, whether through offering a listening ear, providing feedback, or simply being there, has meant the world to me.

This thesis would not have been possible without the collective support of all these remarkable individuals. I am deeply thankful for their contributions, which have been vital to the completion of this work.

Abstract

Automatic Speech Recognition (ASR) is widely used in various applications, enhancing clarity in educational, daily, and cross-cultural interactions. While promising for older adults, ASR systems often struggle with their speech due to physiological and cognitive changes. This study addresses this challenge by fine-tuning ASR models with older adults' speech data and employing data augmentation techniques. Focusing on Welsh, a low-resource language, the research demonstrates that fine-tuning the XLSR model reduced word error rate (WER) from 62.19% to 57.64%. Further improvements were achieved using advanced techniques such as speed perturbation with a factor of 0.9, reducing WER to 54.30%. These results underscore the potential for enhancing ASR performance for older adults through tailored augmentation methods, contributing to more inclusive speech technology for low-resource languages.

Contents

1 Introduction

Automatic Speech Recognition (ASR) has become widely utilized in various real-life applications, offering benefits to diverse populations. These applications range from real-time course transcriptions to mobile voice assistants in smartphones or cars, and machine translation, all of which play significant roles in educational contexts, daily interactions, and cross-cultural communication. While these technological advancements are accessible to all, they hold particular promise for older adults, particularly those facing age-related deterioration in voice quality. As individuals age, physiological changes in the respiratory system, larynx, vocal tract, and cognitive functions can impact the acoustic properties of their speech, potentially leading to reduced intelligibility compared to younger speakers (Torre & Barlow, 2009). ASR serves as a solution by providing accurate transcriptions of speech, thereby enhancing clarity for listeners. However, previous research indicates that ASR performance for older adults' speech has often fallen short, with word error rates (WER) for individuals over 60 years old typically surpassing those of younger adults (Loizou & Pantzaris, 2023). The reasons why ASR for older adults has worse performance are multifaceted. Firstly, the acoustic properties of speech change with age due to physiological factors such as reduced lung capacity, changes in vocal fold elasticity, and slower speech rates, which can affect the clarity and consistency of speech signals. Secondly, older adults often exhibit greater variability in speech patterns, including pronunciation, prosody, and speaking style, which ASR systems, trained predominantly on younger voices, struggle to accommodate. Additionally, cognitive factors, such as memory and processing speed, may lead to more frequent hesitations, repetitions, and fillers in older adults' speech, further complicating accurate transcription. Finally, the limited representation of older adults in ASR training datasets means that these systems are less likely to have encountered and learned to accurately transcribe the nuances of older adult speech patterns. Hence, there is a pressing need to focus on enhancing ASR capabilities specifically tailored to the needs of older adults.

On the other hand, concurrently, low-resource languages face challenges in contemporary speech technology development, risking marginalization. While high-resource languages like English, French, Spanish, Mandarin, and Japanese dominate the commercial market, speakers of low-resource languages risk economic and social exclusion (Cooper et al., 2019). Ensuring the availability of speech technology for small languages is crucial. Welsh, a minority language primarily spoken in Wales, United Kingdom, with speakers often bilingual in English, serves as an example. Despite over half a million Welsh speakers, com prising about 19% of the Welsh population, the dominance of English in the UK's technological landscape has marginalized Welsh in speech technology development (Cooper et al., 2019). Previous literature has made strides in ASR development for Welsh, including the collection of Welsh speech data, the creation of real-time Welsh ASR models, and the development of Welsh natural language toolkits (Cooper et al., 2019; Cunliffe et al., 2022; Jones, 2022; Knight et al., 2021; Vangberg et al., 2023).

However, no research has specifically addressed ASR for older Welsh speakers. Considering data augmentation techniques' effectiveness in improving ASR performance for low-resource languages, such as minority languages, dysarthric languages, and children's speech, using pretrained multilingual models (Bartelds et al., 2023; Harvill et al., 2021; Patel & Scharenborg, 2024; Wang et al., 2020), this study aims to explore the feasibility and effectiveness of data augmentation techniques in improving ASR model performance for Welsh older adults' speech.

The structure of the thesis is the following: subsection 1.1 introduces the research question. Section 2 provides an extensive literature review that frames the research question and hypothesis in the state-of-the-art, and further raises hypothesis on the outcome of the research. In section 3, the methodology is covered and the underlying models used are explained. Then, section 4 describes the results obtained and compares them to the baseline. In section 5, I discuss the previously-mentioned results in detail. Lastly, section 6 summarizes the thesis and presents the conclusions drawn, along with recommended future work.

1.1 Research Question

In light of the preceding discussion, the research question at the core of this study can be formulated as follows:

> **How effectively can the pre-trained multilingual XLS R model, when fine-tuned and supplemented with data augmentation techniques, improve speech recognition accuracy for older Welsh speakers**

From which the following subquestions are derived:

- Can the fine-tuned model achieve lower word error rate (WER) than the baseline model?
- Can data augmentation techniques improve the fine-tuned model performance?
- How does different data augmentation techniques affect the modeol performance?

2 Literature Review

2.1 Development of ASR

The research on speech recognition dates back to the 1950s, when systems could only recognize isolated digits. In the 1960s, IBM developed the ”Shoebox” system, which expanded the recognition scope to 16 words and 10 digits (Juang & Rabiner, 2005).

Dynamic Time Warping (DTW) is a template matching algorithm based on dynamic programming, introduced into speech recognition in the 1970s, becoming a classic algorithm of the time. Additionally, technologies like Linear Predictive Coding (LPC) were applied to speech recognition. In 1971, the Defense Advanced Research Projects Agency (DARPA) of the U.S. Department of Defense launched a speech understanding project. This led to the development of the ”Harpy” system by Carnegie Mellon University, capable of recognizing 1,000 words, equivalent to a three-year-old child’s vocabulary (Juang & Rabiner, 2005). Thus, researchers began exploring and solving the problem of large vocabulary continuous speech recognition.

In the mid-1980s, statistical modeling methods gradually replaced the original template matching methods, particularly Hidden Markov Model (HMM)-based statistical methods for acoustic modelling (Juang & Rabiner, 2005). Scholars from Carnegie Mellon University proposed using HMM to model the acoustic dynamics of speech, using Gaussian Mixture Models (GMM) to model the likelihood of each HMM state. This method established a GMM-HMM-based speech recognition framework, marking a significant breakthrough in the field and laying the foundation for future technological developments. Until the end of the last century, research in speech recognition focused on refining and expanding this framework, proposing a series of methods. Despite the gradual improvement and maturity of speech recognition technology, it still could not meet daily usage requirements. Speech recognition entered a period of relatively stable development until the rise of deep neural networks, which brought new performance leaps.

In 2009, scientists first proposed using Deep Neural Networks (DNN) to replace GMM in traditional acoustic models. Subsequently, Microsoft researchers proposed using context-dependent triphone states as the modeling targets for neural networks, i.e., the DNN-HMM model, which achieved a 20% improvement compared to the GMM-HMM model (Juang & Rabiner, 2005). The open-source tool Kaldi, developed by Daniel Povey and others, provided a series of tools and algorithms for the research and application of DNN-HMM technology. With the rapid development of deep learning, more complex neural networks, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and their variants, also showed different advantages in speech recognition systems. Through the efforts and optimizations of numerous researchers, speech recognition systems have reached a usable level in some practical applications (X. Huang et al., 2014). However, these systems were still based on the HMM framework, requiring separate training for acoustic and language models and the construction of dictionaries based on linguistic expertise.

In recent years, end-to-end speech recognition systems have become a research hotspot, receiving widespread attention and discussion (X. Huang et al., 2014). End-to-end systems can directly model the mapping relationship between input speech sequences and output text sequences. These generally include models like Connectionist Temporal Classification (CTC), Attention-based Encoder-Decoder (AED) models, and Recurrent Neural Network Transducer (RNN-T) models. The CTC model, first proposed by Graves and others, addresses the alignment problem between input sequences and output labels through dynamic programming. However, a limitation of the CTC model is its reliance on conditional independence assumptions, leading Graves and others to propose the RNN-T model to enhance CTC with language modeling capabilities. Chan and others proposed the attention-based encoder-decoder model, which consists mainly of an encoder, an attention module, and a decoder. These models share the common trait of moving beyond the HMM framework, using a complete neural network as the speech recognition model, integrating the functions of acoustic and language models into one unit. Hence, building and training end-to-end models are more accessible and efficient. However, end-to-end models are highly data-dependent, usually requiring a large amount of labeled data for training to achieve satisfactory results, which is a core bottleneck for their deployment.

Transformers, a type of neural network model utilizing attention mechanisms, comprise an encoder and a decoder (Han et al., 2021). They can capture the context information of entire sentences when processing each sequence unit, initially excelling in machine translation and showing strong modeling capabilities across various tasks. Currently, models based on the Transformer paradigm have achieved the best results in various standard speech recognition tests. Google's Conformer model (Gulati et al., 2020) combines the local modeling capabilities of CNNs with the long-term dependency modeling capabilities of attention mechanisms, achieving the best results in the Librispeech audiobook English benchmark test and significantly reducing the error rate of the test set. Additionally, Microsoft's conversational speech recognition system built in 2021 uses Conformer as the encoder, Long Short-Term Memory (LSTM) as the decoder, and integrates language models, achieving human-level performance in the Switchboard conversational English benchmark test (Y. Liu et al., 2021).

In summary, the performance of speech recognition is gradually improving, and the recognition process is becoming more simplified. However, as models become larger, more data is needed for training to achieve better results

2.2 ASR framework

The task of automatic speech recognition (ASR) is to output a corresponding text sequence for a given speech signal. Generally, this is done by using statistical methods to find the mapping relationship between speech and text, abstracting it into the following mathematical problem:

$$W^* = \underset{W}{\arg\max} P(W \mid O)$$

Where $O = \{o_1, o_2, \ldots, o_T\}$is the observation sequence of the input speech, and $w = \{w_1, w_2, \ldots, w_T\}$ represents all possible word sequences corresponding to it, with T representing the length of the speech sequence and N representing the length of the word sequence. The goal of speech recognition is to maximize the probability \(P(W—O) \), finding the most matching word sequence. This probability is typically not directly calculable, but using Bayes' theorem, it can be expanded as: Where

$O=\{o_1,o_2,\ldots,o_T\}$is the observation sequence of the input speech, and $w=\{w_1,w_2,\ldots,w_T\}$ represents all possible word sequences corresponding to it, with T representing the length of the speech sequence and N representing the length of the word sequence. The goal of speech recognition is to maximize the probability of P$(W|O)$, finding the most matching word sequence $W*$. This probability is typically not directly calculable, but using Bayes' theorem, it can be expanded as:

$$W^* = \arg\max_W \frac{P(O \mid W)P(W)}{P(O)} = \arg\max_W P(O \mid W)P(W)$$

Where P(O) is the prior knowledge of the observation sequence and is unrelated to W, so the objective is to maximize the product of P(W) andP$(O|W)$. P$(W)$is the probability of the word sequence itself appearing in natural language, also known as the prior probability. P$(O|W)$is the likelihood probability of the given word sequence Wcorresponding to the observation sequence O. In traditional speech recognition, P$(O|W)$ is determined by the acoustic model, and P(W) is obtained through statistical language modeling.

The traditional speech recognition process is illustrated in Figure 1. As shown, the traditional speech recognition process generally includes four modules: feature extraction, acoustic model, language model, and decoding. These modules are designed and optimized independently, whereas end-to-end speech recognition integrates the pronunciation dictionary, acoustic model, and language model together, directly modeling the mapping relationship from speech signals to text sequences, thus eliminating the need for a manually constructed pronunciation dictionary.

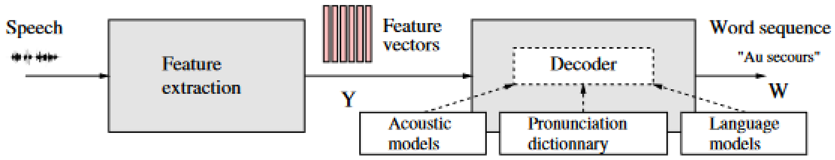

Figure 1: Components of Classical ASR

The process of traditional speech recognition is depicted in Figure 1. It can be seen that the traditional speech recognition process typically includes four modules: feature extraction, acoustic model, language model, and decoding. These modules are designed and optimized independently, whereas end-to-end speech recognition integrates the pronunciation dictionary, acoustic model, and language model, directly modeling the mapping relationship from speech signals to text sequences without the need for manually constructing a pronunciation dictionary.

2.2.1 Preprocessing and Feature Extraction

Feature extraction is the process of converting time-domain speech waveforms into vectors that represent the characteristics of speech for model processing. Since human ears distinguish sounds through frequency rather than waveform, Fourier Transform can be used to convert speech waveforms into spectra. Based on human auditory perception, a set of filters can be designed to extract features related to the spectrum. Two commonly used features are Mel Filter Bank (Fbank) features and Mel-Frequency Cepstral Coefficients (MFCCs). This section first introduces the Fbank feature extraction process.

1. Preprocessing: Before feature extraction, preprocessing is usually required, mainly including three steps:

(a) Pre-emphasis: Passing the speech signal through a first-order high-pass filter to enhance the energy of the high-frequency part, retain high-frequency speech information, and filter out some noise unrelated to speech.

(b) Framing: Although the overall speech signal is non-stationary, it can be considered quasi-stationary within 10 ms to 30 ms. Based on this property, the speech is divided into frames for processing. Generally, the frame length is set to 25 ms, and the frame shift is set to 10 ms, ensuring continuity between frames through overlap.

(c) Windowing: To avoid spectral leakage when transforming to the frequency domain using Fourier Transform, each frame is multiplied by a window function, typically a Hamming or Hanning window.

2. Perform Short-Time Fourier Transform on each frame of the preprocessed signal to obtain the spectrum X(k):

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nkfN}, \quad 0 \leq k \leq N-1$$

where N is the number of points in the Fourier Transform, usually equal to the window length. Then, calculate the power spectrum $|\mathrm{X}(k)|^2$.

3. Pass the power spectrum through a Mel filter bank, calculate the energy within each filter band, and take the logarithm. The energy of each filter bank is the Fbank feature coefficient. Typically, 40 or 80 filter banks are set. Fbank retains the original features, suitable for neural networks with strong modeling capabilities; MFCCs, derived from Fbank using discrete cosine transform, remove the correlation between feature dimensions, making them more suitable for GMM modeling. Usually, features need further normalization, such as cepstral mean and variance normalization.

2.2.2 Acoustic Model

The acoustic model is the core component of speech recognition. In traditional speech recognition, the DNN-HMM model is the mainstream framework for acoustic modeling. Here is a brief introduction:

The acoustic model is used to model the posterior probability of a word sequence W given an observation sequence O. However, the lengths of speech feature sequences and word sequences are inconsistent. Traditional methods use HMM to model the dynamic changes of speech signals. HMM introduces an intermediate state sequence$S = \{s_1, s_2, \ldots, s_N\}$ so that the number of states and the number of observations are independent of each other. Thus, P$(O|W)$ can be represented as:

$$P(O \mid W) = \sum_{s \in S} P(O, s \mid W) P(s \mid W)$$

This includes two stochastic processes: one for state transitions and another describing the randomness between states and observations, corresponding to the randomness in human speech content and pronunciation variability.

Since states are unobservable and can only be inferred through observation sequences, they are termed hidden states. Additionally, HMM includes two basic assumptions:

- Homogeneous Markov Assumption: The current state only depends on the previous state.

- Observation Independence Assumption: Any observation at a given time only depends on the state at that time.

In speech recognition, a set of HMMs represents the acoustic model, where each speech modeling unit (e.g., phoneme) is bound to one hidden state of HMM. Each state's transition can only proceed left-to-right. To train the acoustic model, three basic problems of HMM must be addressed:

1. Probability Estimation Problem: Given a model and observation sequence, calculate the probability of the observation sequence using the forward-backward algorithm.

2. Optimal State Sequence Problem: Given a model and observation sequence, find the most probable corresponding hidden state sequence using the Viterbi algorithm.

3. Model Parameter Training Problem: Given the observation sequence, estimate the model parameters to maximize the conditional probability of the observation sequence using the Expectation-Maximization (EM) algorithm.

DNN is a perceptron network with multiple hidden layers, where each layer consists of several neurons connected to the next layer through weights and biases, undergoing nonlinear transformations. In speech recognition, DNN uses a fixed-length speech feature sequence as input, with each output unit corresponding to one hidden state of HMM. The output vector's dimension equals the number of HMM hidden states, and the output value represents the posterior probability of observing the state P$(O|S)$. Using Bayes' formula, the posterior probability can be converted to the observation probabilityP$(O|S)$. DNN requires frame-level annotation for supervised training. Thus, in practical applications, alignment information must first be obtained using a well-performing GMM-HMM model to establish the correspondence between hidden states and feature sequences.

GMM model, composed of a linear combination of multiple sub-Gaussian probability density functions, directly fits the observation probability. Within the HMM framework, GMM and DNN play similar roles, with the difference being that GMM assumes independence between input frames and uses unsupervised learning for training. DNN's main advantage is its powerful modeling capability, enabling learning deep nonlinear feature transformations and leveraging the structure information between adjacent speech frames. After training a good GMM-HMM model, DNN can replace GMM while retaining parts like transition probabilities and initial probabilities of HMM. Beyond GMM-HMM, the DNN-HMM model achieves significant performance improvements, enabling efficient training even with randomly initialized parameters and offering better generalization, thus becoming the mainstream framework in traditional speech recognition.

2.2.3 Language Model

The language model is an important component of the speech recognition framework, used to describe the probability of a word sequence and predict the next word given the context. This provides semantic modeling capability to the speech recognition model, improving recognition results and eliminating possible ambiguities. The classic language model is the n-gram language model, which states that the occurrence of each word in a sentence depends only on the previous words.

A drawback of this method is that when the training text is relatively small, the frequency estimates for some word combinations may be zero. To address this issue, data smoothing techniques are usually adopted, such as Laplace smoothing and Kneser-Ney smoothing. Therefore, the n-gram model has a notable disadvantage: the usable contextual information is limited, and it has a weak ability to model long sentences. With the development of neural networks, RNNs are also widely used for language modeling. Compared to traditional n-gram models, RNN models have better long-term dependency modeling capabilities and are often used to re-score recognition results from an initial n-gram model decoding, providing more accurate evaluation and ranking.

2.2.4 Decoding

Decoding refers to the process of recognizing speech using a trained speech recognition model. Traditional speech recognition is based on Weighted Finite-State Transducers (WFST) representing HMMs, pronunciation dictionaries, and language models. These networks are composed to build a complete decoding graph. The word transition probabilities from the language model are integrated with the state transition probabilities of the HMM and the transition probabilities of the phonetic units to form the inherent parts of the decoding graph, while the log observation probabilities of the acoustic features calculated by the acoustic model form the acoustic part. The decoder's function is to input the sequence of speech features into the acoustic model to obtain the acoustic scores, search through the decoding graph, and get the most likely word sequence path. The search algorithm commonly uses the Viterbi algorithm combined with beam search to optimize search efficiency.

2.3 End-to-End Speech Recognition Models

From the previous section, we know that traditional speech recognition uses HMM to a certain extent to solve the modeling problem, but the training process is relatively complex, requiring separate optimization of the acoustic model and language model and the construction of dictionaries and decoding graphs. Considering that both acoustic and language models can be modeled using neural networks, some researchers proposed the concept of end-to-end modeling, using a complete neural network for speech recognition. The end-to-end model views speech recognition as a sequence-to-sequence problem, directly calculating the probability from the input speech sequence to the text sequence as the training objective, greatly simplifying the training process. End-to-end models can consist of a complete neural network without relying on pronunciation dictionaries, allowing flexible choice of modeling units and better scalability and adaptability. Additionally, end-to-end models can omit the construction of language models, making the entire speech recognition process simpler and more efficient. In recent years, end-to-end models have developed significantly, surpassing traditional speech recognition performance in many scenarios through large-scale data training and are widely used in industrial production.

Currently, mainstream end-to-end models are mainly divided into three categories: the CTC model, the Attention-based Encoder-Decoder (AED) model, and the RNN-T model. Among them, CTC and AED models are the most commonly used in the field of speech recognition and are adopted in this paper.

2.3.1 CTC Model

CTC is the earliest model proposed with end-to-end modeling capabilities. It can automatically learn the alignment between the input speech feature sequence and the corresponding text sequence. Since the length of the real labels is much shorter than the number of input speech frames, the CTC mechanism allows repeated occurrences of output labels and introduces a blank label to represent gaps, ensuring that the output length matches the input speech frame sequence. After de-duplication and blank removal of the output label sequence, the final text sequence is obtained.

CTC uses a forward-backward algorithm similar to the HMM model, calculating the loss function through dynamic programming.

From the above, it is clear that CTC is essentially a training objective that can be directly applied to the output layer of the acoustic model. However, the CTC model does not directly model the correlations between labels, so it does not have strong language modeling capabilities. Typically, CTC models need to be jointly decoded with a language model to achieve better performance.

2.3.2 Attention-Based Encoder-Decoder Model

The AED model mainly includes: an Encoder module, an Attention layer module, and a Decoder module. Among these, the encoder is the most important part, similar to the role of the acoustic model. It encodes each frame of the input speech feature sequence into an intermediate representation vector of a fixed dimension. The attention layer performs a weighted average on the intermediate representation vector, indicating the emphasis on different moments of the input sequence. The decoder adopts an autoregressive approach to predict the target text step by step, similar to an autoregressive language model. At each decoding step, it uses the previously generated prediction text as an additional input and interacts with the intermediate representation vector output by the encoder through the attention layer, aligning the input sequence and the output sequence.

a) Attention Mechanism

The concept of the attention mechanism comes from the selective attention phenomenon in humans when processing information. When receiving audio or visual signals, more attention is paid to the key parts to obtain more details and reduce or ignore less important information. Speech signals are long, redundant, and complex. Similarly, if the key information needed for recognition can be identified, it will help improve the model's information processing capability and interpretability.

The core part of the attention layer is the attention calculation, which mainly includes two methods: dot-product and additive. This paper primarily uses scaled dot-product attention, which can be divided into self-attention and cross-attention based on different calculation areas.

b) Transformer

The Transformer model was initially proposed by Google for machine translation tasks. It relies entirely on attention mechanisms and feedforward neural networks, integrating many advantages of the attention mechanism. Today, it is widely used in the field of deep learning. The Transformer model used in this paper is the same as the original model in core parts, consisting of an encoder and a decoder. However, the absolute position encoding module is removed, and convolutional neural networks are introduced. The following describes the core parts of the Transformer:

The Transformer uses multi-head scaled dot-product attention, integrating multiple attention modules, each called a head. Multi-head attention provides multiple subspace representations, enhancing the model's ability to focus on different positions.

1. Encoder and Decoder

The encoder of the Transformer is formed by stacking multiple identical submodules. Each submodule contains two sub-layers: a Multi-head Self-attention layer and a fully connected Feed-forward Neural Network (FFN) layer. The feed-forward neural network is composed of two linear transformations and uses the ReLU activation function.

The structure of the decoder is similar to that of the encoder, as both are composed of multiple identical submodules stacked together. The key difference is that each submodule in the decoder includes an additional Cross-attention layer. This layer facilitates information interaction between the current layer input of the decoder and the output representation of the encoder. The calculation process of the cross-attention layer is similar to that of the self-attention layer. In cross-attention, the key and value vectors are replaced by the encoder's output, while the query vectors come from the output of the lower layer's multi-head self-attention layer. Moreover, during inference, the decoder operates in an autoregressive manner, meaning that the input at each time step depends on the output from the previous time step.

2. Residual Connections and Normalization

Each sublayer in both the encoder and decoder implements residual connections and layer normalization (LayerNorm). Residual connections aim to solve the training difficulties of deep neural networks, ensuring that information from the previous layer is accurately passed to the next layer. Layer normalization normalizes the input vectors along the channel dimension.

3. Positional Encoding

For text or speech sequences, the position information of the input is meaningful. However, the attention mechanism operates in parallel without considering the input position information. Therefore, positional encoding (PE) is added to enable the model to better handle contextual information. The original Transformer uses absolute positional encoding, which adds a fixed vector to the input vectors. The positional encoding vector can be generated by a set of sine and cosine functions at different frequencies, creating a unique encoding vector for any position in the sequence.

4. Complete Transformer Network Structure

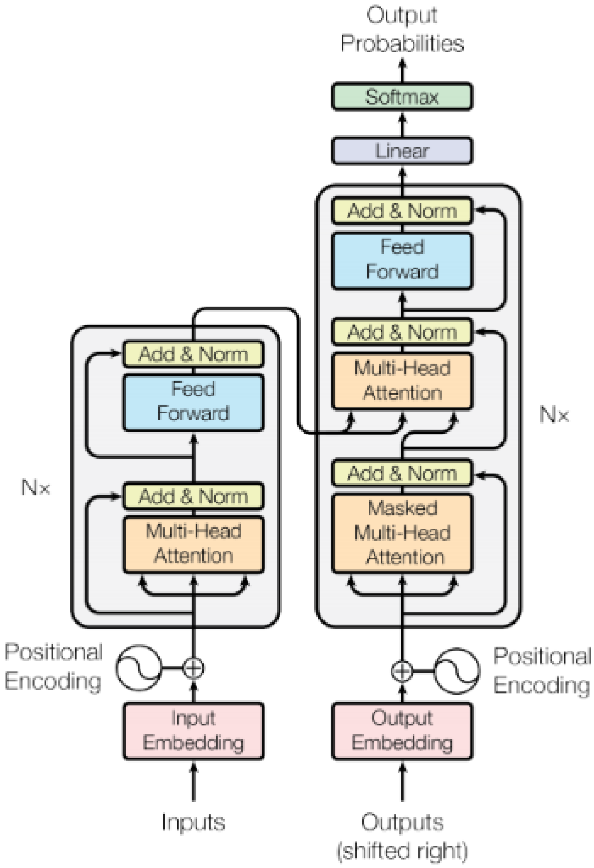

Figure 2: The structure of Transformer network

The complete Transformer network structure is shown in Figure 2. The Transformer can access all positions in the sequence through the attention mechanism, making it more adept at modeling long-term dependencies in sequences. The multi-head mechanism enhances the model's expressive ability and allows for parallel computation. The Transformer encoder has become the mainstream structure for end-to-end speech recognition modeling.

c) Combined CTC-Attention Multi-task Learning

The encoder-decoder model based on the attention mechanism has certain limitations. Unlike HMM and CTC, the attention mechanism does not have a monotonic constraint, making alignment between speech and text susceptible to noise in the speech and more challenging for long sequences. Therefore, researchers proposed a method combining CTC and attention for multi-task learning. The loss function of the encoder-decoder model typically uses cross-entropy loss, calculated at the decoder output. This method adds a linear layer at the encoder output to simultaneously compute the CTC loss. The multi-task learning training strategy effectively improves the convergence of the AED model and alleviates alignment learning issues.

d) Convolutional Neural Networks (CNNs)

1. CNN Structure

CNNs mainly consist of convolutional layers, pooling layers, and fully connected layers.

Convolutional layers extract feature information from the input feature map using convolutional kernels. The kernels slide over the input feature map with a certain stride, performing dot products with the elements in the local region covered by the kernel while keeping the kernel weights constant. Therefore, convolutional layers have two characteristics: local connections and weight sharing. Each convolutional layer usually contains multiple convolutional kernels, each corresponding to a set of linear projections to capture features at different scales, producing outputs with multiple channels. Zero-padding is an important operation in convolutional layers, where zeros are padded around the edges of the input feature map to expand its size. As pure convolution operations reduce the size of the output feature map, zero-padding helps maintain its size, allowing the convolution operation to better utilize edge information.

Pooling layers are usually used after convolutional layers to downsample the dimensions, reducing redundant information and computational complexity. Common pooling operations include max pooling and average pooling.

Typically, a fully connected layer follows at the end of the convolutional network, utilizing information from all dimensions of the output feature map. Non-linear activation functions such as ReLU and Sigmoid are introduced in the fully connected layer.

2. CNN in Speech Recognition

Convolutional Neural Networks excel at capturing local features, making them suitable for combining with attention mechanisms to enhance the model's ability to learn features at different scales. CNNs were first proven to perform excellently in processing 2D signals. When dealing with 1D speech signals, the spectrogram features are typically used as input. CNNs capture correlations in the time and frequency domains, forming acoustic models in combination with DNN or RNN networks.

Additionally, 1D convolutional networks can directly model the waveform in the time domain, with the convolutional kernel sliding only along the time dimension according to the stride. This convolution operation acts like a filter, with different convolutional kernels corresponding to different filters. Another important function is to downsample the speech sequence, reducing computational complexity and enhancing information density.

Through multiple convolutional layers, a series of feature vectors can be obtained. These feature vectors are similar to traditional spectral features and can be used to build end-to-end models with performance comparable to speech recognition models built using spectral features as input. Since convolutional operations express relative position information and have translation invariance, convolution can also serve as a means of positional encoding. To obtain sufficient contextual information, larger convolutional kernels are usually used.

2.3.3 Wav2vec2

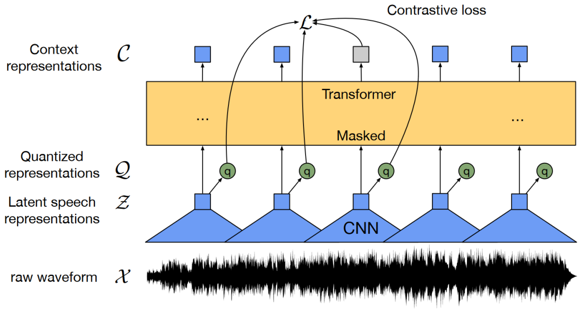

Figure 3: The structure of Wav2vec 2.0

After reviewing the frameworks of automatic speech recognition (ASR), this thesis transitions to discussing one modern end-to-end approaches in ASR.

Wav2Vec 2.0, developed by Facebook AI Research (FAIR), is a cutting-edge end-to-end ASR model that has demonstrated exceptional performance in transcribing speech to text (Baevski et al., 2020). Unlike traditional ASR systems, which rely on manual feature engineering and separate modeling stages, Wav2Vec 2.0 takes raw audio waveforms as input and learns to extract meaningful representations directly from the audio signal.

At the core of Wav2Vec 2.0 is its novel architecture, which combines self-supervised learning with transformer-based models. Through pre-training on large amounts of unlabeled speech data, the model learns to capture contextual information and long-range dependencies within the audio signal. This pre-training phase enables Wav2Vec 2.0 to generate robust representations of speech features without the need for manual annotations.

Following pre-training, Wav2Vec 2.0 can be fine-tuned on specific downstream tasks, such as speech recognition, using labeled data. Fine-tuning allows the model to adapt its representations to the target task, leading to improved performance and accuracy in speech transcription.

In the context of this thesis, the baseline model used for research is fine-tuned on Wav2Vec 2.0.

By leveraging the capabilities of Wav2Vec 2.0, the thesis aims to explore and advance the state-of-the-art in speech recognition, particularly in the domain of low-resource languages.

2.4 ASR for low-resource languages

The main challenge in ASR for low-resource languages is the lack of annotated data resources. If sufficient data were available, models could be built using conventional speech recognition methods. How to improve speech recognition technology under limited resource conditions to achieve performance comparable to that obtained with larger datasets has attracted the attention of many institutions and scholars. Two representative techniques to address the lack of data are data augmentation and model knowledge transfer.

2.4.1 Data Augmentation

Data augmentation refers to the expansion of the diversity and scale of training data through certain transformations to improve the generalization performance of models. A classic method is to apply a series of transformations to the original audio signals, including speed perturbation (Ko et al., 2015), volume perturbation, adding reverberation, and adding noise (Ko et al., 2017). These methods are easy to implement and simple yet effective. By mixing the transformed data with the original data for training, the robustness of the model can be improved. Some researchers have proposed methods to transform audio signals on the spectrum, such as channel length normalization (Jaitly & Hinton, 2013), which simulates changes in channel length by stretching or compressing components within different frequency ranges. Google proposed SpecAugment (Park et al., 2019) for speech recognition tasks, which randomly distorts and masks features in the time and frequency dimensions to effectively mitigate overfitting problems. This method has been widely used in recent years. Additionally, the mixup method (H. Zhang et al., 2018) linearly weights the audio and text of two speech samples to create a new sample. However, the new sample created by weighting discrete text sequences may have semantic ambiguities or grammatical errors. MixSpeech extends the mixup method to the frequency domain by mixing features of two different speech samples, trying to recognize the corresponding two text sequences separately and weighting them at the loss function level.

Generative models are also used for data augmentation. For example, voice conversion models generate new speech data by altering the timbre, speech rate, and speaker information of the original audio as a data augmentation strategy. In one study (Zalouk et al., n.d.), the authors used CycleGAN to convert American English spoken by native speakers into African American Vernacular English. Similarly, another study (Singh et al., n.d.) used CycleGAN to convert adult speech into child speech, combining adversarial loss and cycle consistency loss for training. In yet another study (Baas & Kamper, 2022), researchers attempted to improve data scarcity in low-resource languages through cross-language voice conversion, using Tacotron combined with HiFi-GAN vocoder as the voice conversion model.

Some researchers adopt text-to-speech (TTS) technology to augment training datasets, a semi-supervised data augmentation method. Ideally, a TTS model can simulate native speakers. As long as there is sufficient domain text data, a large amount of speech-text paired data can be generated, thus solving the problem of data insufficiency. Li et al. (2018) used a TTS dataset with three speakers to train a

Tacotron2 model alone, validating the effectiveness of the generated paired data in improving ASR training on the Librispeech dataset. Another study (Rosenberg et al., 2019) proposed a hierarchical variational encoder to model prosodic features and pointed out the domain mismatch problem between synthetic data and real data. However, these methods rely on a certain amount of additional paired data to train the TTS model. In one study (Rossenbach et al., 2020), the researchers trained a TTS model using only a limited ASR dataset, demonstrating the complementarity between synthetic speech-based data augmentation and other data augmentation methods. Similarly, another study (Laptev et al., 2020) trained TTS and ASR models separately using limited data, comparing the impact of different amounts of synthetic data on the performance of low-resource speech recognition.

2.4.2 Knowledge Transfer

Neural networks can be viewed as tools for extracting representations from original data. The lower-layer neurons extract low-level features, while higher-layer neurons extract more abstract knowledge representations, with the output layer providing the corresponding classification results. For a class of similar tasks, the feature representations learned by the intermediate layer neurons also exhibit similarity and can be shared across different tasks. Therefore, transferring the representations learned from high-resource data (source data) to the target model for low-resource data (target data) can help improve its performance and generalization ability. This is known as transfer learning (Huang et al., 2013). Transfer learning can be combined with multi-task learning methods, where multiple related tasks are trained in parallel by sharing network parameters and optimizing shared parameters using the error functions of multiple tasks, thereby achieving information sharing between different tasks. These methods are collectively referred to as knowledge transfer and can be classified into three categories based on whether the source data is labeled.

2.4.2.1 Supervised Learning

To improve the speech recognition performance of resource-scarce languages, many researchers attempt to leverage commonalities between different languages to optimize the model's learning of fundamental speech features. Research methods include model unit sharing (such as phonemes, subwords), feature sharing (such as bottleneck features, posterior probabilities), and acoustic model parameter sharing (partial or complete sharing). In one early study (Schultz & Waibel, 2001), the authors mapped seven languages to a common phoneme set based on the International Phonetic Alphabet, pre-training a GMM-HMM model using multilingual data and fine-tuning it with target language data, achieving better results than training with a single language. Another study (Huang et al., 2013) proposed a multilingual DNN model with shared hidden layers, sharing the input and hidden layers across multiple languages to learn common features in parallel, with separate output layers for each language. Similarly, Ni et al. (2017) proposed multilingual bottleneck features, using a bottleneck layer with fewer nodes than other layers, training the bottleneck layer with multilingual data and extracting features from it to train the low-resource language data.

When the source data is in the same language, multi-task learning methods are typically used to improve the recognition performance of the main task. For example, one study (Chen et al., 2014) used subword modeling as an auxiliary task for phoneme modeling, leveraging the high correlation between the two tasks to share the same features, thereby enhancing the model's generalization ability in low-resource scenarios. If the target domain is a specific scenario with insufficient data for training, existing data can be used to train a source domain model, which is then adapted to the target domain using a small amount of target domain data. In another study (Samarakoon et al., 2018), the authors inserted domain tags at the beginning of the label sequence, jointly optimizing the domain classification and speech recognition tasks, introducing orthogonal hidden layers and domain-related gating mechanisms to improve recognition performance in the target domain. However, the aforementioned supervised methods still rely on additional collected labeled data and cannot guarantee good generalization performance.

2.4.2.2 Unsupervised Learning

Typically, it is not the speech data itself that is scarce but the annotated paired data. To address the low-resource problem, some researchers explore the use of additional unlabeled speech data. Unsupervised pre-training methods have become a popular research direction in recent years. These methods use large amounts of unlabeled data for self-supervised training, followed by fine-tuning with a small amount of labeled data, effectively improving the performance of low-resource speech recognition.

Unsupervised pre-training methods can be broadly divided into two categories based on the pre-training task: those based on contrastive loss functions, which maximize the similarity between positive sample pairs and minimize the similarity between negative sample pairs; and those based on reconstruction loss functions, which learn feature representations by reconstructing the input data. Regarding research based on contrastive loss functions, Oord et al. (2019) first proposed contrastive predictive coding, encoding the original audio signal into low-dimensional latent representations and using an autoregressive model to predict future latent representations. The model learns the inherent structure of the speech signal by calculating the difference between predicted and actual representations using a contrastive loss function. A representative pre-training model based on contrastive loss functions is wav2vec 2.0 (Baevski et al., 2020), which extracts shallow features of the original speech through several layers of CNN, randomly masks these shallow features, and then inputs them into a context network, using contextual information to predict the masked parts.

Shallow features are discretized into a finite set of codewords. For the unmasked part, the contrastive loss function between the calculated predicted output and the true code is computed. Experimental results in reference (Baevski et al., 2020) indicate that using large-scale unlabeled data pretraining of a big model, with only 10 minutes of labeled data fine-tuning, achieves 100 hours of labeled data training performance in LibriSpeech benchmark testing. In one related research (Chung et al., 2019), the authors proposes autoregressive predictive coding, predicting future frames of information through autoregression to enable the model to learn the capability of predicting future speech frames. Additionally, Liu et al. (2020) proposes a Mockingjay pretraining model, which randomly masks input speech, reconstructs masked part information using past and future frame contextual information, and adopts the L1 norm loss. Another study (Hsu et al., 2021) proposes the Hubert pretraining model. Although it uses the same encoder structure as wav2vec 2.0, it obtains pseudo labels through clustering as training targets and uses reconstruction loss training to enable the model to learn discretized speech hidden units. With the same order of magnitude of model parameters, Hubert achieves better results on multitask speech tasks than wav2vec 2.0.

Some researchers explore the use of pure text data in end-to-end models. The most common approach is to train a language model using pure text data, then jointly decode it with recognition models through methods like shallow fusion (Kannan et al., 2017), cold fusion (Sriram et al., 2017), and deep fusion (Gulcehre et al., 2015). These methods only separately train the language model, increasing computational costs during inference. Therefore, some researchers introduce language models during training of speech recognition models and fuse the knowledge of language models through knowledge distillation (Bai et al., 2021; Futami et al., 2020; Kubo et al., 2022). Bai et al. (2021) uses pretraining BERT language models as teacher models, guiding the training of recognition models through knowledge distillation, achieving better performance than shallow fusion methods. Futami et al. (2020) utilizes RNN language models to provide soft labels and constrains the output probability distribution of the decoder and language model through KL divergence (Kullback-Leibler divergence). Kubo et al. (2022) uses BERT's representation vectors as the source of language knowledge and trains the decoder output vectors and BERT representation vectors' distance as an additional training target through multitask learning. Additionally, some researchers use pure text data for pretraining (Devlin et al., 2019; Gao et al., 2022; Yi et al., 2021; Z. Zhang et al., 2023). Yi et al. (2021) combines unsupervised pretraining acoustic models with pretrained language models, using monotonic alignment modules to adapt acoustic and language representations. Zhang et al. (2023) proposes joint pretraining of speech and text, extracting acoustic discrete units through Hubert models and phoneme units through text encoders, training explicitly unifying both speech and text modalities. Gao et al. (2022) uses artificially constructed vectors to replace the output of the decoder and pretrains the transformer decoder with a large amount of pure text data. The above training methods are multi-stage. Some researchers have proposed simpler and more efficient training methods directly learning additional text information during model training. Sainath et al. (2020) simultaneously trains a decoder task with language model tasks and applies it to rescore recognition results. Wang et al. (2021) proposes a multitask learning approach, jointly training decoder with language model tasks and speech recognition tasks without adding additional modules. Since speech synthesis and speech recognition are dual tasks, some researchers propose methods based on cycle consistency to jointly train ASR and TTS models to improve the performance and robustness of speech recognition tasks (Tjandra et al., 2017, 2018b, 2018a). Specifically, the ASR model converts speech into text, while the TTS model reconstructs speech from text. By iteratively optimizing these two processes, the ASR model learns additional text information.

2.4.2.3 Semi-supervised Learning

Semi-supervised learning is an effective approach that utilizes unlabeled data. One of the most common methods is self-training, where a pre-trained model is used to generate pseudo-labels for unlabeled data, which are then combined with the original data for retraining. The error rate of pseudo-labels is usually high and often requires filtering. Khurana et al. (2021) proposed using the uncertainty of dropout to eliminate pseudo-labels with low accuracy. Google proposed the Noisy Student Training (NST) strategy and utilized 60,000 hours of unlabeled data for NST training, achieving new heights in performance on the LibriSpeech benchmark test (Park et al., 2020). This strategy employed a better initial acoustic model as the teacher model and trained a language model with a large amount of text data. Both models generated pseudo-labels for a large amount of unlabeled speech data through shallow fusion and then filtered out relatively accurate pseudo-labels based on a threshold value to add to the training set. Subsequently, a student model was trained on the new training set using data augmentation methods such as SpecAugment and added noise. This process usually needs to be repeated multiple times to achieve better performance.

Xu et al. (2020) pointed out that self-training and unsupervised pre-training methods are complementary, and combining the two can more fully utilize unlabeled data. This method first uses a large amount of unlabeled data for pre-training, then fine-tunes with a small amount of labeled data to obtain an initial teacher model. Next, it adopts the Noisy Student Training method to generate pseudo-labels for unlabeled data and repeats the above steps iteratively to obtain the final model.

2.5 ASR for Older Adults

As people age, various alterations occur in the vocal mechanisms, including the lungs, vocal cords, and vocal tract (comprising the pharynx, mouth, and nose). In the respiratory system, the most significant change is the loss of elasticity (Mahler et al., 1986), accompanied by chest stiffness, weakened respiratory muscles, and reduced diaphragm strength (Tolep et al., 1995). These changes lead to a decrease in forced expiratory volume and lung pressure in older adults, resulting in reduced airflow volume and efficiency (Ramig et al., 2001). In the larynx, aging causes changes such as the hardening of the cartilage to which the vocal cords are attached and the degeneration of intrinsic muscles (Rodeño et al., 1993), making vocal cord adjustments during phonation more challenging (Hirano et al., 1989). Increased stiffness of the vocal fold cover also leads to unstable vocal fold vibrations (Rodeño et al., 1993). Reports indicate that the laryngeal epithelium thickens gradually with age (Sato & Hirano, 1997), which may contribute to a lower fundamental frequency and increased hoarseness in elderly voices.

Changes in the vocal tract include the degeneration of pharyngeal muscles, decreased saliva secretion, loss of tongue strength, and tooth loss (Rother et al., 2002). Degenerative changes also occur in the temporomandibular joint, which controls jaw movement during phonation (Weinstein, 2012). These changes significantly impact speech expression. The dimensions of the vocal tract in older adults also change (Xue & Hao, 2003), potentially affecting their resonance patterns and leading to reduced pronunciation accuracy. However, the extent and rate of vocal aging vary greatly. Vocal aging depends not only on chronological age but also on several other factors such as lifestyle, health status, smoking habits, and occupation.

Despite extensive research on the effects of aging on the voice, studies on how these changes impact the performance of ASR systems are limited. It has been reported that the WER for elderly voices is approximately 9-12% higher than for adult voices (Vipperla et al., 2008). A study on speech recognition for children and the elderly (Wilpon & Jacobsen, 1996) found a significant increase in WER for individuals over 70 years old.

However, research aimed at improving ASR performance for this demographic remains relatively sparse compared to the extensive efforts dedicated to enhancing ASR for children. A bunch of studies have been made to tailor ASR systems to the unique vocal characteristics and linguistic patterns of children. Some of them successfully improved the ASR performance for children speech with traditional data augmentation techniques (Patel & Scharenborg, 2024; Singh et al., n.d.; Wilpon & Jacobsen, 1996). However, similar attention has not been proportionately given to older adults. The limited research on optimizing ASR systems for older adults underscores the need for a more focused investigation into age-related vocal changes and their impact on ASR performance. Addressing this gap is essential for developing more inclusive and effective ASR technologies that cater to users across the entire age spectrum

2.6 Hypothesis

According to previous literature, my hypothesis is that fine-tuning the model with older adults speech data can improve the model performance, so as the data augmentation techniques. The combination of data augmentation techniques can better reduce the WER.

3 Methodology

This section provides an overview of the datasets, the augmentation and normalization techniques, training configurations used, and the experimental setup.

3.1 Database

Common Voice Corpus 17.0: The Common Voice Corpus contains multilingual speech data which was obtained by crowdsourcing. Speakers used the Common Voice Website (https://commonvoice.mozilla.org/en to record their speech data while reading sentences on the screen (Ardila et al., 2019). The Welsh Common Voice Corpus 17.0 comprises a total of 157 hours of speech data, with 124 hours being validated. Notably, 8.68 hours of speech emanate from speakers in their sixties, while an additional 1.2 hours stem from speakers in their seventies. The data was into two subsets: a train set and a test set. In order to guarantee that there was no speaker overlap between the training and test sets, we split the data by speakers to make sure that speakers appear in one set will not appear in the other sets. Finally, 7.9 hours are used for training and the remaining for the test set.

3.2 Augmentation techniques

Speed Perturbation (SP): SP involves altering the original raw speech signal by resampling it, causing a distortion in the timing (Ko et al., 2015). When an audio speech signal s(t) undergoes time warping by a factor , it transforms into the signal s(t). The Fourier transform of s(t) becomes S(/)/. This manipulation in the time domain results in a change in the number of frames, thereby impacting both tempo and pitch.

Spectral Augmentation (SpecAug): SpecAug manipulates the spectrogram by compressing and expanding it locally, a technique that has demonstrated enhanced recognition accuracy across various speech scenarios, including casual conversation and spontaneous speech (Park et al., 2019). Instead of directly altering the raw audio, SpecAug operates on the log mel spectrogram of the input audio. It employs three augmentation strategies: time masking, frequency masking (which blocks consecutive time steps or mel frequency channels), and time-warping, randomly distorting the spectrogram along the time axis. SpecAug is applied with default settings, including a maximum width of 30 for each frequency mask (F), a maximum width of 40 for each time mask (T), two frequency and time masks, and filling the masked regions with the mean value.

3.3 Baseline ASR model

The model utilized in this research will be the most compact publicly accessible pre-trained multilingual XLS-R model (Babu et al., 2021), aiming to reduce computational demands. XLS-R has been pretrained on approximately 436,000 hours of speech across 128 different languages, with the predominant portion of training data originating from Indo-European languages (87%), notably English, which accounts for roughly 70,000 hours. Welsh data, totally 156 hours, is also included.

The architecture and pre-training objective of XLS-R closely resemble those of wav2vec 2.0 (Baevski et al., 2020). This model operates as a unified end-to-end system, comprising a convolutional encoder, a quantizer, and a 24-layer Transformer model. Speech representations are acquired through a contrastive task applied to the quantized encoder representations. Following pre-training, the model can undergo fine-tuning for speech recognition by utilizing transcribed speech. A linear projection is appended to the Transformer network to forecast characters from transcriptions employing connectionist temporal classification (CTC).

3.4 Experimental setup

The baseline model underwent fine-tuning using exclusively sourced speech data from older adults within the corpus, aiming to specialize its parameters to better capture the linguistic nuances prevalent in speech produced by this demographic.

To investigate the impact of augmentation and normalization techniques on Welsh older adults' speech, data augmentation techniques were further applied:

Initially, the speech data of younger Welsh adults (<60 years old) was subjected to SP. The perturbed speech, combined with the original younger adults' speech, was then used to retrain the fine-tuned models, the speech rate of all training data by applying SP factors of 0.9 times and 1.1 times the original speed. This ensures the sound remains natural and is comparable to the actual conversational speed, thereby eliminating the influence of the speaker's speech rate.

Subsequently, the perturbed speech data, combined with the original speech data, were utilized to retrain the baseline models specific to Welsh older adults. These ”SP-augmented” models underwent further retraining with SpecAug applied during training (SP + SpecAug). The performance of SP and SP + SpecAug models was then evaluated on Welsh older adults' speech, with no alterations made to the audio signal of the test sets.

The fine-tuning process of the model was conducted utilizing Google Colab, a cloud-based computational platform provided by Google, specifically leveraging its T4 Graphics Processing Units (GPUs).

The fine-tuning of the model was conducted with careful consideration of various training parameters. Data samples were grouped by length to optimize training efficiency, while a batch size of 16 samples per device was chosen to balance computational efficiency and memory constraints. Gradient accumulation over 2 steps stabilized optimization in the presence of limited GPU memory. Evaluation was performed at regular intervals using the ”steps” parameter strategy. Training extended over 40 epochs to ensure convergence and robust parameter optimization. The use of mixed precision training (FP16) accelerated training and conserved memory resources. Model checkpoints were saved every 100 steps, with evaluations conducted at the same frequency for monitoring training progress. Training progress was logged every 10 steps to track performance metrics and facilitate troubleshooting. The initial learning rate of 3e-4 was chosen based on recommendations for transformer-based models, with a warmup period of 500 steps to gradually increase the learning rate.

3.5 Evaluation

Recognition performance will be reported in word error rate (WER). WER is calculated as the ratio of word or character insertion, substitution, and deletion errors in the recognized transcription to the total number of spoken words or characters in the ground truth transcription. The analysis will examine the estimated warping factors for each speaker group in the younger and older adults' test sets to investigate the link between the estimated warping factors and the recognition performance.

3.6 Ethical considerations

I have not collected any sort of data from human participants. Instead, I used a previously-recorded dataset, which is Mozilla's Common Voice project. It is a multilingual, open-, and crowdsourced corpus that is constantly updated, with support for over 100 languages. The participants in the Common Voice project are informed about their data being collected and they do so voluntarily. The recordings are also validated by the community. The corpus is licensed under CC0[1], therefore any distribution, adaptation, or otherwise may be made freely, without having to credit or mention in any way.

Most of the dataset contains data from speakers whose characteristics are unknown. Therefore, a certain degree of bias might be present in the models trained and evaluated which I mitigate by disclosing clearly and precisely that bias is present.

Objective metrics were used for evaluation which are relevant to the field. Therefore, subjective evaluation methods involving human participants have not been used and are mostly not significant to use in the field of speech recognition. Therefore, there are no concerns regarding the ethics of involving human participants or any other issues that do not align with the ethics of the faculty.

[1] Information about the CC0 license: https://creativecommons.org/share-your-work/public-domain/cc0/

4 Results

4.1 Experiment 1: Fine-tuning

The pre-tuninge-trained wav2vec 2.0 XLSR model provided on Huggingface using was fine tuned with Welsh older adult speech. The results indicate that the original model facebook/wav2vec2-large-xlsr-53 achieved a Word Error Rate (WER) of 63.31% on the development set and 62.19% on the test set. While the fine-tuned model achieved a WER of 59.40% on the development set and 57.64% on the test set.

4.2 Experiment 2: Data Augmentation

Data Augmentation	Hours	WER%
None	-	57.64
SP (110%)	102.70	58.75
SP (90%)	125.53	55.40
SpecAug	114.12	56.33
SP (110%)+SpecAug	102.70	56.81
SP (90%)+SpecAug	125.53	**54.30**

Table 1: Results in WER. The lowest WER is highlited in bold. SP represents Speed purterbation. SpecAug represents spectrum augmentation

The results obtained from the experimentation on data augmentation techniques, as summarized in the provided table, offer valuable insights into the impact of various augmentation strategies on the performance of the speech recognition model.

Introducing Speech Perturbation (SP) alone, with perturbation factors of 0.9 and 1.1, led to varying outcomes. When the data was perturbed with a factor of 1.1, the WER increased to 58.75%. Conversely, a perturbation factor of 0.9, yielded a lower WER of 55.40%. These contrasting results suggest that the magnitude and direction of perturbation can significantly influence the model's performance, with higher perturbation factors potentially enhancing model robustness.

Similarly, the application of SpecAug alone, resulted in a WER of 56.33%. SpecAug introduces variations in the spectrogram features, such as time warping and frequency masking, thereby enhancing the model's ability to generalize to different acoustic conditions.

Combining Speech Perturbation with SpecAug further elucidates the interplay between augmentation techniques. When Speech Perturbation with a factor of 1.1 was combined with SpecAug, the WER slightly increased to 56.81%. Conversely, combining Speech Perturbation with a factor of 0.9 with SpecAug, led to a notable reduction in WER to 54.30%.

The experiments on data augmentation techniques highlights the importance of selecting and combining augmentation methods to enhance the robustness and performance of speech recognition models. Future research endeavors may explore additional augmentation strategies and their synergistic effects to further advance the state-of-the-art in speech processing tasks.

5 Discussion

The results of this study align well with the initial hypothesis and the findings of previous literature. The hypothesis posited that fine-tuning the ASR model with speech data from older adults, coupled with data augmentation techniques, would enhance model performance and reduce WER. The results confirm this hypothesis, demonstrating a significant improvement in the model's accuracy when these methods are employed.

Previous research has consistently shown that ASR systems struggle with accurately transcribing the speech of older adults due to several factors, including changes in acoustic properties, variability in speech patterns, and limited representation in training datasets (Loizou & Pantzaris, 2023). These studies highlighted the need for targeted adaptations in ASR models to address these age-related differences.

Our findings build on this body of work by providing empirical evidence that fine-tuning ASR models with age-specific data can indeed mitigate these challenges. By training the model on a dataset enriched with speech samples from older adults, the system becomes better equipped to handle the unique characteristics of their speech, resulting in lower WER.

5.1 Efficacy of Data Augmentation Techniques

Data augmentation has been widely recognized as an effective strategy to enhance the robustness of machine learning models, including ASR systems. In this study, we explored 2 data augmentation techniques SP and SpecAug . The results indicate that these techniques not only contribute to the model's ability to generalize better to unseen data but also play a crucial role in further reducing WER when used in conjunction with fine-tuning.

5.1.1 Specific Findings on Speech Perturbation

A notable finding from this study is the differential impact of speech perturbation factors. Specifically, we observed that speech perturbation with a factor of 0.9 yielded better performance compared to a factor of 1.1. This outcome aligns with the slower speech rate typically observed in older adults. Perturbing the speech by a factor of 0.9 effectively simulates a slower speaking rate, which seems to make the model more adept at handling the natural pace of older speakers. In contrast, a factor of 1.1, which speeds up the speech, appears less effective and even counterproductive, as it moves further away from the natural speaking rate of the target demographic.

5.2 Implications for ASR Development

The findings of this study have important implications for the development of more inclusive ASR systems. By incorporating fine-tuning and data augmentation into the training pipeline, developers can create models that perform more equitably across different age groups. This approach not only improves accessibility for older adults but also enhances the overall user experience by providing more accurate transcriptions.

5.3 Limitations

This study improves the performance of acoustic models for older adults' Welsh speech recognition. However, despite these efforts, the WER of the recognition system based on the limited corpus remains relatively high and has not reached an applicable level. In future work, improvements can be made in the following directions:

1. Conduct more refined research based on older adults' acoustic characteristics, such as spectrum smoothing and rhythm optimization, considering specific resonance peaks and differences in vowel pronunciation. Furthermore, age-specific acoustic characteristics can be further investigated, conducting experiments grouped by age stages.

2. The available older adults' Welsh speech data is limited, resulting in significant performance gaps compared to models built on large-scale corpora. Therefore, expanding the corpus by seeking additional applicable resources or collecting more recordings of speech is necessary.

3. Incorporate more domain knowledge into older adults' speech recognition, such as utilizing multilingual corpora to study the impact of diverse language datasets on acoustic modeling for children's Mongolian speech. Transfer learning can be employed to further enhance system performance.

4. In low-resource scenarios, adversarial neural networks can be utilized to introduce domain-specific losses, allowing the model to better learn knowledge from other domains and reduce negative transfer effects.

6 Conclusion

This research contributes to reducing the discrepancy in performance between younger and older adult speech recognition, particularly for Welsh who lacking older adult speech and text data for training. Through training the baseline XLSR model using older adult speech, the WER was reduced from 62.19% to 57.64%. This illustrates the adaptability of these methodologies to various age groups, speech styles, and languages. Performance is further improved when employing advanced speed perturbation and spectrum augmentation techniques on younger adult speech, with the combination of speed perturbation at the factor of 0.9 and spectrum augmentation, the model achieved the lowest WER at 54.30% . These outcomes underscore the potential for enhancing end-to-end older adults speech recognition efficacy by: (1) employing state-of-the-art techniques, such as data augmentation techniques proven effective for children speech ASR; (2) strategically considering data availability and training approaches feasibility to refine older adults speech recognition results. This discovery contributes to the advancement of more accessible and inclusive child speech technology applications.

6.1 Future Work

Future research should continue to explore the potential of combining various augmentation techniques and further refine the fine-tuning process. Additionally, expanding the dataset to include a more diverse range of speech patterns from older adults across different dialects and sociolects could provide further improvements. Longitudinal studies examining the performance of these models over time as users age could also offer valuable insights.

In conclusion, this study demonstrates that fine-tuning ASR models with older adults' speech data and employing data augmentation techniques significantly improve model performance. These results are in accordance with the hypothesis and the established findings in previous literature, paving the way for more effective and inclusive speech recognition technologies.

7 References

Baas, M., & Kamper, H. (2022). *Voice Conversion Can Improve ASR in Very Low-Resource Settings* (arXiv:2111.02674). arXiv. http://arxiv.org/abs/2111.02674

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (arXiv:2006.11477). arXiv. https://doi.org/10.48550/arXiv.2006.

Bai, Y., Yi, J., Tao, J., Wen, Z., Tian, Z., & Zhang, S. (2021). *Integrating Knowledge into End-to-End Speech Recognition from External Text-Only Data* (arXiv:1912.01777). arXiv. http://arxiv.org/abs/1912.0177

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). *Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation* (arXiv:2305.10951). arXiv. http://arxiv.org/abs/2305.10951

Chen, D., Mak, B., Leung, C.-C., & Sivadas, S. (2014). *Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition*. 5592–5596. https://doi.org/10.1109/ICASSP.2014.6854673

Chung, Y.-A., Hsu, W.-N., Tang, H., & Glass, J. (2019). *An Unsupervised Autoregressive Model for Speech Representation Learning* (arXiv:1904.03240). arXiv. http://arxiv.org/abs/1904.03240

Cooper, S., Jones, D. B., & Prys, D. (2019). Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. *Information*, *10*(8), Article 8. https://doi.org/10.3390/info10080247

Cunliffe, D., Vlachidis, A., Williams, D., & Tudhope, D. (2022). Natural language processing for under-resourced languages: Developing a Welsh natural language toolkit. *Computer Speech & Language*, *72*, 101311. https://doi.org/10.1016/j.csl.2021.101311

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.181

Futami, H., Inaguma, H., Ueno, S., Mimura, M., Sakai, S., & Kawahara, T. (2020). *Distilling the Knowledge of BERT for Sequence-to-Sequence ASR* (arXiv:2008.03822). arXiv. https://doi.org/10.48550/arXiv.2

Gao, J., Zhang, Z., Zhou, L., Liu, S., Li, H., Ko, T., Dai, L., Li, J., Qian, Y., & Wei, F. (2022). Pre-Training Transformer Decoder for End-to-End ASR Model with Unpaired Speech Data. *Interspeech 2022*, 2658–2662. https://doi.org/10.21437/Interspeech.2022-10368

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). *Conformer: Convolution-augmented Transformer for Speech Recognition* (arXiv:2005.08100). arXiv. https://doi.org/10.48550/arXiv.2005.08100

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., & Bengio, Y. (2015). *On Using Monolingual Corpora in Neural Machine Translation* (arXiv:1503.03535). arXiv. https://doi.org/10.48550/arXiv.1503.03535

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). *Transformer in Transformer* (arXiv:2103.00112). arXiv. https://doi.org/10.48550/arXiv.2103.00112

Harvill, J., Issa, D., Hasegawa-Johnson, M., & Yoo, C. (2021). Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6428–6432. https://doi.org/10.1109/ICASSP39728.2021.9414869

Hirano, M., Kurita, S., & Sakaguchi, S. (1989). Ageing of the vibratory tissue of human vocal folds. *Acta Oto-Laryngologica*, *107*(5–6), 428–433. https://doi.org/10.3109/00016488909127535

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units* (arXiv:2106.07447). arXiv. https://doi.org/10.48550/arXiv.2106.07447

Huang, J.-T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7304–7308. https://doi.org/10.1109/ICASSP.2013.6639081

Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*(1), 94–103. https://doi.org/10.1145/2500887

Jaitly, N., & Hinton, G. E. (2013). Vocal Tract Length Perturbation (VTLP) improves speech recognition. *ICML Workshop on Deep Learning for Audio, Speech, and Language*, *117*.

Jones, D. (2022). Development and Evaluation of Speech Recognition for the Welsh Language. In T. Fransen, W. Lamb, & D. Prys (Eds.), *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022* (pp. 52–59). European Language Resources Association. https://aclanthology.org/2022.cltw-1.8

Juang, B. H., & Rabiner, L. R. (2005). *Automatic Speech Recognition – A Brief History of the Technology Development.*

Kannan, A., Wu, Y., Nguyen, P., Sainath, T. N., Chen, Z., & Prabhavalkar, R. (2017). *An analysis of incorporating an external language model into a sequence-to-sequence model* (arXiv:1712.01996). arXiv. https://doi.org/10.48550/arXiv.1712.01996

Khurana, S., Moritz, N., Hori, T., & Roux, J. L. (2021). *Unsupervised Domain Adaptation for Speech Recognition via Uncertainty Driven Self-Training* (arXiv:2011.13439). arXiv. http://arxiv.org/abs/2011.13

Knight, D., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2021). Developing computational infrastructure for the CorCenCC corpus: The National Corpus of Contemporary Welsh. *Language Resources and Evaluation*, *55*(3), 789–816. https://doi.org/10.1007/s10579-020-09501-9

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. *Interspeech 2015*, 3586–3589. https://doi.org/10.21437/Interspeech.2015-711

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5220–5224. https://doi.org/10.1109/ICASSP.2017.7953152

Kubo, Y., Karita, S., & Bacchiani, M. (2022). *Knowledge Transfer from Large-scale Pretrained Language Models to End-to-end Speech Recognizers* (arXiv:2202.07894). arXiv. https://doi.org/10.48550/arXiv.2

Laptev, A., Korostik, R., Svischev, A., Andrusenko, A., Medennikov, I., & Rybin, S. (2020). You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation. *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 439–444. https://doi.org/10.1109/CISP-BMEI51763.2020.9263564

Li, J., Gadde, R., Ginsburg, B., & Lavrukhin, V. (2018). *Training Neural Speech Recognition Systems with Synthetic Speech Augmentation.*

Liu, A. T., Yang, S., Chi, P.-H., Hsu, P., & Lee, H. (2020). Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6419–6423. https://doi.org/10.1109/ICASSP40776.2020.9054458

Liu, Y., Xu, Z., Wang, G., Chen, K., Li, B., Tan, X., Li, J., He, L., & Zhao, S. (2021). *DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021* (arXiv:2110.12612). arXiv. https://doi.org/10.48550/arXiv.2110.12612

Mahler, D. A., Rosiello, R. A., & Loke, J. (1986). The Aging Lung. *Clinics in Geriatric Medicine*, *2*(2), 215–225. https://doi.org/10.1016/S0749-0690(18)30878-4

Ni, C., Leung, C.-C., Wang, L., Chen, N. F., & Ma, B. (2017). Efficient methods to train multilingual bottleneck feature extractors for low resource keyword search. *2017 IEEE International Confer-*

ence on Acoustics, Speech and Signal Processing (ICASSP), 5650–5654. https://doi.org/10.1109/ICASSP.2017.79

Oord, A. van den, Li, Y., & Vinyals, O. (2019). *Representation Learning with Contrastive Predictive Coding* (arXiv:1807.03748). arXiv. https://doi.org/10.48550/arXiv.1807.03748

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Park, D. S., Zhang, Y., Jia, Y., Han, W., Chiu, C.-C., Li, B., Wu, Y., & Le, Q. V. (2020). Improved Noisy Student Training for Automatic Speech Recognition. *Interspeech 2020*, 2817–2821. https://doi.org/10.21437/Interspeech.2020-1470

Patel, T., & Scharenborg, O. (2024). Improving End-to-End Models for Children's Speech Recognition. *Applied Sciences*, *14*(6), Article 6. https://doi.org/10.3390/app14062353

Ramig, L. O., Gray, S., Baker, K., Corbin-Lewis, K., Buder, E., Luschei, E., Coon, H., & Smith, M. (2001). The Aging Voice: A Review, Treatment Data and Familial and Genetic Perspectives. *Folia Phoniatrica et Logopaedica*, *53*(5), 252–265. https://doi.org/10.1159/000052680

Rodeño, M. T., Sánchez-fernández, J. M., & Rivera-pomar, J. M. (1993). Histochemical and Morphometrical Ageing Changes in Human Vocal Cord Muscles. *Acta Oto-Laryngologica*, *113*(3), 445–449. https://doi.org/10.3109/00016489309135842

Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., & Wu, Z. (2019). *Speech Recognition with Augmented Synthesized Speech* (arXiv:1909.11699). arXiv. https://doi.org/10.48550/arX

Rossenbach, N., Zeyer, A., Schlüter, R., & Ney, H. (2020). *Generating Synthetic Audio Data for Attention-Based Speech Recognition Systems* (arXiv:1912.09257). arXiv. https://doi.org/10.48550/arXiv.1912.09

Rother, P., Wohlgemuth, B., Wolff, W., & Rebentrost, I. (2002). Morphometrically observable aging changes in the human tongue. *Annals of Anatomy = Anatomischer Anzeiger: Official Organ of the Anatomische Gesellschaft*, *184*(2), 159–164. https://doi.org/10.1016/S0940-9602(02)80011-5

Sainath, T. N., Pang, R., Weiss, R. J., He, Y., Chiu, C., & Strohman, T. (2020). An Attention-Based Joint Acoustic and Text on-Device End-To-End Model. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7039–7043. https://doi.org/10.1109/IC

Samarakoon, L., Mak, B., & Lam, A. (2018). *Domain Adaptation of End-to-end Speech Recognition in Low-Resource Settings*. 382–388. https://doi.org/10.1109/SLT.2018.8639506

Sato, K., & Hirano, M. (1997). Age-related changes of elastic fibers in the superficial layer of the lamina propria of vocal folds. *The Annals of Otology, Rhinology, and Laryngology*, *106*(1), 44–48. https://doi.org/10.1177/000348949710600109

Schultz, T., & Waibel, A. (2001). Experiments on cross-language acoustic modeling. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2721–2724. https://doi.org/10.21437/Eurospeech.2001-636

Singh, D. K., Amin, P. P., Sailor, H. B., & Patil, H. A. (n.d.). *Data Augmentation Using CycleGAN for End-to-End Children ASR*.

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2017). *Cold Fusion: Training Seq2Seq Models Together with Language Models* (arXiv:1708.06426). arXiv. https://doi.org/10.48550/arXiv.1708.06426

Tjandra, A., Sakti, S., & Nakamura, S. (2017). *Listening while Speaking: Speech Chain by Deep Learning* (arXiv:1707.04879). arXiv. https://doi.org/10.48550/arXiv.1707.04879

Tjandra, A., Sakti, S., & Nakamura, S. (2018a). *End-to-End Feedback Loss in Speech Chain Framework via Straight-Through Estimator* (arXiv:1810.13107). arXiv. https://doi.org/10.48550/arXiv.1810.131

Tjandra, A., Sakti, S., & Nakamura, S. (2018b). *Machine Speech Chain with One-shot Speaker Adaptation* (arXiv:1803.10525). arXiv. https://doi.org/10.48550/arXiv.1803.10525

Tolep, K., Higgins, N., Muza, S., Criner, G., & Kelsen, S. G. (1995). Comparison of diaphragm strength between healthy adult elderly and young men. *American Journal of Respiratory and Critical Care Medicine*, *152*(2), 677–682. https://doi.org/10.1164/ajrccm.152.2.7633725

Torre, P., & Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, *42*(5), 324–333. https://doi.org/10.1016/j.jcomdis.2009.03.001

Vangberg, P., Farhat, L. S., Jones, D. B., & Kinahan, S. (2023). Developing Live Welsh Speech Recognition Models for a Commercial Product—A case study. *2nd Annual Meeting of the ELRA/ISCA SIG on Under-Resourced Languages (SIGUL 2023)*, 113–115. https://doi.org/10.21437/SIGUL.2023-24

Vipperla, R. chander, Renals, S., & Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. *Interspeech 2008*, 2550–2553. https://doi.org/10.21437/Interspeech.2008-632

Wang, P., Sainath, T. N., & Weiss, R. J. (2021). *Multitask Training with Text Data for End-to-End Speech Recognition* (arXiv:2010.14318). arXiv. https://doi.org/10.48550/arXiv.2010.14318

Weinstein, B. E. (Ed.). (2012). *Geriatric Audiology* (2nd edition). Thieme.

Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, *1*, 349–352 vol. 1. https://doi.org/10.1109/ICASSP.1996.541104

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., & Auli, M. (2020). *Self-training and Pre-training are Complementary for Speech Recognition* (arXiv:2010.11430). arXiv. https://doi.org/10.48550/arXiv.2010.11430

Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study. *Journal of Speech, Language, and Hearing Research: JSLHR*, *46*(3), 689–701. https://doi.org/10.1044/1092-4388(2003/054)

Yi, C., Zhou, S., & Xu, B. (2021). Efficiently Fusing Pretrained Acoustic and Linguistic Encoders for Low-resource Speech Recognition. *IEEE Signal Processing Letters*, *28*, 788–792. https://doi.org/10.11

Zalouk, S., Hotta, H., & Pajot, C. (n.d.). *Data Augmentation for ASR using CycleGAN-VC.*

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *mixup: Beyond Empirical Risk Minimization* (arXiv:1710.09412). arXiv. https://doi.org/10.48550/arXiv.1710.09412

Zhang, Z., Chen, S., Zhou, L., Wu, Y., Ren, S., Liu, S., Yao, Z., Gong, X., Dai, L., Li, J., & Wei, F. (2023). *SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data* (arXiv:2209.15329). arXiv. https://doi.org/10.48550/arXiv.2209.15329