CHALMERS
UNIVERSITY OF TECHNOLOGY

Using language models to improve a speech recognition based maritime emergency call detection system

How the introduction of n-gram language models in wav2vec2.0 decoding affects the quality of transcripts in a low-quality audio domain

Master's thesis in Mathematical Sciences

ERIC JOHANSSON

DEPARTMENT OF Mathematical Sciences

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Using language models to improve a speech recognition based maritime emergency call detection system

How the introduction of n-gram language models in wav2vec2.0 decoding affects the quality of transcript in a low-quality audio domain

ERIC JOHANSSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2022

Using language models to improve a speech recognition based maritime emergency call detection system
How the introduction of n-gram language models in wav2vec2.0 decoding affects the quality of transcript in a low-quality audio domain
ERIC JOHANSSON

Supervisor: Kristoffer Röshammar, Tenfifty
Supervisor: Niklas Zechner, Språkbanken/GU
Examiner: Adam Andersson, Mathematical Sciences

Master's Thesis 2022
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Abstract

Novel applications of the transformers architechture as well as the availability of pre-trained models have drastically reduced the amount of data required to train successful speech-to-text (STT) models. By using the Connectionist Temporal Classification (CTC) algorithm, the process is further simplified as the training data does not have to be pre-segmented. This work aims to improve the performance of such a model developed to detect maritime VHF radio emergency calls by adding a language model to the CTC-decoding. We experiment with language models trained on several different text corpora and apply language models both in the decoding and on the resulting transcripts. The results indicate the importance of large amounts of domain-specific text. The results also show that a reduced Word Error Rate (WER) does not necessarily lead to an improvement in contextual comprehension. Finally, it is shown that relatively large improvements are given by fine-tuning various pre-trained STT-models on a curated dataset.

Keywords: speech to text, automatic speech recognition, natural language processing, NLP, language model, wav2vec2.0, VHF, emergency call detection.

Acknowledgements

I want to thank everyone at Tenfifty for making me feel welcome and for insight into the day-to-day tasks of a company engaged in AI. Special thanks to Niclas Johansson for sharing your knowledge on the subject. I want to thank Niklas Zechner for interesting discussions where you asked questions that often made me aware of what I understood and most importantly what I did not understand. Thank you Adam Andersson for examining the project and Kristoffer Röshammar for supervising. I want to thank the whole AI community for the open source culture that gives a simple man like me the opportunity to experiment with SOTA techniques. Last but not least I want to thank my family. Linn because you put up with me sitting in front of the computer in what feels like an infinite number of evenings and late nights. Nike because you give me everyday joy and always a reason to try to be my best self, even though it has not been the easiest to combine paternity with a master's thesis.

Eric Johansson, Gothenburg, June 2022

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ASR	Automatic Speech Recognition
CER	Character Error Rate
CNN	Convolutional Neural Networks
CTC	Connectionist Temporal Classification
ELMo	Embeddings from Language Models
HMM	Hidden Markov Model
JRCC	Sjöfartsverkets Sjö- och Flygräddningscentral
LM	Language Model
LSTM	Long Short Term Memory
MLE	Maximum Likelihood Estimate
NLM	Neural Language Model
OOV	Out Of Vocabulary
RNN	Recurrent Neural Network
SOTA	State Of The Art
STT	Speech To Text
VHF	Very High Frequency
WER	Word Error Rate

Contents

List of Figures

List of Tables

1

Introduction

The following section will give the background necessary to understand the task of the thesis, define the aim and limitations of the thesis and state a number of questions to be answered during the working process.

1.1 Background

Sjöfartsverkets sjö- och flygräddningscentral (JRCC) is a Swedish state rescue centre for sea- and air rescues [3]. The main task of the naval department is to coordinate sea rescue between a large number of actors. Very High Frequency (VHF) Radio is the main tool for communication at sea and it is used both for ship-to-ship communication and for ship-to-coast communication. The Swedish coastal radio network consists of 56 VHF base stations that together covers the full area of responsibility of JRCC [1]. VHF channel 16 is designated to emergency calls by international agreement and monitored 24 hours a day by rescue leaders from JRCC [4]. However, due to a combination of noisy audio, multiple messages transmitted in parallel and that the reception of a call is completely dependent on the operator's ability to perceive it, there is always a risk that emergency calls will be missed. For this reason and on behalf of JRCC, Tenfifty has developed a speech-to-text (STT) solution utilizing the wav2vec2.0 framework [7] for detecting emergency calls and notify the rescue leaders [21]. The project goes under the name *Heimdall*. At the current state the system monitors transmissions from all of the 56 base stations independently and uses a interface to alert the JRCC rescue leaders if a message is classified as a emergency call [56].

The process of converting speech to text is done in two main steps. First, the audio is transformed into a sequence of speech representations [7]. Second, the speech representations is transformed into a sequence of probable words. Tenfifty considers the transformation of audio into speech representations as well-functioning but that there are deficiencies in the algorithm used to transform the speech representations into words. One possible method of improving this transformation would be to integrate a language model in the inference process, which is also the main purpose of this thesis.

1.2 Problem

There are three main difficulties with the project besides the low quality of the audio data which is more related to the audio-to-speech representation transformation. First, there is a lack of transcripts. This is a problem since language models in general perform better when trained on a larger text corpus [13, 69]. Second, incoming messages are of various spoken languages. Most messages are in Swedish or English, but JRCC also picks up transmissions from Danish, German, Finnish and Norwegian communication. Additionally, some messages consists of various languages in themselves. To handle messages in different languages, an ensemble of STT models trained in English and Swedish messages is used. Third, particular difficulties arise in the transcription of named entities. These are often out of vocabulary (OOV) words, i.e not part of the language model lexicon. Without any measures taken, a language model will thus consider the probability of these words to appear as zero. It is also difficult for an uninitiated individual to perceive these words in the transcription process.

1.2.1 Objectives

The aim of the thesis is to improve the quality of the STT-model's text output in the sense of reducing the Word Error Rate (WER) and Character Error Rate (CER) as well as captioning more of the context of the messages. By extension, this means a higher precision in identifying messages of interest, but also the possibility of enabling more in-depth analysis of incoming messages. A key aspect is to improve the recognition of named entities.

1.3 Limitations

An initial limitation regarding the use of language models is that only n-grams will be considered. This is because of their smaller computational cost and that *Hugging Face*[1] together with *pyctcdecode*[2] made it possible to easily connect n-gram language models to wav2vec2.0. Language models will only be developed for messages in English and Swedish. Thus, no emphasis will be placed on improving the recognition of messages in Norwegian, Danish, German, Finnish or other languages.

1.4 Contribution

Because the thesis is done on a very specific domain, there are no standard benchmarks to compare with. We will instead compare locally with the currently used models in project *Heimdall*. Our main contribution is about the usability of language models in the specific domain and more generally as the availability of training data to the language models is very limited.

[1] https://huggingface.co/blog/wav2vec2-with-ngram
[2] https://github.com/kensho-technologies/pyctcdecode

2
Theory

In the following chapter the theory needed to understand the thesis is presented. The reader is assumed to possess basic knowledge in machine learning and statistics.

2.1 Very High Frequency radio

Very High Frequency (VHF) [65] is radio frequencies between 30-300 Mhz. VHF radio communication is commonly used in maritime communication between different operators such as ships and coastal stations due to its excellent coverage. VHF basically has an optical range, which means that the range increases with the height of the location of the radio transmitter. Civil shipping uses the band 156-174 Mhz distributed along 91 channels where Channel 16 is designated as emergency channel.

2.2 Artificial Neural Networks

This section explains the neural network structures that are directly involved in wav2vec2.0. The reader is assumed to have prior knowledge of Convolutional Neural Networks (CNN) and variations of Recurrent Neural Networks (RNN). To refresh the memory or if prior knowledge is lacking, the textbook of Ian Goodfellow et al. [24] is recommended.

2.2.1 Encoder-Decoders in Sequence Learning

In 2014, Sutskever et al. propose an *Encoder-Decoder* approach to machine translation, a task that can be described as sequence to sequence mapping with non-fixed dimension of input- and output vectors and non-monotonic alignment between inputs and outputs [58]. Their idea was to use an *encoder* to transform the sentence to be translated, $(x_1, x_2, ..., x_T)$, into a vector representation $\vec{c}$ from which the associated translation, $(y_1, y_2, ..., y_{T'})$, is then decoded using a *decoder*. Here T and T' are the number of words in the source sentence and the translation respectively and that their length are not necessarily the same. The encoder consisted of a multilayered Long Short-Term Memory (LSTM) [26] computing hidden states at time t based on the input at time t and the hidden state at time $t-1$,

$$h_t = f(x_t, h_{t-1}).$$

$\vec{c}$ is then taken as the resulting LSTM hidden state after feeding the network with the entire source sentence. Note that the output of the LSTM encoder for each time step is ignored, only the hidden states are of interest throughout the computations. Another LSTM is then used as the decoder to extract the output sequence from $\vec{c}$ by setting $\vec{c}$ as the initial hidden state. The computations done by the decoder can be stated as

$$p(y_{t'} \mid y_1, ..., y_{t'-1}, x_1, ..., x_T) = g(f(y_{t'-1}, h_{t'-1})), \quad t' \in 1, ..., T', \tag{2.1}$$

where f is the decoder LSTM, g is typically a dense layer with the softmax function (2.3) applied to produce probabilities over the vocabulary, $y_{t'-1}$ is the predicted word of the translation at time-step $t'-1$ and

$$h_{t'-1} = \begin{cases} \vec{c} & \text{if} \quad t' = 1 \\ s_{t'-1} & \text{if} \quad t' > 1 \end{cases}.$$

Here $s_{t'-1}$ is the decoder hidden state at time step $t'-1$.
The encoder-decoder pair is trained jointly to maximize the probability of a correct translation given a source sentence

$$p(y_1, y_2, ..., y_{T'} \mid x_1, x_2, ..., x_T).$$

Because the decoder is only conditioned on the context vector this approach can be generalized to any task with the goal of outputting a sequence from some content as long as that content can be summarized in a context vector, such as image captioning [62].

2.2.2 Attention

In 2015 Bahdanau et al. proposed attention [8], a key innovation leading to neural machine translation systems that outperform traditional phrase-based translation systems [54]. A classic sequence to sequence model passes only the last hidden state to the decoder, which is conjectured as a bottleneck in performance [8], while attention allows a pass of all hidden states. The attention mechanism is then used to provide information about which hidden states the decoder should pay attention to at a given time step. This is done by modifying (2.1) into

$$p(y_{t'} \mid y_1, ..., y_{t'-1}, x_1, ..., x_T) = g(y_{t'-1}, f(h_{t'-1}, y_{t'-1}), c_{t'}), \quad t' \in 1, ..., T', \tag{2.2}$$

where $c_{t'}$ is a weighted sum of the encoder hidden states $h_1, ..., h_T$:

$$c_{t'} = \sum_{t=1}^{T} \alpha_{t't} h_t.$$

The weights $\alpha_{t't}$ are obtained from the softmax function

$$\alpha_{t't} = \frac{e_{t't}}{\sum_{t=1}^{T} e_{t't}}, \tag{2.3}$$

where

$$e_{t't} = \Psi(s_{t'-1}, h_t)$$

is a score of how well the encoder hidden state from time step t in the input time domain matches with the decoder hidden state from time step $t' - 1$ in the output time domain and Ψ is a feed forward neural network trained jointly with all other parts of the system. The intended effect of attention is then that the calculation of $y_{t'}$ should be most strongly influenced by the time steps in the input time domain that contribute with contextual information to maximize (2.2).

2.2.3 Transformers

Multiple ways of calculating attention scores have been proposed by various authors, e.g. cosine similarity [28], additive attention as described in section 2.2.2 and dot-product attention $s_t^T h_i$ [42] where s_t and h_i corresponds to the hidden states of positions t and i in the output and input respectively if compared to Section 2.2.2. The authors of [61] proposed the Transformer model, relying on a scaled version of the dot-product attention and completely ignore the use of RNN's and CNN's which at that time was state of the art in sequence to sequence models. They argued that this would be beneficial in computational complexity, that it would allow for more computations to be parallelized and that it would facilitate long range dependencies. The key to all of this is the self attention mechanism.

2.2.3.1 Self attention

As opposed to the attention mechanism described in Section 2.2.2 where attention is paid to the encoder hidden states from the perspective of the decoder, self-attention is applied within one and the same model-component in order to compute a representation of a single sequence by relating different positions of this sequence [61]. The idea is to use learnable weight matrices W_q, W_k and W_v to extract queries (q), keys (k) and values (v) from each word in the input. Intuitively we can think of the queries as a way to send requests and keys as a way to regulate the response. Following the same line of thinking, the value represent the response. For each word in the input, its corresponding query is compared to the keys of all words in the input (including itself) by computing their dot product. The attention scores are then computed by dividing the dot-products by $\sqrt{d_k}$, where d_k is the dimension of the queries and keys, and then normalizing using the softmax function

$$\text{AttentionScores}(q_i, K) = \text{Softmax}(\frac{q_i K^T}{\sqrt{d_k}}). \tag{2.4}$$

Here (2.4) describes the attention scores corresponding to word i of the input. q_i is a one dimensional vector of length d_k and K is a two dimensional matrix where row j corresponds to the key of word j in the input. The reason for dividing by $\sqrt{d_k}$ is to prevent the saturation of softmax and thus the stagnation of the gradients. The i:th self attention output is then the weighted sum of all input values taking the attention scores as weights

$$\text{Attention}(q_i, K, V) = \text{AttentionScores}(q_i, K)V, \tag{2.5}$$

where V is a two dimensional matrix with its rows corresponding to value vectors. Note that d_k, the dimension of keys and queries, must be the same because of the dot product computation in (2.4) but that the dimension of values, d_v, is independent of these according to (2.5). In the transformer architecture the computations are done over all queries simultaneously and the vectorized attention function is defined as

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.6}$$

where Q, K and V are sets of queries, keys and values respectively.

2.2.3.2 Multi-head attention

In [61] it was noted that the averaging of value vectors in the self-attention output prevented the transformer from paying attention to multiple parts of the processed sequence. They therefore introduced multihead attention defined as

$$\text{MultiheadAttention}(Q, K, V) = [\text{head}_1; ..., \text{head}_\text{h}]W_o$$

where

$$\text{head}_\text{i} = \text{Attention}(QW_{i,q}, KW_{i,k}, VW_{i,v}).$$

Here (2.6) is applied h times in parallel with different weight matrices $W_{i,q}$, $W_{i,k}$ and $W_{i,v}$ for each head. This enables each of the heads to find vector representations focusing on different parts of the input. The output of each head is then concatenated and linearly transformed into the appropriate dimension using W_o.

2.2.3.3 Transformer model architecture

The original transformer was built mainly for the purpose of machine translation as a encoder-decoder configuration. Both the encoder and decoder is built up by stacked blocks of self-attention modules with some differences between the two. The encoder block has two sublayers, the first consisting of multi-head attention and the second by a feed forward module. Both sublayers has residual connections around them and both are followed by a layer normalization to improve training performance. The decoder block has three sublayers with the same residual connections and layer normalization as the encoder block. The first sublayer is a masked multi-head attention where future sequence positions are hidden to the self-attention mechanism. The second sublayer is defined as the encoder-decoder multi-head attention where the values and keys are provided by the encoder and the queries come from the previous decoder sublayer. The third sublayer is a feed forward layer. The decoder is then followed by a linear projection and a softmax to predict the next token. The architecture can be seen in Figure 2.1.
Later tranformers come in different configuration, e.g. BERT which is a bidirectional encoder-only configuration [17] and GPT-2 which is unidirectional encoder-only configurated [52].

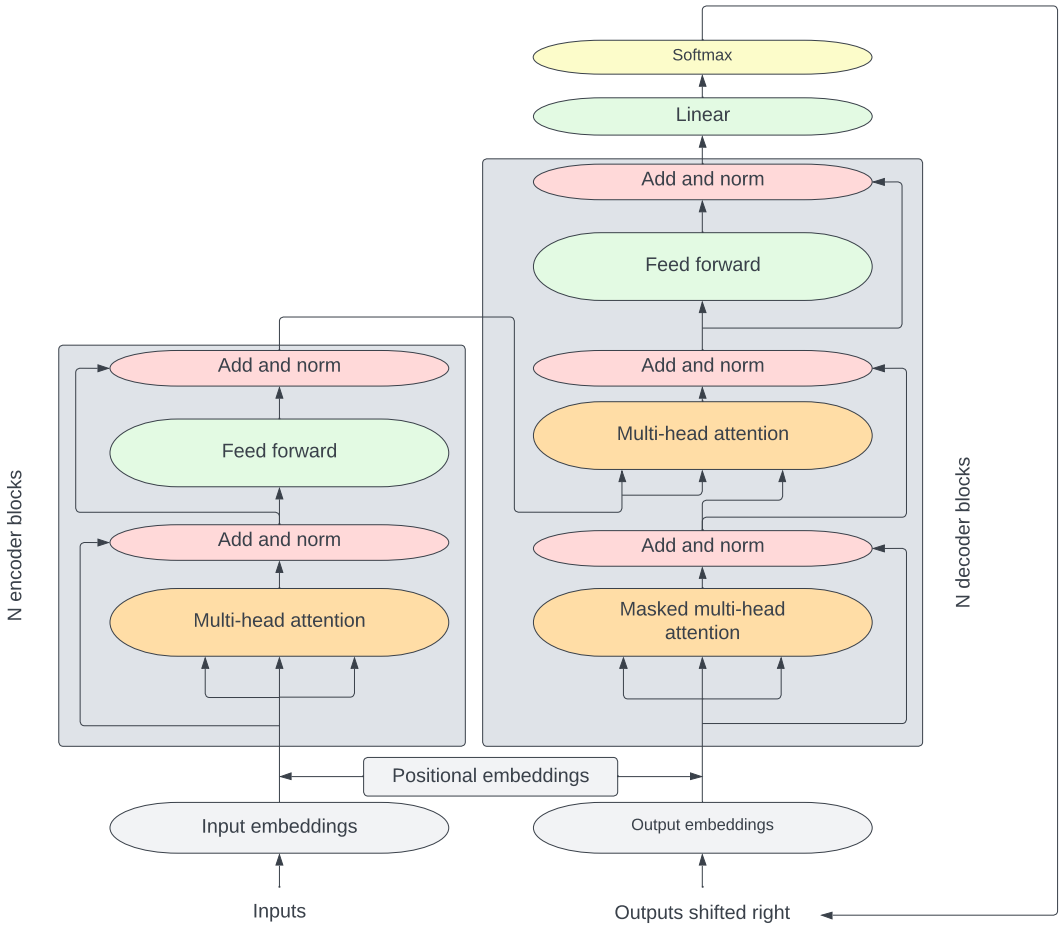

Figure 2.1: Transformers architecture with the encoder block to the left and the decoder block to the right [61]. Arrows forged into three represents queries, keys and values from left to right.

2.3 Speech to text

Speech to text is the task of deriving a sequence of words from audio recordings of speech. The first speech-to-text system is named Audrey and built by Bell industries in 1952 [49]. It was fully analogue and managed to recognize spoken digits with high accuracy. Since the 1980s, the field has been dominated by Hidden Markov Models (HMM) models in different configurations [51] and later by hybrid HMM/ANN models [11]. In the mid 2010s end-to-end deep learning systems began to show competitive results [31]. This despite being simpler and without the need for hand crafted pipelines which is the case for previous systems. Today, unsupervised learning has also established itself in speech to text, which makes it possible to train models with raw audio data without associated transcripts. A contributing factor to this development is described in the following section.

2.3.1 Connectionist Temporal Classification

Many sequence to sequence tasks include the difficulty of different input length and output length. For example, in speech to text, several subsequent input positions often correspond to one and the same phoneme/letter or to silence. We also usually do not know how the associated transcription is aligned to the sound. This alignment can be done by hand but it is time consuming and costly. There are methods to overcome these issues, i.e. the use of HMMs [51] and CRFs [40] but these are sub-optimal e.g. as they require initial understanding of the data before they can be used [27]. Neural networks on the other hand has potential of performing sequence to sequence tasks without any a priori knowledge about the data as long as the input and output is defined together with a suitable objective function. Connectionist Temporal Classification (CTC) [27] was proposed in 2006 as a method to perform aligned labeling of sequences and thus enable the use of neural network-only based models in tasks such as speech recognition and hand-writing labeling.

2.3.1.1 CTC decoder

The general idea with the CTC decoding is to include a special blank token in the vocabulary. The blank token is then used as a marker for when vocabulary tokens should be repeated or collapsed. CTC-networks ends with a softmax layer resulting in a probability distribution over the vocabulary tokens $y_1, \ldots, y_{V+1}$ for each input position. This means that the CTC decoding is applied to a $(V+1) \times T$ sized matrix where V is the vocabulary size without the blank token and T is the input length. The decoded transcription Y^* is then given by

$$Y^* = \underset{Y}{\operatorname{argmax}} P_{CTC}(Y \mid X), \tag{2.7}$$

where $P_{CTC}(Y \mid X)$ is the probability of Y given the input data X and the CTC decoding strategy. The simplest decoding strategy is to greedily choose the vocabulary token with the highest softmax score at each time step. The resulting sequence of tokens is then decoded by collapsing all repeated labels into one instance and then removing all blank labels, i.e. *cccc-aa-t* $\rightarrow$ *cat* and *ccc-c–aaa-t-t* $\rightarrow$ *ccatt* where *'-'* is the blank token. To enable paths other than the one found by greedy decoding, a so called beam search can be performed which will be next described. At each time step the top k candidates are then kept based on their P_{CTC}-score so far. This is done by extending each of the k kept candidates from the previous time step by each vocabulary item into a set of hypotheses and from these save the top k candidates. The parameter k defines the width of the beam [30]. The beam search may also be modified into consider the k best states instead of the k best paths where a state is defined by all alignments resulting in a given output prefix, e.g. *caat* and *ca - t* both decodes to *cat*.

Beam search also enables the insertion of a language model in the CTC decoder according to

$$Y^* = \underset{Y}{\operatorname{argmax}} P_{CTC}(Y \mid X) + P_{LM}(Y)^{\alpha} + L(Y)^{\beta},$$

where $P_{CTC}(Y \mid X)$ is the same conditional probability as in (2.7), $P_{LM}(Y)$ is the language model probability of the proposed label Y, $L(Y)$ is the number of words in Y and α and β are hyperparameters. Language models are described in more detail in Section 2.4 but the purpose of their introduction is to promote pathways that consists of valid words and sentences in the decoding. The reason for including $L(Y)$ is to counteract the restrictive effect of the language model.

2.3.1.2 CTC training

The understanding of CTC training is simplified if we first reconstruct the $(V+1)\times T$ softmax matrix described in Section 2.3.1.1 by removing all rows corresponding to vocabulary tokens that are not in the annotated transcript. We then arrange this table such that it, from top to bottom, has the same sequence of tokens as the transcript. Finally, we add the row corresponding to the blank token between each pair of token as well as to the start and the end of the sequence. The resulting table for the case where the annotated transcription is *cat* can be seen in Figure 2.2. We now have a starting point for using dynamic programming to find all paths corresponding to the correct label. The algorithm used is based on the forward/backward HMM algorithm [51] and can be seen described in detail in [27]. In short, the algorithm is based on defining a forward variable $\alpha_t(s)$ to be the total probability of the labeled sequence $\vec{l}_{1:s}$ up until state s at time t and a backward variable $\beta_t(s)$ to represent the probability of the labeled sequence $\vec{l}_{s:\|\vec{l}\|}$ subsequent to state s at time t. The network weights is then updated to maximize the likelihood of all states corresponding to the correct labeling. The functionality of the CTC algorithm thus entails an ambiguous segmented annotation of each training example based on the sound model's softmax output score.

2.3.2 Wav2vec 2.0

Wav2vec2.0 is a self-supervised framework for learning speech representations from unlabeled data [7]. It is a further development of a number of previous frameworks:

- Wav2vec introduces a unsupervised training technique where raw audio input is embedded in a latent space by a convolutional encoder network. A convolutional context network then combines multiple time steps of the latent embeddings to obtain contextual representations. These contextual representations is then fed to the acoustic model [57].
- Vq-wav2vec utilize the architecture and loss of wav2vec to produce discretized contextual representations which enables the application of advances made in the NLP field, in this case specifically BERT, to improve the quality of the contextual representations to be fed to the acoustic model [6].

Wav2vec2.0 solves the same two tasks as vq-wav2vec combined with BERT but in a end-to-end framework. By the addition of a linear layer and softmax, wav2vec2.0 can then be fine-tuned to map the learned sound representations directly to the vocabulary and thus be used as an end-to-end speech to text model. Wav2vec2.0 reached state of the art performance on various STT-datasets as it was published and proved to yield very good results even with as little as 10 minutes of labeled data available for fine-tuning.

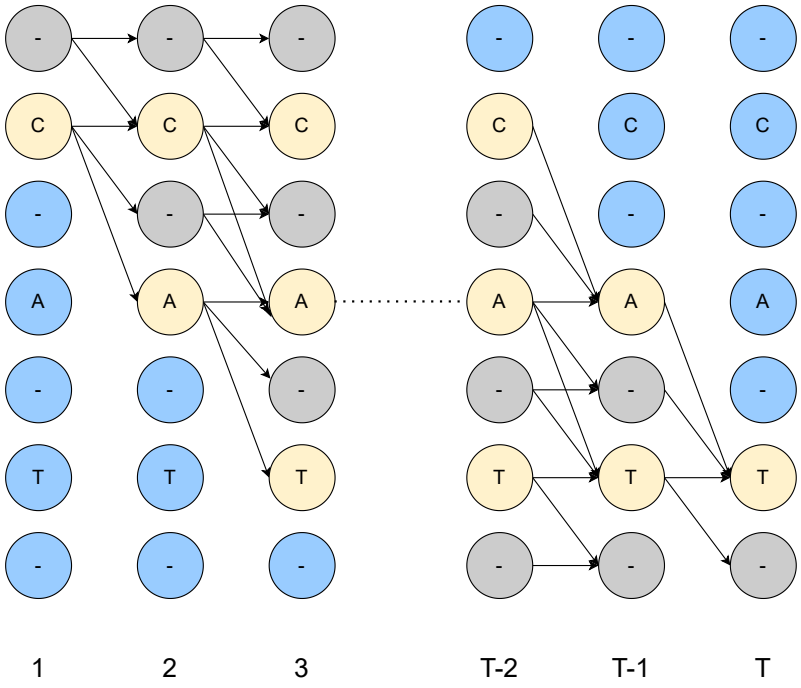

Figure 2.2: CTC forward/backward algorithm setup corresponding to the annotated transcription *cat* with states on the y-axis and time (positions) on the x-axis. Paths must begin in the state of the first letter token of the truth transcript or its preceding blank token. Paths must end in the state of the last letter token of the truth transcript or in its subsequent blank token. Paths are only allowed to go right or down. Paths may skip blank tokens, but not letters. The arrows represent possible directions from each state. States colored in blue represent states that are unfeasible of being part of paths corresponding to the annotated transcript.

2.3.2.1 Model architecture

Wav2vec2.0 has two different shapes depending on whether it is in the pre-training mode or in the fine-tuning/evaluation mode.

In the pre-training mode raw audio $\mathcal{X}$ is fed to a feature encoder composed by a temporal convolutional network mapping the audio to latent speech representations $f : X \rightarrow Z$. The latent speech representations are then fed to a transformer network used to learn contextual representations based on the whole sequence $g : Z \rightarrow C$. In parallel, the latent speech representations are sent through an quantization module $h : Z \rightarrow \mathcal{Q}$ where $\mathcal{Q}$ are quantized and used as targets which enables the pre-training to be self-supervised. The pre-training process is explained in more detail in section 2.3.2.3. The model architecture for pre-training can be seen in Figure 2.3.
When fine-tuning the model, the quantization module is removed and the contextualized transformer-outputs is fed to a fully connected layer ending in a softmax converting its inputs to probabilities over the vocabulary for each time step. The feature encoder module is frozen during fine-tuning, the rest of the network is trained by minimizing the CTC-loss. Note that the fine-tuning thus require annotated data. The model architecture for fine-tuning can be seen in Figure 2.4.

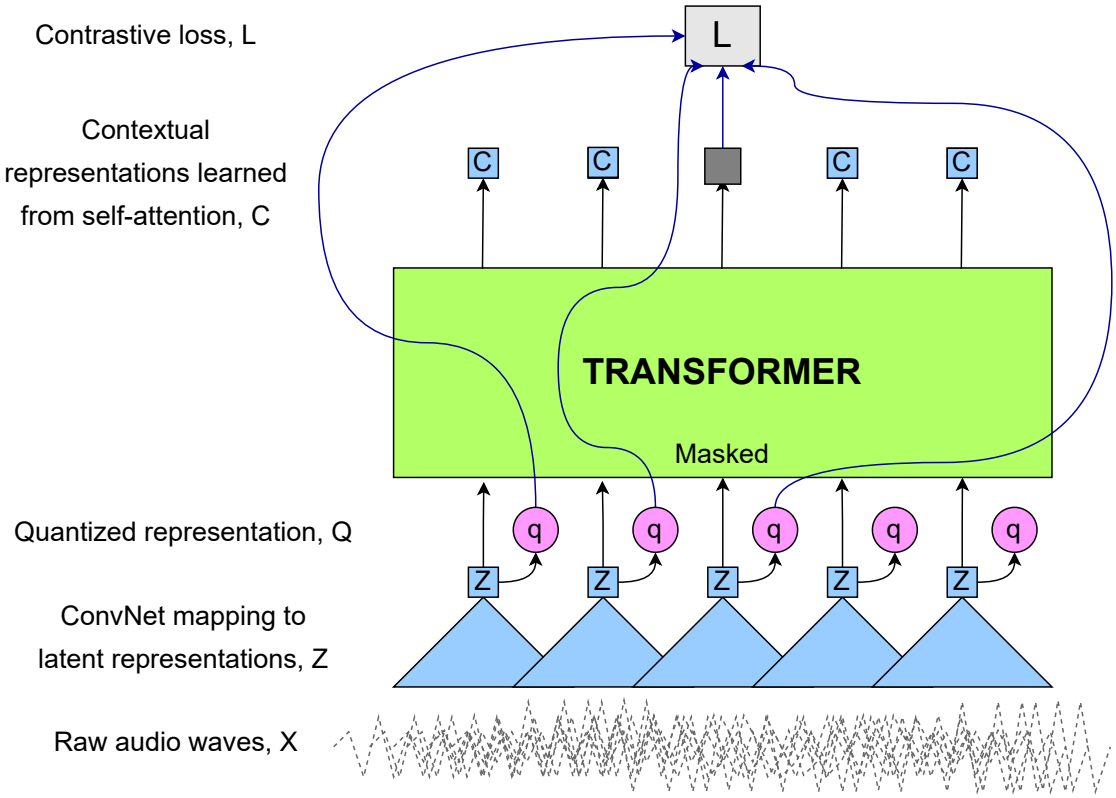

Figure 2.3: Wav2vec2.0 model architecture for pre-training. About 49% of the transformer network inputs are masked [7] which is utilized in (2.9).

2.3.2.2 Quantization module

The quantization module is applied to transform the latent speech representations $\mathcal{Z}$ into quantized representations $\mathcal{Q}$ which is then to be used as targets in the self-supervised training. This is done through product quantification which basically means choosing a single quantized representation e from each of G different look-up tables of dimension $R^{V \times d/G}$ and then concatenating the resulting quantized representations such that $e = [e_1; ...; e_G]$. A linear transformation $f : \mathbb{R}^d \to \mathbb{R}^f$ is then applied to obtain $\boldsymbol{q} \in \mathbb{R}^f$.

To decide what entries in each code-book to be chosen, $\mathcal{Z}$ is mapped to $\boldsymbol{l} \in \mathbb{R}^{G \times V}$ logits which is then fed to the Gumbal softmax function, a differentiable approximation of argmax, defined by

$$p_{g,v} = \frac{\exp((\boldsymbol{l}_{g,v} + n_v)/\tau)}{\sum_{k=1}^{V} \exp((\boldsymbol{l}_{g,k} + n_k)/\tau)}. \tag{2.8}$$

Here τ is a non-negative temperature and $n_v = -log(-log(u))$ where u is sampled from $\mathcal{U}(0, 1)$. During forward pass, code-book entries are chosen according to $i = \underset{j}{argmax}\, p_{g,j}$ and during the backward pass weights are updated according to the gradient of the Gumbel softmax output. The temperature τ is initiated to a large value to favor exploration of the code-books in the beginning of training and then

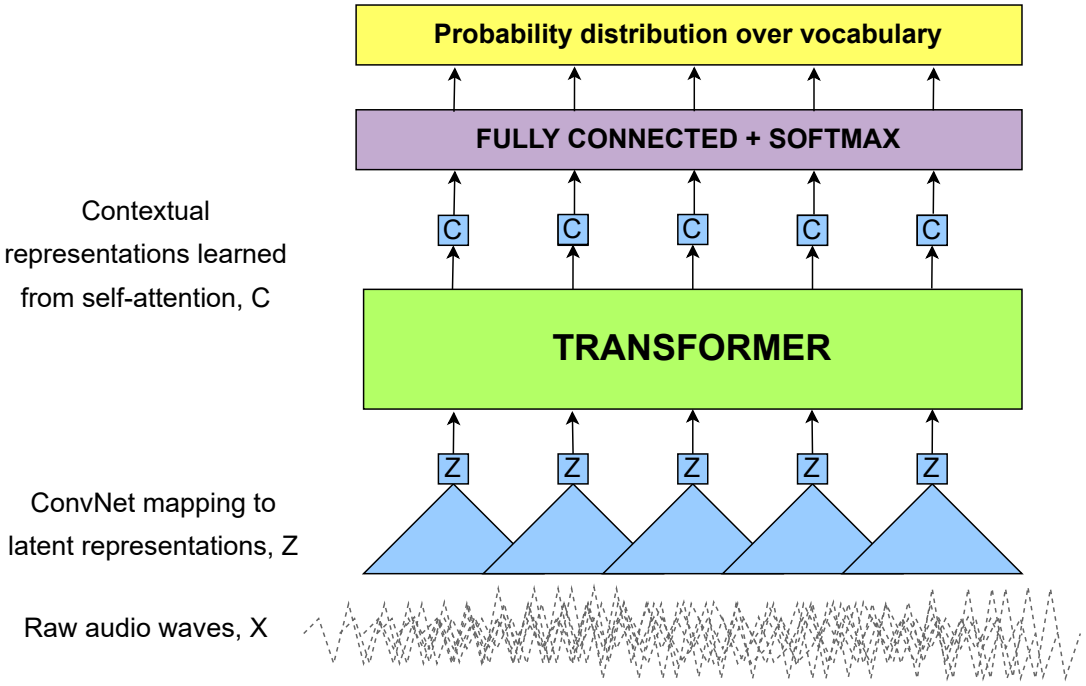

Figure 2.4: Wav2vec2.0 model architecture for fine-tuning and evaluation.

decreased to favor exploitation in later stages.

2.3.2.3 Pre-training

The network is optimized according to

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d$$

where $\mathcal{L}_m$ is a contrastive loss and $\mathcal{L}_d$ is a diversity loss. The contrastive loss is defined as

$$\mathcal{L}_m = -\log \frac{\exp\left(\text{sim}(\boldsymbol{c_t}, \boldsymbol{q_t})/\kappa\right)}{\sum_{\tilde{\boldsymbol{q}} \sim Q_t} \exp\left(\text{sim}(\boldsymbol{c_t}, \tilde{\boldsymbol{q}})/\kappa\right)} \tag{2.9}$$

where $\boldsymbol{c_t}$ is the transformer output corresponding to a masked latent input at time step t. Q_t is a set of $K+1$ candidate quantized representations including $\boldsymbol{q_t}$, the true quantized representation, and K discriminators sampled from other masked time steps. κ is the temperature of the constrastive loss. Each masked position of the transformer is rewarded by its similarity to its corresponding quantized representation as well as its dissimilarity to all negative samples. The similarity is meassured as $sim(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a^T b}/||\boldsymbol{a}||||\boldsymbol{b}||$. The diversity loss is defined as

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1} G \sum_{v=1} V p_{g,v}^{-} \log p_{g,v}^{-} \tag{2.10}$$

where $p_{g,v}^{-}$ is the same as (2.8) but without Gumbel noise and temperature. The role of (2.10) is to maximize the entropy of (2.8) to promote the use of all lines V in each code book G.

2.3.2.4 Fine-tuning

The model is fine-tuned on labeled data using CTC-loss by adding a linear layer mapping the contextual speech representations of each time step to the vocabulary.

2.4 Language modeling

The overall idea of a language model (LM) is to determine the probability of a sequence of words [25]. Typical uses is to predict the next word of a sequence given the previous words or to determine how likely a sequence of words are. Language modeling play a key role in Natural Language Processing (NLP), having a good LM often improves the performance on downstream tasks such as speech recognition, machine translation and text summarizing [35]. Relevant to the task of this thesis, a LM can be seen as a aid guiding the CTC beam search decoder to find paths that make up existing words that fits with the context.

As an example, assume the following text as outputted by a STT-model without using a language model: *I like chokolat*. By analyzing the context and looking in their vocabulary it is clear to most human readers that *chokolat* should actually be *chocolate*. But without a LM, the STT-model has no sense of context or vocabulary and thus it outputs the words corresponding to the phonemes that are assigned the highest probabilities based on the transformation of input audio only. A well functioning LM added to the model pipeline brings with it a vocabulary and the ability to draw conclusions about the context which in this case is exactly what is needed to interpret *choklat* as *chocolate*. As described above the use of a LM in combination with a STT-model will ideally enhance the output text quality, e.g. by correcting misspellings.

2.4.1 Statistical language models

Statistical language models computes the joint probability of a sentence or a sequence of words [9]. As an example the probability of generating the sequence of words $S = w_1, w_2, w_3, w_4$ can by found by the chain rule of probability as

$$P(S) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)P(w_4 \mid w_1, w_2, w_3)$$

or more general for a sequence $S_n = w_1, \ldots, w_n$ of arbitrary length n as

$$P(S_n) = \prod_{i=1}^{n} P(w_i \mid w_1, \ldots, w_{n-1}). \tag{2.11}$$

2.4.2 Neural language models, a short survey

Another approach to LMs is the use of neural networks. The first neural language model (NLM) to make an impact was presented by Bengio et al. [10] in 2001 where vector representations was learned for each word. These vectors would later go under the name word embeddings and the key idea is that words with similar meaning has similar representation in the vector space. Further, the probability distribution of word sequences can then be expressed by these embeddings. The training of word embeddings was refined in 2013 by Mikolov et al. [43, 44] with the *word2vec*-embedding which enabled large scale training. Word embeddings trained on large data led to better and a more wide range of associations between words such as semantic and syntactic patterns [54]. Interestingly, many of these patterns can be represented as linear translations of vector representations such as $\mathbf{vec}(king) - \mathbf{vec}(man) + \mathbf{vec}(woman) = \mathbf{vec}(queen)$ [44]. Word embeddings was a fundamental building block as ANN-models got adopted to the field of NLP [16, 64, 59].

The approach of using pretrained word embeddings suffer from two major limitations. First, all previous knowledge is incorporated in the first layer of the model only. The rest of the network has to be trained from scratch to learn to understand the knowledge that the word embeddings bring as well as to solve the intended task [55]. Second, all specific words have the same embedding, regardless of context.

As a solution to the second limitation, ELMo (Embeddings from Language Models) was introduced in 2018 [48]. ELMo is a pretrained language model based on bidirectional LSTMs that outputs word embeddings which contains information about the word's syntax and semantic properties but which also depends on the context surrounding the word. A specific word can then have different representations depending on the linguistic context.

With inspiration from the transfer learning paradigm in computer vision [18] ULMFiT (Universal Language Model Fine-tuning) was proposed as a corresponding method for NLP in 2018 [32]. The training is done in three steps, (1) general LM pretraining, (2) target domain fine-tuning and (3) target classifier fine-tuning. In the first step the model is trained as a LM using a large general text corpus, in the second step the LM is fine-tuned on a corpus representative of the target task and for the third step a two layer linear feed forward network is added to the LM ending with a softmax activation function generating a probability distribution over the target classes. In order to prevent *catastrophic forgetting* the authors of ULMFiT utilize novel fine tuning techniques such as *discriminative fine-tuning*, *gradual unfreezing* and *slanted triangular learning rate* [32].

2.4.3 N-grams

An obvious problem with a language model such as the one given by (2.11) is that it assigns zero probability to a sequence if any portion of that sequence is previously unseen. Looking back at equation (2.11) and letting n grow large one can imagine

that many such sequences of words S_n are very unlikely to have appeared in the data which the language model has been trained on and thus $P(S_n) = 0$. Given that the amount of data is large enough it is however likely that most portions of small length, say $n < 5$, which make up S_n can be found somewhere in the training data. N-gram language models make use of this very assumption, building upon the Markov assumption that the next word only depends on the $n-1$ previous words [68].

The simplest Markov assumption is a *unigram* model, i.e where $n = 1$. This means that the probability of a sequence of words is approximated by the product of the probability of each word:

$$P(S_k) \approx \prod_{i=1}^{k} P(w_i),$$

where k is the sequence length. Furthermore, the following applies:

$$\sum_{i=1}^{m} P(w_i) = 1,$$

where $w_1, w_2, ..., w_m$ are all words that occur in the training corpus and

$$P(w_i) = \frac{c(w_i)}{\sum_{j=1}^{m} c(w_j)},$$

where $c(w_i)$ is the count of occurrences of word w_i in the training corpus.

The Markov assumption that each word is locally dependent on the previous word is called a *bigram* (2-gram) model:

$$P(S_n) \approx P(w_1) \prod_{i=2}^{n} P(w_i \mid w_{i-1}) \qquad ,$$

$$\sum_{i=1}^{n} P(w_j \mid w_i) = 1, \quad \forall w_j \in V,$$

where V is the vocabulary and

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{\sum_{j=1}^{n} c(w_{i-1}, w_j)}, \tag{2.12}$$

where

$$c(w_{i-1}, w_i), \tag{2.13}$$

is the count of occurences of the bigram w_{i-1}, w_i in the training corpus. (2.12) describes the maximum likelihood estimate (MLE) of a word w_i following a history w_{i-1} estimated from a training set. This estimation is done by counting the times w_i follow w_{i-1} and divide by the number of times the history (w_{i-1}) occurs. The reasoning for n-gram models of higher order (trigram, 4-gram, 5-gram etc.) is analogous to the above. (2.13) may be generalized to any order of n-grams by the short notation $c(w_{i-n+1}^{i})$.

Even though neural LMs have shown SOTA performance on various NLP tasks in the last years[1], n-grams are still widely used due to its simplicity.

2.4.4 N-gram smoothing

The Markov assumptions made for n-gram models reduces the problem with zero probability described in Section 2.4.3, but it is not an absolute solution. Assume a vocabulary of size $|V|$. Theoretically this means $|V|^2$ possible bigrams. Even though this number is generally limited by grammatical and syntactic reasons there is always a risk that unseen n-grams will occur as the model is utilized in practice. In fact, given any finite amount of text or speech data in a given language you should expect to see new words and n-grams in another sample of that language [20]. To avoid this overfitting of the training data, *n-gram smoothing* is introduced. The term smoothing is a collective term for techniques used to spread out the probability distribution over n-grams such that small probabilities (including zero) are adjusted upwards and large probabilities are adjusted downwards. A way of seeing this is that some probability mass from all seen n-grams is transferred to unseen ones. This lowering of non-zero probabilities is also referred to as *discounting*. Smoothing does not just shift the zero probabilities upwards but also attempt to improve the model at whole [15].

2.4.4.1 Additive smoothing

In additive smoothing [33] we pretend that we see each n-gram more often than we actually do. This is done by adding a factor α to each n-gram count. (2.14) show the probability of a trigram using additional smoothing:

$$P_{add}(w_i \mid w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, c_{i-1}, c_i) + \alpha}{c(w_{i-2}, w_{i-1}) + \alpha|V|} \tag{2.14}$$

where α is typically in $[0, 1]$ and $|V|$ is the size of the vocabulary. This technique has been shown not to work well in practice [19] but it is a good introductory method.

2.4.4.2 Good-Turing estimate

Good-Turing smoothing [23] is a method that builds upon the assumption that to appear zero times in a sample is not so different from appearing once in that sample. Thus we can use the count of things we've seen once to help estimate the count of things we've never seen. The Good-Turing estimate says that for any n-gram that occurs r times in the training data we should pretend that it occurred r^* times where

$$r^* = (r + 1)\frac{N_{r+1}}{N_r} \tag{2.15}$$

where N_r is the frequency of frequency r, i.e the number of n-grams that occurred r times in the training data. The estimated Good-Turing probability of such a n-gram

[1] https://github.com/syhw/wer$_a$re$_w$e

is then

$$P_{GT}(w_{i-n+1}^{i}) = \frac{r^*}{N}$$

where $N = \sum_{r=0}^{\infty} N_r r^*$.

As an example, considering the unigram counts of a training set consisting only of the sentence *The brown fox is quick and he is jumping over the lazy dog*:

Unigram	the	brown	fox	is	quick	and	he	jumping	over	lazy	dog
Count	2	1	1	2	1	1	1	1	1	1	1

we get the frequencies of frequencies $N_1 = 9$, $N_2 = 2$, $N_3 = 0$, ...

If we were to use maximum likelihood estimation, the probability mass assigned to

- *unseen words* is then $N_0 P_{MLE}(\text{unseen word}) = 0$
- *words that occur one time* is then $N_1 P_{MLE}(\text{word seen once}) = 9/13$
- *words that occur two times* is then $N_2 P_{MLE}(\text{word seen twice}) = 4/13$

If we were to use Good-Turing estimation, the probability masses of the above cases would instead be

- $N_0 P_{GT}(\text{unseen word}) = N_0 r^*/N = N_0(0+1)\frac{N_1}{N_0}/N = N_1/N = 9/13$
- $N_1 P_{GT}(\text{word seen once}) = N_1 r^*/N = N_1(1+1)\frac{N_2}{N_1}/N = 2N_2/N = 4/13$
- $N_2 P_{GT}(\text{word seen twice}) = N_2 r^*/N = N_2(2+1)\frac{N_3}{N_2}/N = [N_3 = 0] = 0$

Thus, Good Turing works in such a way that the MLE estimated probability mass for words of a certain frequency is moved to words of one step lower frequency. The probabilities from the example above is calculated from a very small toy dataset but it still highlights a major shortcoming of Good-Turing, namely that the probability mass assigned to unigrams of frequency $r = 3$ is zero. Without modifications, this will always be the case for the highest frequency which is explained by analyzing (2.15) and realize that $N_{r_{\max}+1} = 0$. The typical pattern for linguistic data is that the most common frequency of occurrence for a object is once, followed by twice, followed by three and so on. This corresponds to N_r gradually decreasing as r grows. The fact is that N_r is often sparse for large r, i.e there are many zero frequencies of frequencies and few non-zero ones. Again referring to (2.15) it is clear that such sparseness undermines the estimations. The solution for this is to smooth the N_r's, e.g by a best-fit power law for N_r as $r > k$ where k is some integer found e.g. by cross validation [20].

2.4.4.3 Interpolation

Assume the trigram w_{i-2}^{i} having zero counts but the corresponding bigram w_{i-1}^{i} and unigram w_i being present in the training set. It is then a reasonable assumption that information about the bigram and the unigram can be used to estimate the probability of the trigram. Combining different orders of n-grams by interpolating

them is a way to generalize from less context to learn about context not seen in the data [36]. The simplest form of interpolation is linear interpolation with the probability of a given trigram estimated as

$$P(w_i \mid w_{i-2}, w_{i-1}) = \lambda_1 P_{MLE}(w_i) + \lambda_2 P_{MLE}(w_i \mid w_{i-1}) + \lambda_3 P_{MLE}(w_i \mid w_{i-2}, w_{i-1}), \tag{2.16}$$

where $\sum_{i=1}^{3} \lambda_i = 1$ and the λ's are tuned to maximize the likelihood on a validation corpus [15, 34].

2.4.4.4 Backoff

In backoff, if a n-gram has zero counts, we approximate it with the counts of the (n-1)-gram. This is done by discounting the probability for each n-gram, which means that some proportion of the probability mass is relocated to the lower-order n-grams. In general the backoff approximation is calculated by the following scheme

$$P_{BO}(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} P^*(w_i \mid w_{i-n+1}^{i-1}) & \text{if} \quad c(w_i \mid w_{i-n+1}^{i-1}) > 0 \\ \alpha(w_{i-n+1}^{i-1}) P_{BO}(w_i \mid w_{i-1+2}^{i-1}) & \text{otherwise,} \end{cases} \tag{2.17}$$

where P^* is some discounted probability and $\alpha(w_{i-n+1}^{i-1})$ is the backoff weights typically used to make sure that probabilities are normalized. Using the scheme in (2.17) we back off recursively until reaching an n-gram with non-zero counts or alternatively a uniform distribution over unseen unigrams.

2.4.4.5 Stupid backoff

An exception from normalized probabilities is the *Stupid backoff* method with the scheme

$$S(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} P^*(w_i \mid w_{i-n+1}^{i-1}) & \text{if} \quad c(w_i \mid w_{i-n+1}^{i-1}) > 0 \\ \alpha S(w_{i-1} \mid w_{i-1+2}^{i-1}) & \text{otherwise,} \end{cases}$$

where P^* is the MLE, α is set to a fixed value and the $S(.)$-notation emphasizes that we are no longer dealing with a probability distribution but rather a score function. Stupid backoff is inexpensive to calculate and has shown to work very well with large amount of data [12].

2.4.4.6 Katz's backoff

Katz's backoff [38] is given by the following scheme:

$$P_{Katz's}(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} P^*(w_i \mid w_{i-n+1}^{i-1}) & \text{if} \quad c(w_i \mid w_{i-n+1}^{i-1}) > 0 \\ \alpha(w_{i-n+1}^{i-1}) P_{Katz}(w_{i-1} \mid w_{i-1+2}^{i-1}) & \text{otherwise,} \end{cases}$$

where P^* is the Good-Turing discounted probability and $\alpha(w_{i-n+1}^{i-1})$ is a normalizing constant deciding how the discounted probability mass of the current order of n-grams should be distributed among the zero count current order n-grams according to the next lower-order n-gram distribution [15].

In order to understand the logic behind α, it is convenient to define a function β as

$$\beta(w_{i-n+1}^{i-1}) = 1 - \sum_{w_i : c(w_{i-n+1}^{i}) > 0} P^*(w_i \mid w_{i-n+1}^{i-1}).$$

In words, β is the estimated conditional probability discounted from all words w_i for which $c(w_i \mid w_{i-n+1}^{i-1}) > 0$. This is the probability mass we want to distribute among the unseen n-grams, i.e. among w_i such that $c(w_i \mid w_{i-n+1}^{i-1}) = 0$, using the (n-1)-gram distribution $P^*(w_i \mid w_{i-n+2}^{i-1})$. Putting all of this together we arrive at

$$\alpha(w_{i-n+1}^{i-1}) = \frac{\beta(w_{i-n+1}^{i-1})}{\sum_{w_i : c(w_{i-n+1}^{i})=0} P^*(w_i \mid w_{i-n+2}^{i})}.$$

2.4.4.7 Absolute discounting

In Absolute discounting [46] higher- and lower-order n-grams are interpolated by subtracting a fixed discount d from each non-zero count

$$P_{ABS}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c(w_{i-n+1}^{i}) - d, 0)}{\sum_{w_i} c(w_{i-n+1}^{i})} + \lambda_{w_{i-n+1}^{i-1}} P_{ABS}(w_i \mid w_{i-n+2}^{i-1}), \quad (2.18)$$

where

$$\lambda_{w_{i-n+1}^{i-1}} = \frac{d}{\sum_{w_i} c(w_{i-n+1}^{i})} |w_i : c(w_{i-n+1}^{i}) > 0|$$

is the normalizing factor and $|w_i : c(w_{i-n+1}^{i}) > 0|$ is the number of unique words following the history w_{i-n+1}^{i-1} at least once. To determine d, the leaving-one-out method, an extension of cross validation where each word of the training corpus is systematically held out to simulate the effect of not being observed during training, is used arriving at

$$d = \frac{n_1}{n_1 + 2n_2}.$$

Here n_r is the number of n-grams (of the highest order) being observed r times in the training data. By comparing counts and Good-Turing counts it has been seen that $d = 0.75$ is a adequate discount factor [36].

2.4.4.8 Interpolated Kneser-Ney

Kneser-Ney smoothing [39] build upon Absolute discounting with the aim of reducing the bias that arises towards words that are highly conditioned on directly preceding words. An example of this addressed in [39] is the word *dollars* which is

very frequent in the *Wall Street Journal* corpus but it occurs almost only after numbers or sometimes after country names. Because of the high unigram probability assigned to *dollars*, the smoothing estimate of $P(x, dollars)$ will be unreasonably high if $P(dollars)$ is taken for backing off. Another common example of this phenomenon [15] is the word *Francisco* that almost always follow the word *San* due to the context of *San Francisco*. *San Francisco* is a common bigram in many corpora, and thus *Francisco* is a common unigram which results in $P(w, Francisco)$ being assigned a large probability for w's of totally different contexts.

Kneser-Ney smoothing counteracts this by modifying the $c(\bullet)$-terms in (2.18) such that

$$c_{KN}(\bullet) = \begin{cases} c_{KN}(\bullet) & \text{for the highest order n-gram} \\ cc_{KN}(\bullet) & \text{for the lower order n-grams} \end{cases} \tag{2.19}$$

where $c_{KN}(\bullet)$ is the regular $c(w_{i-n+1}^{i})$-term and $cc_{KN}(\bullet)$ is the number of unique contexts after which a n-gram occur [36], i.e.

$$\begin{aligned} cc_{KN}(w_{i-n+2}^{i}) &= |w_{i-n+1}^{i-1} : c(w_{i-n+1}^{i}) > 0| \\ \sum_{w_i} cc_{KN}(w_{i-n+2}^{i}) &= |(w_{i-n+1}^{i-1}, w_i) : c(w_{i-n+1}^{i}) > 0|. \end{aligned}$$

The recursive interpolated Kneser-Ney algorithm is then given as

$$P_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c(w_{i-n+1}^{i}) - d, 0)}{\sum_{w_i} c(w_{i-n+1}^{i})} + \lambda_{w_{i-n+1}^{i-1}} P_{KN}(w_i \mid w_{i-n+2}^{i-1}), \tag{2.20}$$

where the rules of (2.19) is applied to the first term of the RHS.

2.4.4.9 Modified Kneser-Ney

Chen and Goodman [15] introduces a variation of Kneser-Ney smoothing called modified Kneser-Ney smoothing. Unlike (2.20), they have three different parameters d_1, d_2 and d_{3+} where d_r is applied to n-grams with one count, two counts and three or more counts respectively. Modified Kneser-Ney is developed based on that the optimal discount for n-grams with one and two counts differ from that of n-grams with three or more counts.

The recursion scheme for modified Kneser-Ney is given as

$$P_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c(w_{i-n+1}^{i}) - d(c(w_{i-n+1}^{i}), 0)}{\sum_{w_i} c(w_{i-n+1}^{i})} + \lambda_{w_{i-n+1}^{i-1}} P_{KN}(w_i \mid w_{i-n+2}^{i-1}),$$

where

$$d(c(\bullet)) = \begin{cases} 0 & \text{if} \quad c(\bullet) = 0 \\ d_1 & \text{if} \quad c(\bullet) = 1 \\ d_2 & \text{if} \quad c(\bullet) = 2 \\ d_{3+} & \text{if} \quad c(\bullet) \geq 3. \end{cases}$$

Here $\lambda_{w_{i-n+1}^{i-1}}$ is modified such that probabilities are normalized.

2.5 Metrics

2.5.1 Word Error Rate

Word error rate (WER) is a common metric for evaluation of SST systems [45]. It is based on the Levenshtein distance [41] but applied to sequences of words instead of sequences of characters in strings. The basic idea is to arrange the resulting text with the ground truth and compare for substitutions, insertions and deletions in the resulting text where

- a *substitution* is a word that has been transcribed differently than the corresponding word in the ground truth, e.g. *roast* → *rose*,
- a *insertion* is a word that is added but not said in the ground truth, e.g. *it's me* → *it's a me*,
- a *deletion* is when a word is left out of the transcribed text, e.g. *say hi to the family for me* → *say to the family for me*.

The equation for calculating WER is

$$\text{WER} = \frac{S + D + I}{S + D + C}, \tag{2.21}$$

where C is the number of correct words and $S + D + C = N$ is equal to the number of words in the ground truth text. A error-free translation results in a WER of zero since $S = D = I = 0$ and $C = N$. Note that the WER in theory has no upper limit since I has no upper limit.

2.5.2 Character Error Rate

Character error rate (CER) is similar to WER but applied to characters instead of words. It uses the same equation as WER, i.e.

$$\text{CER} = \frac{S + D + I}{S + D + C}, \tag{2.22}$$

where $S + D + C$ is equal to the number of letters in the ground truth.

2.5.3 Manual evaluation

A significant difference between WER and CER is that WER interpret misspelled words as completely incorrect while CER is more forgiving in cases where only a few characters are inaccurate. Therefore, CER may be more in line with a human

ability to interpret the transcribed text while WER may be more consistent with a computers ability to interpret the results, e.g. by searching.

Low WER- and CER-values indicates a well-executed transcript, but it is important to note that neither WER- or CER-values tell anything about the quality of the transcript from a contextual perspective since it doesn't give any information about which words or characters that have been subject to substitutions, insertions and deletions. Therefore it is also a need for manual evaluation to get a qualitative judgment of the transcripts.

2.5.4 String similarity measures

As stated above, WER and CER provides information about the quantitative performance of the transcriptions but they don't tell us much about which words that are left out and what words are poorly transcribed. For the task of this thesis, there are a number of words that are in particularly important to detect such as *mayday* and *pan*. The use of string similarity provides the possibility of detecting keywords that has been misspelled or substituted into a similar word. This is done by drawing the operators attention not only to correctly spelled keywords but also to words that are within some string similarity threshold value from that keyword.

2.5.4.1 Hamming distance

Hamming distance is originally a metric used to compute the similarity between two binary string [29]. The Hamming distance is then the number of positions where the two bits are different, e.g. the binary strings 001 and 100 has a Hamming distance of 2. This can be directly applied to strings of different length as follow

$$\text{Hamming}(a,b) = \begin{cases} |a| & \text{if} \quad |b| = 0 \\ |b| & \text{if} \quad |a| = 0 \\ \text{Hamming}(\text{tail}(a), \text{tail}(b)) + 0 & \text{if} \quad a[0] = b[0] \\ \text{Hamming}(\text{tail}(a), \text{tail}(b)) + 1 & \text{if} \quad a[0] != b[0], \end{cases} \tag{2.23}$$

where $\text{tail}(a)$ is everything but the first character in a and $|a|$ is the length of a. In practice, each position of the strings a and b is examined from the left. If the positions match, we simply move on to the next position, otherwise we move on and add 1 to the distance. If we reach the end of any of the strings, the length of the remaining part of the other string is added to the distance.

2.5.4.2 Levenshtein distance

Compared to the Hamming distance in section 2.5.4.1, Levenshtein distance [41] is more flexible in that the distance between two strings is measured as the minimum number of substitutions, deletions and insertions needed to convert one of the word to the other regardless of the position in which the operations are performed

$$\text{Levenshtein}(a,b) = \begin{cases} |a| & \text{if} \quad |b| = 0 \\ |b| & \text{if} \quad |a| = 0 \\ \text{Levenshtein}(\text{tail}(a), \text{tail}(b)) & \text{if} \quad a[0] = b[0] \\ 1 + min \begin{cases} \text{Levenshtein}(\text{tail}(a), b) \\ \text{Levenshtein}(a, \text{tail}(b)) \\ \text{Levenshtein}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise} \end{cases} \tag{2.24}$$

e.g. the binary strings 001 and 0001 has a Levenshtein distance of 1. As an illustrative example of the difference between (2.23) and (2.24) we take $a =$ book and $b =$ sbook where $\text{Hamming}(a,b) = 5$ and $\text{Levenshtein}(a,b) = 1$ due to $\text{Levenshtein}(\text{tail}(\text{sbook}), \text{book}) = 0$. A drawback of Levenshtein is the higher computational complexity due to deep recursive calls.

3

Datasets

This chapter describes the data used during training of language models and fine-tuning of STT-models. The primary source of data is the availability of large amounts of VHF transmissions from the JRCC. For this thesis, however, the amount of useful data from this dataset is limited to the transmissions that have been transcribed and then mainly to the transmission that have been completely transcribed. The meaning of this will be clear in section 3.1. Furthermore, external data has been collected for the training of language models. This is described in more detail in section 3.2.

3.1 JRCC data

JRCC has provided VHF data recorded at the Swedish coastal radio network [1] from December 2019 to January 2021. In total, there are 2.7 million recordings, which correspond to approximately 11,000 hours of sound. The audio is stored in WAV-format sampled at 16 kHz with one channel (mono) and 16 bits per sample. The dataset contain some duplicates of the same messages since each message can be perceived by several masts, but the quality of these duplicates may differ e.g. depending on the distance from the sender to the perceiving masts.

3.1.1 JRCC transcriptions

The raw audio is in theory useful for pre-training wav2vec2.0 but when this was tried in [21], this resulted in below average performance. As shown in [21], however, there is a great potential in using the data to fine-tune a pre-trained wav2vec2.0-model. For this to work, requirements are placed on the availability of transcripts as these are used by the CTC algorithm to map sound-representations to the vocabulary.

This transcription takes place through an interface where the user listens to an audio file, transcribes the message and labels the audio file with meta information of interest. The transcription is done according to the following protocol:

Protocol 1 JRCC audio data transcription scheme

Inputs. JRCC audio data

Goal. To generate transcripts of the audio data and to label the files with information that describes the type of message and quality of audio
The protocol:

1. **Setup.**
   (a) Transcribe what is heard as verbatim as possible. The focus should be on how the words are said and not on how you as a human being interpret them.
   (b) Numbers must be transcribed in words, i.e. "*sixteen*" and not "*16*".
   (c) No special signs. If someone says "*dollar*", write "*dollar*" and not "*$*". Another example: write "*e*" and not "*é*".
   (d) Don't use abbreviations if they are not explicitly stated.
   (e) If you cannot understand a word being said or a part of a sentence, replace this with a question mark. Therefore, do not use question marks at the end of sentences that are questions.
   (f) To distinguish between different speakers, you can use punctuation or line breaks.
   (g) If you write words you think you hear but are unsure, tag the message as **unsure**.
   (h) If you cannot distinguish any speech at all in the audio, mark the message as **only noise**.
   (i) For messages that you want to return to at a later time, mark these as **TODO**
   (j) For all messages where *Sweden Rescue*, *JRCC*, *Sjöräddningen* etc. is mentioned, mark these as **JRCC**.
   (k) For all types of emergency calls, mark these messages as **Mayday**.
   (l) Mark messages with the language/s spoken.
   (m) A large part of the messages are general broadcasts. These often start with 'all ships all ships' and apply to weather reports, traffic information, etc. Since a large number of this type of message is already transcribed, you do not need to transcribe new instances. However, all these messages should be marked as **Broadcast**.
   (n) If the audio contains communication between different ships, mark message as **Ship2ship**.
   (o) Rate the sound quality from 1-5 where 1 means completely inaudible and 5 means very good quality.

The finished transcription files then contain the following information:
call id - The id number of the call
comment - A comment about the message (optional)
date - The date of the message
langs - The language/s spoken in the message

level - The audio quality of the message
tags - Describes what kind of message it is. E.g. **Broadcast**, **Mayday**, etc.
transcript - The transcribed text
updated - The date and time of transcription
username - Who did the transcription
.

3.1.2 JRCC data analysis

As of 2022-05-05, 4,605 messages had been transcribed. An overview of how tags are distributed between these messages is shown in Figure 3.1. Note that each message can be assigned multiple tags. Interesting observations are the large number of messages marked with **broadcast**, **unsure** and **noise** as well as the low number of emergency calls and messages directed to JRCC. It is also clear that a predominant part of messages are in English with a smaller proportion of messages in Swedish.

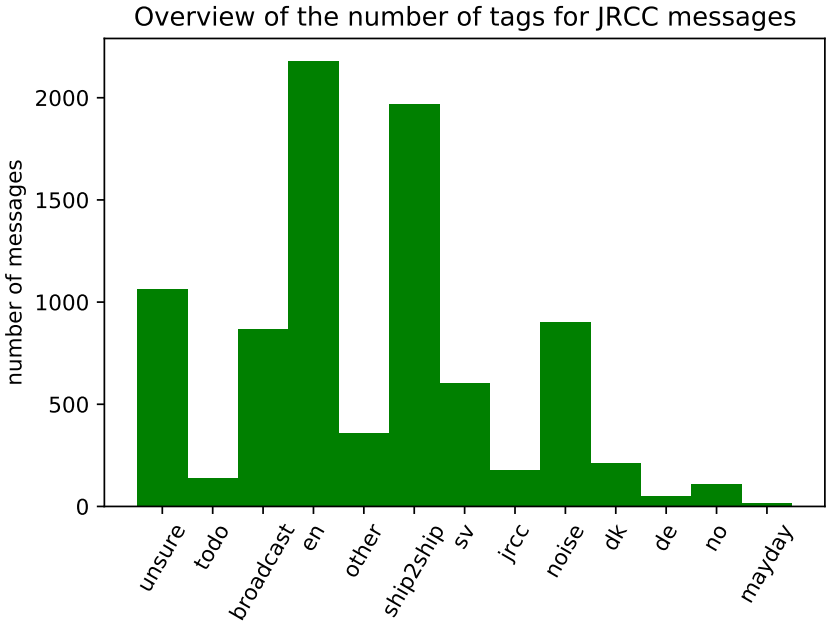

Figure 3.1: The number of occurrences of all tags among the transcribed JRCC-messages.

After removing duplicates and transcriptions where the set of tokens in the transcription equals to "?", we are left with 463 messages in Swedish and 1808 messages in English. The Swedish transcriptions then consists of 6556 words where the number of unique words is 1595. The English transcriptions consists of 33085 words where the number of unique words is 3381. Figures 3.2 and 3.3 show the rank vs. frequency distributions where it can be seen that a small number of words takes up a large part of the word distribution. In Figure 3.4 it can be seen that this agrees reasonably well with Zipf's law[1]. This concordance increases in theory with the size of the corpus. From analyzing the most common words in each language, the

[1]https://en.wikipedia.org/wiki/Zipf%27s_law

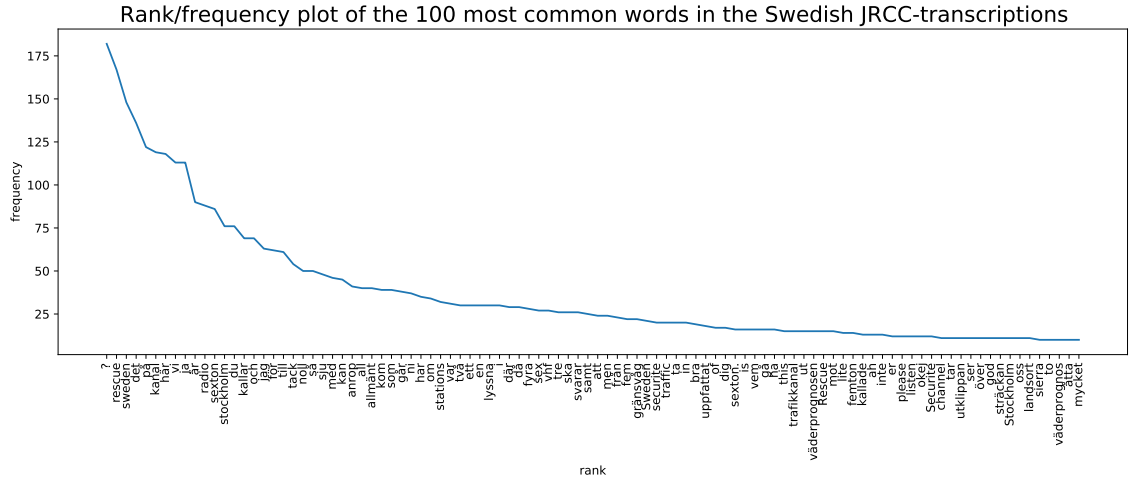

Figure 3.2: Rank-frequency plot of the 100 most common words in the Swedish JRCC-transcriptions.

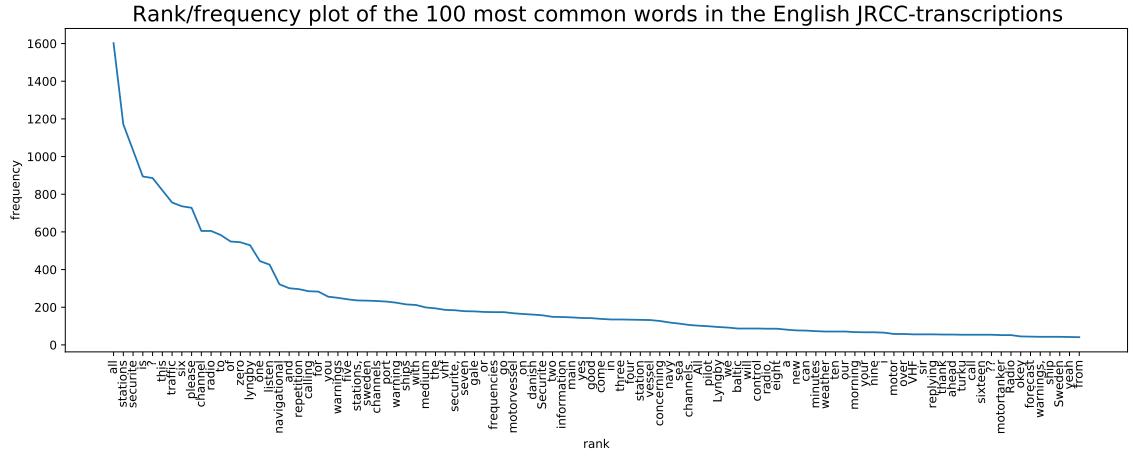

Figure 3.3: Rank-frequency plot of the 100 most common words in the English JRCC-transcriptions.

Swedish transcriptions is dominated by the terms *Sweden rescue*, *Stockholm radio* and *kanal sexton* while the English transcriptions largely consists of broadcasts.
In total there are 16 emergency calls in the transcribed dataset where 2 are in Swedish, 9 in English and 5 in other languages, more particularly in Norwegian and Danish. The majority of these messages are responses to emergency calls sent from ground station and rescue boats.

The transcribed dataset has two uses in the creating of a ASR-model. Firstly, the audio files with corresponding transcriptions are used as training data to fine-tune wav2vec2.0 and secondly, the transcripts by themselves are used as training data in building language models to improve performance. Depending on the task, the data must be seen as useful to different degrees. When it comes to fine-tuning a model, it is of utmost importance that the transcript matches the sound since CTC is trained to map a given sequence (the transcript) to the corresponding audio and a deficient

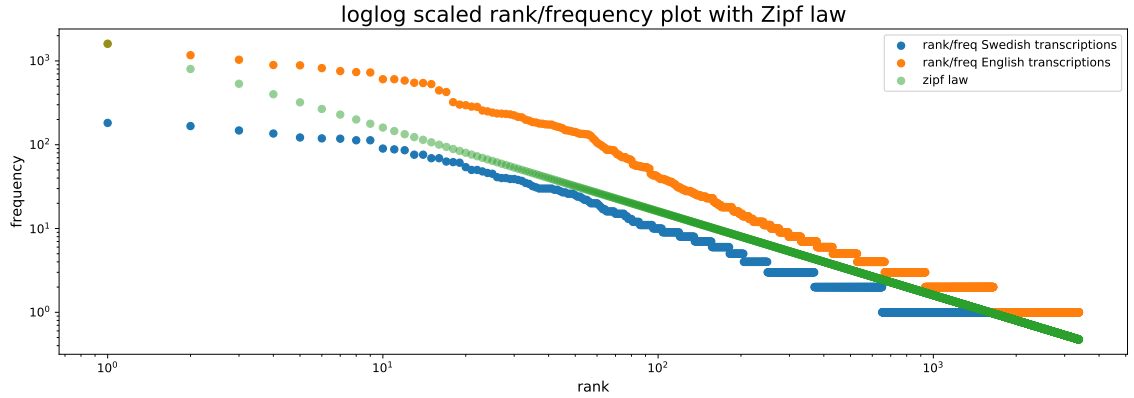

Figure 3.4: Loglog rank/frequency plot of words in Swedish and English transcriptions plotted together with Zipf's law.

transcription would thus bring disorder to this process. Figure 3.5 gives an overview of the completeness of the transcripts where the following types of transcript have been judged to be incomplete:

- transcripts marked as **unsure**
- transcripts marked as **TODO**
- transcripts containing question marks, i.e. where part of the messages is inaudible/uninterpreted according to the transcriber
- transcripts corresponding to audio clips only consisting of noise
- empty transcripts

.

As seen from Figure 3.5 less than half of all transcribed messages are considered as complete.

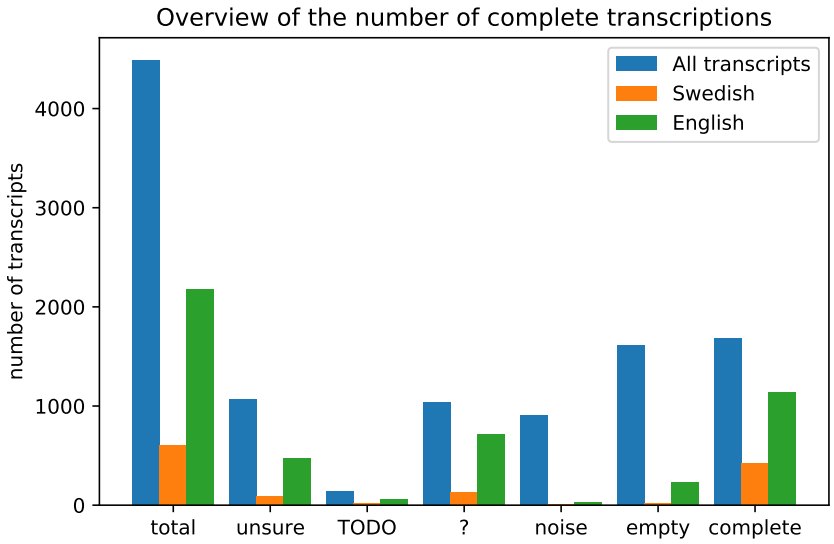

Figure 3.5: Overview of the usefulness of messages to fine-tune a model with CTC.

When it comes to building a language model, flawed transcripts do not play as big a role. As long as a word is perceived as reasonable by the transcriber, it is conceivable

that this word could occur in the same context even if the actual transcription is incorrect given the associated audio file. A more detailed description of how transcripts were processed in the creation of language models is given in Section 3.2.

3.1.3 Curated JRCC data

A curated subset of the JRCC data was extracted according to section 4.5. The details are shown in Table 3.1.

Number of messages	Length in minutes	Number of words	Number of characters
Swedish			
349	38.42	4342	20040
English			
795	112.4	12080	61171

Table 3.1: Quantitative descriptions of the curated Swedish and English datasets.

3.1.4 Test set

In addition to the training data, two test sets in Swedish and English have been transcribed by the author to ensure high quality in the final evaluation. Care has been taken to extract test sets that are varied across the message types perceived during the work on the thesis. Some more details are presented in Table 3.2.

Number of messages	Length in minutes	Number of words	Number of characters
Swedish			
20	3.52	371	1765
English			
22	4.85	541	2552

Table 3.2: Quantitative descriptions of the Swedish and English test sets.

3.2 Language model data

A number of text sources were used to build the corpora for training language models. The framework for most corpora is the JRCC-transcripts, but these have been augmented with texts that may more or less contribute to language knowledge of the current domain.

3.2.1 JRCC text data

Table 3.3 show the number of words and number of unique words in the resulting JRCC texts used as language model training data.

Language	Number of words	Number of unique words
Swedish	6497	1360
English	33447	2805

Table 3.3: The number of words and unique words in the Swedish and English JRCC texts respectively.

3.2.2 Vessel names

The vessel names used are collected from the Vesselfinder[2] database. Two sets of vessels has been created. The first consists of Swedish vessel names only where the addition of new vessel names was limited with respect to the fact that the most recently registered longitude and latitude would be within a box enclosing the Swedish coast and its vicinity. This first set consists of 3206 vessel names and is augmented on the form $vessel\_name_i$ $vessel\_name_i$ $vessel\_name_{random}$ for each $i \in \{1, ..., 3206\}$. The second set consist of all vessel names extracted from Vesselfinder which results in 154,265 vessel names after removing duplicates and those consisting only of digits.

3.2.3 NATO phonetic alphabet (Swedish)

The NATO phonetic alphabet (*Alfa*, *Bravo*, *Charlie*, *Delta*, *Echo*, ...) is the most commonly used radio telephone spelling alphabet [66] and is used in the current domain mainly to clarify letter combinations in vessel names and the International Maritime Organization (IMO) number used to identify registered ships. 100 sentences are generated consisting of 50 phonetic alphabet code words sampled from a uniformly random distribution.

3.2.4 Sjörapporten (Swedish)

Sjörapporten is the magazine of the Swedish Maritime Administration [2]. The text of 22 editions published between 2017-2022 is extracted from pdf's into text files. The cleaning of text is done manually and rule-based to extract body-text only resulting in 242.554 words in total with 29.161 unique words.

3.2.5 Maringuiden (Swedish)

The maringuiden dataset is text collected from Maringuidens online discussion forum. The texts are taken from the sections *Motorsnack*, *Seglarsnack* and *Långfärds-*

[2]https://www.vesselfinder.com/

snack which in English translates to *Motor talk*, *Sailing talk* and *Long-distance talk*. The processed texts consist of 13,384,642 words with 261,170 unique words.

3.2.6 Wikipedia (Swedish)

A Swedish wikipedia dump is used to investigate the effect of a general language corpus. The text is preprocessed by the removal of punctuation marks, numerals being replaced with words and the text being lower cased. The text contains 367.399.551 words with 4.583.549 unique words.

3.2.7 Librispeech (English)

A off-the-shelf language model trained on the Librispeech corpus [47] is used as a general language base-line. Librispeech is a ASR dataset of read English speech from public domain audio books.

3.2.8 ATCO2 (English)

The ATCO2 (Air Traffic Communication) dataset was build for development and evaluation of automatic speech recognizer techniques for English air traffic control data[3]. It consists to large part of NATO phonetic code words. The total number of words in the extracted texts is 10,807 with 765 unique words.

3.2.9 Switchboard (English)

The switchboard dataset [22] is a telephone speech corpus of about 2500 conversations by 500 speakers from around the US. The extracted texts consists of 1,442,441 words with 21,430 unique words.

3.2.10 Silicone (English)

The Silicone corpus [14] consists of 10 datasets of conversational speech. The extracted texts consists of 3,730,365 words with 40,323 unique words.

[3]https://www.atco2.org/data

4
Methods

This chapter describes the methodology used in the thesis.

4.1 Baseline wav2vec2.0 models

As baseline wav2vec2.0 models, we will use the English and Swedish models that are currently used in project Heimdall. These will henceforth be referred to as *en baseline* and *sv baseline*. Both of these are 317M parameter wav2vec2.0 models[1] pre-trained on the VoxPopuli corpus [63] and fine-tuned on JRCC data in the corresponding language.

4.2 Language models

All language models are built using `kenLM`[2] which is a fast and scalable implementation of the modified Kneser-Ney smoothing explained in Section 2.4.4.9. The `kenLM` language models are then supported by the `pyctcdecode`[3] library which in turn is integrated with the wav2vec2.0 decoder[4] in the Hugging Face `transformers` library. Because the linguistic knowledge of language model is entirely based on the text which it is trained on and the JRCC messages mainly consist of conversations within the maritime domain, text is searched for within the maritime- and conversational domains. The most important resource is the JRCC transcripts, but we are also exploring the possibility of augmenting these with texts from other sources. A more detailed description of the datasets used is to be found in Section 3.2

4.2.1 Normalizing of JRCC-transcripts

As mentioned in Section 3.1.2 it is not necessary with the same strict measures for the data when it is to be used for language models as when it is to be used to fine-tune wav2vec2.0. Of course, it would have been desirable for all transcripts to have been completely translated for the language models as well, but given the limited amount of data, we want to make the most of it and thus add all non-empty transcripts to the language model corpus. An initial review of the JRCC transcripts showed that a normalization of these was required. The protocol given in Section

[1] https://huggingface.co/facebook/wav2vec2-large-100k-voxpopuli
[2] https://github.com/kpu/kenlm
[3] https://github.com/kensho-technologies/pyctcdecode
[4] https://huggingface.co/docs/transformers/model_doc/wav2vec2.0

3.1.1 had not always been completely followed, mainly due to numbers typed as numerals. This normalization was performed by a rule-based algorithm to ensure that all protocol points that could be checked directly from the transcript were met. For the language model corpus, the JRCC transcripts were then preprocessed as follow:

- All punctuation marks but question marks and apostrophes (for the English texts) are removed
- Repeated question marks are collapsed to one question mark
- Question marks are replaced by line breaks such that a sequence containing question marks are split into multiple ones where all of them consist of running text
- The text is lower cased
- Numerals are transformed into words, e.g. 16 $\longrightarrow$ *sixteen*

.

4.2.2 Normalizing of texts from external sources

For the texts from external sources, we modify the preprocessing of Section 4.2.1 such that all punctuation marks but apostrophes (for the English texts) are removed. In addition, some of the text from external sources required further actions presented in Table 4.1.

Source	Actions in addition to Section 4.2.1
NATO phonetic alphabet	Text generation
Sjörapporten	Parsing and filtering of PDF documents
Wikipedia	None
Librispeech	None (off-the-shelf model)
Atco2	None
Switchboard	Removal of comments made by transcribers, removal/normalization of interjections
Silicone	Removal/normalizing of interjections, removal of behavior descriptive notes, normalized use of apostrophes such as *don ' t* $\rightarrow$ *don't*

Table 4.1: Short descriptions of the measures taken (in addition to the modification of those in Section 4.2.1) to normalize the texts from external sources.

4.2.3 Combining texts in language model corpora

Various combinations of the above mentioned texts were used as language model training corpus with the aim of creating a enhanced language model in comparison with just using the JRCC-texts. The idea with using text from additional sources is to capture linguistic features of the naval-communication domain that has not been seen in the JRCC-texts. The combinations of texts used is presented in Table 4.2.

	Swedish
1	Swedish JRCC
2	1 + NATO phonetic alphabet
3	2 + augmented vessels
4	2 + Maringuiden + all vessels
5	3 + Sjörapporten
6	5 + Wikipedia
	English
1	Librispeech
2	English JRCC
3	2 + atco2
4	3 + switchboard
5	3 + Silicone

Table 4.2: Combinations of texts used as language model training corpora.

4.2.4 Hyper parameters

As explained in Section 2.3.1.1 the CTC-decoder combined with a language model has three hyperparameters; the beam width k, the language model parameter α and the word insertion parameter β. For our experiments, these parameters are tuned on a evaluation set using grid search.

4.3 Modifying the logits

The influence of the language model on the final transcripts is mainly affected by the settings of the decoder's hyperparameters. In cases where the acoustic model strongly predicts an incorrect vocabulary item and at the same time very weakly predicts the correct vocabulary item it was noticed that the language model did not manage to guide the paths correctly without changing the value of the parameters so that other parts of the transcript were destroyed instead. In an attempt to facilitate the influence of the language model without changing the value of the hyperparameters, more paths in the logits matrix (see 2.3.1.1) were promoted through two naive approaches. First, the top five vocabulary logits corresponding to each time step was set to a fixed positive value while the other logits were set to zero probability. Second, the top five vocabulary logits corresponding to each time step was set to a positive value based on a normalizing factor and the rest were set to zero probability.

4.4 Visualization of the logits

The logit matrices are visualized by plotting them as a heat map with time on the x-axis and vocabulary items on the y-axis. The logits with the highest values are then highlighted to clarify the path choosen by greedy decoding. We also mark the range for each transcribed word together with the accumulated logit-scored in order

to simplify the analysis of where and when the sound model works well and less well. A similar visualization is done for the case with using beam search and a language model. We then add the LM-scores to the visualization but omits the highlighting of the path.

4.5 Curation of JRCC transcripts

In a preliminary evaluation, it was discovered that several of the transcripts that were assumed to be complete were in fact defective. The inaccuracies ranged from misspelled words to the fact that large pieces of spoken sound were missing in the transcripts. To ensure correct transcriptions the messages corresponding to all potentially complete transcripts were listened to and inaccuracies were corrected. Furthermore, a large number of broadcasts were filtered out, mainly from the English transcripts, to reduce the skewness in the distribution. The curation was done as follow:

- All messages with the potential of being fully transcribed were extracted
- All these messages were listened to, duplicates (i.e forecasts/broadcasts by the same announcer and with the same content) were removed, deficient transcriptions were supplemented and audio files considered without potential of being fully transcribed were removed e.g. where half of the message is in some unknown language or where the audio quality is to low to perceive what is said without filling in gaps based on linguistic intuition
- Furthermore, these transcripts were normalized according to 4.2.1

This curation resulted in that the 1649 original messages were reduced to 349 Swedish and 795 English messages. 100 transcriptions were corrected in some way, 15 Swedish and 375 English broadcasts were removed, 32 messages were considered untranscriptable. The resulting dataset contain 30 Swedish and 129 English broadcasts where the reasons for the higher number of kept broadcast in English are a slightly higher variation in broadcast types and spoken language dialects as well as a overall larger number of English messages.

4.6 Fine-tuning wav2vec2.0 on curated dataset

To evaluate the effect of curating the JRCC-datasets these was used to fine-tune a number of pre-trained wav2vec2.0 models. All pre-trained models used are provided by the Hugging Face transformers library and described in Table 4.3 where it should be noted that the Switchboard- and Fischer datasets differ from the rest in that they consist of noisy telephone communication.

model name	pre-trained on	# parameters
	Swedish	
xls-r-300m[5]	436k hours of unlabeled speech, including VoxPopuli [63], MLS [50], Common-Voice [5], BABEL [53], and VoxLingua107 [60]	317M
large-100k-voxpopuli[6]	100k hours unlabeled subset of Voxpopuli	317M
	English	
base-100k-voxpopuli[7]	100k hours unlabeled subset of Voxpopuli	93M
large-100k-voxpopuli[8]	100k hours unlabeled subset of Voxpopuli	317M
large-robust[9]	60k hours Librilight [37], 2.2k hours CommonVoice, 100 hours Switchboard[10], 2k hours Fischer[11]	317M

Table 4.3: Descriptions of the pre-trained wav2vec2.0 models which were fine-tuned on the curated JRCC-dataset during the thesis.

As explained in Section 2.3.2 the fine-tuning is done using the CTC-loss by sending the contextualization module outputs trough a linear layer with a softmax activation function to obtain a probability distribution across the vocabulary for each time step. The vocabulary for the English dataset contains all standard letters (not including å,ä,ö) as well as the *"PAD"*, *"UNK"*, *"blank space"* and *"'"* -tokens resulting in a vocabulary of length 30, where the *"PAD"*-token is the special token used for collapsing sequences of the same vocabulary item as explained in 2.3.1.1. The vocabulary for the Swedish dataset contains all standard letter as well as the *"PAD"*, *"UNK"* and *"blank space"* -tokens resulting in a vocabulary of length 32. The convolutional feature-extraction layer weights are frozen during fine-tuning since these according to [7] are sufficiently trained during pre-training and thus does not need to be updated further. The fine tuning is entirely based on a notebook[12] shared by Hugging Face. For both languages, 10% of the training data is used for validation. Data points where the audio is longer than 30 seconds are ignored because of training time considerations. Special characters such as *"é"* are replaced by their

[5]https://huggingface.co/facebook/wav2vec2-xls-r-300m
[6]/wav2vec2-large-100k-voxpopuli
[7]https://huggingface.co/facebook/wav2vec2-base-100k-voxpopuli
[8]/wav2vec2-large-100k-voxpopuli
[9]facebook/wav2vec2-large-robust
[10]https://catalog.ldc.upenn.edu/LDC97S62
[11]https://catalog.ldc.upenn.edu/LDC2004T19
[12]https://huggingface.co/blog/fine-tune-wav2vec2-english

equivalent in the vocabulary such that *é* → *e* and *å, ä, ö* → *a, a, o* in the case of the English dataset. The training proceeded until no further improvement was made to the WER with respect to the validation data during the last 5 updates. The checkpoint with the lowest WER was then chosen to be the final model. We use an effective train batch size of 8 (train batch size of 2 and gradient accumulation of 4), a learning rate of 0.0001 and 300 warm-up steps (i.e. a linear increase of the learning rate from 0 to 0.0001 during the 300 first update steps). Evaluation and checkpoint storing is done every hundred update step. Henceforth the fine-tuned models will be refered to as (from top to bottom in Table 4.3) *sv xls-r*, *sv voxpop*, *en voxpop base*, *en voxpop large* and *en robust*.

4.7 Auto-correction with and without language model

As a naive baseline approach, auto-correction is applied to the transcripts obtained from using the wav2vec2.0 models from Section 4.1 and 4.5 without a language model in the inference. The `Symspellpy`[13]- and `Norvig`[14] algorithms were utilized to generate candidate substitutions for each word in the transcripts and then the candidate with highest probability given a unigram language model trained on the best performing corpus from Section 4.2.3 were chosen as a replacement. Both `Symspell` and `Norvig` generate candidates based on Levenshtein distance, `Symspell` also offers the possibility to split compounds. As an additional step we use the best performing 3-gram language model from 4.2.3 to decide the best candidate, i.e. we experiment with deciding the best candidate based on words frequency over the whole corpus and on words frequency given the context.

4.8 Evaluation

Here, the methods by which the performance of the different models are evaluated is presented.

4.8.1 WER and CER

The most natural method of evaluation is to use the WER and CER scores. To compensate for the relatively small test set we use the bootstrap method where one bootstrap sample is obtained by sampling n samples from the test set with replacement. The result is then reported as the mean and standard deviation of m bootstrap samples. Regarding the choice of parameter values we let $n = 10000$ and m to be the number of samples in the test set which is 22 for the English and 20 for the Swedish.

[13]https://github.com/mammothb/symspellpy
[14]http://norvig.com/spell-correct.html

4.8.2 Manual review of transcripts

As later discussed in Section 6.1 the WER and CER is not always a true measure of the contextual understanding given by the transcript. Therefore, the transcripts will also be reviewed manually to ensure that the results are not misleading.

4.8.3 Keyword detection

The current use of the application is to detect calls that contain specific keywords and report these to the JRCC operators. It is therefore desired that a generally better performing model, which is the aim of this thesis, will also lead to better recognition of keywords. A simple method will be used to evaluate the keyword recognition and only two keywords will be examined; *rescue* and *mayday*. First extract all transcribed messages that contain *rescue* or *mayday*, then remove all of these messages that is part of any of the training sets. For both keywords then generate all words within a Levenshtein distance of 1 and 2. Pass a sliding window over the corresponding transcriptions inferred by wav2vec2.0 models and count the number of times there is a match. The width of the sliding window depends on the Levenshtein distance used. With a Levenshtein distance of 1, the width of the sliding window equals to the length of the keyword + 1. Similarly with a Levenshtein distance of 2, the width of the sliding window equals to the length of the keyword + 2. In order not to count the same appearance of a keyword multiple times, we raise a flag for the next n number of steps where n depends on the Levenshtein distance and a new count cannot be made before the flag is lowered.

For each model evaluated on keyword detection, the evaluation will be done with the model in three different settings. First, greedy decoding will be used to generate transcripts. Second, the best performing language model of the corresponding language will be added to the decoding. Third, in addition to a language model hotword boosting will be used. Hotword boosting is a feature of *pyctcdecoding*[15] where the decoder is encouraged to look for specific hotwords in interference.

[15]https://github.com/kensho-technologies/pyctcdecode

5

Results

The aim of this thesis is to improve the performance of a STT system applied to VHF-messages in the maritime domain compared to the results in [21]. This section includes the results from evaluating models and language models developed according to Chapter 4. Section 5.1 presents the results from including n-gram language models trained on various corpora in the decoding process. Section 5.2 then presents visualizations of the logit matrices which are the basis of the CTC-decoding. Section 5.3 presents the results from using models fine-tuned on the curated dataset described in Section 4.5. Section 5.4 presents the results from applying auto-correction to transcripts decoded without a language model and Section 5.5 presents the results from a basic keyword detection experiment using various models. The resulting transcripts from all experiments are to be found in Appendices A.1 and A.2 for the Swedish and English messages respectively.

5.1 Combining Wav2vec2.0 with n-gram language model

Tables 5.1 and 5.2 summarizes the results from using n-gram language models in the CTC-decoding of wav2vec2.0 output logit matrices.

For the case of the Swedish model shown in Table 5.1 all experiments have been done using the *sv baseline*-model. The abbreviations in the language model column refers to the text on which the language model has been trained where J is the JRCC transcriptions, N is the NATO phonetic alphabet code words, V is the Swedish vessel names, Va is the full list of vessel names, S is Sjörapporten, M is text from the Maringuiden forum and W is the Swedish Wikipedia dump. Results are presented in terms of the mean and standard deviation of WER- and CER scores obtained from bootstrap-resampling over the test set. According to the experiments, the best results are obtained with a 3-gram language model trained on JRCC, NATO phonetic alphabet and Swedish vessel names.

Swedish		
LM	WER	CER
None	0.594 ± 0.054	0.237 ± 0.032
2-gram JNV	0.487 ± 0.060	0.224 ± 0.037
2-gram JNVS	0.504 ± 0.060	0.223 ± 0.037
2-gram JNVSW	0.511 ± 0.062	0.230 ± 0.040
3-gram J	0.499 ± 0.062	**0.222 ± 0.038**
3-gram JNV	**0.483 ± 0.062**	0.222 ± 0.039
3-gram JNVS	0.514 ± 0.062	0.229 ± 0.037
3-gram JNVSW	0.511 ± 0.065	0.229 ± 0.040
3-gram JNVaM	0.507 ± 0.064	0.225 ± 0.004
4-gram J	0.497 ± 0.062	0.224 ± 0.038
4-gram JNV	0.488 ± 0.063	0.226 ± 0.039
4-gram JNVS	0.503 ± 0.064	0.226 ± 0.039
4-gram JNVSW	0.515 ± 0.067	0.231 ± 0.041

Table 5.1: Resulting mean and standard deviation obtained by bootstrap-resamples of WER- and CER scores from evaluating the original Swedish model using a language model in the ctc-decoding. All language models are based on *J*, the JRCC-transcriptions. *N* is randomly sampled sequences of the NATO phonetic alphabet code words, *V* is Swedish vessel names and *Va* is the full list of vessel names. *S* is the text from 22 editions of *Sjörapporten*, *M* is text from Maringuiden forum and *W* the text from a Swedish Wikipedia dump.

For the case of the English model shown in Table 5.2 all experiments have been done using the *en baseline*-model. The abbreviations in the language model column refers to the text on which the language model has been trained where *L* is the off-the-shelf Librispeech n-gram model, *J* is the JRCC transcripts, *A* is the *ATCO2* transcripts, *Sw* is the transcripts from the *Switchboard* corpus, *Si* is the transcripts from the *Silicone* corpus and *V* is the full set of vessel names. Results are presented in terms of the mean and standard deviation of WER- and CER scores obtained from bootstrap-resampling over the test set. According to the experiments, the best results are obtained with a 3-gram language model trained on *JRCC*, *ATCO2* and *Switchboard*.

English		
LM	WER	CER
None	0.443 ± 0.050	0.192 ± 0.030
3-gram L	0.510 ± 0.059	0.206 ± 0.032
3-gram Si	0.406 ± 0.047	0.183 ± 0.028
3-gram J	0.401 ± 0.047	0.187 ± 0.029
3-gram JA	0.396 ± 0.044	0.185 ± 0.028
3-gram JASw	**0.391 ± 0.042**	**0.180 ± 0.027**
3-gram JASwV	0.396 ± 0.048	0.182 ± 0.027
3-gram JASi	0.401 ± 0.047	0.181 ± 0.027

Table 5.2: Resulting mean and standard deviation obtained by bootstrap-resamples of WER- and CER scores from evaluating the original English model using a language model in the decoding. All models are trained on different variations of text corpora where L is the off-the-shelf Librispeech n-gram model, J is the JRCC transcripts, A is the *ATCO2* transcripts, *Sw* is the transcripts from the *Switchboard* corpus, *Si* is the transcripts from the *Silicone corpus* and V is the full set of vessel names.

5.2 Visualization of Logits

Figures 5.1-5.4 show visualizations of the logit matrices for the JRCC message with the ground truth transcript *Charlotta b Charlotta b east coast pilot channel one six*. In 5.1 and 5.2 the logit matrices are obtained from the *en baseline* and *en robust* model respectively. The figures highlights the best path obtained from greedy decoding, the spans of each transcribed word and the resulting transcript. Figures 5.3 and 5.4 visualizes the same logit matrices but with the transcript obtained from using a language model in the decoding. The logit- and language model scores for each transcribed word is included which show that the misspelled vessel name is awarded a particularly low score in both configurations. Looking at the logit values it seems as *en baseline* is more confident in its predictions whilst *en robust* distributes its logit scores over a larger number of vocabulary items.

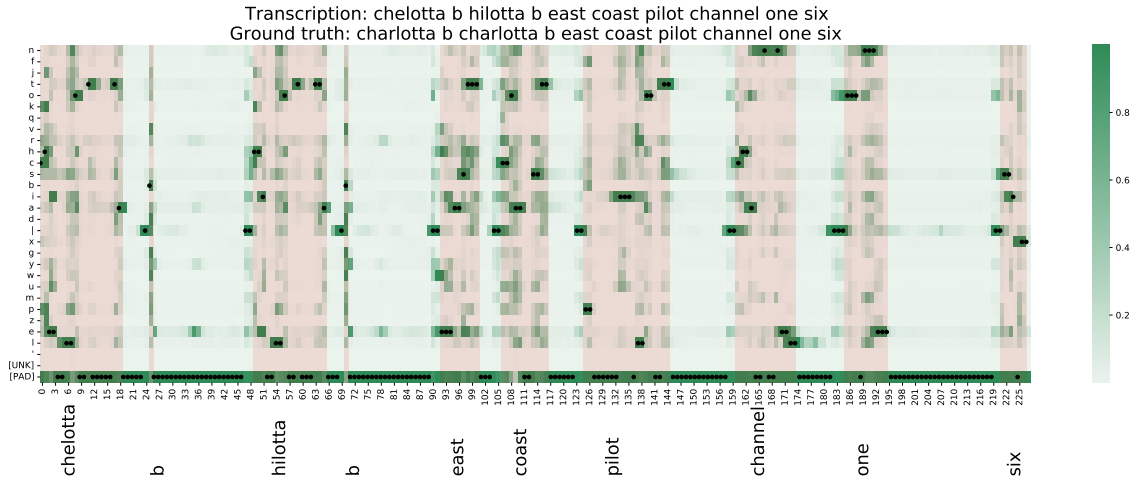

Figure 5.1: Visualization of the resulting logit matrix corresponding to a JRCC message using the *en baseline*-model. The best path from using greedy decoding is marked with black dots. Areas in red highlights spans containing transcribed words.

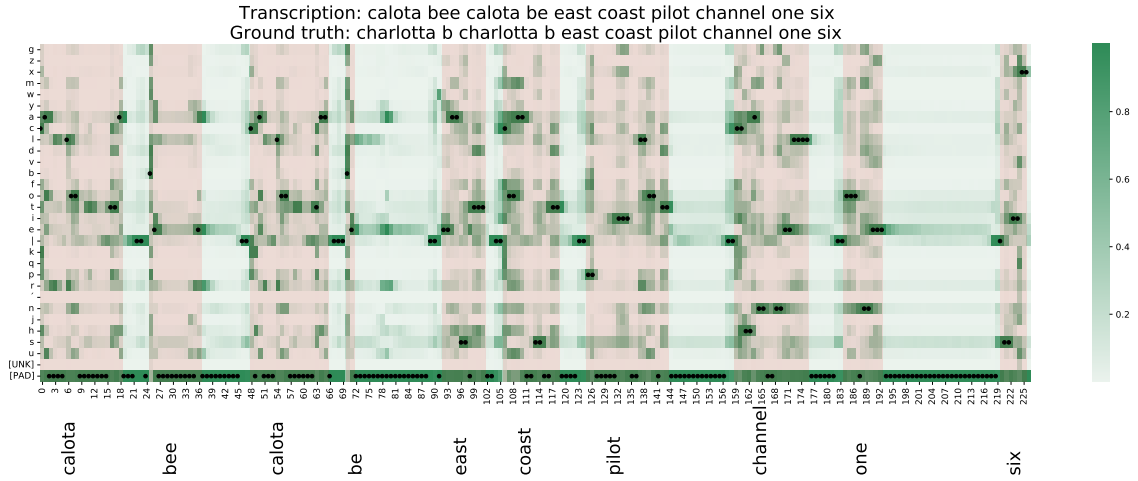

Figure 5.2: Visualization of the resulting logit matrix corresponding to a JRCC message using the *en robust* model. The best path from using greedy decoding is marked with black dots. Areas in red highlights spans containing transcribed words.

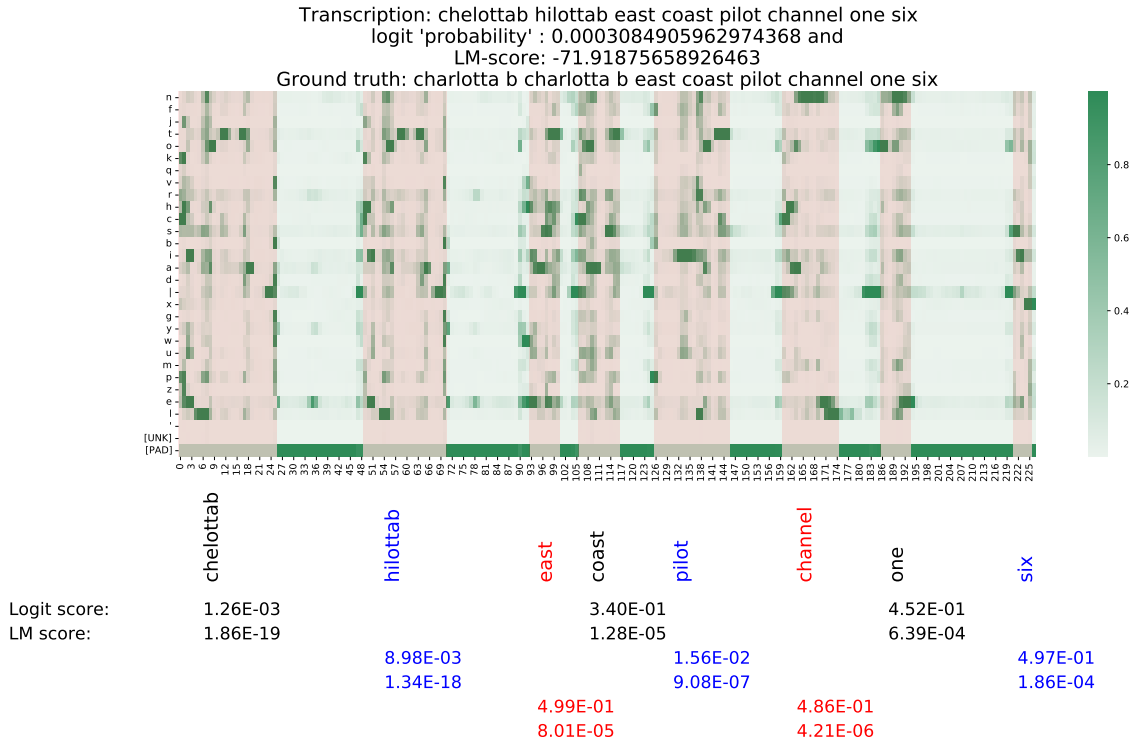

Figure 5.3: Visualization of the resulting logit matrix corresponding to a JRCC message using the *en baseline*-model. Areas in red highlights spans containing transcribed words. The logit- and language model score is shown below the transcribed words.

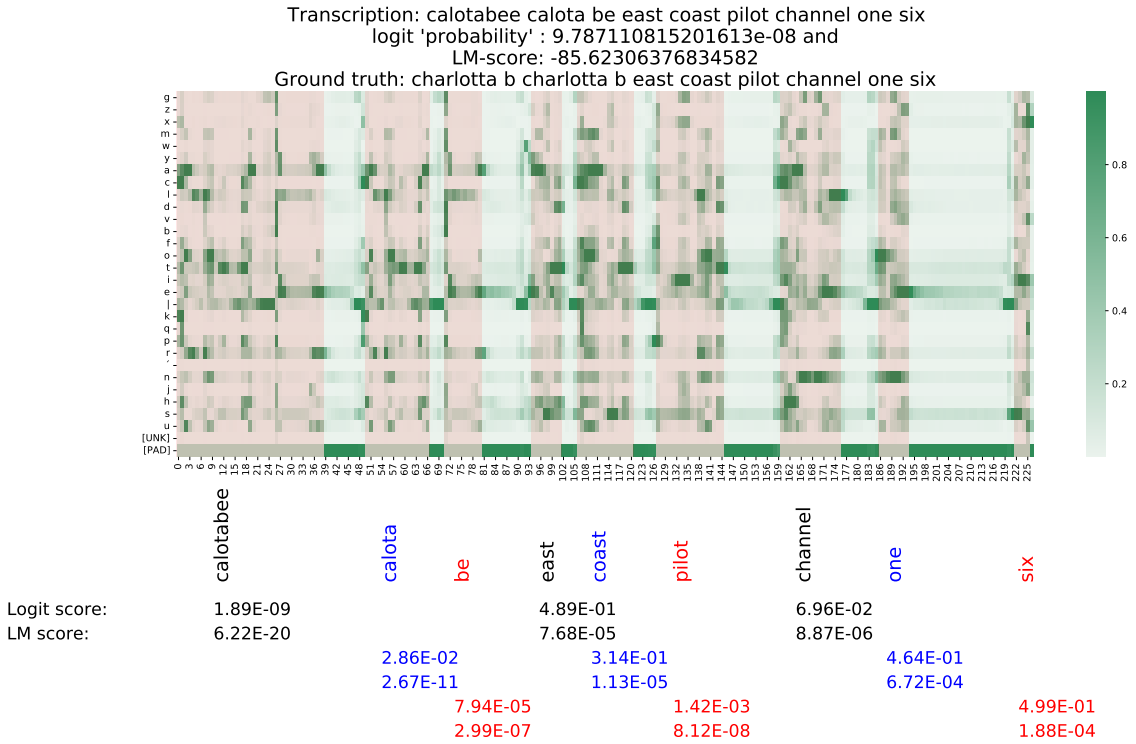

Figure 5.4: Visualization of the resulting logit matrix corresponding to a JRCC message using the *en robust* model. Areas in red highlights spans containing transcribed words. The logit- and language model score is shown below the transcribed words.

5.3 Fine-tuning a model on a curated dataset

As presented in Sections 4.5 and 4.6, a number of pre-trained models taken from Hugging Face were fine-tuned on the curated datasets. The results from the fine-tunied *xls-r-* and *sv voxpop*-models on the Swedish subset of the curated JRCC data is shown in Table 5.3. The results is presented as the mean and standard deviation of WER- and CER scores obtained from bootstrap-resampling over the test set. For the *sv voxpop* model it is shown that the WER-score is decreased by 17% compared to the *sv baseline*-model without using a language model in the decoding and decreased by 19% when using the best performing language model according to Table 5.1. When comparing *sv baseline* without language model and the *sv voxpop*-model with a language model in the decoding, the WER-scored is decreased by 33.5%.

The results from fine-tuning the three different pre-trained models presented in Table 4.3 on the English subset of the curated JRCC data is shown in Table 5.4. Consistently, the *en robust* model performs best both in terms of WER and CER and both with and without a language model. Also the *en voxpop large*-model performs better than *en baseline* while the *en voxpop base* model generally performs worse. Both the *en robust-* and *en voxpop large* models reduce the WER by 3% in comparison to *en baseline* without the addition of a language model. The CER is

Swedish			
model	WER	CER	LM
sv baseline	0.594 ± 0.054	0.237 ± 0.032	None
sv xls-r	0.573 ± 0.056	0.244 ± 0.032	None
sv voxpop	**0.494 ± 0.056**	**0.192 ± 0.031**	None
sv baseline	0.483 ± 0.062	0.222 ± 0.039	3-gram JRCC + NATO alphabet + vessels
sv xls-r	0.460 ± 0.062	0.220 ± 0.038	3-gram JRCC + NATO alphabet + vessels
sv voxpop	**0.395 ± 0.046**	**0.180 ± 0.030**	3-gram JRCC + NATO alphabet + vessels

Table 5.3: Resulting WER- and CER scores from evaluating *sv baseline* and the Swedish models fine tuned on the curated dataset.

reduced by 9% using *en voxpop large* and by 15% using *en robust* compared to the CER using *en baseline*. With the addition of the best performing language model from Table 5.2, *en robust* decreases the WER by 21% and the CER by 20% in comparison to *en baseline* without the addition of a language model.

English			
model	WER	CER	LM
en baseline	0.443 ± 0.050	0.192 ± 0.030	None
en voxpop base	0.673 ± 0.054	0.300 ± 0.035	None
en voxpop large	0.430 ± 0.043	0.174 ± 0.025	None
en robust	**0.429 ± 0.040**	**0.164 ± 0.022**	None
en baseline	0.391 ± 0.042	0.180 ± 0.027	3-gram JRCC + atco2 + switchboard
en voxpop large	0.381 ± 0.038	0.165 ± 0.226	3-gram JRCC + atco2 + switchboard
en robust	**0.352 ± 0.037**	**0.153 ± 0.220**	3-gram JRCC + atco2 + switchboard

Table 5.4: Resulting WER- and CER scores from evaluating *en baseline* and the English models fine-tuned on the curated dataset.

5.4 Auto-Correction

Here, the results from the auto-correction of transcripts described in Section 4.7 are presented. Table 5.5 contains the results for the Swedish test set and Table 5.6 contains the results for the English test set.

Swedish				
model	Auto-correction	Language model	WER	CER
sv baseline	Norvig	X	0.520 ± 0.058	0.234 ± 0.039
sv baseline	Norvig	✓	0.517 ± 0.056	0.237 ± 0.041
sv baseline	Symspell	X	0.519 ± 0.056	**0.232 ± 0.038**
sv baseline	Symspell	✓	0.508 ± 0.056	0.240 ± 0.040
sv voxpop	Norvig	X	0.507 ± 0.076	0.243 ± 0.044
sv voxpop	Norvig	✓	0.497 ± 0.078	0.238 ± 0.045
sv voxpop	Symspell	X	0.511 ± 0.072	0.245 ± 0.042
sv voxpop	Symspell	✓	**0.486 ± 0.074**	0.242 ± 0.043

Table 5.5: Resulting WER- and CER scores from evaluating the effect of auto-correction to transcriptions of the test set using the Swedish models.

English				
model	Auto-correction	Language model	WER	CER
en baseline	Norvig	X	0.402 ± 0.048	0.197 ± 0.030
en baseline	Norvig	✓	0.402 ± 0.047	0.196 ± 0.030
en baseline	Symspell	X	0.399 ± 0.047	0.195 ± 0.029
en baseline	Symspell	✓	0.399 ± 0.047	0.194 ± 0.029
en voxpop large	Norvig	X	0.392 ± 0.045	0.178 ± 0.028
en voxpop large	Norvig	✓	0.390 ± 0.043	0.176 ± 0.027
en voxpop large	Symspell	X	0.393 ± 0.044	0.178 ± 0.027
en voxpop large	Symspell	✓	0.387 ± 0.045	0.174 ± 0.028
en robust	Norvig	X	0.386 ± 0.046	0.161 ± 0.024
en robust	Norvig	✓	**0.383 ± 0.047**	**0.159 ± 0.025**
en robust	Symspell	X	0.387 ± 0.056	0.162 ± 0.024
en robust	Symspell	✓	**0.383 ± 0.046**	0.160 ± 0.024

Table 5.6: Resulting WER- and CER scores from evaluating the effect of auto-correction to transcriptions of the test set using the English models.

5.5 Keyword detection

The keyword detection is explained in Section 4.8.3. Each model is evaluated in three settings; using greedy decoding, using a language model and using a language model + hotword boosting for the keywords. There are 81 occurrences of *rescue* and 3 occurrences of *mayday* in the ground truth. Detection of a keyword with edit distance n is not counted if the keyword is also detected with edit distance $< n$. The results for keyword detection of *rescue* are presented in Table 5.7. It is shown that *sv voxpop* detect 3 more *rescues* than *sv baseline* in the setting with a language model and hotword boosting used in the inference while *en robust* detect 14 more *rescues* than *en baseline* with the same setting. For the keyword detection of *mayday* all

English models in all settings detected all maydays in the evaluation set while none of the Swedish models in any setting detected a single mayday.

Keyword detection out of 81 occurrences of 'rescue'

model	method	Edit distance:	0	1	2
sv baseline	greedy		63	2	0
sv baseline	language model		65	0	0
sv baseline	language model + hotwords		65	0	0
sv voxpop	greedy		67	0	0
sv voxpop	language model		67	0	0
sv voxpop	language model + hotwords		**68**	0	0
en baseline	greedy		49	1	0
en baseline	language model		50	1	0
en baseline	language model + hotwords		52	1	0
en voxpop large	greedy		56	0	0
en voxpop large	language model		60	0	0
en voxpop large	language model + hotwords		62	0	0
en robust	greedy		59	0	0
en robust	language model		63	0	0
en robust	language model + hotwords		**66**	0	0

Table 5.7: The number of 'rescue' detected by various models using greedy decoding, a language model in the inference and a language model + hotword boosting in the inference.

6
Discussion

In this chapter, we discuss the most important results and findings of the thesis. We begin with some discussion about WER as a metric since this will be referred to quite a lot.

6.1 A reduction in WER means better transcripts, right?

One aspect to take into account regarding linking WER to the quality of transcripts is that WER treats all words as equally important. Consider, for example, a case where the transcript without a language model results in a correctly spelled vessel name and two NATO code words that are misspelled but still understandable from the perspective of an operator, e.g. *alpha* and *fostrot*. If the addition of a language model corrects the spelling of the code words but changes the vessel name to something unrecognizable, the WER score will signal for a better transcript even though the contextual understanding of the operator is impaired. Basically the WER/CER score is a measure of how many of the words/characters present in the transcript that are correct given the ground truth. This means that a misspelled word has the same impact on the WER score no matter how misspelled it is. Two given transcripts may have the same WER scores but different CER scores and in such a case it is not hard to realize that the transcript with the lowest CER score is likely to be the easiest to decipher. Another example is when all the words are misspelled with one replaced letter in each word. Despite this resulting in a WER score of 1, the CER would be relatively low (depending on the length of the words) and it would most likely be easy to read the context from the transcript. A third example, which was encountered several times in the evaluation, is closed/open compound words. Two correctly spelled words written together will result in one deletion and one substitution to the WER but just one deletion to the CER. Looking at (2.21) and (2.22) it is clear that a single error has higher impact to WER than CER because the normalizing term in CER is typically much larger than that of WER. From analyzing the test set transcripts (see A.1 and A.2), it was perceived that the CER score had a higher correlation with the readability of transcripts.

On the other hand, there is of course also a correlation between WER and the quality of transcripts. The essence with the above is that the analysis of the performance of different models needs to be more nuanced than just staring blindly at WER.

6.2 Language models

The addition of a language model to the CTC-decoder improves the WER and CER performance compared to the results generated without the use of a language model. The best results are achieved when the language model is trained on JRCC data with domain specific augmentation. For the Swedish language model this corresponds to vessel names and phonetic alphabet and for the English language model it corresponds to *ATCO2* which largely consist of phonetic alphabet code words and *Switchboard* which is transcriptions from conversational speech. The addition of *Maringuiden*, *SFV-magazines* and *Swedish Wikipedia* improves the results compared to the baseline (no language model) but worsen the results compared to when only using JRCC-data. This is probably because the general language skills then dominates the specific language skills instead of supplementing the JRCC texts which was the intention.

The use of a language model is not a quick fix in the sense that it does not contribute much if the performance of the audio model is 'too' defective. The addition of a language model mainly corrects words that are already easy to 'auto-correct' by a human reader, i.e. *chanel* $\rightarrow$ *channel* or *zeroe* $\rightarrow$ *zero*. One example of when the language model auto-correct a word falsely is *bog* $\rightarrow$ *borg* in the Swedish large voxpopuli results. In this case, the interpretation of *bog* is actually correct but the language model changes it to *borg*. The issue thus arise about what is more valuable, the correction of almost correctly spelled words but with possible damage to words such as vessel names, or to keep the transcript as they are and let the transcriber interpret the wrongly spelled words.

When compared to the results of [7] the addition of a n-gram language model (4-gram in [7]) results in significantly larger reductions in WER than in our case. The most comparable results based on WER is the ones they obtain from using 10 minutes of data for fine-tuning. Their model without a language model then achieves a WER of 45.3 on the noisier test set of Librispeech. The addition of a n-gram language model resulted in a WER of 13.1 which is an improvement of almost 70%. A possible explanation of why the impact of a language model differs so much can be found by reviewing the transcriptions of Appendices A.1-A.2 and Appendic C in [7]. It then becomes clear that the raw transcripts (without language model) is much more readable in [7] even though the WER is often approximately the same. This could be explained by the fact that several parts of the JRCC data suffers from extreme noise which make it very hard for the acoustic model to predict something close to the correct vocabulary entries. These errors then propagate to the decoding process which makes it hard for the language model to guide it along a desirable path. Unfortunately there is no mention of the CER scores in [7] but based on its appendix C, they are presumably lower than our results. To model a best case scenario, we used a language model trained directly and solely on the test data transcript. These resulted in a WER score of 35 on the Swedish test set and 28 on the English test set, which is a clear improvement compared to other language models, but they are not close to the results in [7], which strengthens the above reasoning about the deficiency in the acoustic module in combination with partly extraordinary noisy data.

6.2.1 The processing of unknown words in transcripts

As mentioned in Section 3.1.2 sentences with question marks in them are still useful for extracting linguistic knowledge but we don't want to include the question marks in the language model training data. We solved this by splitting the sentences into parts containing running text, i.e. into the part before a question mark and the part following a question mark. Another idea would be to replace all question marks by vessel names since they are often the most difficult-to-interpret words and thus the underlying word of a question mark is likely to be a vessel name. The problem with this approach is that a random vessel name would damage the context in situations where the question mark is not due to a vessel name. One could add a rule to the transcript normalization that a indistinct vessel name should be transcribed as a double question mark, *'??'* while indistinct general words should be transcribed as a single question mark, *'?'*. This way, there would be no issue with adding random vessel names to contexts marked with *'??'*.

6.2.2 The choice of text for the language model corpora

The JRCC data consists of conversational language in the maritime domain. Examples of domain-specific phrases that occur frequently are:

- References to a channel on the vhf band where you want to move the communication.
- Calling for attention by mentioning the called ship name twice followed by the calling ship name once.
- Broadcast announcements in different forms.

It has been difficult to find external text sources that contain these standard phrases, but since at least the first and third type of phrase in the list above are so common, the relatively small amount of text from the JRCC transcripts is considered sufficient to provide a language model knowledge of these. The vessel names can also be arranged by augmentation, the difficulty in this lies in the fact that many of the vessel names extracted from vesselfinder are not relevant to the area around the Swedish coast. The best condition would have been if there was access to a live-updated Automatic Identification System (AIS) from which one could periodically filter out only those ships that are in the area of interest. It is not particularly computationally expensive to create new n-gram language models and it would therefore not be a problem to update the used language model at regular intervals to ensure that it is always trained on the vessel names that are most likely to appear in messages. A simple attempt with this has been made in the thesis by filtering out Swedish boat names based on longitude and latitude, but these boat names are only relevant for the day when the filtering was done. In addition, many vessels lack this information, which means that not all vessels of interest have been included. Additionally, the idea of double question marks from section 6.2.1 could be utilized in combination with updated vessel names to augment new texts to the language model.

In addition to the standard phrases, the message contains mostly conversational speech. We have therefore chosen to augment the language model's corpus with datasets consisting of conversational speech. For example, *Switchboard* consists of

telephone conversations about a predetermined subject. The way of speaking is often similar to that in the JRCC messages with repetitions and ill-considered formulations in comparison with written text, but the conversations are mainly about topics that are completely disconnected from the JRCC domain. The idea behind including texts from *Maringuiden* in the corpus of the Swedish language model is that these are more likely to contain words and expressions more suitable for the domain, even if the text's origins are not from conversations but more or less well thought out forum posts. Lastly, it is considered to be a great need for large amounts of JRCC transcripts as these of course are completely compatible with the domain.

6.3 Fine tuning models on curated data

First of all, the fine-tuning of new models on the cured dataset proved to be more successful than expected. The Swedish model pre-trained on Voxpopuli outperforms the initial model both according to the WER and CER scores and when evaluating the transcripts for contextual content. The same applies to both of the English fine-tuned models, but the difference in contextual content is not as clear. These results demonstrates the importance of clean data. With that said, there are probably errors left in the cured set, but it is obviously of a higher quality than before the curation. The results are also interesting as *en robust* gave the best results among the English models which indicates that the noise in the Switchboard and Fischer data that are part of *en robust*'s pre-training data brings the model knowledge that is utilized in fine-tuning on the JRCC data which is generally of even higher noise.

As shown in Tables 5.3 and 5.4 the total length of the curated datasets are $\tilde{4}0$ minutes for the Swedish subset and $\tilde{1}10$ minutes for the English subset. We must take into account that also the reported lengths contain periods of silence. The exact total length of speech is therefore unknown but as an approximation up to 30-40% of the total sound is silence or none-speech. Figure 6.1 show a comparison of the WER achieved when using a limited amount of resources between the *sv voxpop*, *en robust* and the results from [7]. The comparison is not entirely fair because of the uncertainty of the curated datasets lengths as well as the models being evaluated on different test sets. Also, the WER scores of *sv voxpop* and *en robust* are not really useful as a measure of the models' ability because the test sets are small and therefore does not provide a fair picture of the performance over the entire domain. The figure still shows the potential for improvement with more data. It would therefore be interesting to extract high quality datasets of for example double the size to investigate how the reduction trend of WER continues.

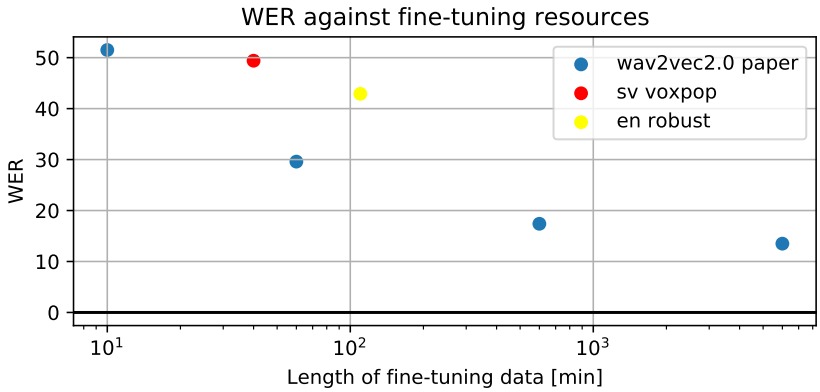

Figure 6.1: Comparison of the WER against the amount of resources for fine-tuning between the best Swedish and Engish models of this thesis and the results from [7].

6.4 Keyword recognition

The successful results for the Swedish models in detecting the term *rescue* corresponds to the analysis in section 3.1.2 which showed that *sweden rescue* was along the most common subjects in the Swedish JRCC data. Even though the above mentioned analysis was performed on the texts of transcripts, this also reflects the composition of the data on which the models are fine-tuned. Likewise, the poor results on the keyword detection of *mayday* using the Swedish models reflects the lack of maydays in the Swedish transcripts. One direction towards the solution of this problem would be to simply augment the language model's corpus with fictitious sentences containing *mayday* or to augment the Swedish corpus with translated English texts. The results for the English models further highlights the benefits of using a language model in the CTC decoding where the addition of a language model and hot word boosting resulted in a 12% increase in the detection of *rescue*'s in the case of the *en robust* model. Further suggestions on this subject is presented in Section 7.

6.5 Ethical considerations

As is often the case with machine learning and other data-driven research areas, it is important to think through whether there are ethical dilemmas that need to be considered.

6.5.1 Sensitive data

Because the content of the JRCC-message can be from people in very stressful situations or even outright life-threatening situations, much of the available data is to be considered sensitive. In principle, anyone can listen to VHF radio, so individual messages can thus be considered openly available. However, it is not as

trivial for individual users to store the type of quantities as done in project Heimdall. These contain information about sea accidents, rescue operations, etc. which in their entirety must be assumed to be the subject of national security.

6.5.2 Bias

The developed models are biased against gender in that the larger part of JRCC messages is transmitted by men and that the models have therefore been trained to a greater extent in transcribing speech by men. The same goes for nationality in that some languages/dialects are easier to understand for Swedish transcribers and thus these languages/dialects are more represented in the training data.

6.5.3 Language model data

The large amounts of texts used in the language models' corpora have not been searched for hate speech or other inconveniences. The Maringuiden texts in particular run the risk of containing inappropriate texts as these are written by ordinary people in an open internet discussion forum.

6.5.4 Interaction between human and machine

The machine-based transcripts are currently used as an aid to human operators to detect emergency calls. It is easy to see the potential usefulness of artificial intelligence in such a system, but there is also a risk that human operators will unknowingly begin to place too much trust in the system's performance. In the worst case, this can lead to operators not being as attentive to incoming calls as they would be without the introduction of the system or to relying more on the system's interpretations than their own. It must therefore be clear from the start who is responsible for potential failures such as missed emergency calls or similar events.

7

Conclusion

This section concludes the thesis, it summarizes the conclusions drawn during the work and propose suggestions for future work.

The addition of language models in the inference of wav2vec2.0 transcripts improve the performance based on WER and CER, but not always regarding contextual understanding. The inference is dependent on the acoustic model performance in that misleading logit values will result in imperfect transcriptions even with a language model applied to the decoding. The main positive impact of a language model is spelling-correction of words, but these words are often perceived correctly by the operator already when reading the greedily decoded transcript. On the other hand, e.g. vessel names often change to something that is further from ground truth than with greedy decoding which makes it more difficult for an operator to guess the correct vessel name. This raises the question of whether it is worth applying a language model for this type of application or whether the risk of losing context is too high as the changes that a language model leads to more or less are beyond human control. However, the above statement depends on the task and the results from Table 5.7 show that keyword recognition is improved with a language model applied as long as the keywords are present in the language model's training data. Therefore, the use of a language model should be well suited to the current main task of project Heimdall, but this requires that the corpus must be augmented in a way so that all keywords are guaranteed to be represented to a sufficiently large extent.

Fine-tuning models on a curated dataset led to great improvements both regarding the WER and CER but also for the contextual understanding. These results show the importance of having high quality data, but perhaps most interesting that the data on which the wav2vec2.0 models are pre-trained plays a big role. The *en robust* model yields the best performance among the English models and the reason for this is most likely that it was pre-trained partly on noisy telephone communication which proved to be useful on the JRCC domain even though it consists of considerable noisier data.

7.1 Suggestions for future work

The most limiting factors for the STT-model to achieve higher quality transcripts are currently assumed to be the limited amounts of training data for the language model and labeled training data for fine-tuning. Proposals for future work are therefore based on the assumption that these areas are the ones of the highest interest to explore further.

7.1.1 Collection of a naval domain corpora

It was shown in section 5.1 that configurations with language models trained only on JRCC transcripts outperformed configurations with language models trained on corpora of much larger scales but from other domains. This clarifies the desirability of a large scale naval domain corpus since this has a high potential to significantly improve results. The availability of open source text within the maritime domain seems to be extremely limited. A number of equivalents in air traffic control have been stumbled upon, but most of these are not open source and have therefore not been applied to this project. The transcription of JRCC messages is a time consuming and often frustrating task as most messages contain indistinct parts due to low sound quality or difficult speech. With enough manpower, large amounts of text could be extracted from all the JRCC-messages available, but this would be unreasonably costly on a large scale. The most desirable thing would have been if the models used in this project produced transcripts good enough to be used in the corpus of the language model, see Section 7.1.2, but this is generally not the case. One possibility would be to create a model that classifies the quality of the transcripts and then use the transcripts that are classified as complete.

7.1.2 Self-training approach

The approach of self-training is to first use a model to label data and then use this 'pseudo-labeled' data for further training. It has been shown that this approach works well for wav2vec2.0 on a variety of labeled data setups [67]. Related to this theses, it was discovered during evaluation that some models performed better than the human transcriber in multiple occasions. This was partly due to the models ability to catch small words such as prepositions but also it's ability to sometimes better interpret vessel names. It was not always the case that the model made a fully correct transcription but it was enough for the human transcriber to re-evaluate its initial interpretation and thus provide a correct transcript. Even though the capacity does not yet exist for successful self-training one could use an intermediate approach to support the human transcribers by including the wav2vec2.0 transcript associated with the audio file to be transcribed. Similarly to the ideas proposed in 7.1.1 a classification of the 'transcriptability' could be used to select JRCC data with high potential to be successfully used in self-training.

7.1.3 Search for messages of interest

The JRCC dataset is saturated with relatively uninteresting messages such as ship-to-ship communication and broadcasts. There is no method in the transcription interface to specifically select the type of messages to be transcribed, so only a very small portion of the transcribed messages contain keywords such as 'mayday' and 'pan pan pan'. Given the results in Section 5.5 it should be possible to select a suitable model and use this to search the JRCC data for messages containing specific keywords. This would then result in more transcribed messages of interest to use in fine-tuning and the training of language models, and thus potentially to models with further increased ability to detect keywords. A related subject to this would be to search the full JRCC data for messages that contains at least one word, i.e. to sort out empty messages, and use the extracted data for pre-training of a wav2vec2.0 model directly on the domain.

Bibliography

[1] Bygg och teknik. https://www.sjofartsverket.se/sv/tjanster/bygg-och-teknik/. Accessed: 2022-02-11.

[2] Sjorapporten. https://www.sjofartsverket.se/sv/om-oss/sjorapporten/. Accessed: 2022-05-06.

[3] Sjö- och flygräddningscentralen jrcc. https://www.sjofartsverket.se/sv/sjo--och-flygraddning/sjo--och-flygraddningscentralen-jrcc/. Accessed: 2022-02-11.

[4] Så hanteras ditt larm via telefon, vhf och rakel. https://www.sjofartsverket.se/sv/sjo--och-flygraddning/sjo--och-flygraddningscentralen-jrcc/kontakta-sjo--och-flygraddningen/. Accessed: 2022-02-11.

[5] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[6] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[9] Jerome R Bellegarda. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108, 2004.

[10] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

[11] Herve Bourlard and Nelson Morgan. Hybrid hmm/ann systems for speech recognition: Overview and new research directions. *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 389–417, 1997.

[12] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. 2007.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[14] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online, November 2020. Association for Computational Linguistics.

[15] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

[16] Yahui Chen. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo, 2015.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

[19] William Gale and Kenneth Church. What's wrong with adding one. *Corpus-based research into language: In honour of Jan Aarts*, pages 189–200, 1994.

[20] William A Gale and Geoffrey Sampson. Good-turing frequency estimation without tears. *Journal of quantitative linguistics*, 2(3):217–237, 1995.

[21] Jonathan Gildevall and Niclas Johansson. Automatic emergency detection in naval vhf transmissions. 2021.

[22] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

[23] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[25] Joshua T Goodman. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434, 2001.

[26] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[27] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[28] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[29] Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

[30] Awni Hannun. Sequence modeling with ctc. *Distill*, 2017. https://distill.pub/2017/ctc.

[31] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition, 2014.

[32] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[33] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[34] Frederick Jelinek. Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*, 1980.

[35] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[36] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[37] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020. `https://github.com/facebookresearch/libri-light`.

[38] Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401, 1987.

[39] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE, 1995.

[40] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[41] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[42] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[43] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[45] Andrew Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. 01 2004.

[46] Hermann Ney, Ute Essen, and Reinhard Kneser. On the estimation of'small'probabilities by leaving-one-out. *IEEE transactions on pattern analysis and machine intelligence*, 17(12):1202–1212, 1995.

[47] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE*

international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE, 2015.

[48] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12, 1802.

[49] Roberto Pieraccini and ICSI Director. From audrey to siri. *Is speech recognition a solved problem*, 23, 2012.

[50] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020.

[51] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[53] Peter Roach, Simon Arnfield, W Barry, J Baltova, Marian Boldea, Adrian Fourcin, W Gonet, Ryszard Gubrynowicz, E Hallum, Lori Lamel, et al. Babel: An eastern european multi-language database. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1892–1893. IEEE, 1996.

[54] Sebastian Ruder. A Review of the Neural History of Natural Language Processing. `http://ruder.io/a-review-of-the-recent-history-of-nlp/`, 2018.

[55] Sebastian Ruder. Nlp's imagenet moment has arrived. *The Gradient*, 2018.

[56] Kristoffer Röshammer. private communication, Jan. 2022.

[57] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[58] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[59] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[60] Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE, 2021.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[62] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[63] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 993–1003, Online, August 2021. Association for Computational Linguistics.

[64] Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 225–230, 2016.

[65] Wikipedia. Marin vhf — wikipedia,, 2021. [Online; hämtad 6-maj-2022].

[66] Wikipedia contributors. Nato phonetic alphabet — Wikipedia, the free encyclopedia, 2022. [Online; accessed 8-June-2022].

[67] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE, 2021.

[68] ChengXiang Zhai. Statistical language models for information retrieval. *Synthesis lectures on human language technologies*, 1(1):1–141, 2008.

[69] Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*, 2020.

A

Appendix 1

A.1 Swedish transcripts

här svarar sweden rescue				
Model	LM	Transcription	WER	CER
sv 7	None	hen svarar sweden rescue	0.250	0.083
curated xls-r	None	här svarar sweden rescue	0.000	0.000
curated voxpop-uli	None	vem svarar sweden rescue	0.250	0.125
sv 7	JNV	men svarar sweden rescue	0.250	0.125
curated xls-r	JNV	här svarar sweden rescue	0.000	0.000
sv 7	JNVS	men svarar sweden rescue	0.250	0.125
sv 7	JNVSW	hen svarar sweden rescue	0.250	0.083
curated voxpop-uli	JNV	vem svarar sweden rescue	0.250	0.125

Table A.1: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

sweden rescue sweden rescue sweden rescue rebecka sexton ? här svarar sweden rescue ja god middag rebecka lämnar kaj två man ombord transport två kom ja men det är uppfattat tack så mycket				
Model	LM	Transcription	WER	CER

sv 7	None	sweden rescue sweden rescue de resscue till becka sexton marda j här svarar sweden rescueja går medda rebecka lännar kar två man mambåhordg fåmsbot på kom an m de uppfattat tack för ket	0.636	0.271
curated xls-r	None	sweden rescue sweden rescue bern rescue rebecka sexton marsda jar svarar sweden rescue ja går midda pls rebecka lenmar kaj två man bovrd famsbort tvåt kom an är de uppfattat tack förm ger det	0.515	0.250
curated voxpop-uli	None	sweden rescue sweden rescue u rescue ebeca sexton masta ähär svarar sweden rescue ja går midda rebecka lännar kaif två mann om boorde transport påt kom aom v dt uppfattat tack fyket	0.545	0.213
sv 7	JNV	sweden rescue sweden rescue rescue till becka sexton marta här svarar sweden rescue ja går med rebecka lännar kar två manmambåhordg fåmsbot på kom ja är uppfattat tack för ket	0.485	0.277
curated xls-r	JNV	sweden rescue sweden rescue bn rescue rebecka sexton marta jag svarar sweden rescue ja gårmidda rebecka lenmar kaj två man bovrd famsbort tvåt kom a är de uppfattat tack för er det	0.485	0.229
sv 7	JNVS	sweden rescue sweden rescue rescue till becka sexton maria här svarar sweden rescue ja går med rebecka lännar kar två man mambåhordgfåmsbot på kom an är e uppfattat tack för ket	0.515	0.282
sv 7	JNVSW	sweden rescue sweden rescue rescue till becka sexton mara här svarar sweden rescue ja går medda rebecka lännar kar två man mambåhordgåmsbot på koman är e uppfattat tack för ket	0.545	0.271
curated voxpop-uli	JNV	sweden rescue sweden rescue rescue rebecka sexton marta här svarar sweden rescue ja går middarebecka lännar kai två man om borde transport på kom om vi dt uppfattat tack fyket	0.485	0.191

Table A.2: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

ja du som kallade stockholm radio på kanal sexton ropa på sweden rescue				
Model	LM	Transcription	WER	CER
sv 7	None	ja de som kallady stckholm radio på kanal sexton ropa på sweden rescue	0.231	0.042
curated xls-r	None	ja dy som kallader stockholm radio på kanal sexton ropa på sweden rescue	0.154	0.028
curated voxpopuli	None	ja di som kallade stockholm radio på kanal sexton ropa på sweden rescue	0.077	0.014
sv 7	JNV	ja det som kallady stckholm radio på kanal sexton ropa på sweden rescue	0.231	0.056
curated xls-r	JNV	ja du som kallade stockholm radio på kanal sexton ropa på sweden rescue	0.000	0.000
sv 7	JNVS	ja det som kalla dy stckholm radio på kanal sexton ropa på sweden rescue	0.308	0.070
sv 7	JNVSW	ja de som kallady stckholm radio på kanal sexton ropa på sweden rescue	0.231	0.042
curated voxpopuli	JNV	ja dit som kallade stockholm radio på kanal sexton ropa på sweden rescue	0.077	0.028

Table A.3: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

lifeguard noll noll fyra rescue adam johan				
Model	LM	Transcription	WER	CER
sv 7	None	licegad noll noll fyra rescue rada me an	0.571	0.190

curated xls-r	None	lifeguard nol noll fyra rescubu harar mohan	0.571	0.167
curated voxpop-uli	None	laifegad noll noll fyira rescue ghara myeaen	0.571	0.286
sv 7	JNV	licegad noll noll fyra rescue rada meran	0.429	0.190
curated xls-r	JNV	lifeguard noll noll fyra rescue harar mohan	0.286	0.095
sv 7	JNVS	licegad noll noll fyra rescue radar med an	0.571	0.214
sv 7	JNVSW	licegad noll noll fyra rescue rada medan	0.429	0.190
curated voxpop-uli	JNV	laifegad noll noll fyra rescue bara men	0.429	0.238

Table A.4: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

lifeguard noll noll fyra rescue adam johan				
Model	LM	Transcription	WER	CER
sv 7	None	licegad noll noll fyra rescue rada me an	0.571	0.190
curated xls-r	None	lifeguard nol noll fyra rescubu harar mohan	0.571	0.167
curated voxpop-uli	None	laifegad noll noll fyira rescue ghara myeaen	0.571	0.286
sv 7	JNV	licegad noll noll fyra rescue rada meran	0.429	0.190
curated xls-r	JNV	lifeguard noll noll fyra rescue harar mohan	0.286	0.095
sv 7	JNVS	licegad noll noll fyra rescue radar med an	0.571	0.214
sv 7	JNVSW	licegad noll noll fyra rescue rada medan	0.429	0.190
curated voxpop-uli	JNV	laifegad noll noll fyra rescue bara men	0.429	0.238

Table A.5: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

rescue elva noll sju sweden rescue sexton				
Model	LM	Transcription	WER	CER
sv 7	None	driske eddva nol sju sweden rescue sexton	0.429	0.171
curated xls-r	None	desue eddva noll sju sweden rescue sext	0.429	0.146
curated voxpop-uli	None	dresc e edva noll sju sweden rescue sextn	0.571	0.098
sv 7	JNV	driske eddva nol sju sweden rescue sexton	0.429	0.171
curated xls-r	JNV	desue edda noll sju sweden rescue sexton	0.286	0.098
sv 7	JNVS	driske eddva nol sju sweden rescue sexton	0.429	0.171
sv 7	JNVSW	driske eddva nol sju sweden rescue sexton	0.429	0.171
curated voxpop-uli	JNV	des e ebba noll sju sweden rescue sex	0.571	0.195

Table A.6: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

ok är det grå rök eh vi ser inga öppna lågor ok då är det kan det va vattenånga som ryker har du sett öppna lågor tidigare har du sett öppna lågor tidigare				
Model	LM	Transcription	WER	CER
sv 7	None	okej ir dig grore ee du se ringa uskna dågor okej då är vi kande braen backelonna sam ruker har du settupp na dågatydigarö har du stett us na då går tydigare	0.879	0.348

curated xls-r	None	okej er dig gror ee vi tr ringa upna någår e okej då är vi kan du pa vasse nolet sam uke r har du stett sa pp ga nolae tydyjare har du sett uthna når gå l tydijärreh	0.939	0.426
curated voxpop-uli	None	okej er vi groree v ser inge uppna noågore okej då ä v kan ni van vakeno som ruker har du setep den dåg tidigare har di stet upna logår tydigare	0.758	0.335
sv 7	JNV	okej dig grore ee du se ringa utkna dågor okej då är vi kande braen backelonna sam ruker har du settupp nan då tidigare har du sett usna då går tidigare	0.758	0.329
curated xls-r	JNV	okej er dig groree vi er ringa upna någår okej då är vi kan du pa vassenolet sam ur har du sett sappga noll tydyjare har du sett thna nor gå tydijärreh	0.727	0.394
sv 7	JNVS	okej dig grore du se ringa utkna dågor okej då är vi kande bran backelonna sam ruker har du settpp na då tidigare har du sett usna då går tidigare	0.758	0.323
sv 7	JNVSW	okej de gror du se ringa ukna dågor okej då är vi kand baen backelonna sam ruker har du sett upp na dåatydigaröhar du sett usna då går tidigare	0.788	0.348
curated voxpop-uli	JNV	okej er vi groree v se inga uppna noågore okej då är v kan ni gan vaenonsom ryker har du sett den då tidigare har vi sett uppna logår tidigare	0.636	0.303

Table A.7: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

stockholm radio stockholm radio segelbåten rosalind kallar du som kallade stockholm radio på kanal sexton kan du gå över till kanal tjugotre två tre kanal tjugotre				
Model	LM	Transcription	WER	CER

sv 7	None	stockhulmradio stockholm radio segelbåste r starin kallar i sän kalvl stckhonama på knal sexton kanme radig till kanal tjugotre två degnkanal jofiue	0.692	0.276
curated xls-r	None	stockholm radio stockholm radiosegelbåter brodsaring kallar ni som kllaer stockhl et va på kanol sexton kan du ä ha ur till kanal tjugot tll d kanal tjuocio	0.615	0.221
curated voxpopuli	None	stockholm radio stockholm radio segelbåter rosaring kallar ni sen kallal stockhl havar påo knas sexton kan dy ru över till kanal tjugotre två de kanal tjuttiou	0.500	0.172
sv 7	JNV	stockholm radio stockholm radio segelbåt rstarin kallar i sun kalel stckhonama på kanal sexton kamedradig till kanal tjugotre två den kanal je	0.500	0.258
curated xls-r	JNV	stockholm radio stockholm radiosegelbåter brodsaring kallar ni som klar stockhol eva på kanal sexton kan du är ha ur till kanal tjugot tll d kanal tjuocoa	0.538	0.215
sv 7	JNVS	stockholm radio stockholm radio segelbåt rstarin kallar i sen kalvl stckhonama på kanal sexton kanmedradig till kanal tjugotre två den kanal ju	0.500	0.258
sv 7	JNVSW	stockholm radio stockholm radio segelbåt starin kallar i sen kalva stckhonama på kanal sexton kamedradig till kanal tjugotre två den kanal joe	0.500	0.252
curated voxpopuli	JNV	stockholm radio stockholm radio segelbåten rosa ring kallar ni sen kall stockholm var på kn sexton kan vi du över till kanal tjugotre två de kanal sjuttio	0.423	0.172

Table A.8: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

ja panta mera sweden rescue kan ta sjuttifyra				
Model	LM	Transcription	WER	CER

sv 7	None	a mantamera sweden rescue kan ta sjutifyra	0.500	0.089
curated xls-r	None	ja pantamera sweden rescue s kan ta sjutiofyra	0.500	0.111
curated voxpop-uli	None	ja pantamera sweden rescue skan ta sjutifyra	0.500	0.067
sv 7	JNV	a mantamera sweden rescue kan ta sjuttifyra	0.375	0.067
curated xls-r	JNV	ja pantamera sweden rescue kan ta sjutiofyra	0.375	0.067
sv 7	JNVS	a mantamera sweden rescue kan ta sjuttifyra	0.375	0.067
sv 7	JNVSW	a mantamera sweden rescue kan ta sjuttifyra	0.375	0.067
curated voxpop-uli	JNV	ja pantamera sweden rescue ska ta sjuttifyra	0.375	0.067

Table A.9: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

polisbåt nittionio tio kallar nordan sexton				
Model	LM	Transcription	WER	CER
sv 7	None	polisbå fäheteolnidi tiggr kallar nordan sexton	0.500	0.279
curated xls-r	None	po lisbåt nryco ninitygr kallar nordan sexton	0.667	0.256
curated voxpop-uli	None	polisbåt niotionio tiehär kallar nordan sexton	0.333	0.116
sv 7	JNV	polisbåfäheteolnidi tiger kallar nordan sexton	0.500	0.279
curated xls-r	JNV	polisbåt nice ninitygr kallar nordan sexton	0.333	0.209
sv 7	JNVS	polisbåfäheteolnidi tiger kallar nordan sexton	0.500	0.279
sv 7	JNVSW	polisbåfäheteolnidi tigger kallar nordan sexton	0.500	0.302
curated voxpop-uli	JNV	polisbåt nitionio tiger kallar nordan sexton	0.333	0.093

Table A.10: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

allmänt meddelande solskär går in i bangestrecket om tre minuter på väg mot stockholm				
Model	LM	Transcription	WER	CER
sv 7	None	allmänt meddelande stor sköär går in i baggerstäket om tre minuter på väg mot stockholm	0.214	0.106
curated xls-r	None	allmänt medderdander sorskör gr in i baggenstäket om tre minuter på väg mot stockholm	0.286	0.129
curated voxpopuli	None	allmänt meddelande sor kär går in i bagngestäket om tre minuter på väg mot stockholm	0.214	0.071
sv 7	JNV	allmänt meddelande storskär går in i baggerstäket om tre minuter på väg mot stockholm	0.143	0.082
curated xls-r	JNV	allmänt medderdandersorskör gå in i baggens stäket om tre minuter på väg mot stockholm	0.357	0.141
sv 7	JNVS	allmänt meddelande stor skär går in i baggerstäket om tre minuter på väg mot stockholm	0.214	0.094
sv 7	JNVSW	allmänt meddelande storskär går in i baggerstäket om tre minuter på väg mot stockholm	0.143	0.082
curated voxpopuli	JNV	allmänt meddelande sorkär går in i bangestäket om tre minuter på väg mot stockholm	0.143	0.059

Table A.11: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

ingen kontakt med piraya klart slut				
Model	LM	Transcription	WER	CER
sv 7	None	ingen kontakt med firaa klarslut	0.500	0.114

curated xls-r	None	ingung kontakt me byrraa klar slut	0.667	0.229
curated voxpop-uli	None	ingent kontakt med ciraia klart slut	0.333	0.086
sv 7	JNV	ingen kontakt med firaaklarslut	0.500	0.143
curated xls-r	JNV	ingung kontakt med byrraa klar slut	0.500	0.200
sv 7	JNVS	ingen kontakt med firaaklarslut	0.500	0.143
sv 7	JNVSW	ingen kontakt med firaaklarslut	0.500	0.143
curated voxpop-uli	JNV	ingen kontakt med ciraia klart slut	0.167	0.057

Table A.12: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

sweden rescue sweden rescue rescue paul lederhausen kanal sexton ? paul lederhausen kom jag då var paul lederhausen och rescue ? tillbaka vid bryggan om två minuter jag är åter hemma vi tackar tack ?				
Model	LM	Transcription	WER	CER
sv 7	None	sweden rescue sweden rescue rescue hålädravsin kanal sexton tacko lägrra av somn koma då var påländernasen ochk rescue uller gårsäll tilbakan v brikan nolm två bräinster ja åter hemma vi tackar hatack sjör de gåt bacgk	0.686	0.427
curated xls-r	None	sweden rescue sweden rescue rescue kalleg dral sem kanal sexton tackkal ollera oh sen kom ja h tdur var påleäd deransen och rescue ullev berställ tilbaken frebrigga nnal två minuteåter hemma vi tacl här tack skä d s gord vak	0.800	0.457
curated voxpop-uli	None	sweden rescue sweden rescue rescue koll drausen kanal sexton tackar påo lirar som kom ja då var pålig derausen ochden rescue ullderbörfsäl tillbaka vibygga nom två minuter återhemma vi tackar tac jt got vak	0.629	0.352

sv 7	JNV	sweden rescue sweden rescue rescue hålädravsin kanal sexton tack olga av som kom ja då var påläderausen och rescue ullrgårsäll tilbakan v brikannom två brunte åter hemma vi tackar tack sjö det går ba	0.600	0.387
curated xls-r	JNV	sweden rescue sweden rescue rescue kalleg drason kanal sexton nacka olleraohsen kom ja då var påledderansen och rescue ullev bertäll tilbaken frebriggannal två minuteåter hemma vi tack här tack ska s god vak	0.657	0.387
sv 7	JNVS	sweden rescue sweden rescue rescue hålädravsin kanal sexton tack lära av som kom ja då var påläderausen och rescue ullegårsäll tilbakan v bricka nom två brinster åter hemma vi tackar tack skördetgåt back	0.543	0.382
sv 7	JNVSW	sweden rescue sweden rescue rescue hålädravsin kanal sexton tackla av som koma då var pålädrasen och rescue ullergårsäll tilbakan v brickan no två brinster åter hemma vi tackar tack sjödegåt back	0.543	0.402
curated voxpopuli	JNV	sweden rescue sweden rescue rescue koldrausen kanal sexton tackar polirar som kom ja då var på ligderausen och rescue ulerbefsäl tillbaka vibyggan nom två minuter åter hemma vi tackar tack ja god vk	0.486	0.302

Table A.13: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

söderarm tärnan ja söderarm ja ska vi ta femton nej jag hörde frågan förut kör på du bara jag ska in till eknö ja då så tack tack själv				
Model	LM	Transcription	WER	CER
sv 7	None	sböderarm härnan ja ht kad laren ja ska pte femton h jag hördeö sågen för dutt jug för pådurbarhet sen dtrkler jah et vat srå tack tac igl	0.862	0.452

curated xls-r	None	söderar fenan ja a vk ar bålarnd ja skju utta femton äh jag hör det sågen för t berdni sert forderborjar print ter ätt ledg ja jat dåt gar ftra tack tack fväll	0.931	0.578
curated voxpop-uli	None	stöderarn tarnan jar sködelaren ja ska vi ta femton a jag hörde svågen förurtig för poderbarha sfingtiäkne ja hät så ar så tack tack sel	0.655	0.341
sv 7	JNV	söderarm härnan ja dt kör laen ja ska te femton jag hörde sågen för ut du för pådurbarhetsen der ja det var så tack tack ill	0.690	0.407
curated xls-r	JNV	söderarm fenan ja vk har bålarnd ja skjuvutta femton äh jag hör det sågen för tbn se forderborjar ping tyr led ja at då går fra tack tack fäll	0.793	0.481
sv 7	JNVS	söderarm är nan ja dt ka lagen ja ska te femton jag hörde sågen för ut ju för pådurbarhetsen de ja det var så tack tack vill	0.759	0.415
sv 7	JNVSW	söderarmrnan ja ht ka lagen ja ska pite femton jag hörde sågen för uttu för pådurbarhetsen dle ja det var så tack tac il	0.793	0.422
curated voxpop-uli	JNV	söderarm tarnan ja ködelaren ja ska vi ta femton jag hörde svågenförurtig för podebarhatingtiäkne ja så är så tack tack ll	0.586	0.326

Table A.14: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

båt på vår barbords bord med utombordare och fiskeutrustning på vraket skisserens position kom				
Model	LM	Transcription	WER	CER
sv 7	None	bo as pvår babåords bog med utom bodarö och frisscue srustning på brake kissterenposition kom	0.857	0.255
curated xls-r	None	bå afk påvåor babords bog meusom bordare ochiktrustning på grake fy kyteren position kom	0.857	0.287

curated voxpop-uli	None	bo på vår babord bog med uton bodare och fiske utrustnin på brakes kisterens position kom	0.643	0.160
sv 7	JNV	boaspvår babord bog med utombodarö och frisscuesrustning på brake krissterenposition kom	0.714	0.245
curated xls-r	JNV	b af på vår babordsbog meusom bordare ochik-trustning på grakefykyteren position kom	0.714	0.255
sv 7	JNVS	bas på vår babordsbog med utom bodarö och fiske rustning på brake kissterenposition kom	0.714	0.181
sv 7	JNVSW	boaspvår babordsbog med utom bodar och fiskerustning på brake kissterenposition kom	0.786	0.213
curated voxpop-uli	JNV	boss på vår babord borg med utonbodare och fiskeutrustning på brakeskisterens posi-tion kom	0.429	0.128

Table A.15: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

ja rescue göta du kan nog sticka den vägen då om bara skarven passar lite grann där om du backar löser jag det ja jag backar och så vänder jag om när de räknar mer ja kom skarven håller han här då				
Model	LM	Transcription	WER	CER
sv 7	None	ja rescue jötar dukarnustikar na väger då rja skal vin passa drjäne o dubakarna sköd difja jag backar oc stå ven det om där ag krom digen ah kon men ska vän hålla an här dåg	0.881	0.429
curated xls-r	None	ja rescue r gjötr du kan usticka nig vägen då omböa skarmen passa dy rig gande om du bakkar na sfördett ja jag bankar o så blän det a om där takdr med ja a fkom oskaven holle an här o ha	0.762	0.342
curated voxpop-uli	None	ja rescue jöta dukar sticka di vägen då om bara skavn passa livgn di oh du bakarna sör dit ja jag vanckar oh tå vän detga om däre haktar midg ja ja kråm mån ska ven halla en här då	0.714	0.327

sv 7	JNV	rescue jötar dukarnustikarna väger då ja ska vi passa drjäne dubakarna sd veja jag backar och så men det om där ag kom igen ah on men ska ven hållaan här då	0.810	0.444
curated xls-r	JNV	ja rescue gjötr du kan sticka nig vägen då om a skarven passa dyrigande om du backar na sör ett ja jag bankar o så blän det a om där akter med ja ja kom oskavenhollean här o ha	0.595	0.332
sv 7	JNVS	rescue jötar dukarnustikarna väger då ja ska vi passa djäne dubakarna södra jag backar och stå men det om där ag kom igen ah kon men ska vän hålla an här då	0.857	0.459
sv 7	JNVSW	rescue jötar dukarnustikarnag väger då ja ska vi passa drjäne dubar södra jag backar och stå men det om där ag kom igen ah kon men ska vän hålla an här då	0.857	0.459
curated voxpop-uli	JNV	ja rescue hör du kan sticka i vägen då om bara ska passar liv de och du backar sör dit ja jag vanckar och ta ven det om där akka mig ja ja krono skarven hallaen här då	0.524	0.342

Table A.16: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

sweden rescue sweden rescue det var rescue åtta femtiosju lämnar bergkvara för ing tre man ombord rescue åtta femtiosju lämnar bergkvara för övning det var taget hos sweden rescue				
Model	LM	Transcription	WER	CER
sv 7	None	sweden rescue sweden rescue det var rescue åtta femtiosju lämna bäg svarar för ökning fre ma ombordrescue åtta femtiosju lämna bärk-vara för övning det var taget håd sweden res-cue	0.379	0.095
curated xls-r	None	sweden rescue sweden rescue det var rescue åtta femtiosju lemrberg svarar för ö n linga man om bord rescue åttar femtiosju lem-snaberg svara förövning det var taget sweden rescue	0.448	0.156

curated voxpop-uli	None	sweden rescue sweden rescue det var rescue åtta femtisju lämdr beäg svarar för öning sträma om bord rescue åtta femtiosju lemena bärgkvarar för övningh det var taget ås sweden rescue	0.414	0.123
sv 7	JNV	sweden rescue sweden rescue det var rescue åtta femtiosju lämna bg svarar förökning tre man ombord rescue åtta femtiosju lämna bärk-vara för övning det var taget hådsweden rescue	0.276	0.084
curated xls-r	JNV	sweden rescue sweden rescue det var rescue åtta femtiosju lemrberg svarar för lingamanom bordrescue åtta femtiosju lesnaberg svara för övning det var taget sweden rescue	0.345	0.145
sv 7	JNVS	sweden rescue sweden rescue det var rescue åtta femtiosju lämna bg svarar för ökning tre man ombord rescue åtta femtiosju lämna bär kvar för övning det var taget råd sweden res-cue	0.276	0.089
sv 7	JNVSW	sweden rescue sweden rescue det var res-cue åtta femtiosju lämna bg svarar förökning frema ombord rescue åtta femtiosju lämna bär kvar för övning det var taget rådsweden rescue	0.379	0.112
curated voxpop-uli	JNV	sweden rescue sweden rescue det var rescue åtta femtiosju lämnar bg svarar för övning sma ombord rescue åtta femtiosju lemena bärgk-varar för övning det var taget sweden rescue	0.276	0.123

Table A.17: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models J is for JRCC, N for NATO phonetic alphabet, V for vessel combinations, S for Sjörapporten and W for Wikipedia.

ingen kontakt med sweden rescue klart slut ingrid ramstedt				
Model	LM	Transcription	WER	CER
sv 7	None	ingen kontakt föå sweden rescue klarlus styegert bramsten	0.556	0.276
curated xls-r	None	igen kontakthå sweden rescue uklartsö inger brate	0.778	0.310

curated voxpop-uli	None	ingen kontakt he sweden rescue klartoslus higerdbramsen	0.556	0.207
sv 7	JNV	ingen kontakt en sweden rescue klarlusstegert ramsten	0.556	0.241
curated xls-r	JNV	ingen kontakt sweden rescue utklartsö inger brate	0.556	0.310
sv 7	JNVS	ingen kontakt få sweden rescue klarlusstegert bramsten	0.556	0.276
sv 7	JNVSW	ingen kontakt få sweden rescue klarustegert ramsten	0.556	0.259
curated voxpop-uli	JNV	ingen kontakt sweden rescue klartoslushigerd-bramsen	0.556	0.241

Table A.18: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

rescue rebecka sweden rescue ska vi ta kanal sjuttiofyra sju fyra				
Model	LM	Transcription	WER	CER
sv 7	None	rescue rebecka sweden rescue strvic ta kanal sjuttifyra sjufyra	0.455	0.092
curated xls-r	None	rescue rebecka sweden rescue strik ta kanal sjuttiofyra sju fyra em	0.273	0.123
curated voxpop-uli	None	rescue rebecka sweden rescue strvixk ba kanal sjuttiofyra sju fyra en	0.364	0.138
sv 7	JNV	rescue rebecka sweden rescue storvik ta kanal sjuttiofyra sju fyra	0.182	0.062
curated xls-r	JNV	rescue rebecka sweden rescue strik ta kanal sjuttiofyra sju fyra en	0.273	0.123
sv 7	JNVS	rescue rebecka sweden rescue storvik ta kanal sjuttiofyra sju fyra	0.182	0.062
sv 7	JNVSW	rescue rebecka sweden rescue storvik a kanal sjuttiofyra sju fyra	0.273	0.077

curated voxpop-uli	JNV	rescue rebecka sweden rescue strvik på kanal sjuttiofyra sju fyra en	0.364	0.138

Table A.19: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö till nordkoster vänern och vättern lyssna v h f trafikkanal samt gränsvåg				
Model	LM	Transcription	WER	CER
sv 7	None	allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö fi ordkoster vänern och vättern lyssna vhf trafikkanal samt gränsvåg	0.217	0.039
curated xls-r	None	allmänt anrop här kallar stockholm radio med väderprognosen för skhallands väderö till nodkoster vänen och vättern lyssna v h f trafikkanal samt gränsvåg	0.130	0.026
curated voxpop-uli	None	allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö fri nordkoster vänern och vättern lyssna v ha f trafikkanal samt gränsvåg	0.087	0.033
sv 7	JNV	allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö nordkoster vänern och vättern lyssna vhf trafikkanal samt gränsvåg	0.174	0.046
curated xls-r	JNV	allmänt anrop här kallar stockholm radio med väderprognosen för skallands väderö till nordkoster vänern och vättern lyssna v h trafikkanal samt gränsvåg	0.087	0.026
sv 7	JNVS	allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö nordkoster vänern och vättern lyssna vhf trafikkanal samt gränsvåg	0.174	0.046
sv 7	JNVSW	allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö nordkoster vänern och vättern lyssna vhf trafikkanal samt gränsvåg	0.174	0.046

curated voxpop-uli	JNV	allmänt anrop här kallar stockholm radio med väderprognosen för hallands väderö fri nordkoster vänern och vättern lyssna v ha f trafikkanal samt gränsvåg	0.087	0.033

Table A.20: Overview of the different model and language models impact on a given transcription. For the model descriptions, *sv 7* is the model used at the time of the thesis and *curated xls-r* is the xls-r pre-trained model and extitcurated voxpopuli is the voxpopuli pre-train model both fine-tuned on the cured set of JRCC data. For the language models *J* is for JRCC, *N* for NATO phonetic alphabet, *V* for vessel combinations, *S* for Sjörapporten and *W* for Wikipedia.

A.2 English transcripts

all ships all ships all ships motorvessel volga traffic information regarding port of landskrona please listen channel ten one zero				
Model	LM	Transcription	WER	CER
en 2	None	all ships all ships all ships morvesskalsa traffic information regarding port asernan korskrona please listen channel ten one zero	0.200	0.130
base V	None	all ships all ships all ships motorvessel alti-crat traffic information regarding pot channel stana please listen channel ten one zero	0.200	0.145
large V	None	all ships all ships all ships motorvessel valsa traffic information regarding or asrantaskona please listen channel ten one zero	0.200	0.084
Robust	None	all ships all ships all ships motorvessel valsa traffic information regarding olst onundas-coma please listen channel ten one zero	0.200	0.099
en 2	L	all ships all ships all shipsovsskalsa trafficinfor-mation regarding portsernankorskrona please listen channel ten one zero	0.400	0.160
en 2	J	all ships all ships all ships motorvessel traf-fic information regarding port senankorskrona please listen channel ten one zero	0.150	0.099
en 2	JA	all ships all ships all ships motorvessel traffic information regarding port sernankorskrona please listen channel ten one zero	0.150	0.099
en 2	JASw	all ships all ships all ships motorvessel traffic information regarding port sernankorskrona please listen channel ten one zero	0.150	0.099

en 2	JASi	all ships all ships all ships motorvessel traffic information regarding port sernankorskrona please listen channel ten one zero	0.150	0.099
Robust	JASw	all ships all ships all ships motorvessel val traffic information regarding olstudascoma please listen channel ten one zero	0.200	0.122

Table A.21: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

? as you heard we cancelled the mayday and we just want to say thank you so much for your assistance				
Model	LM	Transcription	WER	CER
en 2	None	yes nowbric a you ahead we caesen thea mayday and t one te szele thank you so much for ou asistanc	0.619	0.380
base V	None	yis nobbt a dyu reade rea kupseans ve mady and retouns wone tof laie nink you somaug four your sipten	0.905	0.540
large V	None	yes nobit e as you head we katten v mayday and vithas one to sey thank you so much for your assistance	0.476	0.320
Robust	None	yes hog pek e as you aheard the wee a caseln v mayday and e w us one to day thank you so much for your assistance	0.714	0.370
en 2	L	yes nowbric youahead we caesentea mayday andtsone thank you so much for ou asistanc	0.667	0.420
en 2	J	yes nowbric a you ahead we can the mayday and arton ee thank you so much for your assistans	0.524	0.380
en 2	JA	yes nowbric a you ahead we can the mayday and one tel thank you so much for your assistans	0.524	0.380
en 2	JASw	yes nowbric you had we can the mayday and one tele thank you so much for our assistans	0.524	0.370
en 2	JASi	yes now bic you had we case the mayday and tone tele thank you so much for your assistans	0.524	0.350

Robust	JASw	yes hope as you heard the wea case v mayday and one today thank you so much for your assistance	0.524	0.330

Table A.22: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

all ships all ships all ships traffic information concerning port of falkenberg tugboat nadir with barge is entering port of falkenberg tugboat nadir with barge is approaching port of falkenberg				
Model	LM	Transcription	WER	CER
en 2	None	all ships all ships all ships traffic information concerning port of falkenberg tagbout nadi witbargeis entering port of falkenberg tgbout nadi witbarg is approaching port of falkenberg	0.300	0.062
base V	None	all ships all ships all ships traffic information concerning port of faltinber forgeboat nab wilt barg his entering port of falkinbeyg forebot nadb wil barg is a rauching port of falkin begde	0.500	0.180
large V	None	all ships all ships all ships traffic information concerning port of valkenberg fagbot nadir wilbarge is entering port of valkenberg fragbout nadi wilbarge is approching port of valkenberg	0.367	0.093
Robust	None	all ships all ships all ships traffic information concerning port of falkenberg tagbat nadir wilbarge is entering port of falkenberg tagboat nadir wilbarge is approching port of falkenberg	0.233	0.052
en 2	L	all ships all ships all ships traffic information concerning port of falkenberg tagbout nadi witbargeis entering port of falkenberg tgbout nadi witbargis approaching port of falkenberg	0.333	0.067
en 2	J	all ships all ships all ships traffic information concerning port of falkenberg tagbout nadi witbargeis entering port of falkenberg tugbout nadi witbarg is approaching port of falkenberg	0.300	0.057
en 2	JA	all ships all ships all ships traffic information concerning port of falkenberg tagbout nadi wisborg is entering port of falkenberg tagbout nadi witbarg is approaching port of falkenberg	0.267	0.072

en 2	JASw	all ships all ships all ships traffic information concerning port of falkenberg tagbout nadi wisborg is entering port of falkenberg tagbout nadi witbargis approaching port of falkenberg	0.300	0.077
en 2	JASi	all ships all ships all ships traffic information concerning port of falkenberg tagbout nadi wisborg entering port of falkenberg tagbout nadi witbargis approaching port of falkenberg	0.333	0.093
Robust	JASw	all ships all ships all ships traffic information concerning port of falkenberg tagbat nadir wilbarge is entering port of falkenberg tagboat nadir wilbarge is approching port of falkenberg	0.233	0.052

Table A.23: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

all station all station vinterland vinterland departure from husum departure from husum in about ten minutes ten minutes				
Model	LM	Transcription	WER	CER
en 2	None	all stations all station vinterland vinterland departure from husum departure from husum in about ten minutes ten minutes	0.056	0.008
base V	None	all station all station vinteland vintelande-partefrom husun departero husun in ebout can inutes sten minutes	0.667	0.158
large V	None	all stations all station vinteland vinteland departe from heyusumn departef from heysum in about ten minutes ten minutes	0.389	0.108
Robust	None	all station all station vintiland vintiland departuer from hysum departir from hysum in about ten minutes ten minutes	0.333	0.083
en 2	L	all stations all station vinterland vinterland departure from husum departure from husum inabout ten minutes ten minutes	0.167	0.017
en 2	J	all stations all station vinterland vinterland departure from husum departure from husum in about ten minutes ten minutes	0.056	0.008

en 2	JA	all stations all station vinterland vinterland departure from husum departure from husum in about ten minutes ten minutes	0.056	0.008
en 2	JASw	all stations all station vinterland vinterland departure from husum departure from husum in about ten minutes ten minutes	0.056	0.008
en 2	JASi	all stations all station vinterland vinterland departure from husum departure from husum in about ten minutes ten minutes	0.056	0.008
Robust	JASw	all station all station vintiland vintiland departure from husum depart from husum in about ten minutes ten minutes	0.167	0.058

Table A.24: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

baltic tern baltic tern five bravo sierra bravo four east cost pilot on channel sixteen				
Model	LM	Transcription	WER	CER
en 2	None	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057
base V	None	baltic tane baltic tan five bravo yela bravo four lest coast pilot on channel sixtee	0.400	0.149
large V	None	baltic tan baltic tan five bravo sierra gravo four east coast pilot on channel sixteen	0.267	0.069
Robust	None	baltic tan baltic tan five bravo siera bravo four east coast pilot on channel sixteen	0.267	0.069
en 2	L	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057
en 2	J	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057
en 2	JA	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057
en 2	JASw	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057
en 2	JASi	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057

Robust	JASw	baltic tan baltic tan five bravo sierra bravo four east coast pilot on channel sixteen	0.200	0.057

Table A.25: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

aastind aastind aastind callsign zulu delta romeo juliett nine lyngby radio is calling				
Model	LM	Transcription	WER	CER
en 2	None	aursleat arrach arsre callsign zulo delta romeo juliet nine lyngby radio s calling	0.462	0.233
base V	None	aul sleapv ou swagt ous sreapf call sign zono delnta romeo juliet nine dengby radioh calling	1.077	0.407
large V	None	austrikte aostre astreet call sign zolo denta romeo juliet nine lyngby radio is calling	0.615	0.221
Robust	None	aostrev aostrea ostreat call sign zulu delta romio juliea nine lyngby radio is calling	0.538	0.209
en 2	L	aursleat arrach arsre callsign zulo delta romeo juliet nine lyngby radio calling	0.462	0.256
en 2	J	area area are callsign zulo delta romeo juliet nine lyngby radio calling	0.462	0.267
en 2	JA	area area are callsign zulo delta romeo juliet nine lyngby radio calling	0.462	0.267
en 2	JASw	area area are callsign zulo delta romeo juliet nine lyngby radio calling	0.462	0.267
en 2	JASi	area area are callsign zulo delta romeo juliet nine lyngby radio calling	0.462	0.267
Robust	JASw	astrea astrea astrea call sign zulu delta romeo julia nine lyngby radio is calling	0.462	0.186

Table A.26: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

charlotta b charlotta b east coast pilot channel one six				
Model	LM	Transcription	WER	CER

en 2	None	chelotta b hilotta b east coast pilot channel one six	0.200	0.089
base V	None	collostsa ben pilostsabe reast coast filet channe one six	0.700	0.321
large V	None	chelotta b celotta b east coast pilot channe one six	0.300	0.107
Robust	None	calota bee calota be east coast pilot channel one six	0.400	0.161
en 2	L	chelottab hilottab east coast pilot channel one six	0.400	0.125
en 2	J	chelottab hilottab east coast pilot channel one six	0.400	0.125
en 2	JA	chelottab hilottab east coast pilot channel one six	0.400	0.125
en 2	JASw	chelottab hilottab east coast pilot channel one six	0.400	0.125
en 2	JASi	chelottab hilottab east coast pilot channel one six	0.400	0.125
Robust	JASw	calotabee calota be east coast pilot channel one six	0.400	0.179

Table A.27: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

lyngby radio lyngby radio professor logachev come in please yes this is lyngby radio eh you do not read me on the channels eh i don't know why what is you call sign please my call sign is uniform alpha delta zulu uniform alpha delta zulu stand by please ok				
Model	LM	Transcription	WER	CER
en 2	None	lyngby radio lyngby radia prefesseor lga tjhor come in please yes this is lyngby radio you do lot reakd me on dhe channels i dot now wi what is jior call sign please i casl signs uniform oulfa delta zulu yuni form alfa delta zulu stand by please okey	0.440	0.160

base V	None	lyngby radin lingby radio motor pretes r lgathio come ing please yes this is lyngby radio e you oloth e reae me on dhetho channel i two now wive wht eas yiou calt signe please fin coall siges jouni form aoufa zdelta zelo youliform alfer delta zulos steand bive please tokardy	0.720	0.316
large V	None	ahlyngby radio lyngby radioh prefesser wilga chuf come in please yes this is lyngby radio ehh you do lot red mee on ther the channels i dont now wiv what is your call sign please ye call sign uniform alpha delta zulu uniform alfha delta zulu stand byde please okey	0.400	0.148
Robust	None	lyngby radio lyngby radio prefesser ulga cho come in please yes this is lyngby radio e you donot read me on th channels i dont no wiy what is your callsign please my call sign is yuniform alfa delta zulu uniform alfha delta zeulu stand by please okey	0.380	0.102
en 2	L	lyngby radio lyngby radia prefessor lgatjhor come in please yes this is lyngby radio you dolot reakd me ondhe channels i dot now wi what is jior call sign please casl signs uniform oulfa delta zulu uniform alfa delta zulu stand by please okey	0.420	0.156
en 2	J	lyngby radio lyngby radio prefessor lgatjhor come in please yes this is lyngby radio you dont read me on the channels i dont now wi what is jior call sign please i call sign uniform alfa delta zulu uniform alfa delta zulu stand by please okey	0.300	0.125
en 2	JA	lyngby radio lyngby radio prefessor lgatjhor come in please yes this is lyngby radio you dont read me on the channels i dont now wi what is jior call sign please i call sign uniform alfa delta zulu uniform alfa delta zulu stand by please okey	0.300	0.125
en 2	JASw	lyngby radio lyngby radio prefessor lgatjhor come in please yes this is lyngby radio you do not read me on the channels i dont now wi what is jior call sign please i call signs uniform alfa delta zulu uniform alfa delta zulu stand by please okey	0.280	0.113

en 2	JASi	lyngby radio lyngby radio prefessor olga thor come in please yes this is lyngby radio you do not read me on the channels i dont now wi what is jior call sign please i call signs uniform alfa delta zulu uniform alfa delta zulu stand by please okey	0.300	0.117
Robust	JASw	lyngby radio lyngby radio prefesserulgachof come in please yes this is lyngby radio you do not read me on the channels i dont no wi what is your call sign please my call sign is uniform alfa delta zulu uniform alfa delta zulu standby please okey	0.260	0.098

Table A.28: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

yeah i have some small problem monitoring channel two four but pilot for you at gavle pilot station five o clock in the morning zero five zero zero				
Model	LM	Transcription	WER	CER
en 2	None	yeah halsom smol problem maritaring channel two four but pilot for you at javle pilot station five aklasskuin morning zero five zero zero	0.357	0.156
base V	None	yeh halsoms mar trobland morneecalling channel two four bad pitort tujurat yjavle pilot station fie va claso morning zero five zero eigh	0.679	0.347
large V	None	yeh halsom snall problem monitoring channel two four but pilot for yus javle pilot station five acalaskena morning zero five zero ziro	0.464	0.170
Robust	None	yeah havsum small probblem monitoring channel two four but pilot for yw at javla pilot station five oclock in the morning zero five zero six	0.321	0.109
en 2	L	yeah halsomsmol problem maritaring channel two four but pilot for you at javle pilot station fivearklaskuin morning zero five zero zero	0.393	0.163
en 2	J	yeah halsomsmol problem maritaring channel two four but pilot for you at javle pilot station five arklaskuin morning zero five zero zero	0.357	0.156

en 2	JA	yeah halsomsmol problem maritaring channel two four but pilot for you at javle pilot station five arklaskuin morning zero five zero zero	0.357	0.156
en 2	JASw	yeah halsomsmol problem maritaring channel two four but pilot for you at gave pilot station five arklaskuin morning zero five zero zero	0.357	0.156
en 2	JASi	yeah halsomsmol problem maritaring channel two four but pilot for you at gave pilot station five arklaskuin morning zero five zero zero	0.357	0.156
Robust	JASw	yeah have som small problem monitoring channel two four but pilot for you at gala pilot station five oclock in the morning zero five zero six	0.214	0.068

Table A.29: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

stortebeker zulu delta echo golf seven this is this is swedish navy control over swedish navy control motor vessel stortebeker good morning sir question if we proceed in deep in two ports of sweden from karlhamn to holmsund can i use inshore traffic zone south of oland island swedish navy control go to traffic channel twelve one two over one two

Model	LM	Transcription	WER	CER
en 2	None	stor reae zulu delta eco golf seven this is sulien this is swedish navy controll over swedish navvy control motorvessel stor depeter good morning sir atoves can wey iffray proeceding bittn two port of ivden from karral harmoun to halmsuund skena you ixgesihor traffic zound sus of aland ihland swedish navych control go to traffic channel tralego one two over one two	0.508	0.242
base V	None	stort brejhe so dherdelta ecco kall seven this is seden this is swedyh navy control over shoitish nain control movessel strumnd thffacceir vy good morning sir to vassigh chan infrae pocedin belteentwo port ofwidhin prompbakaral harmen to havmsund gena you spin tor traffic on souvs of aland ishland this swhedish navy tholmn control go to traffic channel ber trol tone ywoo over one to	0.721	0.380

large V	None	stor ratier zulu delta ecogolf seven this is zulim this is swedish navy control over swedish navy control motorvessel strond debeter good morning sir to ston efy proceaeding btn two port offfic in from ckarof harmen to falmsund can yus in short traffic zons south of aland ighland swedish navyc control go to traffic channeleh tvel one jw over one two	0.508	0.199
Robust	None	stor rescor zuludelta ecco golf seven this is suln this is swedish navy control over swedish navy control motorvessel sterndte pictor good morning sir todescan if we proced ing btwn to portes of sweden from karols harmund to halsund ka a youcuri shour traffic zound south of aland iland swedish navy trau control go to traffic channel twelv one two over one two	0.443	0.196
en 2	L	stor reae zulu delta eco golf seven this is sulen this is swedish navy controll over swedish navvy control motorvessel stordepeter good morning siratovss canwey iffray proeceding-bittn two port of ivden from karral harmoun to halmsuund skena youiemsihor trafficzound of aland ihland swedish navycho control go to traffic channel tralego one two over one two	0.541	0.245
en 2	J	stor read zulu delta eco golf seven this is sulum this is swedish navy control over swedish navy control motorvessel stordepeter good morning sir at can we iffray proceding bit to port of den from karral harmoun to halmsuund skena you echo traffic sound us of aland island swedish navy control go to traffic channel tralego one two over one two	0.443	0.225
en 2	JA	stor read zulu delta eco golf seven this is sulum this is swedish navy control over swedish navy control motorvessel stordepetery good morning sir at can we if ray preceding bit to port of den from karral harmoun to halmsuund skena you echo traffic sound us of aland island swedish navy control go to traffic channel tralego one two over one two	0.443	0.228

en 2	JASw	sto reader zulu delta eco golf seven this is sulum this is swedish navy control over swedish navy control motorvessel store peter good morning sir arts can we if fry preceding bit to port of den from karral harmoun to halmsuund skena you sensor traffic sound sus of aland island swedish navy control go to traffic channel tralego one two over one two	0.443	0.216
en 2	JASi	sto reader zulu delta eco golf seven this is sulum this is swedish navy control over swedish navy control motorvessel store peter good morning sir arts can we if fray preceding bit to port of widen from karral harmoun to halmsuund skena you sensor traffic sound sus of aland island swedish navy control go to traffic channel tralego one two over one two	0.443	0.216
Robust	JASw	start rescue zulu delta eco golf seven this is so this is swedish navy control over swedish navy control motorvessel store picture good morning sir does if we proceeding been to port of sweden from cars harmond to halmsund kana your sour traffic sound south of aland iland swedish navy trau control go to traffic channel twelve one two over one two	0.410	0.179

Table A.30: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

golfstraum golfstraum north coast pilot channel one six				
Model	LM	Transcription	WER	CER
en 2	None	golf stroaumn golfstroumn north coast pilot channel one six	0.375	0.091
base V	None	goltf stram golstram nort ost pilot channel one six	0.625	0.145
large V	None	golf stroum golf stroum north coast pilot channel one six	0.500	0.073
Robust	None	golf stram golfstraum north coast pilot channel one six	0.250	0.036
en 2	L	golf stroaumn golfstroumn north coast pilot channel one six	0.375	0.091

en 2	J	golf stroaumn golfstroumn north coast pilot channel one six	0.375	0.091
en 2	JA	golf stroaumn golfstroumn north coast pilot channel one six	0.375	0.091
en 2	JASw	golf stroaumn golfstroumn north coast pilot channel one six	0.375	0.091
en 2	JASi	golf stroaumn golfstroumn north coast pilot channel one six	0.375	0.091
Robust	JASw	golf stram golf stram north coast pilot channel one six	0.500	0.073

Table A.31: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

jette teresa jette teresa this is swedish warship channel one six over				
Model	LM	Transcription	WER	CER
en 2	None	yett criessa yett triessa this is swedish warship channel one six over	0.333	0.157
base V	None	yaste tresa yeste tresa this is swedish war ship channel one six over	0.500	0.114
large V	None	yeste tresa yeste tresa this is swedish warship channel one six over	0.333	0.086
Robust	None	yest criesa yest tresa this is swedish wareship channel one six over	0.417	0.157
en 2	L	yett criessa yett triessa this is swedish warship channel one six over	0.333	0.157
en 2	J	yett triessa yett triessa this is swedish warship channel one six over	0.333	0.143
en 2	JA	yett triessa yett triessa this is swedish warship channel one six over	0.333	0.143
en 2	JASw	yett criessa yett triessa this is swedish warship channel one six over	0.333	0.157
en 2	JASi	yett criessa yett triessa this is swedish warship channel one six over	0.333	0.157
Robust	JASw	yest carisa yet tresa this is swedish warship channel one six over	0.333	0.143

Table A.32: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

passenger vessel athena seaways athena seaways this is motortanker azerbaijan calling you please				
Model	LM	Transcription	WER	CER
en 2	None	pacsancjur vessel astena iway astena iway this is motortanke ninsibattian calling yjou please	0.615	0.281
base V	None	asteanier vessel astena huy atena chuy his is motor tanker nake bastiean calling you please	0.769	0.281
large V	None	passanjer vessel atena eaway atena ceaway hits is motor tanker hace bastian calling you please	0.769	0.177
Robust	None	passenjour vessel astena sceiway astena sciway this is motor tanker ance bascia calling good please	0.769	0.250
en 2	L	pacsancjur vessel astena iway astena iway this is motortanke ninsibattian calling jou please	0.615	0.281
en 2	J	pacsancjur vessel astena iway astena iway this is motortanker ninsibattian calling you please	0.462	0.260
en 2	JA	pacsancjur vessel astena iway astena iway this is motortanker ninsibattian calling you please	0.462	0.260
en 2	JASw	pacsancjur vessel astena iway astena iway this is motortanker nisibattian calling you please	0.462	0.250
en 2	JASi	pacsanjur vessel astena iway astena iway this is motortanker insibattian calling you please	0.462	0.240
Robust	JASw	passenjourvessel astena sceway astena sciway this is motor tanker ancebascia calling good please	0.769	0.260

Table A.33: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

station calling karlshamn v t s go ahead god evening sir this is motor tanker kir calling you sir my e t a to anchorage positon a alpha is around one hour sir in one hour time i will be on anchorage position alpha very good thank you for that information please call when you have dropped the anchor ok sir i call you back when we dropped anchor sir very good thank you have a good watch sir channel ten

Model	LM	Transcription	WER	CER
en 2	None	station calling carll sin viti es go ahead ood evening sir this is motor tanker kiif calling you sir ahnd my hit ay to anckoerage position a alfa isaround lone ourt sird in one our time iwl b on ankerage position alfa verry good tank yufo trafc information please comet on your troppitank your okey siar ihave callgyour bukan you ed r ove dankerd ser very gouaod thank you ir good uwahad sire listen channel twen	0.654	0.300
base V	None	station calling carl sungverte as gou aheaden sevening sir this is motor tanker kees calling yousir my hitty e to unker t position ei poufar is ser ont one o wvertser in one o wertin o dhe on unker position poufar verguod thanern an-information please coale woneu ra gopthanker okey sir colber bare koing win d rotlankerd sar ver goud thank you sen eve bood versser listen o channel tennes	0.778	0.422
large V	None	station calling kallshimnd v t s go ahead good evening sir this is motortanker kees calling you sir ahhhhd mye t a to ankerage position a oufa issaround on ouver sird in one our time i wl be on ankurage position oufha very good thank yo fotrafi information please call thon our gropthank yur oke sir i hae call your bark en ou drove the ankoerd sir very good thank you r good wad sir listen t channel ten	0.519	0.256
Robust	None	station calling calsun vits go ahead good evening sir this is motortanker ke est calling jou sir e my t th tankerage position ei alfa its around one ouuer sir in one our time i wil be on tankerage position alfa very good thank yo othif information please call wen yoh drop tanker okey sir i hol call your bairk we you e drop ter anker sir very good thank you hay good wather sir listen channel ten	0.568	0.243

en 2	L	station calling callsin vities go aheadod evening sir this is motortanker kiif calling you sirahnd myitaytoanckoerage position alfa issaroundlone ourtsird in one our time iwldb onankerage position alfa verry good tank yufotrafcinformation please cometon your troppitank yourokey siar ihave callgyour bukhan youedrovedankerd ser very gouaod thank youir gooduwahad sirelisten channel twen	0.778	0.315
en 2	J	station calling call sign vities go ahead good evening sir this is motortanker kiif calling you sir and my it aytoanckoerage position alfa isaroundlone our sir in one our time i on ankerage position alfa very good thank you trafic information please come on your troppitank your okey sir i have call your bukhan you edrovedankerd ser very good thank you good uahad sir listen channel ten	0.593	0.290
en 2	JA	station calling call sign vities go ahead good evening sir this is motor tanker k calling you sir and my it aytoanckoerage position alfa isaroundlone our sir in one our time i on ankerage position alfa very good thank you trafic information please come on your drop tank your okey sir i have call your bukan you edrovedankerd ser very good thank you good uahad sir listen channel ten	0.568	0.288
en 2	JASw	station calling call sign vities go ahead good evening sir this is motor tanker ki calling you sir and my it aytoanckoerage position alfa isaroundlone our sir in one our time i on ankerage position alfa very good thank you trafikinformation please come on your drop tank your okey sir i have call your bukan you edrovdankerd ser very good thank you sir good uahad sir listen channel ten	0.580	0.288

en 2	JASi	station calling call sign vit es go ahead good evening sir this is motortanker ki calling you sir and my it aytoanckoerage position alfa issaroundlone our sir in one our time i on ankrage position alfa very good thank you trafikinformation please come on your drop-pitank your okey sir ihave calligour bukan you edrovedankerd ser very good thank you sir good uahad sir listen channel ten	0.617	0.288
Robust	JASw	station calling carson vts go ahead good evening sir this is motortanker keys calling you sir ah my it the tanker position i alfa its around one our sir in one our time i will be on tanker position alfa very good thank you this information please call wen you drop tanker okey sir i call you back we drop tracer sir very good thank you good weather sir listen channel ten	0.457	0.238

Table A.34: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

tina tina hammershus on channel sixteen				
Model	LM	Transcription	WER	CER
en 2	None	stina stna harmonshus on channel sixteen	0.500	0.154
base V	None	pina ine hamorsus son channel sixteen	0.667	0.179
large V	None	tina tina hammershus on channel sixteen	0.000	0.000
Robust	None	tina tina hammershus on chanel sixteen	0.167	0.026
en 2	L	stina stna harmonshus on channel sixteen	0.500	0.154
en 2	J	stena stna harmonshus on channel sixteen	0.500	0.179
en 2	JA	stina stina harmonshus on channel sixteen	0.500	0.128
en 2	JASw	stina stina harmonshus on channel sixteen	0.500	0.128
en 2	JASi	stina stina harmonshus on channel sixteen	0.500	0.128
Robust	JASw	tina tina hammershus on channel sixteen	0.000	0.000

Table A.35: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

motorvessel aalborg aalborg this is vilnia maersk calling				
Model	LM	Transcription	WER	CER
en 2	None	motorvessel olbor olbor this is vilni emershe calling	0.500	0.193
base V	None	motoer vessel rolbor rolbor this is willni ammersk calling	0.750	0.228
large V	None	motorvessel ollbor ollbor this is wilnieamerce calling	0.500	0.211
Robust	None	motorvessel ulbo olbo this is willne mersce calling	0.500	0.263
en 2	L	motorvessel olbor olbor this is vilni emershe calling	0.500	0.193
en 2	J	motorvessel olbor olbor this is vilni emershe calling	0.500	0.193
en 2	JA	motorvessel olbor olbor this is vilni emershe calling	0.500	0.193
en 2	JASw	motorvessel olbor olbor this is vilni mere calling	0.500	0.175
en 2	JASi	motorvessel olbor olbor this is vilni mere calling	0.500	0.175
Robust	JASw	motorvessel ulbo olbo this is wine mere calling	0.500	0.263

Table A.36: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

good evening sir please can we go to channel sixty four channel six four				
Model	LM	Transcription	WER	CER
en 2	None	de liza denvig to cam ve go to channel sixty-four channel six four	0.571	0.306
base V	None	yerni inter lsicke com mydot cannel sixtyoe fourt channel six four	0.786	0.403
large V	None	do lnednka this ou can we go to channel sixty forvcht channel six four	0.357	0.319
Robust	None	e vinis danish can we go to channel sixtyfoure channel six four	0.429	0.250

en 2	L	delizadenvi to camve go to channel sixtyfour channel six four	0.571	0.319
en 2	J	elida eni to com we go to channel sixtyfour channel six four	0.500	0.292
en 2	JA	elida eni to com we go to channel sixtyfour channel six four	0.500	0.292
en 2	JASw	elida envy to cam we go to channel sixtyfour channel six four	0.500	0.292
en 2	JASi	eliza devi to cam we go to channel sixtyfour channel six four	0.500	0.306
Robust	JASw	is danish can we go to channel sixty four channel six four	0.286	0.264

Table A.37: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

arietta arietta supply boat server				
Model	LM	Transcription	WER	CER
en 2	None	aenta arlenta repplyboart server	0.800	0.265
base V	None	aieta aieta supply boat serve	0.600	0.147
large V	None	arietta aljetta supply boat server	0.200	0.059
Robust	None	arietta arietta supplyboat server	0.400	0.029
en 2	L	aenta alenta repplyboart server	0.800	0.294
en 2	J	aenta alenta supplyboart server	0.800	0.235
en 2	JA	aenta alenta supplyboart server	0.800	0.235
en 2	JASw	aenta alenta supplyboart server	0.800	0.235
en 2	JASi	aenta alenta supplyboart server	0.800	0.235
Robust	JASw	arietta arietta supplyboat server	0.400	0.029

Table A.38: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

eh vadero islander moezelborg moezelborg vadero islander good afternoon sir in two miles south i will drop anchor because i can't pass it to much there i have a drop seven point four eh thats not so good like then you have to eh we are not able to bunker you in that position in this weather but come to channel seven seven please seven seven

Model	LM	Transcription	WER	CER
en 2	None	vadero hillander mserbugboselbu madero highlander o good eveteing sir ina to minedes south ville drobp enker because ican tresseg to muche dir yo have draft seven point four tat northtor gut pliyd chan you have to run-nolr teybal to bunker you on dhet bortition ing bystadl but coptor channel seven seven please seven seven	0.652	0.342
base V	None	vad eref helander melrbog moterl vary wand er ighlandes eh god adentsir a in to mintes a south eh hel derop enker ficcoas eccen passic to mixeder accawy derapft seven point four hh des motsorgod slye san yea two eee roa nof te bol two buink ker un thest of sitton ing ys graddof bat capsto channel seven sevean please seven zevo	0.894	0.447
large V	None	vadero hihlander meseburg vesselburg vadero highlander ehh good aften sir ehh ia two mila south ih willa dro enker picos i can procet to myeh dher i chave drapt seven point four e juas nothser good ply chen you have to e rono taboal to bunker jun thet pocristend ing bys pridlot put pop to channel seven seven please seven seven	0.652	0.322
Robust	None	vadero highlander mazeborgvoselborg vadero highlander eh good avedning sir a in a two mils south i wil drop teanker becoas i can paceve two mutce ther i hav v draft seven point four ah jusc not so good playge than you hav twoe wer not taibel to bunker yun thic potition in the sprader but come to channel seven seven please seven seven	0.576	0.263

en 2	L	vadero hillander mserbugboselbu madero highlanderogood eveteing sir ina to minedes south ville drobp enker because ican tresseg to muche dir have draft seven point fourtat northtor gutpliyd chan youahave to runnor teybal to bunkeryouon etbortitioning bystadlbut coptor channel seven seven please seven seven	0.758	0.363
en 2	J	vadero hillander merbugboselbur matero highlander good evening sir in to minedes south ville drop enker because i can tresseg to much dir have draft seven point four that north to got pliyd can you have to runner seybal to bunker you on det porttioning ystad but captor channel seven seven please seven seven	0.591	0.322
en 2	JA	vadero hillander merbugboselbur matero highlander good evening sir in to minedes south ville drop enker because i can present to much der have draft seven point four that north to got play can you have to runner bal to bunker you on det porttioningbystadl but captor channel seven seven please seven seven	0.591	0.319
en 2	JASw	vadero hillander merbugboselbur matero highlander good evening sir in to minedes south ville drop enker because i can pressing to much dir have draft seven point four that north to got pl can you have to runner seybal to bunker you on that portion in ystad but captor channel seven seven please seven seven	0.561	0.310
en 2	JASi	vadero hillander merbugboselbur matero highlander good evening sir in to minedes south ville drop enker because i can pressing to much dir have draft seven point four that north to go play can you have to runner seybal to bunker you on that portion in ystad but coptor channel seven seven please seven seven	0.561	0.307

Robust	JASw	vaderoc highlander mazoborgvoselborg vadero highlander eh good avedning sir ah in a two miles south i will drop tanker becoas i can pace two much there have draft seven point four ahh just not so good play than you have two e were not abe to bunker you this potitionintspader but come to channel seven seven please seven seven	0.500	0.246

Table A.39: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frquencies on v h f traffic channels				
Model	LM	Transcription	WER	CER
en 2	None	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frequencies an vhf traffic channels	0.147	0.018
base V	None	securite securite securite all stations all stations all stations this is sweden traffic wi repetition aof navigational warnings and baltic sea vether forecast please listen to medim frequencies or v h f traffic channels	0.176	0.036
large V	None	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea wether forecast please listen to medium frequencies or v h f traffic channels	0.088	0.014
Robust	None	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea wethr forecast please listen to medium frequencies wint we h f traffic channels	0.118	0.036
en 2	L	securite securite securite all stations all stations all stations this is sweden traffic with repetitionof navigational warnings and baltic sea weather forecast please listen tomedium frequencies an vhf traffic channels	0.265	0.027

en 2	J	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frequencies and vhf traffic channels	0.147	0.023
en 2	JA	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frequencies and vhf traffic channels	0.147	0.023
en 2	JASw	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frequencies and vhf traffic channels	0.147	0.023
en 2	JASi	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frequencies and vhf traffic channels	0.147	0.023
Robust	JASw	securite securite securite all stations all stations all stations this is sweden traffic with repetition of navigational warnings and baltic sea weather forecast please listen to medium frequencies win we h f traffic channels	0.088	0.023

Table A.40: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

server videborg videborg server yes good morning sir channel six nine six nine				
Model	LM	Transcription	WER	CER
en 2	None	server ibobo te degoye server yes good morning sir channele six nine six nine	0.308	0.154
base V	None	zeurfour zebo cort jis do golo selver yes good morlning sir channella six nine six nunin	0.769	0.346
large V	None	server slepo boring e deborio server yes good morning sir channel six nine six nine	0.308	0.154
Robust	None	server chico pori t deboi server yes good morning sir channel six nine six nine	0.308	0.141

en 2	L	server ibobo degoye server yes good morning sir channele sixnine six nine	0.385	0.154
en 2	J	server b t degoye server yes good morning sir channel six nine six nine	0.231	0.154
en 2	JA	server iba t degoye server yes good morning sir channel six nine six nine	0.231	0.141
en 2	JASw	server bob te deg server yes good morning sir channel six nine six nine	0.231	0.154
en 2	JASi	server bob te deg server yes good morning sir channel six nine six nine	0.231	0.154
Robust	JASw	server hipoporitdeboi server yes good morning sir channel six nine six nine	0.154	0.128

Table A.41: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

ronne port this is express one express one ronne port to channel twelve one two				
Model	LM	Transcription	WER	CER
en 2	None	ronnie port this is express one express ronn ronne port channel swils one two	0.267	0.127
base V	None	honne port this is sexprs one expralos ronne ronne port channel tfive one two	0.400	0.190
large V	None	ronne port this is express one express ronne ronne port channel twivetw one two	0.200	0.114
Robust	None	ronne port this is express one expres rone ronneport channel sweve one two	0.400	0.101
en 2	L	ronnie port this is express one express ronn ronne port channel swils one two	0.267	0.127
en 2	J	ronnie port this is express one express one ronne port channel this one two	0.200	0.114
en 2	JA	ronnie port this is express one express one ronne port channel swiss one two	0.200	0.114
en 2	JASw	ronnie port this is express one express one ronne port channel swiss one two	0.200	0.114
en 2	JASi	ronnie port this is express one express one ronne port channel swiss one two	0.200	0.114

Robust	JASw	ronneport this is express one express one ronneport channel seve one two	0.400	0.101

Table A.42: Overview of the different model and language models impact on a given transcription. For the model descriptions, *en 2* is the model used at the time of the thesis and *V* is short for Voxpopuli. For the language models *L* is for Librispeech, *J* for JRCC, *A* for atco2, *Sw* for Switchboard and *Si* for Silicone.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY