1 Data Preparation

? Do you see anything missing in the text transcription? How would it affect our end goal of performing Speech Translation?

The transcript:

> CHAPTER ONE MISSUS RACHEL LYNDE IS SURPRISED MISSUS RACHEL LYNDE LIVED JUST WHERE THE AVONLEA MAIN ROAD DIPPED DOWN INTO A LITTLE HOLLOW FRINGED WITH ALDERS AND LADIES EARDROPS AND TRAVERSED BY A BROOK

Comparing the audio sample, and its corresponding transcript reveals, misrepresentation of vocal words. For example, the word *Mrs.* is misspelled as *MISSUS*. A non-standard spelling like *MISSUS* might not be as well-represented in the training data, which could lead to less accurate recognition if the system is not robust to such variations. Moreover, ASR systems use language models to predict the likelihood of sequences of words. If "MISSUS" is not common in the language model's training data, the system might be less likely to recognize and transcribe it correctly.

? Can you explain the shape you see?

Figure 2 shows a spectogram. The spectrogram depicts frequency on the vertical axis and time on the horizontal axis. A higher amplitude at a specific time and frequency is represented by a brighter color in the spectrogram. Time intervals (0-50, 350-500) without bright colors signify pauses in speech. Conversely, intervals with brighter colors indicate active speech. The content of the speech can be inferred based on the amplitudes of frequencies over time. Figure 1 illustrates an example where the spoken content is inferred.

? Can you record your own voice, extract the features and plot it in the same manner?

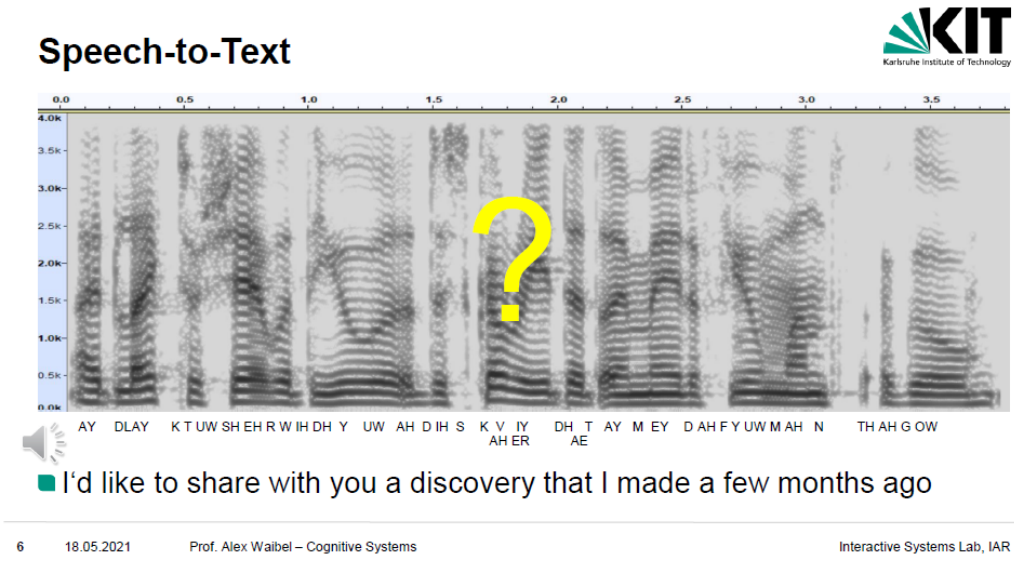


Figure 1: Spectogram with inference of the spoken language. Waibel [2021]

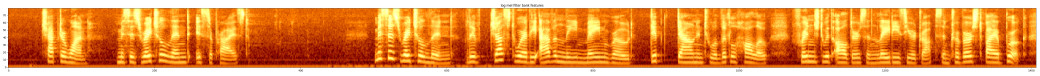


Figure 2: Log Mel Filter Bank Filters

1.1 Pattern Observation

1.2 Challenges of Using Whole Words

2 Training

? Why do we need a much larger max-tokens than when training a Machine Translation model?

? Can you again tell from the training log: the number of encoder layers, the number of decoder layers, embedding size, number of trainable parameters...?

2.1 Encoder

Number of encoder layers: 12 x TransformeEncoderLayers
Embeddings size: 256

2.2 Decoder

Number of decoder layers: 6 x TransformerDecoderLayerBase
Embeddings size: 256
Output feature size: 900

In total, there are 27,206,656 shared parameters in the model.

? What are the difference comparing to the Machine Translation model that we have seen in the previous lab? Can you tell why?

2.3 Sentence Pairs in Training and Dev Sets

2.4 Encoder and Decoder Layers

2.5 Vocabulary Size for Source and Target Languages

2.6 Embedding Table Dimensions

2.7 Model Parameters

2.7.1 Training Objective

2.7.2 Learning Rate Adaptation

2.7.3 Intuition Behind Learning Rate Adaptation

3 Evaluation

? Is WER suitable for languages such as Chinese or Janpanese?

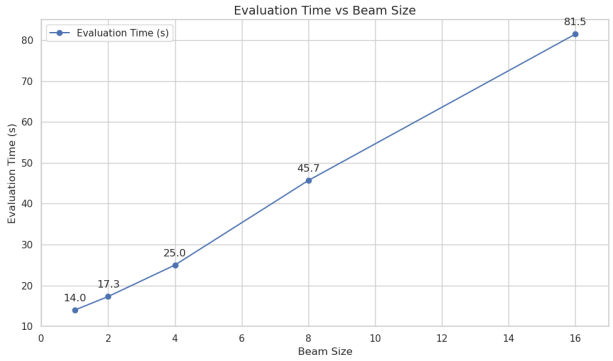
WER can be applied to languages like Chinese and Japanese, but it's important to note that the effectiveness of WER as a metric may vary depending on the specific characteristics of the language. Chinese and Japanese, for example, are character-based languages, and the segmentation of words is not as straightforward as in space-separated languages like English.

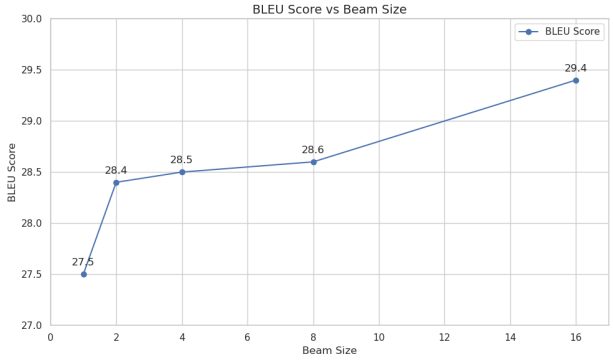
In Chinese, for instance, words are not separated by spaces, and characters represent meaningful units. Therefore, evaluating word-level error rates may not be as meaningful as it is in languages where words are clearly delimited by spaces.

For character-based languages like Chinese and Japanese, you might also encounter metrics like Character Error Rate (CER) or Subword Error Rate (SER), which can provide a more fine-grained analysis at the character or subword level. These metrics take into account the correctness of individual characters or subword units, which is often more appropriate for languages with complex writing systems.

In summary, while WER can be applied to Chinese and Japanese, it's essential to consider language-specific characteristics and, if necessary, explore alternative metrics such as CER or SER for a more accurate evaluation of ASR system performance.

Beam Size	BLEU Score	Evaluation Time
1	27.5	14.0s
2	28.4	17.3s
4	28.5	25.0s
8	28.6	45.7s
16	29.4	81.5s





4 Further Considerations

? Given the two models you've had by now (ASR and MT), what can you do next to have a working ST system?

For having a working ST system, next a cascaded speech translation could be applied. A cascaded ST system might involve several steps, such as:

Speech Recognition: The first stage involves converting spoken language into written text. This is known as speech recognition, and it's the process of transcribing spoken words into a textual representation.

Machine Translation: The transcribed text is then translated from the source language to the target language using machine translation techniques. This could involve neural machine translation (NMT) or other methods for language translation.

? Is there any drawbacks with this approach?

Common drawbacks associated with cascaded speech translation systems are:

Error Propagation: Errors made in one stage of the cascade can propagate to subsequent stages, leading to cumulative inaccuracies. For example, if the speech recognition stage misinterprets a spoken word, the translation and synthesis stages will be based on this error.

Latency: Cascaded systems may introduce additional latency as each stage requires processing time. This can be a concern in real-time applications where low latency is crucial, such as in live conversations or during simultaneous interpretation.

Complexity: Cascaded systems tend to be more complex to design, implement, and maintain compared to end-to-end systems. Managing multiple components increases the likelihood of issues arising, and integration challenges may be present.

Training Data Mismatch: Each stage of the cascade may be trained independently, leading to potential mismatches between the training data for different components. Mismatches can arise in terms of linguistic patterns, accents, or other characteristics, impacting overall performance.

Resource Intensive: Training and running multiple models for different stages of the translation process can be resource-intensive, requiring more computational power and storage compared to end-to-end systems.

4.1 Impact of ASR Errors on MT

4.2 Adaptation for Audio Inputs

5 Notes

5.1 Unusual Vocabulary Sizes

5.2 Tokens in Test and Dev Sets

5.3 Mismatch in Transformer Input Size

5.4 Discrepancy in Token Counts

5.5 Underutilized BPE Tokens

5.6 Extreme Train-Test Split

References

Prof Alex Waibel. Cognitive systems slides, 2021.