

Finstreder: simple and fast spoken language understanding with finite state transducers using modern speech-to-text models

Daniel Bermuth, Alexander Poeppel, Wolfgang Reif

Angaben zur Veröffentlichung / Publication details:

Bermuth, Daniel, Alexander Poeppel, and Wolfgang Reif. 2022. "Finstreder: simple and fast spoken language understanding with finite state transducers using modern speech-to-text models." arXiv. arXiv.
<https://doi.org/10.48550/arXiv.2206.14589>.

Finstreder: Simple and fast Spoken Language Understanding with Finite State Transducers using modern Speech-to-Text models

Daniel Bermuth, Alexander Poeppel, Wolfgang Reif

University of Augsburg, Institute for Software & Systems Engineering
{daniel.bermuth, alexander.poeppel, reif}@informatik.uni-augsburg.de

Abstract

In *Spoken Language Understanding* (SLU) the task is to extract important information from audio commands, like the intent of what a user wants the system to do and special entities like locations or numbers. This paper presents a simple method for embedding intents and entities into *Finite State Transducers*, and, in combination with a pretrained general-purpose *Speech-to-Text* model, allows building SLU-models without any additional training. Building those models is very fast and only takes a few seconds. It is also completely language independent. With a comparison on different benchmarks it is shown that this method can outperform multiple other, more resource demanding SLU approaches.

Index Terms: spoken language understanding, speech to intent, offline voice assistant, finite state transducer decoding

1. Introduction

When building models for *Spoken Language Understanding* (SLU), there are two alternative approaches: Using two separate stages of transcribing the spoken command to text (*Speech-to-Text*, STT) and then extracting the useful information out of the transcribed sentence (*Natural Language Understanding*, NLU), or using a direct SLU approach which combines the two parts into one single model. The first has the benefit, that the two models can be trained independently and the STT-model can often be used across multiple different domains, while the NLU-model can be trained relatively quickly. The second approach on the other hand is often more accurate, because it does not have the problem that errors from the STT transcription are propagated into the NLU module.

Recent systems for NLU or SLU parsing usually build upon neural networks as feature extractors. While they generally achieve a high recognition accuracy, the downside is that training those networks can take a lot of time. In this work a completely different approach is investigated, which does not require any special training, and therefore allows very fast creation of SLU models. It uses *Finite State Transducers* (FSTs) instead of neural networks for SLU parsing. The speech recognition part still uses a neural network, but this only has to be trained once per language on general-purpose STT tasks. In this work a *Quartznet* [1] model from previous work in *Scribosermo* [2], as well as a *Conformer* [3] model from *NeMo* [4], converted to *tensorflow-lite* with *Scribosermo*, are used. Both models, after conversion to *tfLite* and quantization, can run faster than real-time on a RaspberryPi4.

A comparison on multiple benchmarks shows that the performance of this approach is highly competitive with other solutions, despite the fact that the concept is very simple and the models are built in a few seconds.

The source-code of the presented method for training-free SLU parsing, named *finstreder*, as well as the models from *Scribosermo*, can be found at: <https://gitlab.com/Jaco-Assistant>

Using FSTs in speech recognition tasks is not a new idea and was quite common some years ago [5, 6]. FSTs are used by *Kaldi* [7], where a neural network outputs phonemes which are then decoded to sentences with a combination of multiple FSTs. Some works already explored the usage of FSTs for parsing NLU information by adding semantic tags into the FSTs, either from textual inputs [8, 9] or from speech transcription hypotheses of a hidden Markov model [10]. The authors of [11] explored how to use grammar fragments with embedded tags to improve word confusions. In [12] a dialog system is described which transforms textual user utterances into response sentences using weighted FSTs, with the goal to be able to run a full back-and-forth dialog with the users. It was extended by [13] to accept n-best hypotheses from a tri-phone model acoustic model, which were combined with an additional 3-gram language model, as input. *Eesen* [14] introduced FST-decoding to models outputting character-based *Connectionist Temporal Classification* (CTC) [15] labels, similar to the *Quartznet* model of *Scribosermo*. *Alexa* also uses FSTs for its skill kit, but keeps separate models for STT and NLU [16]. This work follows a very similar decoding approach as *Eesen*, which allows using recent CTC-based STT models (in difference to [10, 11, 13]), but alters the *Grammar-FST* (explained in the next chapters) to embed NLU information into it, similar to the semantic tagging of [9, 10, 11], which allows combining the two distinct STT+NLU models into a single SLU decoder. *OpenFST* [17] is used as library for handling the FSTs.

2. Foundations

In general, the language models used for decoding speech features with FSTs can be split into different parts:

(1) A *Grammar-FST*, conventionally denoted as G , which stores the information of complete sentences. Figure 1 shows a very simple grammar that can accept exactly two different sentences.

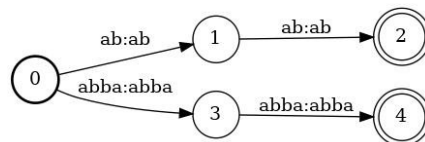


Figure 1: *Simple Grammar-FST* which accepts the sentences “*ab ab*” or “*abba abba*”. The part in the transition on the left side of the colon denotes the input which is required to make the transition and the one on the right side the output which is received afterwards.

(2) A *Lexicon-FST*, denoted as L , which builds words out of characters, as shown in Figure 2. For later optimization of the FST, a disambiguation symbol is required, which ensures that the path for “ ab ” doesn’t also accept “ $abab$ ”. Instead of introducing a special symbol, each word has to end with a space.

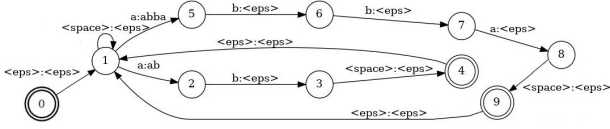


Figure 2: The Lexicon-FST for the words “ ab ” and “ $abba$ ”.

(3) In the case of CTC-Inputs, a *Token-FST* (T) which converts the frame-level CTC-labels to characters. CTC-labels are commonly used in STT-systems and contain a special $\langle blank \rangle$ label (here also written as “-”) besides the default alphabet characters. To generate normal text from those labels, all repeated tokens are merged into one single character, except they are separated by a $\langle blank \rangle$ label, which is removed after the merging step. Figure 3 shows a simple FST for the alphabet “ $\langle space \rangle, a, b, \langle blank \rangle$ ”. To allow the usage of *sentence-piece*-style labels [18] from the *Conformer* model, this FST was extended in a way that the pieces are split into single characters.

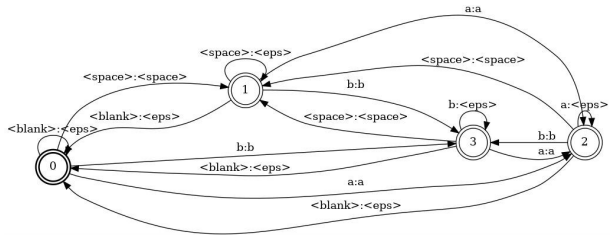


Figure 3: The Token-FST which merges CTC-labels like “ $aaab ab-b$ ” to the characters “ $ab abb$ ”.

Instead of using three single FSTs while decoding, they can be *composed* together. Afterwards an optimization step can be applied, which restructures the graph to remove unnecessary or duplicated transitions. Figure 4 shows the composition of the Lexicon-FST with Grammar-FST. The result can then be composed with the Token-FST, similar to the approach in *Eesen* [14], so that the composition can handle CTC-labels as input.

The input of the combined *TLG* model also has to be in form of a FST. Thus the step-by-step labels are converted into a state-by-state FST with one transition for each character between the two states of a timestep (Figure 5). The label probabilities in the range $[0-1]$, where higher is better, have to be converted so that a lower value is better. The normalization approach of [19] is used for this. Additionally, an extra timestep is added as last timestep, which has the space character as its highest probability, to ensure that the last word ends with a space (the disambiguation symbol of the lexicon).

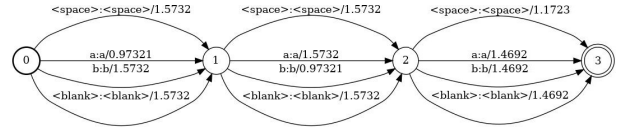


Figure 5: An Input-FST with the most probable path “ ab ”, ending with an extra space.

The reason behind the probability conversion is that after composition, the resulting FST includes all possible (acceptable) combinations of words that can be created with the given input. Since only the best matching result is desired, the shortest path algorithm from *OpenFST* can be used to search the path with the lowest transition weights (which came from the CTC-labels). The result is a new FST, which accepts the best matching input words (Figure 6). The decoding algorithm can then traverse this FST state-by-state and extract the output of each transition, which can then be merged into a returnable sentence.

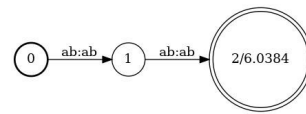


Figure 6: The Output-FST which accepts the sentence “ $ab ab$ ” as most probable (shortest) path, with the total path length in the last state.

3. Finstreder’s approach

In *Spoken Language Understanding*, the goal is to extract important features from a spoken utterance. In the case of users speaking voice commands, which is a common area for such applications, the software developers are mainly interested in the *intent*, the main goal of the command, and in special *entities/slots*, like names, locations or numbers.

This chapter shows how intent and entity information can be included directly in the Grammar-FST, with only slight adjustments. The decoder is then able to transcribe the text and extract the required information in one single step, thus removing the time-consuming step of training a specialized NLU-model.

3.1. Building the models

Building the new Grammar-FST is separated into four different steps. As input a *json*-file following the syntax structure of *Jaco* [20] is required:

```
{
  "intents": {
    "get-looks": [
      "(is a|are) [---] (animal) cute"
    ]
  },
  "lookups": {
    "animal": [
      "whitemargin stargazer",
      "atlantic stargazer",
      "aye aye",
      "(hairy frogfish)->striated frogfish"
    ]
  }
}
```

In the first step, a new text file for each intent is created, including all example sentences. The files are then converted to FSTs, either using *n-grams* or as *fixed* grammar. All final

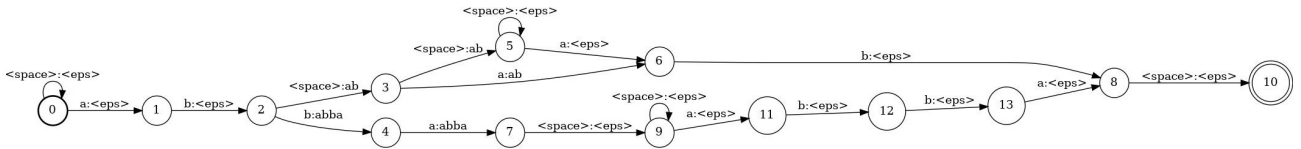


Figure 4: Optimized LG-FST which has characters as input and the sentences from the simple grammar above as output.

states are then forwarded using an *epsilon* transition to include the intent name directly into the FST, as shown in Figure 7.



Figure 7: Example Intent-FST with placeholder for entity values and the intent name at the end.

In the next step one FST for each slot is created. It also includes special symbols as slot markers and handles synonym replacements, by adding the synonym value after the actual value into the graph. See Figure 8.

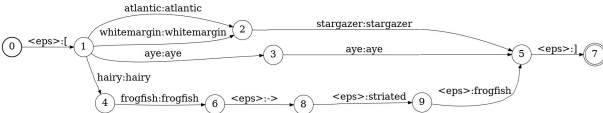


Figure 8: Example Slot-FST with special markers and synonym replacements (for the frogfish).

In the third step, the Slot-FSTs are inserted directly before their placeholders in the Intent-FSTs. The placeholder itself is kept, because it is later needed to determine the corresponding slot type. At last the merged Intent-FSTs are each composed with the Lexicon-FST.

3.2. Decoding

Decoding works as follows: First an Input-FST is built using the CTC-labels. Then it is composed with the Token-FST. Afterwards the resulting IT-FST is composed with each LG-FST of the different intents. The shortest path in each resulting FST is calculated and the overall shortest path is returned as result.

To improve and speed up decoding, several optimizations were implemented. Instead of uniting the LG-FSTs for each intent and composing it with the Token-FST into a single large TLG-FST, they are kept separate. Using multiple FSTs for the intents allows automatic parallelization with the IT-FST if there are many intents. This also allows including or excluding specific intents for each request. To exclude less necessary input characters, one parameter allows removing labels where the CTC-probability is below the *top-k* best values of each timestep. Another one removes all labels where the probability is below the mean value of the *k*-most-probable character over all timesteps (*mean-k*). Both parameters can greatly speed up decoding time with a very small impact on accuracy. A well working combination is *top-k* = 5-12 and *mean-k* = 21. For reweighing the input labels, a parameter was introduced that executes exponential scaling of the probabilities in a timestep, which increases or decreases the relative difference between label probabilities. A second parameter can be used for linear

scaling of the input weights versus the grammar weights from the *ngram* model.

4. Experiments

The performance of *finstred* was evaluated in multiple benchmarks. The first three benchmarks were reused from previous work in the *Jaco* project [20]. Since the *Quartznet* model was already used in *Jaco*, combined with NLU models trained with *Rasa* [21], the experiments with it allow a direct comparison between the two semantic parsing methods. The STT models (*QuartzNet-15x5/ConformerCTC-L*) reach a greedy *Word Error Rate* (WER) of 4.6/2.2% on *LibriSpeech* [22] in English and 17.5/8.0% on *CommonVoice* [23] in French. All of the benchmark code is released in the same repository as this work.

4.1. Spoken Language Understanding

The *Barrista* benchmark was published by *Picovoice* [24] and consists of 620 commands of different people ordering coffee in English. The audio is mixed with different volume levels of background noise from cafe and kitchen environments. An example of a command would be: “i’d like a [medium roast] [large] [mocha] with [lots of cream] and [a little bit of brown sugar]”. A command is correctly detected if the intent, as well as all the slots, could be retrieved by the assistant. This metric will be used for the following benchmarks as well.

The results in Figure 9 show that *finstred* with the *Quartznet* model performs as well as *Jaco* and *Watson*, and, like *Jaco*, has some problems with noisy backgrounds, which is most likely related to poorer recognition in the acoustic model. In comparison to that, using the *Conformer* model improves the results, especially in noisy environments.

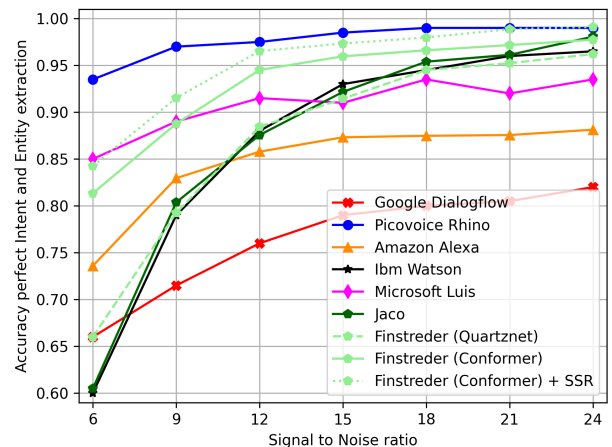


Figure 9: Benchmark coffee orders with noisy backgrounds. The results of *DialogFlow*, *Watson*, *Luis* and *Rhino* have been taken from [24], the results of *Alexa* and *Jaco* from [20].

Because the SLU model is built only from text files, it is very simple to add frequent transcription errors as synonyms into the dialog definition. After adding similar sounding synonyms as replacements (+SSR) for the three most common errors (*(rose|roast)*, (*ons|ounce*), (*ice moka|iced mocha*)), and rebuilding the SLU model, the accuracy improved notably.

The *SmartLights* benchmark from *Snips* [25] tests the capability of controlling lights in different rooms. It consists of 1660 requests which are split into five partitions for a 5-fold evaluation. A sample command could be: “*please change the [bedroom] lights to [red]*” or “*i’d like the [living room] lights to be at [twelve] percent*”. The benchmark results are presented in Table 1. The performance of *finstreder* with the *Quartznet* model is on-par with the two-step approach of *Jaco* in this benchmark, as well as with the E2E-SLU approach of *AT-AT* [26], and outperforms them with the *Conformer* model. *Snips* [27], which was a voice assistant, but is not available anymore, uses *Kaldi* as STT module and their own NLU module. The models were trained on online servers and could then be downloaded, which enabled the assistant to decode the voice requests without an internet connection. *Lugosch et al.* [28] use the features of a pretrained STT-network and add a SLU-decoder on top of it. Their model is then finetuned with a combination of real and synthetic speech data. *AT-AT* [26] uses two single encoder networks for audio and text inputs, but combines them in a single decoder network, with the advantage that the decoder can be finetuned with text only data and the network can also apply the learned knowledge to audio inputs.

Table 1: Results on *SmartLights* dataset.

	Accuracy	WER
<i>Google</i> [25]	0.793	–
<i>Snips</i> [25]	0.842	–
<i>Alexa</i> [20]	0.792	–
<i>Houndify</i> [20]	0.545	0.108
<i>Jaco</i> [20]	0.854	0.108
<i>Finstreder (Quartznet)</i>	0.848	0.107
<i>Finstreder (Conformer)</i>	0.880	0.061
<i>AT-AT</i> [26]	0.849	–
<i>Lugosch et al.</i> [28]	0.714	–

The *SmartSpeaker* benchmark tests the performance of reacting to music player commands in English as well as in French. The benchmark is from *Snips* [25], too, and is the only one that could be found which includes a language other than English. It has the difficulty of containing many artist or music tracks with uncommon names in the commands, like “*play music by [a boogie wit da hoodie]*” or “*I’d like to listen to [Kinokoteikoku]*”. As shown in Table 2, using *finstreder*’s SLU approach greatly improves accuracy compared to *Jaco*’s STT+NLU concept. This could partially be explained by a different handling of the artist names in the *ngram* language models. The *ngrams* of *Jaco* directly include the names, which allows the parser to leave out some parts of the names or mix them up, whereas the ones of *finstreder* only have a placeholder which then has to be matched exactly.

In the *TimersAndSuch* benchmark [29] common use-cases involving numbers are tested. It includes commands like “*set an alarm for 9:24 a.m.*” or “*compute 12.15 plus 26.9*”. The main difficulty is to recognize many different numbers, there are only very few command prefixes (like “*set an alarm for*”, ...), there-

Table 2: Accuracy on *SmartSpeaker* dataset.

	English	French
<i>Snips</i> [25]	0.687	0.751
<i>Google</i> [25]	0.478	0.423
<i>Jaco</i> [20]	0.627	0.480
<i>Alexa</i> [20]	0.455	0.889
<i>Finstreder (Quartznet)</i>	0.776	0.778
<i>Finstreder (Conformer)</i>	0.804	0.783

fore a *fixed* grammar model is used instead of a *2-gram* model, which showed a better performance in the other benchmarks. The results in Table 3 show that *finstreder* performs generally well in this task too, and could outperform the benchmark’s large baseline SLU model, which was trained specifically for this task, as well as the model from [30], which used the baseline’s architecture but included additional unsupervised training for the model’s encoder. A test with *Alexa* was skipped, because *Alexa*’s built-in number entity did not understand numbers with decimals.

Table 3: Accuracy on *TaS* dataset.

<i>TaS-baseline</i> [29]	0.816
<i>SpeechBrain</i> [30]	0.940
<i>Finstreder (Quartznet)</i>	0.900
<i>Finstreder (Conformer)</i>	0.954

FluentSpeechCommands [31] tests simple voice assistant requests. It includes commands like “*turn up the [bathroom] temperature*”, “*switch the lights on*” or “*go get me my [shoes]*”. The benchmark is run with a *fixed* grammar model, too. The results can be found in Table 4. *Kim et al.* [32] are combining a textual *BERT* model with a *vq-wav2vec-BERT* model and a *DeepSpeech2* acoustic model to a large SLU end-to-end network. This followed the idea of knowledge distillation from the text model to the speech encoder during training.

Even though the focus of this work is on training-free SLU, it is of course possible to finetune the acoustic model with *Scribosermo* on the used dataset. After a short training (about 1:20 h on a single RTX2070) the model (+*AMT*) reaches state-of-the-art performance.

Table 4: Accuracy on *FSC* dataset.

<i>Alexa</i>	0.987
<i>FSC-baseline</i> [29]	0.988
<i>Cao et al.</i> [33]	0.990
<i>FANS</i> [34]	0.990
<i>Reptile</i> [35]	0.992
<i>Finstreder (Quartznet)</i>	0.992
<i>Saxon et al.</i> [36]	0.994
<i>AT-AT</i> [26]	0.995
<i>Finstreder (Conformer)</i>	0.995
<i>Borgholt et al.</i> [37]	0.996
<i>Seo et al.</i> [38]	0.997
<i>Qian et al.</i> [39]	0.997
<i>Kim et al.</i> [32]	0.997
<i>Finstreder (Quartznet) + AMT</i>	0.997

4.2. Textual inputs

Instead of decoding CTC-labels which were predicted from an audio input, it is also possible to use the generated LG-FSTs for

NLU extraction from textual inputs. A very simple approach, which was tested here, is to convert the textual input to CTC-labels. Table 5 shows that NLU parsing with *finstredet* can outperform traditional approaches on simple datasets (in *SmartSpeaker* only the artist needs to be extracted), but falls behind if the datasets are more complex (see next chapter for explanation).

Table 5: NLU only test with textual inputs.

	SmartSpeaker		SmartLights	
	Accuracy	WER	Accuracy	WER
<i>Jaco (Rasa)</i>	0.977	—	0.960	—
<i>Finstredet</i>	0.994	0.153	0.889	0.054

Instead of assigning a probability of 1 to the actual character and 0 to the rest, a very high probability around 0.99 is used and the other characters get a very low probability around 0.001. The smaller value includes some random noise, which is important for the *top-k* optimization, to ensure that not always the same characters are chosen. Another important note is that between every character a timestep containing the *blank* symbol as highest probability is added. This has two reasons. First, it ensures that repeated characters are not merged into one, and second, it greatly improved the accuracy in some experiments. An explanation might be that it allows the model to slightly change words or invent new ones, if the input sentence doesn't match the training examples very well.

4.3. Limitations

The FST-based approach of *finstredet* also has some limitations which should be mentioned. First, it does not work with open questions or commands and can only recognize predefined lookup values. The performance also decreases if there is a large difference between the textual training examples and the test sentences. These limitations can be seen in the *Spoken Language Understanding Resource Package (SLURP)* benchmark [40], which is currently the largest and most complicated SLU benchmark and includes multiple different domains and open questions like “give me the weather forecast for this week”, “who won the presidential election this year” or “if you had to kill someone to save three people would you do it and if so why”. The results in Table 6 show that *finstredet* only understands less than a half of the questions, which is much less than the baseline presented with the benchmark. The model from *SLURP* uses state-of-the-art STT and NLU models which were finetuned on this dataset. Testing *Alexa* was planned as well, but was not possible, because the total number of intents and slots included in the dataset was too high and raised an error message when trying to build the skill.

Table 6: Results on SLURP dataset.

	ScenAct-F1	Entity-F1	SLU-F1
<i>Multi-SLURP</i> [40]	0.783	0.642	0.708
<i>Finstredet (Quartznet)</i>	0.432	0.313	0.380
<i>Finstredet (Conformer)</i>	0.531	0.395	0.452

A second limitation is that the decoding process slows down with growing datasets. Small benchmarks like *Barrista* run about $8/4 \times$ (*Quartznet/Conformer*) faster than real-time on a standard desktop CPU (AMD-3700X) using the *tfLite*-runtime, but the large *SLURP* run is only $1.6/1.7 \times$ faster, and uses a

lower *top-k* decoding parameter. The decoding speed is influenced by two characteristics of the STT model: The computation complexity of the model itself, which is much higher for the *Conformer* model compared to the *Quartznet* model, as well as the number of CTC-timesteps in the model outputs, where the *Conformer* has half the length of the *Quartznet* model, which has a growing impact on larger FST models. For future development, two options could be interesting to improve the decoding speed for large models: Testing decoders from the *Kaldi* project [7], which are optimized on large FSTs, and splitting up the skills into sub-grammars, similar to the approach of *Alexa*, which requires an activation phrase to start a skill after speaking the wake-word. This could be implemented with a small FST, which differentiates between the skill activation word, and enables to select only the Intent-FSTs required for this specific skill for the following decoding step. Such splitting should also have a positive effect on recognition accuracy.

5. Discussion and Conclusion

In this paper a simple method for direct *Spoken Language Understanding* without training is presented under the name of *finstredet*. As basis *Finite State Transducers* are used and optimized for the task of intent and entity extraction. In multiple benchmarks a performance greater than multiple direct SLU or two-step STT+NLU approaches could be achieved.

The main advantage of *finstredet* over other approaches is that no extra training of the SLU model is required. Only text files describing possible requests are needed, which are easy to create or adjust, and could be distributed, similar as in the voice assistants *Jaco*, *Alexa* or others, through shareable skills.

Building the proposed SLU-FST models for the investigated benchmarks took between 1 second (*FluentSpeechCommands*, *Barrista*, *SmartLights*) and about 30 seconds (*SLURP*) on a standard desktop computer (AMD-3700X). Compared to *Jaco* [20] this is much faster, training the NLU models of *Rasa* on the same hardware took about 15 minutes for *Barrista* and about 20-25 minutes for *SmartLights* and *SmartSpeaker*, even though the hyperparameters were optimized for training speed.

The models can also be built and used on edge-devices like a RaspberryPi 4. Building the model for *FluentSpeechCommands*, in which the request are most similar to a simple voice assistant, took 2.2 seconds (compared to 0.7 seconds on the desktop computer). The benchmark itself then runs about $1.1/1.5 \times$ (*Quartznet/Conformer*) faster than real time using the quantized *tfLite* models. As mentioned in the last chapter, it can be seen here that on the RaspberryPi, the reduction of the number of CTC-steps outweighs the slower inference speed of the *Conformer* model.

Most other approaches did not mention training times or the used hardware, except for the *SpeechBrain* run on *Timers-AndSuch* [30], where published logs indicate that the training took about 2-3 hours on unknown hardware, as well as the approach of *Kim et al.* on *FluentSpeechCommands* [32], which required $8 \times$ Nvidia-V100 GPUs to train its network.

The presented method is especially interesting for use-cases that have frequent domain changes, relatively small datasets or restricted training possibilities. For example this could be the case in customizable smart home assistants running on edge-devices or on smartphones, where using a cloud service is undesirable, either to be independent of unstable internet connections or due to privacy concerns.

6. References

- [1] S. Krivan, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.
- [2] D. Bermuth, A. Poeppl, and W. Reif, "Scribosermo: Fast Speech-to-Text models for German and other Languages," *arXiv preprint arXiv:2110.07982*, 2021.
- [3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [4] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Krivan, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.
- [5] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.
- [6] T. Hori and A. Nakamura, "Speech recognition algorithms using weighted finite-state transducers," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–162, 2013.
- [7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [8] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Interspeech 2007-8th Annual Conference of the International Speech Communication Association*, 2007.
- [9] T. Homma, A. S. Arantes, M. T. G. Diaz, and M. Togami, "Maximizing SLU performance with minimal training data using hybrid RNN plus rule-based approach," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 2018, pp. 366–370.
- [10] C. Raymond, F. Béchet, R. De Mori, and G. Damnati, "On the use of finite state transducers for semantic interpretation," 2005.
- [11] G. Tür, J. H. Wright, A. L. Gorin, G. Riccardi, and D. Hakkani-Tür, "Improving spoken language understanding using word confusion networks," in *Interspeech*, 2002.
- [12] C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, "Recent advances in WFST-based dialog system," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.
- [13] N. Kimura, C. Hori, T. Misu, K. Ohtake, H. Kawai, and S. Nakamura, "Expansion of wfst-based dialog management for handling multiple asr hypotheses," in *International Workshop on Spoken Dialogue Systems Technology*. Springer, 2010, pp. 61–72.
- [14] Y. Miao, M. Gowayyed, and F. Metze, "EESSEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.
- [15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [16] A. Kumar, A. Gupta, J. Chan, S. Tucker, B. Hoffmeister, M. Dreyer, S. Peshterliev, A. Gandhe, D. Filiminov, A. Rastrow *et al.*, "Just ASK: building an architecture for extensible self-service spoken language understanding," *arXiv preprint arXiv:1711.00549*, 2017.
- [17] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *International Conference on Implementation and Application of Automata*. Springer, 2007, pp. 11–23.
- [18] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.
- [19] M. Jansche and A. Gutkin, "Sampling from Stochastic Finite Automata with Applications to CTC Decoding," *arXiv preprint arXiv:1905.08760*, 2019.
- [20] D. Bermuth, A. Poeppl, and W. Reif, "Jaco: An Offline Running Privacy-aware Voice Assistant," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 2022, pp. 618–622.
- [21] R. T. Inc, "Rasa," 2021, [Online, accessed 17-June-2021]. [Online]. Available: <https://rasa.com/>
- [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [24] Picovoice, "Barrista Benchmark," 2021, [Online, accessed 24-March-2022]. [Online]. Available: <https://github.com/Picovoice/speech-to-intent-benchmark>
- [25] A. Saade, A. Coucke, A. Caulier, J. Dureau, A. Ball, T. Bluche, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone *et al.*, "Spoken language understanding on the edge," *arXiv preprint arXiv:1810.12735*, 2018.
- [26] S. Rongali, B. Liu, L. Cai, K. Arkoudas, C. Su, and W. Hamza, "Exploring Transfer Learning For End-to-End Spoken Language Understanding," *arXiv preprint arXiv:2012.08549*, 2020.
- [27] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.
- [28] L. Lugosch, B. H. Meyer, D. Nowrouzezahrai, and M. Ravanelli, "Using speech synthesis to train end-to-end spoken language understanding models," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8499–8503.
- [29] L. Lugosch, P. Papreja, M. Ravanelli, A. Heba, and T. Parcollet, "Timers and Such: A Practical Benchmark for Spoken Language Understanding with Numbers," *arXiv preprint arXiv:2104.01604*, 2021.
- [30] SpeechBrain, "SLU recipes for Timers and Such v1.0," 2021, [Online, accessed 24-March-2022]. [Online]. Available: <https://github.com/speechbrain/speechbrain/tree/develop/recipes/timers-and-such>
- [31] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.
- [32] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage Textual Knowledge Distillation to Speech Encoder for Spoken Language Understanding," *arXiv preprint arXiv:2010.13105*, 2020.
- [33] Y. Cao, N. Potdar, and A. R. Avila, "Sequential End-to-End Intent and Slot Label Classification and Localization," in *Proc. Interspeech 2021*, 2021, pp. 1229–1233.
- [34] M. Radfar, A. Mouchtaris, S. Kunzmann, and A. Rastrow, "FANS: Fusing ASR and NLU for On-Device SLU," in *Proc. Interspeech 2021*, 2021, pp. 1224–1228.
- [35] Y. Tian and P. J. Gorinski, "Improving End-to-End Speech-to-Intent Classification with Reptile," *arXiv preprint arXiv:2008.01994*, 2020.

- [36] M. Saxon, S. Choudhary, J. P. McKenna, and A. Mouchtaris, "End-to-End Spoken Language Understanding for Generalized Voice Assistants," in *Proc. Interspeech 2021*, 2021, pp. 4738–4742.
- [37] L. Borgholt, J. D. Havtorn, M. Abdou, J. Edin, L. Maaløe, A. Søgaaard, and C. Igel, "Do we still need automatic speech recognition for spoken language understanding?" *arXiv preprint arXiv:2111.14842*, 2021.
- [38] S. Seo, D. Kwak, and B. Lee, "Integration of Pre-trained Networks with Continuous Token Interface for End-to-End Spoken Language Understanding," *arXiv preprint arXiv:2104.07253*, 2021.
- [39] Y. Qian, X. Bianv, Y. Shi, N. Kanda, L. Shen, Z. Xiao, and M. Zeng, "Speech-language pre-training for end-to-end spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7458–7462.
- [40] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A Spoken Language Understanding Resource Package," *arXiv preprint arXiv:2011.13205*, 2020.