Speech recognition for conversational Finnish

Master's thesis research plan v1

Anssi Moisio
Master's programme in Computer, Communication and Information Sciences
Signal, Speech and Language Processing major

December 14, 2020

1 The ASR task

Automatic speech recognition (ASR) is the task of converting speech into text. The difficulty of the task depends on how varied the speech audio signals are. The restricted problem of recognising a few different words pronounced clearly by one speaker recorded in noise-free conditions was solved years ago. Speech recognition becomes more difficult when the speech is continuous and recorded in differing noise conditions from many speakers. Current state-of-the-art ASR systems are nearing the human-level recognition accuracy also in continuous speech recognition tasks if the speech is planned and pronounced clearly, as it is, for example, in broadcast news or parliament discussions. However, spontaneous, informal conversations remain a challenging type of speech to transcribe automatically, and the gap between human and machine accuracy is still very large. This thesis explores methods for improving ASR for conversational Finnish.

2 The basic structure of an ASR system

The conventional ASR system includes two main component systems: the acoustic model (AM) and the language model (LM). The LM generates an *a priori* probability distribution over possible word sequences. For example, the transciption "en minä tiedä" should probably be assigned a larger probability than "en sinä tiedä" even before any speech audio is processed. The AM outputs *a posteriori* probabilities for phoneme sequences based on the speech audio. A dictionary is used to map the phoneme sequences to words, or more accurately, the same grapheme units that the LM uses, which can also be subword units or characters, for instance. The probabilities of the LM and AM are combined to estimate the most likely transcription of the speech audio.

In the past few years, end-to-end (E2E) speech recognition systems have achieved promising results. An E2E system dispenses with the division to an LM and an AM, and instead learns a mapping from (preprocessed) audio straight to the transcription. This makes the training procedure simpler since only one model is trained instead of multiple. However, it has been shown that E2E models can still benefit from, for example, incorporating an external language model (Toshniwal et al., 2018) or speaker embeddings (Rouhe et al., 2020), into an E2E system, making it arguably no longer a pure E2E model, depending on how "E2E" is defined. Results such as these indicate that pure E2E systems will not completely supplant conventional ASR systems, or systems that include multiple separately trained models, any time soon although they benefit from the simplified training procedure. The state-of-the-art results are still obtained with the conventional systems in many ASR tasks, and the

thesis explores methods in this paradigm.

3 Related work and methodology

The purpose of the thesis is to experiment with some of the latest acoustic and language modelling methods to improve upon the previous best results obtained for an informal, spontaneous Finnish conversation speech data set. Both the speech data set used in this thesis and the previous best results are described by Enarvi et al. (2017). In their work, the acoustic models are trained on 85 hours of speech using the Kaldi toolkit. A first pass of large-vocabulary decoding and word lattice generation is done using an n-gram language model trained on a conversational Finnish text corpus collected by Enarvi et al. (2013). A second pass of rescoring the lattices and generating transcripts is done using a recurrent neural network language model trained on the same text corpus. Subword-vocabulary language models based on statistical segmentation of words (Creutz and Lagus, 2002, 2007) were found to perform better than a word vocabulary.

The baseline system is the same in this thesis, and the work begun by replicating the previous results. The Kaldi toolkit includes acoustic model training pipelines, called "recipes", that are tuned to achieve optimal results for a particular speech data set. In the past three years after the above mentioned previous best results were achieved, the Kaldi recipes have been developed further, and the latest machine learning algorithms have been implemented in the toolkit. By applying the latest Kaldi recipes for the Finnish speech data used in this thesis, the previous best results can be improved. Other acoustic modelling experiments of this work include modelling the speaker and channel variability using i-vectors (Dehak et al., 2010) and x-vectors (Snyder et al., 2018).

Vaswani et al. (2017) introduced a neural network architecture for language modelling called the Transformer, which is based solely on (self-)attention mechanisms (Bahdanau et al., 2014). Since then, the state-of-the-art language models have been of the transformer model type. One of the advantages of attention mechanisms is that they are able to exploit parallel computing, unlike the commonly used recurrent neural networks, since their hidden states do not depend on the hidden states of previous time steps. In this thesis, a Transformer-XL (Dai et al., 2019) LM is trained and evaluated in the ASR task, in a similar manner as described by Jain (2020).

Other language modelling experiments in this work include evaluating word and subword vocabularies, tuning the hyperparameters of the language models (including constant- and variable-order (Siivola et al., 2007a) n-gram models, RNN LMs as well as Transformer-XLs), and experimenting with topic modelling (see e.g., Xiong et al. (2018); Xing et al. (2016)).

4 Schedule

- May: Learn to use Kaldi, replicate the acoustic model (Enarvi et al., 2017) and first-pass decoding with n-grams.
- June: Train the LSTM LM, and apply latest Kaldi recipes to improve on the baseline.
- July: Experiments with subword vocabulary segmentation and using the subwords in ASR.
- August: Train the Transformer-XL and experiment with its hyperparameters.
- September: Writing the thesis background section, experiments with i-vectors.
- October - November: speaker embeddings for the AM, topic modelling for the LM, writing the thesis experiments section.
- December: Wrapping up the experiments, writing the results and conclusion sections.

Below is a list of references used so far in the thesis–not a "preliminary" list since the thesis should soon be finished already.

References

Alam, M. J., Gupta, V., Kenny, P., and Dumouchel, P. (2014). Use of multiple front-ends and i-vector-based speaker adaptation for robust speech recognition. *Proc. of REVERB Challenge*, pages 1–8.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). Openfst: A general and efficient weighted finite-state transducer library. In Holub, J. and Žďárek, J., editors, *Implementation and Application of Automata*, pages 11–23, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1137–1140. IEEE.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Campbell, W. M., Sturim, D. E., Reynolds, D. A., and Solomonoff, A. (2006). Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In *2006 IEEE International conference on acoustics speech and signal processing proceedings*, volume 1, pages I–I. IEEE.

Chen, G., Xu, H., Wu, M., Povey, D., and Khudanpur, S. (2015). Pronunciation and silence probability modeling for ASR. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *Harvard Computer Science Group Technical Report TR-10-98*.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. *arXiv preprint cs/0205057*.

Creutz, M. and Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

Crystal, D. and Potter, S. (2020). English language. *Encyclopædia Britannica*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Enarvi, S., Kurimo, M., et al. (2013). Studies on training text selection for conversational finnish language modeling. In *Proceedings of the 10th International Workshop on Spoken Language Translation (IWSLT 2013)*, pages 256–263.

Enarvi, S., Smit, P., Virpioja, S., and Kurimo, M. (2017). Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2085–2097.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98.

Jain, A. (2020). Finnish language modeling with deep transformer models. *arXiv preprint arXiv:2003.11562*.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing (3rd ed. draft, 16th Oct 2019)*. Web access: `https://web.stanford.edu/~jurafsky/slp3/`.

Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 14:28–29.

Kenny, P., Boulianne, G., and Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE transactions on speech and audio processing*, 13(3):345–354.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE.

Lember, J., Koloydenko, A., et al. (2008). The adjusted Viterbi training for hidden Markov models. *Bernoulli*, 14(1):180–206.

Miao, Y., Zhang, H., and Metze, F. (2015). Speaker adaptive training of deep neural network acoustic models using i-vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1938–1949.

Mohri, M., Pereira, F., and Riley, M. (2008). Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer.

Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafiát, M., Kombrink, S., Motlíček, P., Qian, Y., et al. (2012). Generating exact lattices in the wfst framework. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4213–4216. IEEE.

Povey, D., Kuo, H.-K. J., and Soltau, H. (2008). Fast speaker adaptive training for speech recognition. In *Ninth Annual Conference of the International Speech Communication Association*.

Pylkkönen, J. (2006). Lda based feature estimation methods for lvcsr. In *Ninth International Conference on Spoken Language Processing*.

Rath, S. P., Povey, D., Veselỳ, K., and Cernockỳ, J. (2013). Improved feature processing for deep neural networks. In *Interspeech*, pages 109–113.

Rouhe, A., Kaseva, T., and Kurimo, M. (2020). Speaker-aware training of attention-based end-to-end speech recognition using neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7064–7068. IEEE.

Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., and Makhoul, J. (1985). Context-dependent modeling for acoustic-phonetic recognition of continuous speech. In *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 1205–1208. IEEE.

Shinoda, K. (2005). Speaker adaptation techniques for speech recognition using probabilistic models. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 88(12):25–42.

Shinoda, K. (2011). Speaker adaptation techniques for automatic speech recognition. *Proc. APSIPA ASC*, 2011.

Siivola, V., Creutz, M., and Kurimo, M. (2007a). Morfessor and varikn machine learning tools for speech and language technology. In *Eighth Annual Conference of the International Speech Communication Association*.

Siivola, V., Hirsimaki, T., and Virpioja, S. (2007b). On growing and pruning Kneser–Ney smoothed n-gram models. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1617–1624.

Smit, P., Virpioja, S., Kurimo, M., et al. (2017). Improved subword modeling for wfst-based speech recognition. In *INTERSPEECH*, pages 2551–2555.

Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE.

Somervuo, P., Chen, B., and Zhu, Q. (2003). Feature transformations and combinations for improving asr performance. In *Eighth European Conference on Speech Communication and Technology*.

Toshniwal, S., Kannan, A., Chiu, C.-C., Wu, Y., Sainath, T. N., and Livescu, K. (2018). A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, pages 369–375. IEEE.

Trmal, J., Zelinka, J., and Müller, L. (2010). On speaker adaptive training of artificial neural networks. ISCA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Viikki, O. and Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3):133–147.

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W.-Y. (2016). Topic aware neural response generation. *arXiv preprint arXiv:1606.08340*.

Xiong, W., Wu, L., Zhang, J., and Stolcke, A. (2018). Session-level language modeling for conversational speech. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2764–2768.

Young, S. J. (1992). The general use of tying in phoneme-based hmm speech recognisers. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 569–572. IEEE Computer Society.

Young, S. J., Odell, J. J., and Woodland, P. C. (1994). Tree-based state tying for high accuracy modelling. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.