Speech recognition for conversational Finnish

Master's thesis research plan v2

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

December 14, 2020

Automatic speech recognition (ASR) is the task of converting speech into text. The direct application of ASR is useful in many situations, for example to transcribe patient notes dictated by medical doctors. ASR has also an increasing number of use cases in systems with longer pipelines that achieve some speech-initiated task, such as *voice user interfaces* or *speech-to-speech translation* systems. The ASR piece can often be a bottleneck in the pipeline, determining to a large degree the accuracy of the whole system.

1 The basic structure of an ASR system

In the conventional ASR system, the task is divided into subtasks. The first subtask, called feature extraction, is to divide the audio signal into T segments, and convert the segments into feature vectors, also called observations, $\boldsymbol{O} = \boldsymbol{o}_1, ..., \boldsymbol{o}_T$. The observations are a compressed representation of the audio signal. The task is then to find $\operatorname{argmax}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{O})$, where $\boldsymbol{w} = w_1, ..., w_N$ is a word sequence. This probability is not practicable to compute directly, but by Bayes' rule it can be expanded to

$$\operatorname*{argmax}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{O}) = \operatorname*{argmax}_{\boldsymbol{w}} \frac{P(\boldsymbol{w})P(\boldsymbol{O}|\boldsymbol{w})}{P(\boldsymbol{O})} = \operatorname*{argmax}_{\boldsymbol{w}} \{P(\boldsymbol{w})P(\boldsymbol{O}|\boldsymbol{w})\} \tag{1}$$

The probability of the observations $P(\boldsymbol{O})$ in the denominator is not relevant in finding the best transcription ($\operatorname{argmax}_{\boldsymbol{w}}$) for the observations, which leaves the product in the numerator to be estimated. This product includes the two most significant subtasks: *acoustic modelling* and a *language modelling.* A language model (LM) generates an *a priori* probability distribution $P(\boldsymbol{w})$ over possible word sequences. For example, the transcription "the god of thunder was Zeus" should probably be assigned a larger probability than "the god of thunder was juicy" even before any speech audio is processed. An acoustic model (AM) outputs likelihoods of observations conditional on phoneme sequences. The phoneme sequences are mapped to words by a *lexicon* (also called a *pronunciation dictionary*), yielding $P(\boldsymbol{O}|\boldsymbol{w})$. To avoid numerical underflow, the probabilities are converted to the logarithmic domain. A scalar weight λ is added to determine how significant the LM probabilities are compared to the AM probabilities.

$$\underset{\boldsymbol{w}}{\mathrm{argmax}}\{P(\boldsymbol{w})^{\lambda}P(\boldsymbol{O}|\boldsymbol{w})\} = \underset{\boldsymbol{w}}{\mathrm{argmax}}\{\lambda\log\{P(\boldsymbol{w})\} + \log\{P(\boldsymbol{O}|\boldsymbol{w})\}\} \tag{2}$$

Acoustic and language modelling are achieved using machine learning models, primarily deep neural networks (DNNs), which are estimated based on training data. To model acoustics a parallel corpus of speech and text is needed, whereas to model language only text is needed.

2 Related work

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.

The difficulty depends also on the language. The most obvious factor is the availability of training data. The state-of-the-art ASR systems are based on DNNs that require large training data sets. For languages such as Finnish, the resources are more limited than for the most widely spoken languages in the world, which makes the ASR task harder. Some inherent idiosyncratic properties of Finnish should also be taken into account when developing a Finnish ASR system. Finnish is a morphologically rich language. Suffixes and other conjugations perform grammatical functions, such as cases, which in other languages, e.g. English, would be denoted by separate words. This makes the number of word types in the vocabulary large, requiring a lot of computational resources as well as a lot of text to train the LM. This problem can be avoided by segmenting words into smaller units, referred to as *subwords* (Hirsimäki et al., 2006).

In colloquial spoken Finnish, however, the morphology is often different than in formal language. Some of the suffixes and other inflections can be omitted or changed to an incorrect one. For example, it is common to not use the first person plural inflections for verbs, and instead use the passive voice verb inflection (”me ollaan” instead of the correct form ”me olemme”) or the incorrect singular form when the third person plural form should be used (”ne on” instead of the correct ”ne ovat”). Other common habits of informal Finnish include shortening words (e.g., from ”minä” to ”mä”) and/or combining words (e.g., ”oliko se” to ”olikse”). Differences between formal and informal Finnish pose difficulties when modelling informal language. Text from formal sources such as books or newspapers does not resemble colloquial Finnish very closely. Another relevant feature of the Finnish language is its phonemic orthography, i.e., the fact that one letter generally corresponds to one phoneme, and vice versa. Because of this fact, it is possible to also *write* colloquial Finnish as it is spoken. It is therefore possible to find written text that resembles the spoken language, and use this to model colloquial Finnish. The above mentioned incorrect inflections are examples of informal Finnish that is also written in the incorrect way, as it is pronounced.

Colloquial Finnish is written, for example, on online forums, especially in direct-message conversations. Enarvi et al. (2013) collected a conversational Finnish text corpus from Internet forums by searching for key phrases that indicate informal conversations. This text corpus is used also in this thesis to model conversational Finnish. Enarvi et al. (2017) developed and evaluated different ASR systems for the conversational data.

2.1 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. They used a simple *n-gram* language model for generating the first hypothesis transcriptions of the speech and rescored the hypotheses using a more complex neural network language model (NNLM). 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.

In the previous decade, i-vectors (Dehak et al., 2010) have been used to model speaker and channel variability and to add this information to the features. Snyder et al. (2018) described how also DNNs can be utilised to create speaker embeddings, called x-vectors, that model differences between speakers. In this work, i-vectors and x-vectors are evaluated and compared on the conversational Finnish data. The aim is to explore how these speaker embeddings affect the ASR accuracy on the used data set.

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 previously common recurrent neural networks (RNN), since their hidden states do not depend on the hidden states of previous time steps. In this thesis, Transformer-XL (Dai et al., 2019) language models are trained and evaluated on the conversational Finnish data. The goal is to determine whether the transformer models achieve better results than the RNN models on the conversational Finnish data. Other language modelling experiments in this work include evaluating word and subword vocabularies for transformers as well as other LMs.

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)).

3 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*.

Bahl, L., Brown, P., De Souza, P., and Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52. IEEE.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bourlard, H. A. and Morgan, N. (2012). *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media.

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.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. 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*.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Enarvi, S. et al. (2018). Modeling conversational finnish for automatic speech recognition.

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.

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

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

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Gupta, M. R. and Chen, Y. (2011). *Theory and use of the EM algorithm*. Now Publishers Inc.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., and Pylkkönen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to finnish. *Computer Speech & Language*, 20(4):515–541.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Irie, K., Zeyer, A., Schlüter, R., and Ney, H. (2019). Language modeling with deep transformers. *arXiv preprint arXiv:1905.04226*.

Iskra, D., Grosskopf, B., Marasek, K., Heuvel, H., Diehl, F., and Kiessling, A. (2002). Speecon-speech databases for consumer devices: Database specification and validation.

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

Jiang, H. (2010). Discriminative training of hmms for automatic speech recognition: A survey. *Computer Speech & Language*, 24(4):589–608.

Juang, B.-H., Hou, W., and Lee, C.-H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio processing*, 5(3):257–265.

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.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association.*

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

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

Lennes, M. et al. (2009). Segmental features in spontaneous and read-aloud finnish. *Phonetics of Russian and Finnish general description of phonetic systems: experimental studies on spontaneous and read-aloud speech.*

Martínez, A. M. and Kak, A. C. (2001). Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233.

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.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612.*

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Povey, D. (2005). *Discriminative training for large vocabulary speech recognition.* PhD thesis, University of Cambridge.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

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*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

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

Pylkkonen, J. and Kurimo, M. (2012). Analysis of extended baum–welch and constrained optimization for discriminative training of hmms. *IEEE transactions on audio, speech, and language processing*, 20(9):2409–2419.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

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

Rosenberg, A. E., Lee, C.-H., and Soong, F. K. (1994). Cepstral channel normalization techniques for hmm-based speaker verification. In *Third International Conference on Spoken Language Processing*.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

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.

Rownicka, J., Bell, P., and Renals, S. (2019). Embeddings for DNN speaker adaptive training. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 479–486. 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.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

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*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.

Vesel̀y, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, pages 2345–2349.

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

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Woodland, P. C. and Povey, D. (2002). Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, 16(1):25–47.

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., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Ragni, A., Valtchev, V., Woodland, P., and Zhang, C. (2015). *The HTK Book (version 3.5a)*.

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*.