INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Automatic Construction of the Finnish Parliament Speech Corpus

André Mansikkaniemi, Peter Smit, Mikko Kurimo

Department of Signal Processing and Acoustics, Aalto University, Finland
andre.mansikkaniemi@aalto.fi, peter.smit@aalto.fi, mikko.kurimo@aalto.fi

Abstract

Automatic speech recognition (ASR) systems require large amounts of transcribed speech data, for training state-of-the-art deep neural network (DNN) acoustic models. Transcribed speech is a scarce and expensive resource, and ASR systems are prone to underperform in domains where there is not a lot of training data available. In this work, we open up a vast and previously unused resource of transcribed speech for Finnish, by retrieving and aligning all the recordings and meeting transcripts from the web portal of the Parliament of Finland. Short speech-text segment pairs are retrieved from the audio and text material, by using the Levenshtein algorithm to align the first-pass ASR hypotheses with the corresponding meeting transcripts. DNN acoustic models are trained on the automatically constructed corpus, and performance is compared to other models trained on a commercially available speech corpus. Model performance is evaluated on Finnish parliament speech, by dividing the testing set into seen and unseen speakers. Performance is also evaluated on broadcast speech to test the general applicability of the parliament speech corpus. We also study the use of meeting transcripts in language model adaptation, to achieve additional gains in speech recognition accuracy of Finnish parliament speech.

Index Terms: automatic speech recognition, speech-to-text alignment, DNN acoustic models, parliament speech data, transcribed speech corpus

1. Introduction

Advancements in automatic speech recognition (ASR) have been enabled by statistical data-driven methods. In recent years, progress has mainly been driven by the use of deep neural networks (DNNs) in acoustic modeling [1]. The successful training of well-performing DNN models requires large amounts of transcribed speech data, which is still a scarce and usually expensive resource for many domains and languages.

Finnish is not an under-resourced language within the ASR context, but access to all large general-domain training data is limited. The web portal of the Parliament of Finland [1] gives access to a vast resource of transcribed speech data in Finnish. Since 2008, all the TV recordings of the open plenary sessions have been available online. In total, there are over 2000 hours of video recordings, accompanied by handwritten and speaker tagged meeting transcripts. A corpus of this size, could be useful for the research community and also benefit commercial applications.

The main challenges of harnessing this open resource as acoustic model training data are the following: the correct alignment of long audio files with text and the selection of the most accurate speech-transcript segment pairs. Meeting transcripts for the open plenary sessions in the Finnish parliament have been manually transcribed to be easily readable. Hesitations, repetitions, and false starts are not transcribed. Some stylistic changes, such as changing of word order and grammatical corrections, have also been made to improve the readability. Some parts are also left untranscribed, often related to meeting instructions communicated by the Speaker of the Parliament.

Methods for the alignment of long audio files with text have been studied in previous works. In [2], speech-to-text alignment was implemented using universal phone models. A phonetic recognizer was used in [3], to produce phoneme sequences for audiobooks and parliamentary speeches. Alignments to the transcripts were performed using a matrix of approximate sound-to-grapheme mappings. A common framework for aligning long audio recordings with transcripts is to use first-pass ASR output [4] [5]. Anchor points in relation to the transcript are found using dynamic text alignment methods. In [6], alignments between conversational Arabic speech and transcripts were produced by aligning first-pass ASR output with the transcriptions using the Levenshtein algorithm. In [7], a similar approach was used for creating the free LibriSpeech corpus from publicly available audiobooks, containing over 1000 hours of English speech data.

Automatic transcription systems for parliamentary sessions have also been a focus of previous research. A system for Japanese parliament speech was proposed in [8], that focused on training a statistical machine translation (SMT) model between spoken language and official meeting transcripts. The SMT model was used for generating training data for a biased LM. Reliable acoustic model training data could then directly be retrieved from first-pass ASR output.

In this work, our aim is to automatically produce a transcribed speech corpus from the data available at the web portal of the Parliament of Finland, and to make it publicly available on Kielipankki, the Language Bank of Finland [2]. We implement a system, based on free and publicly available tools, for aligning parliamentary sessions with corresponding meeting transcripts (Figure 1). The long audio recording is first automatically split into shorter segments. A biased LM is trained on the meeting transcript, and first-pass ASR hypotheses are generated for all segments. The first-pass ASR hypotheses are merged and aligned with the meeting transcript. The Levenshtein algorithm is used to find anchor points between the two texts. Based on these, shorter speech-transcript segment pairs can be produced. Forced alignment is used for cleaning unreliable segments, where spoken audio and transcript are mismatched.

A DNN acoustic model is trained on the constructed speech corpus. The performance of the model is evaluated on new parliament speech data, and compared with a model trained on a commercially available speech corpus. Furthermore, we will study how much the speech recognition accuracy of parliament

This work was financially supported by the Academy of Finland under the grant number 251170. Computational resources were provided by the Aalto Science-IT project.

[1] https://www.eduskunta.fi

[2] https://www.kielipankki.fi

http://dx.doi.org/10.21437/Interspeech.2017-1115

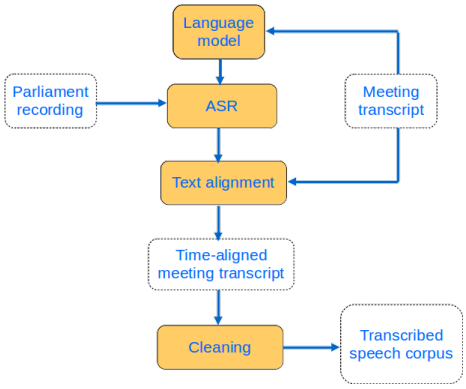

Figure 1: *An overview of the automatic construction of the parliament speech corpus. A biased LM is trained on the meeting transcript, for the first-pass recognition. The ASR output is aligned with the meeting transcript. Cleaning is performed to retrieve the best matching speech-transcript segment pairs, which are added to the final transcribed speech corpus.*

speech can be improved, by adapting the background LM with transcripts of previous meetings. The acoustic models will also be evaluated on speech data from a different domain, broadcast news, to test the general applicability of the parliament speech corpus. To the authors' knowledge, this is the first attempt to construct a transcribed speech corpus from data available at the web portal of the Parliament of Finland.

2. Methods

2.1. Retrieval and preprocessing steps

In the first step, over 2000 hours of parliament sessions with their corresponding meeting transcripts were downloaded from the web portal. Each individual video recording was converted into audio, and segmented into shorter segments using an MFCC-based speaker diarization algorithm [9].

Meeting transcripts are in HTML format. Speaker turns and the related speech transcripts are stored in the HTML files using special XML tags. A preprocessing script was written to extract both the text and speaker information from each meeting transcript.

2.2. Speech-to-text alignment

The goal of speech-to-text alignment is to find the exact time frames when words (or phonemes) are spoken in the audio recording. For shorter segments, forced alignment based on hidden Markov models (HMMs) can be used, but for longer recordings this approach is usually computationally too expensive and prone to errors.

In this work, speech-to-text alignment was performed based on first-pass ASR output. For the first recognition run, a biased LM was trained on a text set consisting of general news texts and the retrieved meeting transcript. First-pass hypotheses were generated for all speech segments. The output, including word hypotheses and timestamps, of the recognized segments were merged into one output file.

Alignment between the first-pass ASR output and the meeting transcript was performed using the Levenshtein distance algorithm [10]. An implementation of the algorithm which is found in the NIST scoring toolkit [3] was used.

The Levenshtein distance measures the difference between two sequences, taking into account substitutions, deletions, and insertions. It is usually used in ASR to calculate word or letter error rates. We used the alignment that the algorithm produces, to find anchor points between the meeting transcript and ASR hypothesis (Figure 2). Timestamp information from words in the ASR output were passed over to their aligned counterparts in the meeting transcript. In the case of a deletion in the ASR output, when a word in the meeting transcript has no counterpart, the time was split and redistributed between the current and previous word.

MEETING TRANSCRIPT: KULUTTAJAT OSTAVAT ympäristötietoisemmin
** mutta SIINÄ on hyvin paljon ongelmia

ASR OUTPUT: KULUTTAJALLE NOSTAVAN ympäristötietoisemmin
ON mutta * on hyvin paljon ongelmia

Figure 2: *Example alignment based on Levenshtein distance, between meeting transcript and ASR output. The first two words are examples of substitutions. The character * marks insertions in the meeting transcript, and deletions in the ASR output.*

The time-aligned sentences in the meeting transcript, were used as reference points for splitting the recording into shorter segments. Speaker information was also stored for each sentence. HMM-based forced alignment, was used to select the best matching speech-transcript segment pairs. All the segments which did not align under a certain threshold beam, were excluded from the acoustic model training set.

A transcribed speech corpus is often produced to be as balanced as possible between individual speakers, gender, and age groups. Because there is great variance how often different Members of Parliament speak during open plenary sessions, we applied an optional step of filtering where the maximum contribution of individual speakers was limited to a maximum length of time.

3. Experiments

3.1. System

First-pass speech recognition hypotheses were generated using the Aalto ASR system [11], which utilizes crossword triphone GMM-HMM acoustic models. The forced alignment filtering of training data, was also performed using Aalto ASR GMM-HMM models. After the segmentation and data division a completely independent speech recognition system was trained using the Kaldi toolkit [12]. At first a GMM-HMM was trained and this model was used to do another cleaning and segmentation step. This cleaning step consisted of doing a recognition pass with a biased language model trained only on the applicable transcript and retaining those segments of speech where the original transcription matched the recognition output. This cleaning step was done not only for the new parliament corpus, but also for the commercial corpus which we compare to.

Decoding was done in two passes. The first pass used a 2-gram language model. The second pass rescored the generated lattices with the full language model. Minimum Bayes Risk decoding [13] was used for generating the final hypotheses.

[3] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Subword language models were trained. The subword segmentations were trained with the Morfessor toolkit [14, 15] which uses the Minimum Description Length principle to create a segmentation in an unsupervised data-driven manner. An n-gram LM was trained on the segmented text using the VariKN toolkit [16]. This toolkit is able to create high-order n-gram which are important for subword models [11].

In the language model adaptation experiment, the background and in-domain data n-gram models were interpolated. The interpolation weight was optimized to minimize perplexity on the utterances of the development recognition set.

The final recognition systems were time-delay neural networks (TDNN) trained using a pure sequential discriminative criterion [17]. These models differ both in topology and frame rate from conventional GMM-HMM. They use a single state for each phone, and utilize only every third data-frame during recognition. For adaptation, ivectors were used as a secondary input to the model. In addition to normal TDNNs, we also trained TDNN-LSTM models which combine a normal TDNN with a number of recurrent layers. These models have a high data-requirement to be beneficial over normal TDNN models.

3.2. Data

Parliament speech data was retrieved from the web portal. In total, 2269 hours of speech was downloaded. After alignment and extraction of speech-transcript segment pairs, forced alignment filtering was applied on the extracted data. In the first step, all segments were selected which aligned under a log probability threshold of -1600. The selected segments were split into a 1559 hour training set (Table 1) and 11 hour test set (Table 2).

The initial training set *parl-all*, was split into smaller subsets. All segments which aligned under a log probability threshold of -400 were selected into a separate training set *parl-400*. To smoothen speaker distributions, two additional subsets were created from the *parl-400* set. In the subset *parl-30min*, the maximum length of time for each speaker was limited to 30 minutes. In the subset *parl-60min*, the maximum length of time was limited to 60 minutes. All the selected training sets were additionally cleaned using the Kaldi toolkit.

The Parliament test set was divided into sets of seen and unseen speakers. In the *parl-unseen* set, none of the speakers were represented in the training set. In the *parl-seen* set, all of the speakers were represented in the training set. Training data was filtered, to only include seen speaker segments from session dates that were not present in *parl-seen*.

Besides the created Parliament datasets, an acoustic model was also trained on the Speecon corpus [18]. This corpus contains read aloud sentences from a large text corpus, recorded with lapel microphones. We also evaluated combining the Parliament and Speecon data sets (*all*), by training an acoustic model on both *parl-all* and *speecon*. For evaluation two different sets were used besides the parliament evaluation sets. The utterances from the Speecon datasets contain read, planned speech. The YLE dataset contains utterances from broadcast news shows. The broadcast data is not marked with speaker information; each utterance was treated as if it is a new speaker.

The background LM in this work was trained on the Kielipankki corpus [19], which is a 150M word corpus containing books, magazines, and newspaper articles. For the LM adaptation experiments, a 20M word in-domain text corpus was extracted from the meeting transcripts. Transcripts of unseen speakers were removed, as were all transcripts from sessions present in the test data.

Table 1: *Training data sets used in the experiments. Data sets are described in terms of number of speakers (**Speakers**), size of the corpus before cleaning (**Hours**), size of the corpus after cleaning (**Cleaned**), and standard deviation of individual speaker data lengths (**Speaker std.,** σ).*

Dataset	Speakers	Hours	Speaker std., σ	Cleaned
parl-30min	357	154.3	0h 8min	137.0
parl-60min	357	284.0	0h 18min	252.9
parl-400	357	893.0	2h 35min	794.2
parl-all	357	1559.4	4h 10min	1395.2
speecon	425	148.6	0h 1min	105.9
all	782	1708.0	3h 22min	1501.1

Table 2: *Test data used in the experiments, divided into development and evaluation sets.*

Dataset	Dev		Eval	
	Hours	Speakers	Hours	Speakers
parl-seen	2h 37min	9	2h 54min	11
parl-unseen	2h 45min	10	2h 48min	10
speecon	57min	20	1h 12min	25
yle	5h 24min	∼5-10	5h 35min	∼5-10

4. Results

4.1. Acoustic model training

In the first experiments, the performances of DNN acoustic models trained on the different data sets, were compared on the Parliament test sets. Results are in Table 3. In general, all the Parliament models outperform the Speecon models with around 10-15 percentage units. The relative improvement gained by the Parliament models is close to 50%. In terms of absolute performance, there is not much difference between the seen and unseen speaker sets. Also comparing the gains relative to the Speecon models, the differences are small.

Table 3: *Results of ASR experiments on the Parliament test sets. Performances of acoustic models trained on different training sets are compared. Results are reported in word error rate (WER [%]).*

Acoustic model	parl-seen		parl-unseen	
	dev	eval	dev	eval
parl-30min	15.98	12.24	15.59	10.96
parl-60min	15.57	11.74	15.25	10.76
parl-400	14.52	10.98	14.17	10.23
parl-all	13.65	10.36	13.62	9.83
parl-all-lstm	12.60	9.42	12.43	8.98
speecon	26.78	20.60	25.90	19.05
speecon-lstm	26.65	20.69	25.98	19.35
all	13.66	10.39	13.63	9.77
all-lstm	**12.18**	**8.99**	**12.13**	**8.82**

Results of the ASR experiments on the Speecon and YLE broadcast news sets are presented in Table 4. The Speecon models perform better than the Parliament models on Speecon test

sets. On the YLE test sets, the Parliament models give better performance. The *parl-all* model achieves a 15-18% relative improvement over the Speecon model.

Best overall performance, on all test sets except *yle-eval*, is achieved with the *all-lstm* model. Relative WER reductions compared to the single corpus models are over 10% on the Speecon test data, and around 2% on Parliament and YLE data. The best results on the Speecon and YLE test sets also represent the current state-of-the-art performance when compared to earlier published work, with a WER of 13.3% for *speecon-eval* reported in [20], and a WER of 30.5% for *yle-eval* reported in [21].

Table 4: *Results of ASR experiments on the Speecon and YLE broadcast news test sets. Results are reported in word error rate (WER [%]).*

Acoustic model	speecon		yle	
	dev	eval	dev	eval
parl-30min	8.72	11.13	19.36	21.15
parl-60min	9.10	11.43	19.26	20.71
parl-400	8.67	11.78	18.75	20.87
parl-all	9.84	11.76	18.88	20.40
parl-all-lstm	12.58	14.51	17.81	**18.56**
speecon	6.96	8.99	23.22	24.09
speecon-lstm	7.79	8.83	24.27	24.59
all	6.60	8.47	18.35	19.90
all-lstm	**5.96**	**7.97**	**17.60**	18.73

4.2. Language model adaptation

In the last set of experiments, the effect of using in-domain data for LM training and adaptation was evaluated on the Parliament test sets. In the first setup, a language model was trained only on parliament meeting transcripts. Results of the ASR experiments using the in-domain LM are in Table 5. Another LM was estimated by interpolating the background and in-domain LMs. Results using the interpolated LM are in Table 6. Results show that using in-domain text data improves recognition accuracy. The relative WER improvement is between 30-40%. There is little difference between the in-domain and interpolated models.

Table 5: *Results of ASR experiments on the Parliament test sets using an **LM only trained on in-domain data**. Results are reported in word error rate (WER [%]).*

Acoustic model	parl-seen		parl-unseen	
	dev	eval	dev	eval
parl-30min	9.65	6.83	9.88	5.95
parl-60min	9.22	6.69	9.70	5.75
parl-400	8.99	6.44	9.32	5.68
parl-all	8.51	6.11	9.00	5.41
parl-all-lstm	8.23	5.94	8.67	5.35
speecon	16.06	11.71	16.12	10.42
speecon-lstm	16.40	11.84	16.28	10.35
all	8.65	6.01	9.02	5.49
all-lstm	**7.94**	**5.91**	**8.52**	**5.17**

Table 6: *Results of ASR experiments on the Parliament test sets using an **interpolated background/domain LM**. Results are reported in word error rate (WER [%]).*

Acoustic model	parl-seen		parl-unseen	
	dev	eval	dev	eval
parl-30min	9.55	6.81	9.92	5.96
parl-60min	9.24	6.60	9.63	5.80
parl-400	9.00	6.38	9.33	5.64
parl-all	8.55	6.05	9.03	5.45
parl-all-lstm	8.26	5.93	8.70	5.42
speecon	16.38	11.72	16.20	10.60
speecon-lstm	16.89	11.97	16.41	10.45
all	8.59	**5.89**	9.07	5.49
all-lstm	**7.95**	**5.89**	**8.58**	**5.30**

5. Discussion

The results of the experiments clearly indicate the usefulness of the automatically constructed Finnish Parliament speech corpus, especially for improving recognition accuracy on in-domain data. The corpus also improves recognition accuracy on broadcast news data. The best performance is achieved when using the entire data set. The uneven speaker distribution did not seemingly affect performance. There was not much difference between the seen and unseen speaker sets, both in terms of absolute error rates and relative gains compared to Speecon models, which did surprise us a little. Speaker differences within the seen data set might explain this to some degree. Some speakers were represented with several hours of data while others only had a few minutes. Results also showed the benefits of using the Parliament data as a complement to existing speech corpora. Best recognition accuracies for all test sets, including the Speecon test data, were achieved using the model trained on both Parliament and Speecon data.

The strength of in-domain data was also clear in the LM adaptation experiments. The in-domain LM, trained on an eight times smaller corpus, managed on its own to clearly outperform the background LM.

A point of interest in the experiments was also the clear difference in error rates between the Parliament development and evaluation sets. The authors assume this difference is simply attributed to chance by more disfluent and hesitant speakers being selected into the development sets.

6. Conclusions

In this work, we implemented a system for automatically aligning recordings and meeting transcripts retrieved from the web portal of the Parliament of Finland. The DNN models trained on the constructed Parliament corpus clearly outperform models trained on a commercial speech corpus, when tested on parliament and broadcast news data. Further improvements in speech recognition accuracy on parliament speech was gained by using an in-domain LM trained on meeting transcripts. The Finnish Parliament speech corpus constructed here will be published through the Language Bank of Finland to provide free access for research use. We also intend to improve the retrieval and alignment algorithms which can be found online [4].

[4] https://www.github.com/aalto-speech/finnish-parliament-scripts

7. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[2] S. Hoffmann and B. Pfister, "Text-to-speech alignment of long recordings using universal phone models," in *Proceedings of Interspeech*, Lyon (France), Sep 2013, pp. 1520–1524.

[3] X. Anguera, J. Luque, and C. Gracia, "Audio-to-text alignment for speech recognition with very limited resources," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1405–1409.

[4] J. Robert-Ribes, R. G. Mukhtar, and A. C. S. Crc, "Automatic generation of hyperlinks between audio and transcript," in *Proceedings Eurospeech '97*, 1997.

[5] P. J. Moreno, C. F. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments." in *ICSLP*. ISCA, 1998.

[6] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "Automatic long audio alignment and confidence scoring for conversational arabic speech," in *LREC*. European Language Resources Association (ELRA), 2014, pp. 3062–3066.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[8] T. Kawahara, "Transcription system using automatic speech recognition for the Japanese parliament (diet)," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012, pp. 2224–2228.

[9] A. Ojeda, "Speaker diarization," Master's thesis, Aalto University, 2014.

[10] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[11] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Trans. Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 724–732, 2009.

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

[13] H. Xu, D. Povey, L. Mangu, and J. Zhu, "An improved consensus-like method for minimum bayes risk decoding and lattice combination," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4938–4941.

[14] M. Creutz and K. Lagus, "Unsupervised discovery of morphemes," in *Proceedings of the ACL 2002 Workshop on Morphological and Phonological Learning*, ser. MPL '02, vol. 6. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 21–30. [Online]. Available: http://www.aclweb.org/anthology/W/W02/W02-0603.pdf

[15] S. Virpioja, P. Smit, S.-A. Grönroos, and M. Kurimo, "Morfessor 2.0: Python implementation and extensions for Morfessor Baseline," Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland, Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, 2013.

[16] V. Siivola, T. Hirsimäki, and S. Virpioja, "On growing and pruning Kneser-Ney smoothed n-gram models," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 5, pp. 1617–1624, 2007.

[17] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Interspeech 2016*, 2016, pp. 2751–2755. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-595

[18] D. J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation." in *LREC*, 2002.

[19] CSC - IT Center for Science, "The Helsinki Korp Version of the Finnish Text Collection," 1998. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2016050207

[20] T. Alumäe, "Neural network phone duration model for speech recognition." in *INTERSPEECH*, 2014, pp. 1204–1208.

[21] M. Kurimo, S. Enarvi, O. Tilk, M. Varjokallio, A. Mansikkaniemi, and T. Alumäe, "Modeling under-resourced languages for speech recognition," *Language Resources and Evaluation*, 2016.