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

Building an ASR corpus using Althingi's Parliamentary Speeches

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, Jón Guðnason

Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland

ingarun@ru.is, r@kjaran.com, annabn@ru.is, jg@ru.is

Abstract

Acoustic data acquisition for under-resourced languages is an important and challenging task. In the Icelandic parliament, Althingi, all performed speeches are transcribed manually and published as text on Althingi's web page. To reduce the manual work involved, an automatic speech recognition system is being developed for Althingi. In this paper the development of a speech corpus suitable for the training of a parliamentary ASR system is described. Text and audio data of manually transcribed speeches were processed to build an aligned, segmented corpus, whereby language specific tasks had to be developed specially for Icelandic. The resulting corpus of 542 hours of speech is freely available on http://www.malfong.is. First experiments with an ASR system trained on the Althingi corpus have been conducted, showing promising results. Word error rate of 16.38% was obtained using time-delay deep neural network (TD-DNN) and 14.76% was obtained using long-short term memory recurrent neural network (LSTM-RNN) architecture. The Althingi corpus is to our knowledge the largest speech corpus currently available in Icelandic. The corpus as well as the developed methods for corpus creation constitute a valuable resource for further developments within Icelandic language technology.

Index Terms: Icelandic, speech corpus, text normalization, automatic speech recognition

1. Introduction

The main hurdle for training automatic speech recognition (ASR) systems for under-resourced languages is the lack of acoustic data. In order to produce the technology, not only must acoustic data be gathered, but existing speech recordings must be developed to make them suitable for speech recognition. Many troves of transcribed speech may exist for an under-resourced language. These can be transcribed radio or television programmes, court or trade records, or governmental and parliamentary speeches. There are two problems with these kind of data sources. The first problem is that the transcriptions often don't reflect the speech recordings accurately enough to be useful for ASR training. The reason for this may simply be that the transcription is meant for publication and text that is completely true to spoken language with its hesitation words and repetitions is rarely suitable for reading. The second problem is that these speech recordings are too long to be suitable for ASR training. Often, audio recordings like these are 10s of minutes in length but the optimal audio segment length for ASR training is below 35 s [1].

The Icelandic parliament, Althingi, is a source of this kind of data. Althingi is in session between 800 and 900 hours a year and their speeches are recorded, transcribed and published. It is therefore immensely important to be able to harvest this data source for Icelandic ASR development, especially when other sources of data are scarce. Althingi's speeches, however, suffer from the two aforementioned problems. The transcriptions need to be normalized to fit the speech better and the speeches need to be segmented.

1.1. Related work

Using approximate transcripts is not new. It has been shown how captioned multimedia speech can be made into a speech corpus [2]. Text normalization is explained in [3, 4]. Other methods include for example lightly supervised methods [5].

1.2. Existing speech corpora in Icelandic

Two open speech corpora already exist for Icelandic, Hjal [6] and Málrómur [7]. The Hjal speech corpus consists of voice samples from 2000 individuals, where each individual was asked to read up words, phrases and sentences written on one sheet of paper. The content on each sheet had been carefully chosen to represent words and phrases likely to be used in ASR applications, phonetically rich sentences and isolated letters. 1000 distinctive sheets were generated, hence each sheet was read by two individuals. In relation to the Hjal project, a pronunciation dictionary was built. That dictionary is now freely available on http://www.malfong.is and contains 65 thousand SAMPA and IPA transcribed words. It was used as a bases for our pronunciation dictionary.

The Malromur corpus is an open source corpus of about 120,000 Icelandic voice samples, from 592 individuals, which sum up to around 255 hours of data. Half of the corpus comes from news stories. The other half consists of rare tri-phones, location and proper names, urls, numbers, names of days, months and times of day, simple questions and greetings.

2. Althingi speeches

The amount of available acoustic data from Althingi is potentially very large. The average total speech time in the parliament is around 600 hours a year. This is about 30% greater than the amount at other Nordic parliaments and the German Bundestag, and a bit under the French and the English parliaments. Althingi is, however, quite unusual in that it is only composed of 63 members, so that each member contributes 9.5 hours on average compared to one to three hours in the German, French and the British parliaments [8].

Althingi's speeches have been systematically recorded and published for many years, stored recordings date back to 2005. The final text and the recording of each speech has then been made available on the Althingi website[1]. The manual transcription process has been done in two stages. First an initial manuscript is obtained from a (usually) contracted transcription service. This manuscript is supposed to reflect the spoken record as well as possible. However, minor editing is usually done during this process. The initial manuscript is then edited

[1] http://www.althingi.is

http://dx.doi.org/10.21437/Interspeech.2017-903

by specialists at Althingi's Information and Publications Department. The purpose of this editing is to make the text fit for publication. The text is modified to increase its clarity and enrich its context without changing its meaning. For example a proper noun might be substituted for a pronoun or references added.

3. Audio Alignment

Speech recognition methods are developed to work on short audio segments, and perfectly transcribed reference texts. This is rarely the case when working with data not produced for speech recognition. Hence, some effort has to be made to get the data into a usable state. That includes text normalization procedures like removing punctuations and writing all numbers and abbreviations out in full length. Depending on the data, there might be considerable more things that will need rewriting, before the data is clean enough for speech recognition training.

The data obtained from the Icelandic parliament consists of 6600 recordings of parliament speeches, dating from 2005 to 2016. Each recording contains one speech, and comes with the two sets of text files, described in Section 2. The recordings usually begin and end with the house speaker introducing the current and next speech, respectively, which is not reflected in the transcripts.

The total length of the recordings, without trimming non-speech material away, is 666 hours and 38 minutes. The mean speech length is 6 minutes, but the range is from just under one minute and up to roughly half an hour. After expanding numbers, the corpus contains 5,097,612 word tokens and 116,753 word types. Speakers are 197, of which 105 are male and 92 female.

3.1. Text processing and normalization

The text normalization is done in a few steps. The text is put in lower case and everything in the text files which is not represented in the speech files is removed. For example, annotations that indicate what is happening in the congress room and labelled references are removed from the transcripts.

Some units in the transcript, such as regulation numbers, decimal numbers and time, have to be rewritten before number expansion, so that the transcripts reflect the speech.Parliament speakers often reference regulations which are usually written on the format "law-number/year/institution". So, for example, "54/1996/ESB" which appears in the transcript has to be rewritten to "54 1996 ESB" before number expansion. Some ambiguity will still remain, since sometimes the speakers say "54 from 1996 ESB". After rewriting, the remaining punctuations are removed and spelling corrected, using the words in the final version of the speeches as a reference dictionary.

Since Icelandic is an inflected language, the expansion of abbreviations and numbers is a non-trivial procedure. The OpenGrm Thrax Grammar Development Tools [9, 10] were used, almost exclusively, to generate all possible expansions of a number or an abbreviation. The Thrax tools compile a weighted finite state transducer out of grammars expressed as regular expressions and context-dependent rewrite rules.

A language model, which contains expanded abbreviations and numbers, is needed to select the correct expansions. The Althingi data was not sufficient for that task since most abbreviations and numbers are rarely expanded. Thus, alongside the parliament speeches, a corpus of 10 million sentences (167M word tokens) from the Leipzig database for Icelandic[11] was used. The corpus consists of data collected from Icelandic websites. Sentences, containing no tokens which could be abbreviated, were filtered away. Inflected languages require much data for building language models, and in our case more data would have been beneficial. Icelandic words can have up to 24 forms. Adjectives have four cases, three genders and two numbers which can all be different in theory, but the number is usually lower. Other types of words have lower theoretical maximum of forms. Still, this poses a problem for rare words whose forms are unlikely to appear in a smaller corpora. This means that the ability to predict all forms of rare words is diminished if the corpus too small.

The parliament often uses very specific abbreviations that naturally did not appear in the language model based on the Leipzig database. The language model was therefore unable to deduct which case and gender to use for those abbreviations and (very typically) gender and case for ordinal numbers. Hence, a manual correction of expansions, was applied to 100 hours of data from the years 2013 to 2016. This expanded data was then added to the language model, resulting in a much better representation of these words and subsequently more expansion success. The resulting language model was used to expand the rest of the parliament speeches, dating from 2005-2012.

Table 1 shows results of an expansion test performed on 100 speeches, chosen at random from the 2005-2012 data. Numbers and abbreviations in the texts were expanded using the Thrax tools, and three language models with different n-gram orders. The language models contain both parliament data, with expanded numbers and abbreviations from the 2013-2016 data, plus the Leipzig data. The expansions were compared to a corrected expansion of the 100 speeches. The table shows the error rates obtained, first for 7 parliament specific abbreviations, and then for general abbreviations and numbers, which have multiple expansions. For this work, the bigram language model was used for the expansion of our texts. Preliminary results on 18 speeches did not reveal the improvement of higher order language models which Table 1 shows. The main aim of this work is to build an ASR corpus and we believe the expansion errors do not affect the acoustic model considerably nor the language model used in training the ASR. The phones in the incorrectly expanded words, appear correctly in numerous other unexpanded words and, in majority of cases the expansion is correct. Hence, we consider the main effect of these errors to be an overestimate of the ASRs word error rate (WER). We will however use the higher order language model in future work.

Table 1: *Expansion error rates for different language model's n-gram orders. The upper line shows results for 7 common parliament abbreviations, i.e. "hv.", "hæstv.", "þm.", "þskj.", "gr.", "mgr." and "tölul.". The abbreviations are expanded 393 times, in 100 randomly chosen parliament speeches. The lower line shows results for up to 53 general abbreviations and numbers, that have multiple expansions. They are expanded 279 times.*

	%Incorrect		
	2-gram	**3-gram**	**5-gram**
Althingi abbreviations	10.6	10.1	9.1
General abbr. + numbers	10.6	10.1	10.2

Sometimes an abbreviation or a number can be correctly expanded to two different set of words based on the surrounding context. The expansion can then result in an error, even if it was

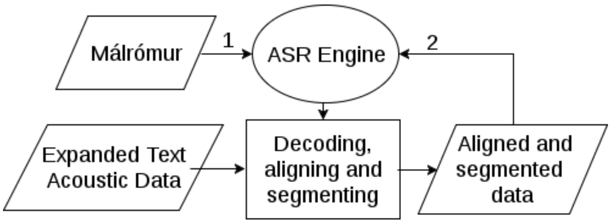

Figure 1: *The flowchart shows how open vocabulary recognizer was used to align and segment a part of the parliament data. That data was then in turn used to train a small in-domain recognizer which could be used to align and segment the rest of the data. Arrow 1 denotes the initial training of an ASR engine from the Málrómur speech corpus. Arrow 2 denotes the in-domain training of 60 hours of aligned and segmented parliament data into an ASR engine. The final output is the aligned and segmented Althingi speech corpus. Outcomes from ASR engines, trained on the resulting corpus, can be seen in Sec. 5.3.*

correctly executed. For example, in the parliament transcripts, both “það er" and “það er að segja" have been abbreviated as “þ.e." and they are interchangeable. It is also not apparent in the text whether the speaker said “two point five" or “two and a half" from a transcript with the text “2.5". These errors will increase the word-error-rate estimate of the ASR but are unlikely to affect training drastically.

3.2. Segmentation and alignment of data

The segmentation and alignment of the database is done in two stages as depicted in Fig. 1. Two speech recognizers are used for this purpose: an experimental version of a large vocabulary ASR system for Icelandic, and a domain specific recognizer trained on the results of the first segmentation round. The LVSR system for Icelandic (in development) was trained on 170 hours from the Málrómur corpus [7]. The acoustic model was a conventional GMM system, built on a standard 13-dimensional cepstral mean-variance normalized Mel-frequency cepstral coefficients (MFCC). Linear discriminative analysis (LDA) was used for dimension reduction, followed by a maximum likelihood linear transform (MLLT) [12]. An extended and modified version of the Icelandic pronunciation dictionary was used (see Section 1.2) along with a trigram language model trained on the Leipzig database. The aligned data was processed into segments of maximum 15 seconds of audio, using a procedure based on the first alignment stage of the Librispeech corpus creation [1, 12]. The main parts of the segmentation procedure are the following: The long audio is split into 30 second long segments with 5 second long overlaps. One decoding graph is built for each utterance segment. The utterance segments are then decoded, and information is obtained about the begin time, duration of each word in the utterance and word-error-rate. This is then used to work out a new segmentation. Segments with a 90% word-error-rate or higher were discarded. To filter the data further, we calculate the uttered words per second for each segment. Average values and percentiles were calculated for each speaker and the values were used to discard segments where the word per second count was outside the speaker's tenth and ninetieth percentile.

A new speech recognizer was trained using 60 hours of the newly aligned parliamentary speeches. We used a triphone model trained with LDA and MLLT, as before, but we added speaker adaptive training and boosted maximum mutual information (MMI) parameter estimation (with boost weight 0.1).

During the first alignment round about 25% of the data could not be aligned and was filtered away. During the 2nd round that number was 18.6%.

4. Database and corpus structure

The aligned data set contains of 6493 recordings with 196 speakers. The recordings consist of 199,614 segments, with average duration of 9.8 s. The total duration of the data set is 542 hours and 25 minutes of data and it contains 4,583,751 word tokens. This is a reduction of 10.0% in text and 18.6% in audio from the original data set.

The aligned Althingi data set was split up into a training-, development- and an evaluation set. The training set was obtained from speeches from 2005 to 2015, with a total duration of 514.5 hours. It contains 192 speakers, 104 men and 88 women. However, even though the women are 46% of the speakers, they only speak 35% of the words in the training set. The speeches from 2016 were split evenly between the development- and evaluation sets, with 14 hours in duration each. There is an overlap between speakers in the training and test sets, with only four speakers exclusively in the test sets. The development and the evaluation set contain 59 speakers, 29 men and 30 women, and here women speak 49% of the words.

The training data was re-segmented, selecting only the more accurate matches between audio and transcripts, using a method, which is similar to the one described in Section 3.2, except it also does minor modifications of transcripts, such as allowing repetitions. The re-segmentation process filtered away roughly 57 hours of unclean data. However, preliminary results show no advantage by training on only the cleaned-up data.

Some effort was put into manually cleaning the development and evaluation test sets. The audio was decoded, and without looking at what caused the errors, the numbers of correct words, insertions, deletions and substitutions, were extracted for each segment. If the number of correct words divided by the sum of correct, deleted, inserted and substituted words, was lower that 0.7 and 0.75 for the development and evaluation sets, respectively, the segment was listened to, and eventual transcript or segmentation errors were fixed. The most significant sources of mismatch in the corpus stem from pre-existing transcript edits, inaccurate text normalization and noisy recording circumstances.

5. Speech Recognition

The main aim of this work was to compile a speech recognition database for Althingi's parliamentary speeches. We did however also assess four standard speech recognition architectures implemented in the Kaldi ASR toolkit [12], using the dataset. A language model was also built for the assessment using transcripts of Althingi's speeches dating back to 2003.

5.1. Acoustic models

A hidden Markov model was trained using Gaussian mixture models (HMM) [13, 14], applying LDA and MLLT, and using features, which were speaker adapted with feature-space Maximum Likelihood Linear Regression (fMLLR). This model is marked as GMM-SAT in Table 2.

Three deep neural network (DNN) architectures were also evaluated. All of them are a HMM-DNN Tandem architectures that predict the probabilities of the context-dependent state of a HMM. The first two are feed-forward neural networks and the last one is a recurrent neural network (RNN).

The first architecture is described in [15]. It is a maxout network, which means the non-linearity is dimension-reducing. It contains 4 hidden layers, with p-norm non-linearities, where $p = 2$. The p-norm input dimension is 2000 and the output dimension is 400. The output layer is a soft-maxout layer, whose output dimension equals the number of context-dependent states in the system, 3451 in our case. This model is marked as DNN in Table 2.

A feed-forward deep neural network architecture, called time-delay deep neural network (TD-DNN) [16] was also evaluated. The setup is described in [17]. However, we used the Wall-Street Journal Kaldi recipe, which varies slightly from the one in the paper. We used, for example, a rectified linear unit (ReLU), with dimension 450, instead of a p-norm non-linearity, and different splicing indices were applied. The TD-DNN architecture is capable of learning long temporal dependencies in the data by letting each layer operate at a different temporal resolution. This implementation uses 100 dimensional iVectors [18] as input to the neural network, to make instantaneous speaker and environment adaptations to the network. This model is marked as TD-DNN in Table 2.

A long-short term memory implementation, as described in [19], was evaluated. It trains a RNN acoustic model, using the cross-entropy objective. The Switchboard LSTM recipe in Kaldi was used. The LSTM RNN contains 3 LSTM layers, with different LSTM-delays at each layer, and a label-delay of 5 frames. Each layer contains 1024 memory cells. The memory cells contain input, forget and output gates with sigmoidal non-linearities, cell input and output activations with tanh non-linearities, and 256 dimensional recurrent and non-recurrent projection layers, with linear activation units. The output dimension is equal to the number of context dependent states in the system, or 3451. Like in the TD-DNN case, iVectors are used as inputs to the neural network. This model is marked as LSTM-RNN in Table 2.

Finally, the TD-DNN architecture was retrained by synthesizing more training data using speed and volume perturbations [20]. Volume perturbations were applied to a copy of the original training data, by scaling the volume of each segment with a uniform random variable between 0.125 and 2.0. Then two extra versions of the training data were synthesized by changing the speed by a factor 0.9 and 1.1. This model is marked as TD-DNN w/sp in Table 2.

5.2. Language Models

The language model training material consists of speech transcripts, from the years 2003 to 2011, scraped from the Althingi website, approx. 30 million word tokens, as well as the parliament training data, up to and including 2015. There is therefore no overlap between the these transcripts and the texts in the development and evaluation sets.

The lexicon used is based on the pronunciation dictionary from the Hjal project [6], available at Málföng[2]. We added words from the language model training data, which appeared three times or more. Words that included foreign letters, numbers, punctuations or no vowels were excluded. A list was created of all the words whose first four letters did not fit with a four letter n-gram and manually checked. About 75% of those misfitted words were kept in the dictionary. The out of vocabulary token rate in the test sets is 1.4%, but 0.96% in the training data.

Phonetic transcriptions were created for the words that did not already have a transcription by using the Sequitur G2P toolkit [21]. The pronunciation of Icelandic is fairly transparent as the projection from the spelling to the pronunciation obeys very consistent rule. Thus, the pronunciation generation of Sequitur is quite accurate.

Modified Kneser-Ney smoothed 3- and 5-grams [22, 23] were built using the MIT Language Modeling (MITLM) toolkit [24]. A pruned 3-gram language model was built using KenLM [25], before the LSTM decoding, to enable faster decoding. The 5-gram language model is used for re-scoring decoding results.

5.3. Results

Table 2 shows the word-error-rate for the four different acoustic models and the TD-DNN model trained with speed perturbation. The table shows that the best results, of 14.76%, are achieved by using the LSTM-RNN architecture. It also shows that the speed perturbation does not increase the performance of the TD-DNN much. It is quite possible that the adding synthesized data to the training data might be of limited value to this problem since the acoustic environment of the Althingi constrains the problem enough, but further studies are needed to test that hypothesis.

Table 2: *Word error rate for different acoustic models, using all available speech training data. All results are obtained by rescoring with a 5-gram language model. "sp" stands for speed perturbations.*

Acoustic model	Development set	Evaluation set
GMM-SAT	22.61	22.24
DNN	17.48	17.28
TD-DNN	16.71	16.38
TD-DNN w/sp	16.44	16.20
LSTM-RNN	15.17	14.76

The WER for the different speakers varied from 7.50% to 26.89%, with a median of 14.75% for the LSTM RNN system. The differences stem from how similar the speakers' spoken language is to written language and how clearly he/she speaks. The use of filler words, repetitions, speaking fast or not clearly, all result in a lower score for that speaker.

6. Conclusions

This paper presents a corpus of aligned and segmented Icelandic parliamentary speeches which is suitable for training speech recognition systems. The challenges involved were text normalization of an inflected language, dealing with edited transcripts and aligning (sometimes noisy) speech recordings. The preliminary results, however, show that the resulting corpus is well suitable for training speech recognition systems. The corpus is made available on the Málföng website (http://www.malfong.is) with an open CC-BY 4.0 license and releasing the Kaldi recipes are also planned.

7. Acknowledgements

This project was made possible through the support of Althingi's information and publications departments. The authors would like to thank Solveig K. Jónsdóttir, Þorbjörg Árnadóttir and Ingvi Stígsson for their valuable help.

[2] http://www.malfong.is

8. References

[1] 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, pp. 5206–5210.

[2] P. J. Jang and A. G. Hauptmann, "Improving acoustic models with captioned multimedia speech," in *Multimedia Computing and Systems, 1999. IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 767–771.

[3] G. Adda, M. Adda-Decker, J.-L. Gauvain, and L. Lamel, "Text normalization and speech recognition in french," *training*, vol. 3, pp. 4–0, 1997.

[4] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer speech & language*, vol. 15, no. 3, pp. 287–333, 2001.

[5] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[6] E. Rögnvaldsson, "The Icelandic speech recognition project Hjal," *Nordisk Sprogteknologi. Årbog*, pp. 239–242, 2003.

[7] J. Guðnason, O. Kjartansson, J. Jóhannsson, E. Carstensdóttir, H. H. Vilhjálmsson, H. Loftsson, S. Helgadóttir, K. Jóhannsdóttir, and E. Rögnvaldsson, "Almannarómur: an open Icelandic speech corpus." in *SLTU*, 2012, pp. 80–83.

[8] H. Bernódusson, "Filibustering in the Alþingi." Lusaka: Interparliamentary Union, March 2016, Association of Secretaries General of Parliaments.

[9] T. Tai, W. Skut, and R. Sproat, "Thrax: An open source grammar compiler built on OpenFst," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[10] B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, and T. Tai, "The OpenGrm open-source finite-state grammar software libraries," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 61–66.

[11] U. Quasthoff, D. Goldhahn, and E. Hallsteinsdóttir, "Technical report series on corpus building (vol. 4)." Leipzig: University of Leipzig, 2013.

[12] 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. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[13] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1137–1140.

[15] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.

[16] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.

[17] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *INTERSPEECH*, 2015, pp. 3214–3218.

[18] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černockỳ, "ivector-based discriminative adaptation for automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 152–157.

[19] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[20] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH*, 2015, pp. 3586–3589.

[21] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434–451, 2008.

[22] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 181–184.

[23] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 310–318.

[24] B.-J. P. Hsu and J. R. Glass, "Iterative language model estimation: efficient data structure & algorithms." in *INTERSPEECH*, 2008, pp. 841–844.

[25] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 187–197.