Automatic Speech Recognition for ILSE-Interviews: Longitudinal Conversational Speech Recordings covering Aging and Cognitive Decline

Ayimunishagu Abulimiti, Jochen Weiner, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany

ay.abulimiti@uni-bremen.de

Abstract

The *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) was initiated with the aim to investigate satisfying and healthy aging. Over 20 years, about 4200 hours of biographic interviews from more than 1,000 participants were recorded. Spoken language is a strong indicator for declining cognitive resources, as it is affected in early stage. Hence, various research topics related to aging like dementia, could be analyzed based on data such as the ILSE interviews. The analysis of language capabilities requires transcribed speech. Since manual transcriptions are time and cost consuming, we aim to automatically transcribing the ILSE data using Automatic Speech Recognition (ASR). The recognition of ILSE interviews is very demanding due to the combination of various challenges: 20 year old analog two-speaker one-channel recordings of low signal quality, emotional and personal interviews between doctor and participant, and repeated recordings of aging, partly fragile individuals. In this study, we describe ongoing work to develop hybrid Hidden Markov Model (HMM)- Deep Neural Network (DNN) based ASR system for the ILSE corpus. So far, the best ASR system is obtained by second-pass decoding of a hybrid HMM-DNN model using recurrent neural network based language models with a word error rate of 50.39%.

Index Terms: automatic speech recognition (ASR), conversational speech, ILSE, Cognitive Aging

1. Introduction

Around 50 million people worldwide are suffering from dementia and every year 10 million of new dementia cases are added [1]. Not only the people with dementia, but also their families and friends are affected. One way to alleviate the serious effects of dementia is to start therapy early, which requires to diagnose dementia as early as possible. To allow for wide-spread casual testing and support of clinical diagnosis, several automatic methods leveraging speech have been investigated [2, 3, 4, 5]. As speech and language require cognitive resources, they are affected in the early stages of dementia, and thus can be strong indicators for detecting dementia. Studies using acoustic or linguistic features for classification tasks which are initiated to differentiate speakers with dementia from healthy speakers have shown promising results [6, 7, 8, 4, 9, 10].

The *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) was initiated at the beginning of the 1990s with the aim to investigate satisfying and healthy aging in middle adulthood and later life [11]. In a time span of over 20 years, about 4200 hours of biographic interviews from more than 1,000 participants were recorded at four measurement times. As the biographic interviews covered a wide range of topics, such as personal life, family, hobby, career, social relationships, satisfaction with the physical and political environment, ILSE enables aging related research in many disciplines, including but not limited to geriatrics, psychology, ecological gerontology, sociology, history, and linguistics [12].

These disciplines mostly work on syntactic and semantic representations of spoken conversation and thus require orthographic transcriptions of the interviews. In the past 20 years of the ILSE project, about 442 hours of interviews have been manually transcribed. To speed up this process and make the data available to disciplines listed above and to improve automatic dementia screening, we aim to fully automatically process and transcribe these data. At the beginning of the ILSE project, automatic speech processing was not anticipated. Hence, little attention was focused on the signal quality, noise conditions, and overall recording setup. Because of the interview topics, the participants were emotional during the major part of the interviews. Due to the one-channel setup, many cross-talk effects occur. Furthermore, participants came from parts of Germany with distinct dialects. The manual transcripts were provided with no time alignment between audio recording and transcripts. The joint occurrence of these challenges complicate the development of a good ASR system.

In this paper, we describe our latest work towards developing ASR for ILSE interviews, which leverages off our previous work [13]. Our new developments include (1) a more robust hybrid Hidden Markov Model-Deep Neural Network (DNN) based speech recognition system, (2) an advanced and extended version of the pronunciation dictionary, which includes pronunciation variants of pseudonymized items to limit the mismatch between the speech signal and word-level transcripts, and (3) an improved acoustic model using larger amounts of manual transcripts and improved speaker diarization and more appropriate segment splitting. In addition, we selected text data similar to ILSE from large amounts of Common Crawl Data[1] based on Term Frequency–Inverse Document Frequency (tf-idf) for the training of Recurrent Neural Network (RNN) based language models.

This paper is organized as follows: In the next section, we introduce the ILSE study and data collection (Section 2) as well as the challenges for developing ASR for ILSE (Section 3). After describing the speech and text corpora used for ASR development (Section 4), we introduce ASR developing experiments in Section 5, discuss the results in Section 6 and conclude in Section 7.

2. The ILSE study

ILSE is a longitudinal study and was initiated with the aim to explore various conditions for achieving physically and mentally healthy aging in middle adulthood and later life [14, 12]. The biographical data was collected from more than one thousand participants from east and west Germany in two cohorts.

[1] http://commoncrawl.org/

Participants were born between 1930-1932 and 1950-1952. The participants were equally distributed by sex and cohort. So far, four measurements have been conducted and more than 4200 hours of interviews were recorded. The biographic interviews were conducted in semi-standardized manner, where participants were asked to give detailed elaborations on the standardized open questions of the interviewers. During the interview, the participants had enough time to think about their answers. The average duration of the interview became shorter with each measurement time, since the gathered biographic information accumulated over time and the interviewer questions were reduced over time. Further details about the ILSE corpus design and procedures can be found in Martin et al. (2000) [12].

The speech recordings of the first two measurements were stored on tapes. From third measurement, digital recording devices were used for the recording. For the speech and language analysis, all interviews have been digitized using a sampling rate of 16kHz, a quantization of 16 bit resolution, and a linear uncompressed encoding in PCM format.

From the interviews of more than 1000 participants, as of today approximately 442 hours of interviews from 74 participants have been manually transcribed. The manual transcripts were provided either per interview (ca. 45 min corresponding to the duration of analog tapes) or part of the interview (ca. 0.3 - 2 hours) and there were no further alignments between recordings and word-level transcripts. In addition, the transcription quality varies between interviews.

3. Challenges for ASR development

Due to the fact that automatic speech processing was not anticipated in the original ILSE study, not much effort was taken to record clean speech [13]. The recordings include interviewers' and participants' speech but also periods with no speech. Quite a lot of background noise from the interview room occurs such as paper shuffling, table tapping etc. and noises from the clinics environment, like ambulance siren, traffic, and alike. Due to the interview guidelines and topics, there is cross-talk between participants and interviewers and major parts of the speech were emotional.

Regarding the manual transcripts, the quality ranges from reliable verbatim to summary script. Moreover, since no further alignments between audio and transcripts within an interview were provided, only the alignments between one side of a cassette of the interview (about 45 minutes) and the corresponding transcripts were available for acoustic training. For the reason of privacy protection, personally identifiable information such as proper names etc., were substituted with generic place holders. The combination of these factors in the ILSE data makes the task of ASR very challenging.

4. ILSE Data for ASR development

After our previous work described in Weiner et al. [13], transcripts of 17 additional speakers have been manually transcribed. These new transcriptions were created with a time alignment at the sentence level and cross-talk was identified.

4.1. Data division

As more interviews have been transcribed after our previous work [13], we re-distributed the data into training, development and test set, but still followed the original constraints [13]. The number of participants, audio duration and ratio of audio length in each set is summarized in Table 1.

Table 1: *Number of participants, audio duration [hrs:min] and data proportion of the ILSE training, development, and test sets*

	Training	**Development**	**Test**
Participants	68	12	11
Audio duration	355:49	43:11	44:24
Proportion	80.24%	9.74%	10.02%

In order to improve the training of ASR, we also put effort into the alignment between the existing word-level transcriptions and reasonable lengths of audio segments . The total duration of segments over different length-ranges is shown in Table 2. Currently, 80% of the segments are shorter than 30 minutes, but 20% of segments (about 86 hours) are longer than 30 minutes.

Table 2: *Total duration of segments (in [hrs:min]) over length ranges (in minutes) for ILSE training, development, and test set*

Segment length	**1 - 10**	**10 - 15**	**15 - 30**	**30 - 60**
Training	155 : 11	145 : 39	22 : 46	31 : 04
Development	13 : 23	3 : 01	2 : 23	24 : 23
Test	7 : 00	2 : 16	4 : 52	31 : 38
Total duration	175 : 34	150 : 56	30 : 01	86 : 05

4.2. Text corpus

4.2.1. In-domain text data

As mentioned in Weiner et al. [13], the manual transcripts were unstructured and used different formats. The newly obtained manual transcripts were pre-processed with a similar procedure as in our previous work. Personally identifiable information has been substituted with generic place holders. Digits, abbreviations and annotations like pauses and hesitations have been normalized. The statistics of word types (number of distinct words in the transcriptions), word tokens (total number of words in the transcriptions) and utterances in each data set are shown in Table 3.

Table 3: *Amount of word types, word tokens and utterances in the ILSE training, development and test sets*

	Training	**Development**	**Test**
Word types	$56\,k$	$16\,k$	$16\,k$
Word tokens	$2.98\,M$	$339\,k$	$368\,k$
Utterances	$41,270$	$8,122$	$3,623$

4.2.2. Out-domain text data

For the training of Recurrent Neural Network-based language models (RNNLM), external text data are used in addition to the ILSE training text data. Since ILSE consists of auto-biographic interview, more emphasis is put on selecting data which has possibly similar content to biographic interviews. Fortunately, freely accessible text corpora like Common Crawl are available. We use the semi pre-processed Common Crawl data in German

from Buck et al. [15] as data source (about 640 GB). Before selecting data similar to the ILSE corpus, the Common Crawl data was normalized. Digits were converted to words; commonly used abbreviations were substituted with their full form and common misspellings in German were corrected; all punctuation were removed and all words were lowercased.

As similarity measure between text data, we applied the Term Frequency–Inverse Document Frequency (tf-idf). The following selection procedure was applied: Firstly, tf-idf scores of the 5-gram-terms of the normalized ILSE data were calculated and 500 5-gram-terms with highest tf-idf score were selected as more interesting terms for the ILSE corpus. Then sentences which contained any of the selected 5-gram-terms were selected from the normalized Common Crawl data. With these steps, about 150 GB data has been selected and used for RNNLM training.

4.3. Pronunciation dictionary

In the step of building a pronunciation dictionary, all words in the training data are extracted and included in the pronunciation dictionary. As mentioned in Section 4.2.1, proper names, city and street names were substituted with generic place holders in transcripts, which consequently causes a mismatch between speech and corresponding word-level transcriptions. To reduce potential errors from these mismatches, pronunciation variants for generic place holders are defined. For example, all female name are anonymized as *[anonymised_firstname_female]* in the transcripts. For this dictionary entry, we added pronunciations of real female names as pronunciation variants. The Out-Of-Vocabulary (OOV) rate of the resulting pronunciation dictionary on the development set is 1.64%.

To reduce the OOV-rate, most frequent words are extracted from external corpora, such as the *Biographische und Reiseerzählungen* [16], *Pfeffer-Korpus* [17] and *Quaero*[2]. These corpora contain over 2M word types and about 194M word tokens. Those most frequent 10k words of the external corpora, which did not appear in the ILSE training data, are added to the pronunciation dictionary. The pronunciations of these words are generated using Sequitur G2P [18], which was trained with the ILSE-only pronunciation dictionary. After adding the words from the mentioned external corpora, the final pronunciation dictionary consists of 61.7k words and gives an OOV-rate on the development of 1.50%.

5. ASR development

Recently, speech recognition technology has gained significant improvements and matches human performance for selected tasks. However, data sparsity is still an issue for training reliable acoustic and language models, especially DNN-based models. In our previous work [13], we developed a hybrid HMM-DNN based ASR for ILSE data. However, the Word Error Rate (WER) of the system was very high. From the experience of developing ASR in previous work and considering the specialty of ILSE corpus, we steered our attention toward the following issues: In our previous experiments, hybrid HMM-DNN based ASR systems outperformed HMM-Gaussian Mixture Model (GMM). However, for training a reliable DNN based acoustic model, the available ILSE data for ASR development was not enough. For that reason we continue our effort to manually transcribe and add additional interviews, so that more data can be used for supervised ASR training. In the meantime, state-of-the-art neural network architectures, like CNN-TDNNf for acoustic training and RNN for language modeling are used. Another point for improving ASR performance could be an audio segmentation into suitable segments. Since unaligned transcripts for audio recordings with duration of at least 45 minutes were provided, the segments for training acoustic model were too long. Therefore, we put effort to segment the data into appropriate segment lengths (Section 4). Compared to the previous work, most of the segments have shorter length in this work.

Moreover, RNN-based language models have shown promising results [19, 20, 21]. In this study, RNN based language models are trained and used to perform lattice rescoring.

Right after our previous study [13], a hybrid HMM-DNN based ASR system was developed as baseline system with similar fashion as in this study. This baseline system is used for experiments with force alignment.

5.1. Acoustic modeling

Two sets of acoustic modeling experiments are conducted: one uses our baseline system [13] to produce initial forced alignments, the other uses flat-start for HMM-GMM initialization. All hybrid HMM-DNN based ASR systems are developed using the open-source Kaldi ASR toolkit [22].

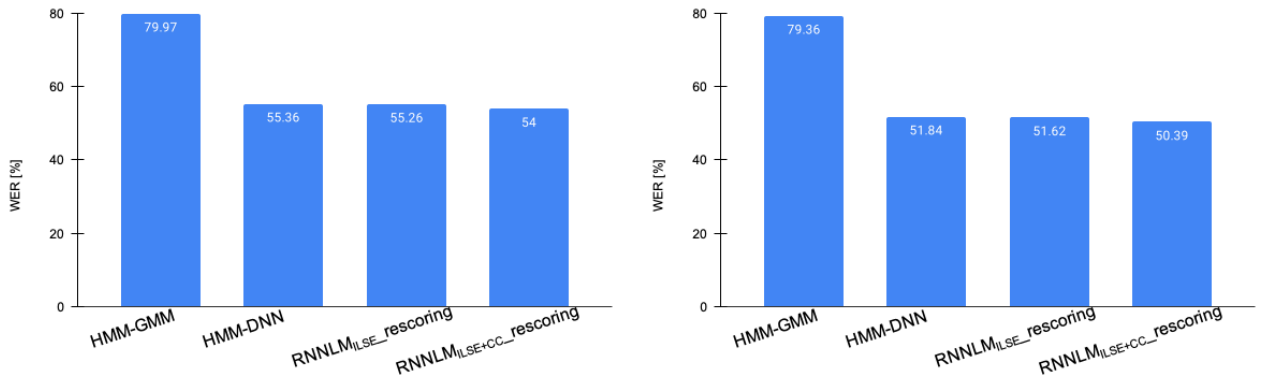
For training DNN-based acoustic models, HMM-GMM based recognition systems are first developed to get alignments. The HMM-GMM based acoustic models are built using 39-dimensional stacked Mel-Frequency Cepstral Coefficients (MFCCs). We apply Cepstral Mean and Variance Normalization (CMVN) with context size of 7 frames. Subsequently a Linear Discriminant Analysis (LDA) + Maximum Likelihood Linear Transform (MLLT) model is generated. Finally, Speaker Adaptive Training (SAT) is performed using an affine transform, feature space Maximum Likelihood Linear Regression (fMLLR). Based on the Switchboard (SWBD) recipe that has been released together with Kaldi, we tune the system to optimize the parameters for numbers of states and Gaussians. In all experiments, fMLLR models outperform the monophone, triphone and LDA+MLLT models. Hence, the alignments from fMLLR models are used for DNN training.

For DNN-based acoustic training, a Factored Time Delay Neural Network with additional Convolutional layers (CNN-TDNNf) is used. 40-dimensional cepstral truncation, 3-dimensional pitch features and 100-dimensional iVectors [23] are given as input to the neural network. The structure of hidden layers and training epoch are tuned based on the WSJ recipe. The current best acoustic model is trained with 19 hidden layers (9 CNN layers followed by 10 TDNNf). The first TDNNf layer following just after the CNN layer consists of 256 bottleneck units and the other 9 TDNNf layers have 1024 nodes and 128 bottleneck units. All acoustic models are trained with the same hyper-parameters, i.e., initial learning rate of 0.005, final learning rate of 0.00005, and minibatch-size of 128, 64. The best acoustic model is trained with 7 epochs.

5.2. Language modeling

For first-pass decoding, we use a traditional 3-gram language model (LM), which consists of an interpolated language model based on a 3-gram LM trained with ILSE data only interpolated with an LM trained from the *Biographische und Reiseerzählungen*, *Pfeffer-Korpus* and *Quaero* corpora. Both 3-gram LMs are trained by applying modified Kneser-Ney discounting [24] and interpolated with the best interpolation weight optimized on a held-out set. For training and evaluat-

[2]http://www.quaero.org/



(a) (b)

Figure 1: *Performance of ASR systems, with (a) forced-aligned acoustic model initialization and (b) flat start*

ing language models, we used the SRILM toolkit [25]. The 3-gram perplexity (PPL) of the interpolated language model on the development set is 103.5.

Two RNNLMs are trained using the ILSE training data only ($\text{RNNLM}_{\text{ILSE}}$) and selected Common Crawl data in addition to ILSE training data ($\text{RNNLM}_{\text{ILSE+CC}}$) as described in section 4.2.2. The RNNs are a mixture of three TDNN layers with two Long short-term memory (LSTM) layers with 800 dimensions. The vocabulary of the pronunciation dictionary is used as vocabulary of the language models. RNNLM training is performed with the open-source Kaldi ASR toolkit [22]. The perplexity of the $\text{RNNLM}_{\text{ILSE}}$ is 109.7. The perplexity of the $\text{RNNLM}_{\text{ILSE+CC}}$ is 159.7.

6. Experimental results

Two hybrid HMM-DNN recognition systems are created in this study, one with acoustic models initialized with forced alignment and one with acoustic models initialized with flat start. The WER of baseline hybrid HMM-DNN model, which is used to provide the forced alignment achieves 56.59% WER on the development set.

Decoding was performed in a two-step approach. In the first pass, the 3-gram LM was used to produce lattices for the second-pass decoding. In the second-pass decoding, the lattices are rescored using 5-gram approximation of the RNNLMs. We present the first-pass decoding performances of (1) the HMM-GMM model which provided forced alignments, (2) the resulting HMM-DNN model, (3) the second-pass decoding of the HMM-DNN model using the $\text{RNNLM}_{\text{ILSE}}$, and (4) using the $\text{RNNLM}_{\text{ILSE+CC}}$. Figure 1 compares the evaluation results of these four ASR recognition systems, one time initialized with forced alignment (Figure 1a) and another time with flat start (Figure 1b), respectively.

Comparing the four ASR systems we observe – not too surprisingly – that the HMM-DNN based models outperform the corresponding HMM-GMM models by more than 30% relative in the first-pass decoding. In the second-pass decoding the RNNLMs achieve a small WER reduction compared to first-pass decoding. The $\text{RNNLM}_{\text{ILSE+CC}}$ outperforms the $\text{RNNLM}_{\text{ILSE}}$ in terms of WER, although the latter gave a much lower perplexity.

Compared to the first-pass decoding results, the $\text{RNNLM}_{\text{ILSE+CC}}$ achieves on average a 2.62% relative WER improvement, and only 0.81% relative WER improvement is gained when using the $\text{RNNLM}_{\text{ILSE}}$.

Finally, the initialization variation results in significant performance difference of the hybrid HMM-DNN model. Comparing the right and left side of Figure 1 shows that the flat start initialization outperforms models resulting from the forced alignment, indicating issues with our baseline system from prior work. Our currently best performing ASR system achieves 50.39% WER on the development set of the ILSE data.

7. Conclusions

We presented our effort to develop an ASR system that automatically transcribes biographic ILSE interviews. The resulting transcripts will be provided to researchers engaged in research on adult development and aging. Furthermore, we aim to apply our work to the automatic extraction of linguistic features for casual low-cost and widespread screening of age-related cognitive decline from conversational speech.

We took several steps forward to overcome challenges of the ILSE data. In particular, we added pronunciation variants for anonymized items to mitigate the mismatch between speech and transcripts. We retrieved topic-relevant data from a large corpus in order to train more reliable RNN based language models. We improved the acoustic models by increasing the amount of data for supervised training, split speech segments into more appropriate length, and studied the impact of model initialization on the HMM-DNN models. After applying the RNNLMs for the second-pass decoding, the final ASR system for conversational ILSE interviews achieves a WER of 50.39%. This performance clearly leaves room for improvements. Among the open challenges, we plan to address regional dialect variations, cohort specific modeling, and to look into the details of the poor audio quality and noise robust acoustic models. This will also pave the road for unsupervised or semi-supervised training using all of the available speech data.

8. Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project "ALMED - acoustic and linguistic features for early prediction of cognitive deficits" (403605461).

9. References

[1] W. H. Organization, *Dementia*, 19 September 2019 (accessed May 3, 2014). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[3] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in alzheimer patients," *Brain and language*, vol. 17, no. 1, pp. 73–91, 1982.

[5] J. Weiner and T. Schultz, "Selecting Features for Automatic Screening for Dementia Based on Speech," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Cham: Springer International Publishing, 2018, pp. 747–756.

[6] S. Wankerl, E. Nöth, and S. Evert, "An analysis of perplexity to reveal the effects of alzheimer's disease on language," in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[7] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[8] C. Frankenberg, J. Weiner, T. Schultz, M. Knebel, C. Degen, H.-W. Wahl, and J. Schröder, "Perplexity - a new predictor of cognitive changes in spoken language?: - results of the interdisciplinary longitudinal study on adult development and aging (ilse)," *Linguistics Vanguard*, vol. 5, 06 2019, s2. [Online]. Available: https://www.degruyter.com/view/j/lingvan.2019.5.issue-s2/lingvan-2018-0026/lingvan-2018-0026.xml

[9] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[10] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 9, 2015.

[11] C. Sattler, H.-W. Wahl, J. Schröder, A. Kruse, P. Schönknecht, U. Kunzmann, T. Braun, C. Degen, I. Nitschke, W. Rahmlow, P. Rammelsberg, J. S. Siebert, B. Tauber, B. Wendelstein, and A. Zenthöfer, *Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE)*. Singapore: Springer, 2017, pp. 1213–1222.

[12] P. Martin, M. Grünendahl, and M. Schmitt, "Persönlichkeit, kognitive leistungsfähigkeit und gesundheit in ost und west: Ergebnisse der interdisziplinären längsschnittstudie des erwachsenenalters (ilse)," *Zeitschrift für Gerontologie und Geriatrie*, vol. 33, no. 2, pp. 111–123, 2000.

[13] J. Weiner, C. Frankenberg, D. Telaar, B. Wendelstein, J. Schröder, and T. Schultz, "Towards automatic transcription of ilse—an interdisciplinary longitudinal study of adult development and aging," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 718–725.

[14] C. Sattler, H.-W. Wahl, J. Schröder, A. Kruse, P. Schönknecht, U. Kunzmann, and A. Zenthöfer, "Interdisciplinary longitudinal study on adult development and aging (ilse)," *Encyclopedia of geropsychology*, pp. 1–10, 2015.

[15] C. Buck, K. Heafield, and B. Van Ooyen, "N-gram counts and language models from the common crawl." in *LREC*, vol. 2. Citeseer, 2014, p. 4.

[16] C. Fandrych, E. Frick, H. Hedeland, A. Iliash, D. Jettka, C. Meißner, T. Schmidt, F. Wallner, K. Weigert, and S. Westpfahl, "User, who art thou? user profiling for oral corpus platforms," 2016.

[17] A. Pfeffer, W. F. Lohnes, and W. D. Ortmann, *Textkorpora 1: Grunddeutsch. Texte der gesprochenen deutschen Gegenwartssprache. Überregionale Umgangssprache aus der Bundesrepublik Deutschland, der DDR, Österreich und der Schweiz*. Walter de Gruyter, 2011, vol. 28.

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

[19] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 1045–1048.

[20] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[21] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[22] 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.

[23] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[24] F. James, "Modified kneser-ney smoothing of n-gram models," *Research Institute for Advanced Computer Science, Tech. Rep. 00.07*, 2000.

[25] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.