Description of the UEDIN System for German ASR

Joris Driesen, Peter Bell, Mark Sinclair, Steve Renals

Center for Speech Technology Research, University of Edinburgh, UK

{jdriesen,peter.bell,s.renals}@inf.ed.ac.uk, M.Sinclair-7@sms.ed.ac.uk

Abstract

In this paper we describe the ASR system for German built at the University of Edinburgh (UEDIN) for the 2013 IWSLT evaluation campaign. For ASR, the major challenge to overcome, was to find suitable acoustic training data. Due to the lack of expertly transcribed acoustic speech data for German, acoustic model training had to be performed on publicly available data crawled from the internet. For evaluation, lack of a manual segmentation into utterances was handled in two different ways: by generating an automatic segmentation, and by treating entire input files as a single segment. Demonstrating the latter method is superior in the current task, we obtained a WER of 28.16% on the dev set and 36.21% on the test set.

Index Terms: Light supervision, Segmentation, Acoustic Model Training

1. Introduction

In ASR, good acoustic models are an important prerequisite for high recognition accuracies. The quality of these models is determined by both the quality and the quantity of the data on which they were trained. Such data consists of speech as well as accurate orthographic transcriptions. Since the latter must be manually created by human transcribers, which is a slow and expensive process, it can be difficult to obtain training data in sufficiently large quantities. In languages or domains where resources are scarce, i.e., where no large amounts of dedicated transcribed training is available, acoustic models can still be obtained from untranscribed or poorly transcribed data, using unsupervised or lightly supervised training methods [1, 2, 3, 4, 5]. Since German ASR has historically received little attention at UEDIN, there are very few resources available for it on site. Therefore, even though German is by no means an under-resourced language, we have been compelled to treat it as such, collecting large amounts of publicly available data and processing it with the lightly supervised training methods mentioned above. Although this methodology is not strictly necessary for German, it can in theory be applied to unlock other, truly under-resourced languages, for which no alternative training methods exist. The available resources used for acoustic model training are discussed below in section 2. The lightly supervised training is explained fully in section 4. Acoustic model training is finalised by training a Deep Neural Network (DNN) in a hybryd setup with a traditional context-dependent tri-phone based Hidden Markov Model (HMM), as explained below, in section 6.

Aside from acoustic modelling, the proposed system has state-of-the-art language modelling. In a first phase, text corpora are collected, containing in total almost 10^9 words. Based on the cross-entropy with the evaluation domain, as proposed in [6], the top 30 percentile of this data is selected and 4-gram language models, as well as Recurrent Neural Network Language Models (RNNLM) are trained on it [7]. Details of this setup can be found below, in section 5.

Since no manual segmentation for the evaluation set is provided, it is necessary to produce a segmentation automatically. Alternatively, ASR can be performed on entire talks, treating them as a single segment. There is an inherent trade-off between these approaches, since each has its own advantages and disadvantages. A segmentation that is generated automatically may contain erroneous segment boundaries, which can easily lead to recognition errors. When segmentation is avoided, on the other hand, recognition could be performed on non-speech segments, generating unpredictable erroneous outputs. In section 6, evaluation is performed comparing both approaches.

2. Available Resources for Acoustic Modelling

The data on which an ASR system is trained determines to a large extent its eventual performance. Several properties of the training data are important. Firstly, its domain must be matched as closely as possible to the domain of the evaluation set. Even when using techniques like fMLLR [8] to adapt acoustic models to the test domain, any mismatch will significantly reduce recognition accuracies. Also accurate orthographic transcriptions of the training data are necessary. Even small amounts of transcription errors can significantly reduce recognition performance, e.g. [9]. Lastly, the size of the training set plays an important role. Although there is no such thing as a direct linear relation between training set size and recognition performance, having more training data does usually lead to better results. Several tens of hours is believed to be a minimum for acoustic model training, depending on

This work has been funded by the European Union as part of the Seventh Framework Programme, under grant agreement no. 287658 (EU-BRIDGE), and by EPSRC Programme Grant grant EP/I031022/1, *Natural Speech Technology*.

the size and complexity of the models being trained.

2.1. Globalphone

One of the suitable speech corpora accessible to us is GlobalPhone [10]. It is a multi-lingual corpus, covering a selection of the world's most widely spoken languages, one of which is German. For each language, it contains speech from about 100 adult native speakers, reading a number of articles taken from a local newspaper. For German, this adds up to about 18 hours of speech. Only 14 hours of this can be used as training data, since the rest is divided over a dev set and a test set. In the context of this paper, the GlobalPhone corpus is less suitable for acoustic model training, due to its small size and its large domain mismatch with the IWSLT evaluation data. However, the German lexicon that is included in the corpus is invaluable to us, since it is the only lexicon we have at our disposal. It contains 36994 unique words, with 39520 pronunciations, indicating that a relatively large number of words is listed with more than a single pronunciation variant. Furthermore, a 3-gram language model for this data is available to us. It is the same language model that was used in [11], and is specifically tuned to the domain of news articles. Using this LM is not our only option though, since we have the option to train our own, more tuned to the domain of TED-talks, see section 5.

2.2. Europarl

The second set of data was obtained by crawling the website of the European Parliament [12], which has committed itself to making its plenary sessions publicly available online, along with their transcripts. These sessions contain speech in a wide variety of languages, German among them. Although, generally speaking, the transcriptions do not match the spoken content of the speech perfectly, techniques for lightly supervised acoustic model training may be employed to circumvent this. We will elaborate on this below in section 4. In this work, we downloaded all parliamentary sessions of the years 2008, 2009, and 2010. This is about 990 hours of audio data. This data contains 23 audio streams in parallel: one stream with the raw unaltered recordings, and one additional stream for each of the 22 languages of the European Union. In these audio streams, speech in any other language than the target language is replaced with its on-the-fly translation, done in real-time by professional interpreters. For each parliamentary speech, there is only a single start and end time given, shared over all 22 parallel versions of that speech. Since translations may take longer than the original speech, or may be shifted in time, the audio segments delineated by these boundaries are usually 10–20 seconds longer than the speech they contain, and tend to overlap each other. Adding the lengths of all these segments together therefore leads to an overestimate of the available data, but can nonetheless be a useful indication. The total amount of speech data we counted like this, is 733 hours. One must be cautious in using all this data directly, however, since it contains directly recorded speech from German-speaking MEP's, as well as interpreters' speech. There are very distinct differences between these types of speech: e.g. whereas MEP's speak more spontaneously, often with an accent, interpreters tend to speak clearly, with long pauses, and very few corrections and repetitions. Since these types of speech may not be equally well matched to the target domain, we have treated them separately. We identified the speeches that were originally spoken in German, by comparing the German audio stream with the raw unaltered audio. Based on the same rough count as before, this adds up to about 95 hours of speech. Since there is no lexicon available with this data, we reuse the GlobalPhone lexicon, to which the out-of-vocabulary words are added using Sequitor Grapheme-to-Phoneme conversion [13].

3. Text Tokenisation

Although the GlobalPhone lexicon does contain 373 numbers, this list is far from exhaustive. Numbers in the evaluation data are therefore very likely to be OOV. To prevent this from happening, we defined rewrite rules to convert any number that is OOV into its constituent parts, most of which do occur in the lexicon, or are easily added to it. For instance, if "1,234" is encountered, it is rewritten as "1,000 2 100 4 und 30". This way, with no more than 33 lexical entries, we are able to handle any number between 1 and $9,999 \cdot 10^6$. Special exception rules are provided to deal with such things as times, dates, years, and IP-addresses. Measures of distance, length, and volume are fully expanded, as well as currencies, e.g. 'km' is written as 'kilometer', '$' is written as 'dollar', etc. Because of time constraints, handling of abbreviations in our system is rudimentary. Basically, any word that either consists of two or more capitalized letters, or of letters separated by full stops is recognized as an abbreviation. They are then written in a consistent form, namely as uncapitalized letters separated by full stops, and then added to the lexicon using grapheme-to-phoneme conversion. There are several ways in which this methodology is suboptimal. For one, it disregards the possibility of abbreviations being pronounced as words, rather than sequences of separate letters, e.g. the pronunciation of "NATO" as /nato/ rather than /ɛnɑteo/. More importantly, the GlobalPhone lexicon, on which we trained the grapheme-to-phoneme conversion, contains far too few examples to enable accurate pronunciation predictions. As a result, abbreviations in training and evaluation data are expected to reduce the performance of our system.

4. Lightly Supervised Acoustic Model Training

To perform acoustic model training and evaluation, the acoustic data is preprocessed as follows. First, it is converted towards mono-channel 16kHz WAVE-files. MFCC-

coefficients are determined within 25 ms frames which are shifted in increments of 10 ms. Cepstral Mean Normalisation is then applied to the resulting 13-dimensional feature vectors. For each frame, the features within a context window of 9 frames, 4 to the left, 4 to the right, are stacked and projected down to 39 dimensions using LDA-MLLT.

4.1. Training an Initial Model on GlobalPhone

We train an initial GMM-HMM acoustic model from scratch on the GlobalPhone corpus. This model contains 3000 context-dependent states and 48000 Gaussians. It was evaluated on three different evaluation sets: the GlobalPhone dev set, where it resulted in a WER of 12.68%, the GlobalPhone eval set, on which it gave a WER of 19.92%, and the IWSLT dev set, on which it yielded a WER of 56.18%. The language model used in each of these evaluations was the GlobalPhone-specific one, introduced in section 2.1.

4.2. Further Training on Europarl

Acoustic model training on Europarl data cannot be done straightforwardly, since the transcriptions we have of it do not match the acoustics perfectly. There is a variety of light supervision techniques, however, with which this problem may be circumvented, e.g. [14, 1]. Here, we used the greedy matching approach described in [5]. We first bias the GlobalPhone LM towards the Europarl domain by interpolating it with a small LM trained on the imperfect transcriptions. This LM, in combination with the acoustic model trained above in section 4.1, is then used to make a recognition of the Europarl training data. By comparing the recognition result with the imperfect transcription, and greedily collecting the longest sequences that occur in both, a new in-domain training set is constructed. From this, a new acoustic model with the same number of states and Gaussians is trained and the whole process is repeated. This iterative process is illustrated in figures 1 and 2. With each iteration, the accuracy of the ASR transcription is expected to rise, and hence more training data is collected for the iteration after that. Also, with each iteration, the models are expected to get more tuned towards the Europarl domain. In this work, we first apply this technique for 10 iterations on the subset with 95 hours of direct MEP recordings, discussed in section 2.2, and evaluated on the IWSLT dev set in each iteration. The result is shown in the leftmost columns of table 1. The initial WER of 46.36% is obtained with the GlobalPhone acoustic model. The reason why this result is different from the 56.18% reported in section 4.1 is that another LM was used in these evaluations, namely the one that is biased towards Europarl data. Looking at the WER's, we can see that the quality of the acoustic models doesn't improve with each new iteration. If anything, the opposite is true, although the statistical significance of these differences may be questionable. This lack of improvement is probably caused by a slight domain mismatch between Europarl and the TED talks in the IWSLT dev set. An interesting experiment would be to evaluate the models in each iteration on an evaluation set in the Europarl domain. Unfortunately, no such evaluation set is available to us. When doing the same experiment on the entire Europarl corpus, MEP speech and interpreters' speech put together, the results become as shown in the rightmost columns of table 1. The acoustic model obtained in iteration 10 of the previous experiment is used here as the initial acoustic model. Although the WER drops about 1% absolute with the inclusion of the interpreters' speech, the results are otherwise comparable to those of the previous experiment. The drop in WER is very likely due to the increase of the training set from 46.98 hours to 67.18 hours. The best performance, a WER of 39.95%, is achieved in the third iteration. Therefore, the training set obtained in that iteration is used for all acoustic model training in further experiments, see section 6.

	MEP		All	
iter	hours	WER(%)	hours	WER(%)
init	NA	46.36	46.98	41.12
1	45.91	41.13	67.15	40.22
2	46.64	41.20	70.28	40.09
3	46.69	41.36	70.80	**39.95**
4	46.80	41.25	70.83	40.01
5	46.89	41.10	70.92	40.27
6	47.00	41.36	70.93	40.28
7	47.07	41.55	70.99	40.26
8	47.01	41.49	70.95	40.12
9	47.00	41.28	70.89	40.50
10	46.98	41.12	70.94	40.35

Table 1: *The data set sizes and WER rates obtained on the IWSLT dev set in each iteration of lightly supervised training.*

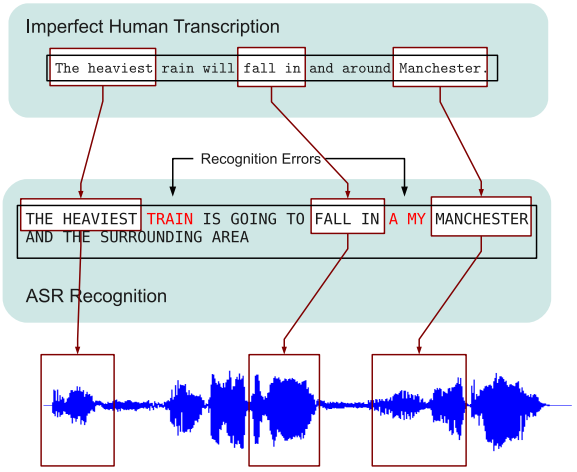

Figure 1: *The longest word sequences occurring both in the approximate transcription and in the ASR output are identified.*

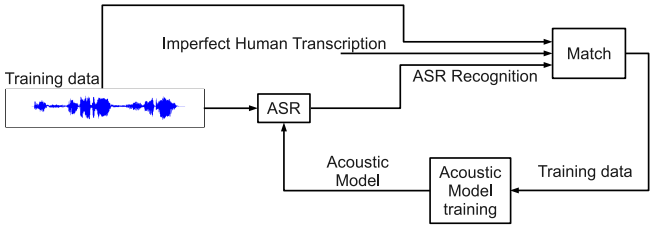

Figure 2: *Illustration of the iterative process, in which training data is collected to obtain acoustic models, which are in turn used to collect a better set of training data.*

name	# words ($\cdot 10^6$)
europarl_v7	47.37
europarl_crawl	2.86
news_crawl_2007	31.47
news_crawl_2008	107.86
news_crawl_2009	101.56
news_crawl_2010	45.89
news_crawl_2011	252.85
news_crawl_2012	319.73
news_comment	4.45
total	914.05

Table 2: *The text resources used for LM training.*

5. Language Modelling

For language model training, we used the resources listed in table 2. All of these were obtained through links on the IWSLT website, except 'europarl_crawl', which consists of the imperfect transcriptions of the Europarl data from section 2.2. All text was first depunctuated and tokenised as described in section 3. From each of these texts, 30% is selected that best matches the domain of the IWSLT dev set, according to the cross-entropy criterion proposed in [6]. Language models are trained on this subset only, disregarding the remaining 70%.

5.1. N-gram Language Models

After winnowing them down to 30%, each of the text corpora is used to train a 3-gram LM, using the MITLM language modelling toolkit [15]. In this training, modified Kneser-Ney smoothing [16] is used with parameters optimised on the IWSLT dev set. These language models are then linearly interpolated with interpolation weights optimised in the same way. The 1-grams in the resulting interpolated model are then written out in decreasing order, according to their smoothed 1-gram probability. Choosing the top-N words from this list allows us to optimally define a dictionary of size N for further LM training. We then repeated the previous procedure, training 3-gram LM's on the whittled down text corpora, with a limited vocabulary of N words, and linearly interpolating them. Finally the same was done with

N	OOV rate(%)	3-gram ppl.	4-gram ppl.
100000	4.18	252.63	246.36
150000	3.32	278.24	263.37
200000	2.78	283.25	275.97
250000	2.52	289.73	282.43
300000	2.37	294.24	**286.86**
350000	2.29	297.03	289.74
400000	2.17	300.30	292.97

Table 3: *The perplexities and OOV rates of the 3-gram and 4-gram LM's on the IWSLT dev set*

4-grams. The OOV rate and perplexity on the dev set for a range of values for N is shown in table 3. As expected, the 4-gram models achieve lower perplexities than 3-gram models. Based on these results, we choose the 4-gram LM with vocabulary size 300000 for the evaluations in section 6, since this yields a good trade-off between word coverage and perplexity. Any of these 300000 words that do not occur either in GlobalPhone or in the crawled Europarl data is added to the lexicon. Using a LM of such size for LVCSR (Large Vocabulary Continuous Speech Recognition) is very demanding in terms of memory and processing power. Therefore, we make a reduced version of this LM, pruning it with a probability threshold of 10^{-7}. The pruned LM is much smaller in size than the original, but this comes at the price of a higher perplexity, which rises from 286.86 to 413.62. Due to its smaller size, it can easily be used to generate word lattices on the evaluation data, which are rescored afterwards using the full unpruned LM. To demonstrate the extent to which they may affect the WER in practice, we perform an ASR evaluation on the IWSLT dev set using the pruned LM, before and after rescoring with the unpruned LM. The acoustic model in this experiment is the optimal model as established in section 4.2. The pruned LM yields in this evaluation a WER of 37.02%, a slight improvement over 39.95%, obtained in section 4.2, with a different LM. Rescoring with the full LM brings the WER further down to 33.69%.

5.2. Recurrent Neural Net Language Models

From a concatenation of all the whittled down text corpora of section 5.1, we train a Recurrent Neural Net Language Model, using the RNNLM toolkit [7]. Due to computational limitations, the vocabulary size for this model is reduced to 50000. The number of nodes in the hidden layer is set to 30. From the final rescored word lattices in section 5.1, N-best lists are generated, with N=100. For each of these 100 recognition hypotheses, the RNNLM is used to calculate a LM score S_{RNNLM}, which is interpolated with the original 4-gram LM score, resulting in the modified score S'.

$$S' = (1-\alpha) \cdot S_{ngram} + \alpha \cdot S_{RNNLM} \qquad (1)$$

This modified score is used to re-rank the N-best list, often changing which hypothesis is considered as the 'best'. The

interpolation factor α was optimised on the dev set, yielding a value of 0.25. Applying this RNNLM rescoring on the word lattices of section 5.1, yields an improvement in WER from 33.69% to 33.17%.

6. ASR System Setup

At this point, we have all the resources to build a finalised system: a large set of transcribed speech for acoustic model training, determined in section 2.2, and a large LM, optimised as described in section 5. The lay-out of our system is depicted in figure 3. All experiments performed with this

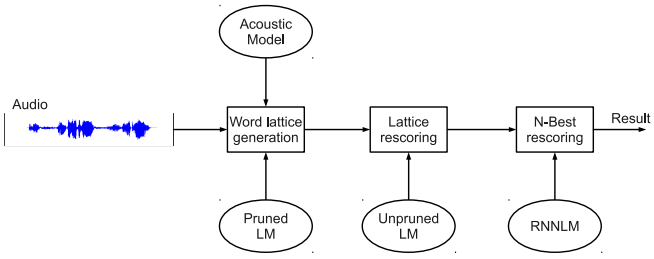

Figure 3: *A schematic overview of the adopted system.*

system, including the evaluations above and those that follow, have been performed using the KALDI Speech Recognition Toolkit [17]. For acoustic modelling, we first train up a GMM-HMM with 3000 context dependent states and 48000 Gaussians, using Speaker Adaptive Training (SAT), where fMLLR is used as the adaptation technique. In principle, it would be possible to assign multiple speeches to a single speaker, since the speaker's identity is given on the Europarl website. This only applies, however, to directly recorded speeches, i.e. untranslated ones. When the speaker is an interpreter, there is no trivial way to ascertain his/her identity. Therefore, we have made the simplifying assumption that each speech in the training data comes from a unique speaker. A feed-forward deep neural network is then trained in a DNN-HMM hybryd configuration, similar to the one used in [18]. This DNN has 6 hidden layers, each containing 2048 nodes. The softmax output layer of this network produces posterior probabilities over the 3000 context-dependent states of the HMM. The input at each time t consists of a stacking of the features in the context window $[t-5, t-4, \ldots t, \ldots t+4, t+5]$. Except for the addition of speaker adaptation, the features in each frame are produced as explained in section 4. Since the IWSLT test set is provided without segmentation into utterances, one can either generate a segmentation automatically, or perform recognition on entire TED-talks without segmentation. For the automatic segmentation, we use a voice activity detection system trained on 70 hours of English conversational speech from the AMI Meetings Corpus [19]. Speech and silence frames are modelled with diagonal covariance GMMs. A minimum duration constraint of 50ms is applied to each segment. For the segmentationless recognition, we use the same technique

	dev2012	tst2013	tst2013\E06
manual segment	27.02	35.27	29.18
auto segment	✗	39.28	33.58
no segment	28.16	36.21	30.24

Table 4: *The resulting WER's in % for several different evaluation sets, both when they are manually segmented, automatically segmented, or recognised in full (not segmented).*

as in [5], where we split an entire talk into overlapping segments, perform ASR on them, and dynamically merge the results into a single long recognition. In this case, segments are 40 seconds long and have an overlap of 20 seconds with each other. The results are listed in table 4. For the development set, no automatic segmentation was performed, since the manual segmentation was available for the official evaluation. There is one talk in the IWSLT test set, namely "E06_Nach-und-doch-so-Fern-Thomas-Mo", that is of very low quality. It has been recorded with a far-range microphone across a reverberant room, and contains quite a bit of non-speaker noise, e.g. coughing, rustling of paper and clothing, etc. Our system has not been designed to deal with such conditions, nor has it been tuned to them in any way, since the development set does not contain similar recordings. We therefore argue that this file unfairly skews the average test results. In tabel 4, the column "tst2013\E06" lists the results when this file is excluded from the evaluation. These error rates are more in line with those obtained on the dev set. The results in this table suggest that for TED talks, in the absence of a manual segmentation, a recognition performed on the whole talk is preferrable to using an automatically generated segmentation. We suspect, however, that this conclusion is fairly domain-specific. An automatic segmentation is essential for files with more music, jingles, applause, laughter, and other non-speaker noise.

7. Conclusion

We have presented the various components in the German ASR system, how they were set up, trained, and combined, to obtain accurate recognitions on the various data sets of the IWSLT evaluation task. Worthy of note is the acoustic model training, which was done almost entirely on publicly available data, without expert human transcriptions, using a lightly supervised training technique. Final evaluation on the unsegmented test set was performed in two different ways. Once with an automatically generated segmentation, and once without segmentation at all. It was found that, even though an oracle segmentation leads to optimal recognition results, avoiding segmentation altogether is preferrable to using an automatically generated one, when an oracle segmentation is not available.

8. References

[1] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, September 2010, pp. 2222–2225.

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

[3] P. Placeway and J. Lafferty, "Cheating with imperfect transcripts," in *Proc. ICSLP*, vol. 4, 1996, pp. 2115–2118.

[4] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. ICASSP 2009.*, 2009, pp. 4869–4872.

[5] J. Driesen and S. Renals, "Lightly supervised automatic subtitling of weather forecasts," in *Proc. Automatic Speech Recognition and Understanding Workshop*, Olomouc, Czech Republic, December 2013.

[6] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proc. ACL*, Uppsala, Sweden, July 2010.

[7] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, Makuhari, Japan, September 2010.

[8] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, 1998.

[9] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration," in *Proc. Interspeech*, Lyon, France, 2013.

[10] T. Schultz, "GlobalPhone: A multilingual speech and text database developed at karlsruhe university," in *Proc. Interspeech*, Denver, Colorado, USA, 2002.

[11] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[12] "The website of the european parliament." [Online]. Available: http://europarl.europa.eu

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

[14] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, January 2011.

[15] "Iterative language model estimation: Efficient data structure & algorithms," in *Proc. Interspeech*, Brisbane, Australia, September 2008.

[16] S. F. Chen, , and J. Goodman, "An empirical study of smoothing techniques for language modeling." in *Proc. ACL*, Santa Cruz, USA, June 1996.

[17] 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 *Proc. ASRU*, Big Island, Hawaii, US, December 2011.

[18] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[19] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.