The 2015 KIT IWSLT Speech-to-Text Systems for English and German

Markus Müller, Thai-Son Nguyen, Matthias Sperber, Kevin Kilgour, Sebastian Stüker and Alex Waibel

Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
{m.mueller|thai.nguyen|matthias.sperber}@kit.edu
{kevin.kilgour|sebastian.stueker|waibel}@kit.edu

Abstract

This paper describes our German and English *Speech-to-Text* (STT) systems for the 2015 IWSLT evaluation campaign. This campaign focuses on the transcription of unsegmented TED talks. Our setup includes systems from both Janus and Kaldi. We combined the outputs using both ROVER [1] and confusion network combination (CNC) [2] to archieve a good overall performance. The individual subsystems are built by using different front-ends, (e.g., MVDR-MFCC or lMel), acoustic models (GMM or modular DNN) and phone sets and by training on different sets of permissible training data. Decoding is performed in two stages, where the GMM systems are adapted in an unsupervised manner on the combination of the first stage outputs using VTLN, MLLR, and cMLLR.

The combination setup produces a final hypothesis that has a significantly lower WER than any of the individual subsystems. For English, our single best system based on Kaldi has a WER of 13.8% on the development set while in combination with Janus we lowered the WER to 12.8%.

1. Introduction

The 2015 *International Workshop on Spoken Language Translation* (IWSLT) offers a comprehensive evaluation campaign on spoken language translation. The evaluation is organized in different evaluation tracks covering automatic speech recognition (ASR), machine translation (MT), and the full-fledged combination of the two of them into speech translation systems (SLT). The evaluations in the tracks are conducted on TED Talks (http://www.ted.com/talks), short 5-25min presentations by people from various fields related in some way to Technology, Entertainment, and Design (TED) [3].

The goal of the TED ASR track is the automatic transcription of fully unsegmented TED lectures. The quality of the resulting transcriptions are measured in word error rate (WER).

In this paper we describe our English and German ASR systems with which we participated in the TED ASR track of the 2015 IWSLT evaluation campaign. Our English and German systems are based on our previous years' evaluation systems [4]. In addition to our Janus[5] based systems, we also built a system based on Kaldi[6] for English. For this, we used the recipe provided in the Kaldi repository for the TEDLIUM corpus [7]. The Janus system setup uses multiple complementary subsystems that employ different phone sets, front ends, acoustic models or data subsets.

The rest of this paper is structured as follows. Section 2 describes the data that our system was trained and tested on. This is followed by Section 3 which provides a description of the acoustic front-ends used in our system and Section 4 which describes our segmentation setup. An overview of the techniques used to build our acoustic models is given in Section 5. We describe the language model used for this evaluation in Section 6. Our decoding strategy and results are then presented in sections 7 and 8. The final Section 8 contains a short conclusion.

2. Data Resources

2.1. Training Data

The following data sources have been used for acoustic model training of our English systems:

a) 200 hours of Quaero training data from 2010 to 2012.

b) 18 hours of various noise data, such as snippets of applause, music or noises from microphone movement.

c) 158 hours of data downloaded from the TED talks website, without disallowed talks.

d) 203 hours of TED talks from the TEDLIUM v2 release [7], excluding disallowed talks.

The Quaero training data is transcribed manually. The noise data consists only of noises and is tagged with specific noise words to enable the training of noise models. The TED data comes with subtitles provided by TED and the TED translation project. The TEDLIUM dataset is provided by Laboratoire d'Informatique de l'Université du Maine (LIUM).

For German we used the following data sources:

Data set	# Talks	# Utts	Dur.	Avg. dur.
tst2013 (manual)	28	2246	3.9h	6.3s
tst2013 (auto)	28	2353	4.0h	6.1s
tst2014 (auto)	15	801	2.2h	9.7s
tst2015 (auto)	12	1013	2.2h	7.7s

Table 1: *Statistics of the English development sets ("tst2013") and the English evaluation sets ("tst2014" and "tst2015"), including the total number of talks (*# Talks*), the total number of utterances (*# Utts*), the overall speech duration (*Dur.*), and average speech duration per utterance (*Avg. dur.*). "tst2014" and "tst2015" have been segmented automatically. Properties of the automatic segmentation of "tst2013" are displayed alongside with those of the manual segmentation.*

a) 180 hours of Quaero training data from 2009 to 2012.

b) 24 hours of broadcast news data

c) 160 audio from the archive of parliament of the state of Baden-Württemberg, Germany

For language modeling and vocabulary selection, we used most of the data admissible for the evaluation, as summarized in Tables 2 and 3.

2.2. Test Data

For this year's evaluation campaign, two evaluation test sets ("tst2014" and "tst2015"), as well as development test sets ("tst2013") were provided for both English and German. Table 1 lists these 3 test sets along with relevant properties for English.

All development test sets were used with the original presegmentation provided by the IWSLT organizers. Additionally, "tst2013" has been segmented automatically in the same way as the evaluation test sets.

3. Feature Extraction

Our systems are built using several different front ends. The two main input variants, each using a frame shift of 10ms and a frame size of 32ms, are the mel frequency ceptral coefficient (MFCC) minimum variance distortionless response (MVDR) (M2) features that have been shown to be very effective when used in bottleneck features [8] and standard lMEL features which generally outperform MFCCs when used as inputs to deep bottleneck features. These standard features are often augmented by tonal features (T). For the extraction of those, we use a pitch tracker [9] and fundamental frequency variation [10]. In [11] we demonstrate, that the addition of tonal features not only greatly reduces the WER on tonal languages like Vietnamese and Cantonese but also results in small gains on non-tonal languages such as English.

3.1. Deep Bottleneck Features

The use of bottleneck features greatly improves the performance of our GMM acoustic models, but also our Hybrid systems benefit from it as well. Figure 1 shows a general overview of our deep bottleneck features (BNF) training setup. 13 frames (+-6 frames) are stacked as the DBNF input which consists of 4-5 hidden layers each containing 1200-1600 units followed by a 42 unit bottleneck, a further 1200-1600 unit hidden layer and an output layer of 6000 context dependent phone states for the German systems and 8000 for the English systems. Layer-wise pretraining with denoising autoencoders is used for the all the hidden layers prior to the bottleneck layer. The network is subsequently finetuned as a whole [12]. For network training, we used a framework based on Theano ([13], [14]).

The layers following the bottleneck are discarded after training and the resulting network can then be used to map a stream of input features to a stream of 42 dimensional bottleneck features. Our experiments show it to be helpful to stack a context of 13 (+-6) bottleneck features and perform LDA on this 630 dimensional stack to reduce its dimension back to 42.

4. Automatic Segmentation

In this evaluation, the test set for the ASR track was provided without manual sentence segmentation, thus automatic segmentation of the target data was mandatory. We utilized an approach to automatic segmentation of audio data that is SVM based. This kind of segmentation is using speech and non-speech models, using the framework introduced in [15]. The pre-processing makes use of an LDA transformation on DBNF feature vectors after frame stacking to effectively incorporate temporal information. The SVM classifier is trained with the help of LIBSVM [16]. A 2-phased post-processing is applied for final segment generation.

We generated the segmentations for both English and German using this SVM based segmentation. The parameters for the SVM segmenter were chosen on a per language basis after preliminary experiments.

5. Acoustic Modeling

5.1. Data Preprocessing

For the English TED data in dataset c) only subtitles were available so the data had to be segmented prior to training. In order to split the data into sentence-like chunks, it was decoded by one of our development systems to discriminate speech and non-speech and a forced alignment given the subtitles was performed where only the relevant speech parts detected by the decoding were used. The procedure is the same as the one that has been applied in [17]. The TEDLIUM data did not require any special preprocessing, except for removing all disallowed talks.

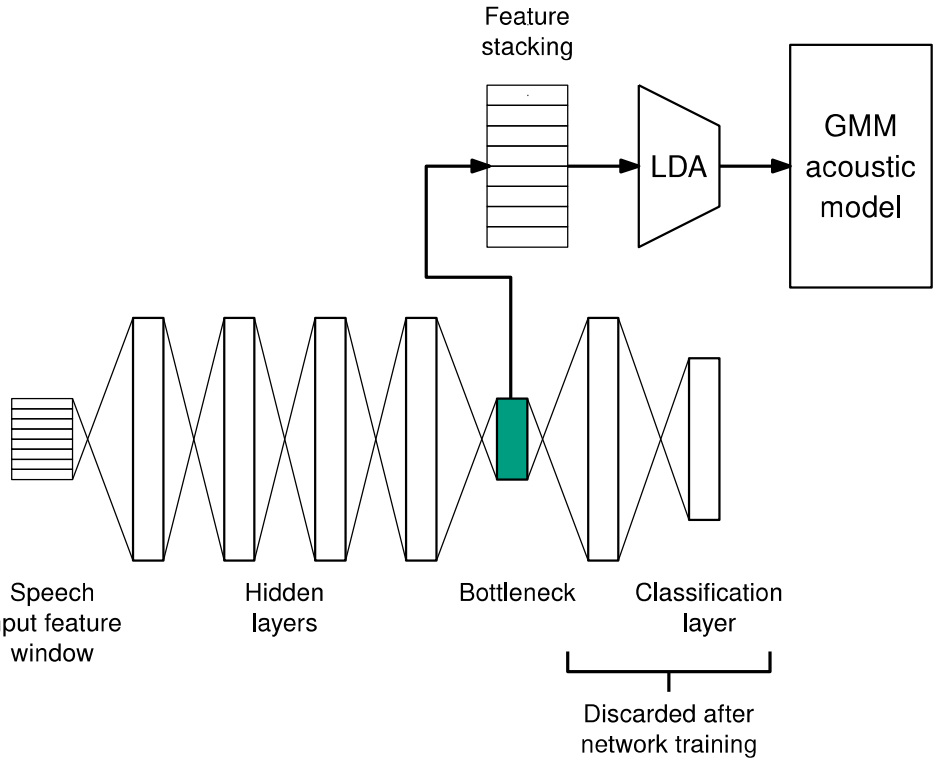

Figure 1: *Overview of our standard DBNF setup.*

5.2. GMM AM Training Setup

All systems use context-dependent quinphones with three states per phoneme and a left-to-right HMM topology without skip states. The English acoustic models use 8000 distributions and codebooks derived from decision-tree based clustering of the states of all possible quinphones. The German acoustic models use 6000 distributions and codebooks.

The GMM models are trained by using incremental splitting of Gaussians training (MAS) [18], followed by optimal feature space training (OFS) which is a variant of *semi-tied covariance* (STC) [19] training using a single global transformation matrix. The model is then refined by one iteration of Viterbi training. All German models use vocal tract length normalization (VTLN), for English it is used where indicated (V).

In order to improve the performance of our GMM based acoustic models Boosted Maximum Mutual Information Estimation training (BMMIE) [20], a modified form of the Maximum Mutual Information (MMI) [21], is applied at the end. Lattices for discriminative training use a small unigram language model as in [22]. After lattice generation, the BMMIE training is applied for three iterations with a boosting factor of b=0.5. This approach results in about 0.6% WER improvement for 1st-pass sytems and about 0.4% WER for 2nd-pass systems.

We trained multiple different GMM acoustic models by combining different front-ends and different phoneme sets. Section 7 elaborates the details of our system combination.

5.3. Hybrid Acoustic Model

As with the GMM systems we trained our hybrid systems on various front-ends and phoneme sets. Our best performing hybrid systems are based on a modular topology which involves stacking the bottleneck features, described in the previous section over a window of 15 frames, with 4-5 1600-2000 unit hidden layers and an output layer containing 6016 context dependent phonestates for German and 8156 context dependent phonestates for English. The deep bottleneck features were extracted using an MLP with 5 1600 unit hidden layers prior to the 42 unit bottleneck layer. Its input was 40 lMel (or MVDR+MFCC) and 14 tone features stacked over a 13 frame window. Both neural networks were pretrained as denoising autoencoders.

We trained neural network acoustic models for English on various input features and with different topologies using the same techniques described in the deep bottleneck layer section. Our best setup uses deep bottleneck features stacked over a window of 15 frames, with 5 1600 unit hidden layers and an output layer containing 8156 context dependent phone states. The deep bottleneck features were extracted using an MLP with 5 1600 unit hidden layers prior to the 42 unit bottleneck layer. Its input was 40 lMEL and 14 tonal

features stacked over a 15 frame window.

The German hybrid system is based on a modular topology which involves the stacking bottleneck features from three separate bottleneck extraction networks (MFCC+MVDR+T, lMEL+T & MFCC+MFCC+lMEL+T) over a window of 13 frames leading to a 1638 (=3 * 42 * 13) neuron bottleneck stack, followed by 4 hidden layers containing 2000 neurons each and an output layer containing 6016 context dependent phonestates. The deep bottleneck features were extracted using an MLP with 5 2000 unit hidden layers prior to the 42 unit bottleneck layer. Their inputs were 40 lMel and 14 tone features for the lMEL+T network, 20 MFCC, 20 MVDR and 14 tone features for the MFCC+MVDR+T network and 20 MFCC, 20 MVDR, 40 lMEL and 14 tonal features for the MFCC+MFCC+lMEL+T MLP.

5.4. Kaldi

For system combination we also trained a system using Kaldi [6]. We trained the acoustic model (AM) on the TED-LIUM corpus release 2 [7] using the tedlium receipe (s5). The AM utilizes a neural network taking bottle neck features extracted from combined filterbank and pitch features that are then fMLLR adapted as input. After optimizing its cross-entropy on the training data, the network is refined using sequence training optimizing the sMBR criteria. For the language model we used the cantab-tedlium tri-gram language model [23].

5.5. Pronunciation Dictionary

For English, we used the CMU dictionary[1]. This is the same phoneme set as the one used in last year's systems. It consists of 45 phonemes and allophones. We used 7 noise tags and one silence tag each. For the CMU phoneme set we generated missing pronunciations with the help of FESTIVAL [24].

Our German system uses an initial dictionary based on the Verbmobil Phoneset [25]. Missing pronunciations are generated using both Mary [26] and FESTIVAL [24].

6. Language Models and Search Vocabulary

For language model training and vocabulary selection, we used the subtitles of TED talks, or translations thereof, and text data from various sources (see Tables 2 and 3). Text cleaning included tokenization, lowercasing, number normalization, and removal of punctuation. Language model training was performed by building separate language models for all (sub-)corpora using the SRILM toolkit [27] with modified Kneser-Ney smoothing. These were then linearly interpolated, with interpolation weights tuned using held-out data from the TED corpus. For German, we split compounds similarly as in [28].

For the vocabulary selection, we followed an approach proposed by Venkataraman et al.[29]. We built unigram language models using Witten-Bell smoothing from all text sources, and determined unigram probabilities that maximized the likelihood of a held-out TED data set. As our vocabulary, we then used the top 150k words for English, and 300k words for German.

For our English Kaldi system, we used the TEDLIUM language model from Cantab Research[23]. It contains 155,290,779 tokens and is based on the '1 Billion Word Language Model Benchmark'[2].

Text corpus	# Words
TED	3.6m
News + News-commentary + -crawl	4,478m
Euronews	780k
Commoncrawl	185m
GIGA	2323m
Europarl + UN + multi-UN	829m
TEDLIUM dataselection	155m

Table 2: *English language modeling data after cleaning. The total number of words was 7.8 billion, not counting Google Books.*

Text corpus	# Words
TED	2,685k
News+Newscrawl	1,500M
Euro Language Newspaper	95,783k
Common Crawl	51,156k
Europarl	49,008k
ECI	14,582k
MultiUN	6,964k
German Political Speeches	5,695k
Callhome	159k
HUB5	20k
Google Web	(118m n-grams)

Table 3: *German language modeling data after cleaning and compound splitting. In total, we used 1.7 billion words, not counting Google Ngrams.*

7. Decoding Setup

For our English submission we trained 3 different DBNF GMM acoustic models in total by combining different feature front-ends (M2 and lMEL), with and without using VTLN adaptation. We also trained one DNN hybrid system using lMEL front-ends and another one with DBNF features. In addition to these systems, we also included a Kaldi based system this year, using the standard recipe for the TEDLIUM dataset. The first CNC was created using the outputs from 3

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] http://www.statmt.org/lm-benchmark

System	tst2013	tst2015	Sub.
lMEL+T+V	17.7	-	
M2+T+V	17.6	-	
lMEL+T	18.1	-	
lMEL+T-DBNF-hyb+V	16.0	-	
lMEL+T-hyb	16.4	-	
CNC 1	14.7	-	
lMEL+T+V+adapt	15.3	-	
M2+T+V+adapt	15.0	-	
lMEL+T+adapt	14.9	-	
CNC 2	14.4	10.9	C 1
Kaldi	15.6	-	
ROVER 1	13.2	-	
Kaldi rescored	13.8	10.4	C 2
ROVER 2	12.8	10.0	Pri

Table 4: *Results for English on 'tst2013' development and 'tst2015' evaluation test sets. Both contrastive systems (**C 1**) and (**C 2**) are shown, as well as the primary submission (**Pri**).*

different DBNF GMM based systems in combination with the output from 2 hybrid systems. Based on this first CNC, the GMM based systems were adapted. Combining the output from the adapted systems and the hybrid systems to another CNC. This second CNC is our first contrastive submission. It contains only output from Janus based systems. The output from our Kaldi setup is incorporated in the first and second ROVER. In the first ROVER, we combined the output from Kaldi, out two hybrid systems and the two best adapted GMM based systems. This result is then included in a second ROVER, where we combined it with the re-scored output from Kaldi and the output from the second CNC. This is our primary condition.

The German setup consists of a DBNF GMM system and a modular Hybrid system. A CNC is performed on the outputs of both systems and used to adapt the DBNF GMM AM. A final CNC is then performed using the adapted GMM output in lieu of the unadapted output.

8. Results

The English systems have been evaluated on the test set "tst2013". The results are listed in Table 4. Based on these results, we decided our decoding strategy for the evaluation. The first CNC results in a WER of 14.7%. Including the output from Kaldi, the WER decreases to 12.8%.

9. Conclusions

In this paper we presented our English and German LVCSR systems, with which we participated in the 2015 IWSLT evaluation. All systems make use of neural network based front-ends, HMM/GMM and HMM/DNN based acoustics models. The decoding set-up of all languages makes extensive use of system combination of single systems obtained by combing different phoneme sets, feature extraction front-ends and acoustic models.

10. References

[1] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings the IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, USA, Dec. 1997, pp. 347–354.

[2] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[3] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 10th iwslt evaluation campaign," in *Procedings of the 10th Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

[4] Kevin Kilgour, Michael Heck, Markus Mller, Matthias Sperber, Sebastian Stker, and Alexander Waibel, "The 2014 KIT IWSLT Speech-to-Text Systems for English, German and Italian," in *International Workshop on Spoken Language Translation (IWSLT)*, Dec. 2014.

[5] M. Woszczyna, N. Aoki-Waibel, F. D. Bu, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel, "Janus 93: Towards spontaneous speech translation," in *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia, 1994.

[6] 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.

[7] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks," in *Proc. of LREC*, 2014, pp. 3935–3939.

[8] K. Kilgour, I. Tseyzer, Q. B. Nguyen, and A. Waibel, "Warped minimum variance distortionless response based bottle neck features for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6990–6994.

[9] K. Schubert, "Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung," Master's thesis, Universität Karlsruhe (TH), Germany, 1999, in German.

[10] K. Laskowski, M. Heldner, and J. Edlund, "The Fundamental Frequency Variation Spectrum," in *Proceedings of the 21st Swedish Phonetics Conference (Fonetik 2008)*, Gothenburg, Sweden, June 2008, pp. 29–32.

[11] F. Metze, Z. A. W. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," in *Proceedings of the 10th Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

[12] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE*, 2013.

[13] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010, oral Presentation.

[14] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[15] M. Heck, C. Mohr, S. Stker, M. Mller, K. Kilgour, J. Gehring, Q. Nguyen, V. Nguyen, and A. Waibel, "Segmentation of telephone speech based on speech and non-speech models," in *Speech and Computer*, ser. Lecture Notes in Computer Science, M. elezn, I. Habernal, and A. Ronzhin, Eds. Springer International Publishing, 2013, vol. 8113, pp. 286–293.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[17] M. Heck, K. Kubo, M. Sperber, S. Sakti, S. Stker, C. Saam, K. Kilgour, C. Mohr, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, "The KIT-NAIST (contrastive) english ASR system for IWSLT 2012," in *Proceedings of the International Workshop on Speech Translation (IWSLT 2012)*, Hong Kong, December 2012.

[18] T. Kaukoranta, P. Fränti, and O. Nevalainen, "Iterative split-and-merge algorithm for VQ codebook generation," *Optical Engineering*, vol. 37, no. 10, pp. 2726–2732, 1998.

[19] M. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[20] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah, "Boosted mmi for model and feature-space discriminative training," in *ICASSP 2008*, 2008, pp. 4057–4060.

[21] Bahl L.R., Brown P.F, de Souza P.V., and L.R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP 1986*, 1986, pp. 49–52.

[22] V. Valtchev, J. J. Odell, P.C. Woodland, and S.J. Young, "MMIE training of large vocabulary recognition systems," in *Speech Communication 22*, 1997, pp. 303–314.

[23] W. Williams, N. Prasad, D. Mrva, T. Ash, and T. Robinson, "Scaling recurrent neural network language models," *arXiv preprint arXiv:1502.00512*, 2015.

[24] A. Black, P. Taylor, R. Caley, and R. Clark, "The festival speech synthesis system," 1998.

[25] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The karlsruhe-verbmobil speech recognition engine," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1. IEEE, 1997, pp. 83–86.

[26] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[27] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*, 2002.

[28] Kevin Kilgour, Christian Mohr, Michael Heck, Quoc Bao Nguyen, Van Huy Nguyen, Evgeniy Shin, Igor Tseyzer, Jonas Gehring, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel, "The 2013 KIT IWSLT Speech-to-Text Systems for German and English," in *International Workshop on Spoken Language Translation (IWSLT)*, Dec. 2013.

[29] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," in *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003, pp. 245–248.