The 2014 KIT IWSLT Speech-to-Text Systems for English, German and Italian

Kevin Kilgour, Michael Heck, Markus Müller,
Matthias Sperber, Sebastian Stüker and Alex Waibel

Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany

{kevin.kilgour|heck|m.mueller}@kit.edu

{matthias.sperber|sebastian.stueker|waibel}@kit.edu

Abstract

This paper describes our German, Italian and English *Speech-to-Text* (STT) systems for the 2014 IWSLT TED ASR track. Our setup uses ROVER and confusion network combination from various subsystems to achieve 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 various subsets of the 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.

1. Introduction

The 2014 *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) [1].

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 Italian, German and English ASR systems with which we participated in the TED ASR track of the 2014 IWSLT evaluation campaign. While our German and English ASR systems are based on our previous years' evaluation systems [2] our Italian system is a completely new system that was developed from scratch. Our general 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, 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 all our English systems:

- 200 hours of Quaero training data from 2010 to 2012.
- 18 hours of various noise data, such as snippets of applause, music or noises from microphone movement.
- 158 hours of data downloaded from the TED talks website that was released before the cut-off date of December 31st 2010.

The Quaero training data is manually transcribed. 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.

For German we used the following data sources:

- 180 hours of Quaero training data from 2009 to 2012.
- 24 hours of broadcast news data
- 160 audio from the archive of parliament of the state of Baden-Württemberg, Germany

Set	#talks	#utt	dur	dur/utt
dev2010	8	887	1.5h	6.2s
dev2012	10	1144 (545)	1.7h (1.8h)	5.4s (12.2s)
tst2010	11	1664	2.5h	5.3s
tst2013	28	1388	4.2h	10.8s
tst2014	15	718	2.2h	11.0s

Table 1: *Statistics of the development sets ("dev2010", "tst2010" and "dev2012") and the evaluation sets ("tst2013" and "tst2014"), including the total number of talks (*#talks*), the total number of utterances (*#utt*), the overall speech duration (*dur*), and average speech duration per utterance (*dur/utt*). "tst2013" and "tst2014" have been segmented automatically. Properties of the automatic segmentation of "dev2012" is described in brackets.*

The training database for our Italian system contains a total of 100 hours of audio. It is based on the data from Quaero Period 4 (54 hours) and Quaero Period 5 (46 hours). The audio consists of recordings from radio and TV broadcasts. The data is manually transcribed and split into segments of varying length, ranging from one sentence to multiple minutes. The textual transcriptions contain annotations for distinct acoustic events as well. We incorporated them as markers for noises in general and for noises originating from humans.

Due to the lack of Italian data, we used additional English data for the neural network training. This data consisted of 426 hours, based on a selection of TED talks, stanford lectures, euronews broadcasts and recordings from videolectures.

For language modeling and vocabulary selection, we used most of the data admissible for the evaluation, as summarized in Tables 2, 3, and 4.

2.2. Test Data

For this year's evaluation campaign, two evaluation test sets ("tst2013" and "tst2014") were provided, as well as three development test sets ("dev2010", "tst2010" and "dev2012"). The test set "dev2012" has preferably been used for system development and parameter optimization. Table 1 lists these five test sets along with relevant properties.

"tst2013" is last year's evaluation set and is solely comprised of TED talks newer than December 2010. This set serves as a progress test set to measure the system improvements with respect to last year's IWSLT ASR track. "tst2014" is a collection of TED talks that have been filmed between early 2012 and late 2013. All development test sets were used with the original pre-segmentation provided by the IWSLT organizers. Additionally, "dev2012" has been segmented automatically, as well this year's evaluation test set.

For the German and Italian systems only a single test each set "dev2013" and "dev2014" was available.

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 BNFs [3] 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). In [4] 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.

For bootstrapping our systems we employed log Mel features with 13 coefficients and a frame size of 16ms. We stacked the individual frames using a context of seven frames to each side.

3.1. Deep Bottleneck Features

The use of bottleneck features greatly improves the performance of our GMM acoustic models. Figure 1 shows a general overview of our deep bottleneck features 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 [5].

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.

For Italian, we used an additional approach by training a neural network using data from more than one language. We re-used a neural network that has been trained using English data. In one setting, we used it directly without any re-training and in another setting, we re-added the discarded output layers after the bottleneck and re-trained them using Italian data.

4. Automatic Segmentation

As was the case for last year's 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 three different approaches to automatic segmentation of audio data, which are:

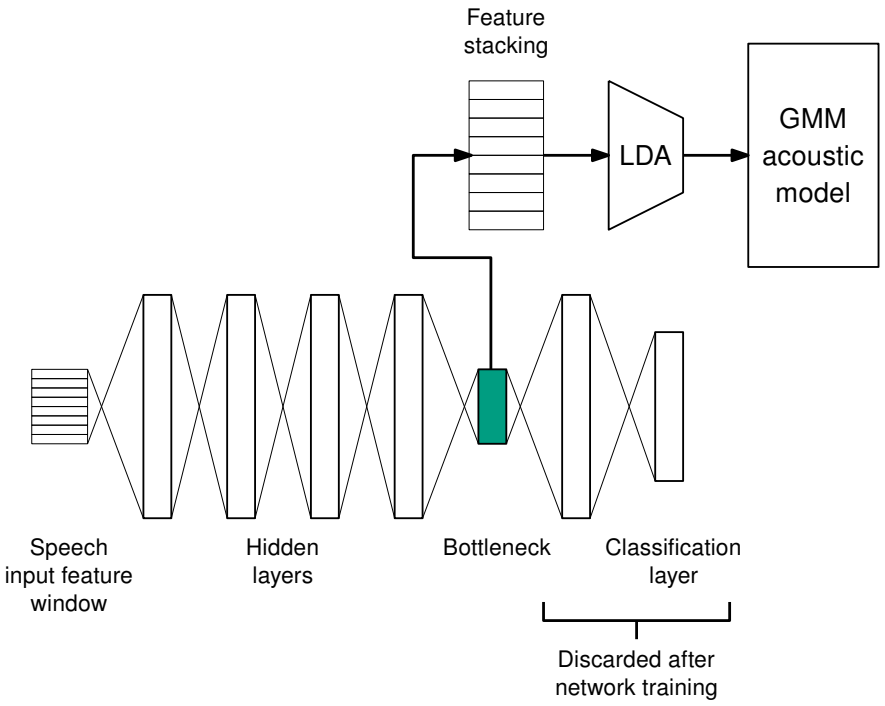

Figure 1: *Overview of our standard DBNF setup.*

a) *Decoder based* segmentation on hypotheses. A fast decoding pass with one of our development systems was done to determine speech and non-speech regions as in [6]. Segmentation is then performed by consecutively splitting segments at the longest non-speech region with a minimal duration of at least 0.3 seconds. *b*) *GMM based* segmentation using speech, non-speech and silence models. This method uses a Viterbi decoder and MFCC GMM models for the three aforementioned categories of sounds. The general framework is based on the one in [7], which was likewise derived from [8]. In contrast to the previous work, we made use of additional features such as a zero crossing rate. *c*) *SVM based* segmentation using speech and non-speech models, using the framework introduced in [7]. 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 [9]. A 2-phased post-processing is applied for final segment generation.

We generated the segmentation of the English data with the decoder based approach. Our German data was segmented with the help of the SVM based segmentation. The data for the Italian track was pre-processed using the GMM framework. The decisions for the respective segmenters have been made in accordance to previous experiments and successful usages within the frame of various projects.

5. Acoustic Modeling

5.1. Data Preprocessing

For the TED data 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 [10].

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 and Italian 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) [11], followed by optimal feature space training (OFS) which is a variant of *semi-tied covariance* (STC) [12] training using a single global transformation matrix. The model is then refined by one iteration of Viterbi training. All models further use vocal tract length

normalization (VTLN).

In order to improve the performance of our acoustic model Boosted Maximum Mutual Information Estimation training (BMMIE) [13], a modified form of the Maximum Mutual Information (MMI) [14], is applied at the end. Lattices for discriminative training use a small unigram language model as in [15]. 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.

In contrast to our systems for English and German, we did not have an existing system for Italian, hence we bootstrapped our acoustic model using a flatstart training technique to acquire the initial models.

5.3. Hybrid Acoustic Model

As with the GMM systems we trained our hybrid systems on variance 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 13 frames, with 4-5 1600-2000 unit hidden layers and an output layer containing 6016 context dependent phonestates. 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.

5.4. Pronunciation Dictionary

For Italian, we used a pronunciation dictionary which is based on SAMPA, including consonant geminates and pronunciation variants. It contains 55 phonemes including noises and consists of the 100k words from the search vocabulary.

For our English systems we used two different phoneme sets. The first one is based on the CMU dictionary[1] and is the same phoneme set as the one used in last year's system. It consists of 45 phonemes and allophones. The second phoneme set is derived from the BEEP dictionary[2] and contains 44 phonemes and allophones. Both sets use 7 noise tags and one silence tag each. For the CMU phoneme set we generated missing pronunciations with the help of FESTIVAL [16], while for the BEEP dictionary we used Sequitur [17] instead. Both grapheme to phoneme converters were trained on subsets of the respective dictionaries.

Our German system uses an initial dictionary based on the Verbmobil Phoneset [18]. Missing pronunciations are generated using both Mary [19] and FESTIVAL [16].

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

[2] ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

5.5. Grapheme System

In addition to systems with a phoneme-based dictionary, we also built grapheme-based recognition systems for both German and Italian. By using a different set of phones, grapheme based systems are an additional source of information when doing system combination. Such systems do not require a pronunciation dictionary, as a 1:1 mapping approach between letters and sounds is used. Depending on the language, the resulting system suffers in performance as this naive approach of letter to sound mapping does not reflect any pronunciation rules.

As the pronunciation of Italian is known to be close to a 1:1 mapping, the Italian system performed only slightly worse compared to the phoneme-based system and including it into system combination resulted in overall gains. The German grapheme systems had about a 1% absolute lower WER than an equivalent phoneme system.

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, 3, and 4). Language model training was performed by building separate language models for all (sub-)corpora using the SRILM toolkit [20] with modified Kneser-Ney smoothing. These were then linearly interpolated, with interpolation weights tuned using held-out data from the TED corpus. For Italian, we attempted to compensate for the small amount of data by using a more elaborate language model with data selected via Moore's method [21], but observed no significant improvement in terms of word error rate. For German, we split compounds similarly as in [22].

For the vocabulary selection, we followed an approach proposed by Venkataraman et al.[23]. 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, 300k words for German, and 100k words for Italian.

7. Decoding Setup

For the evaluation, we built four final systems for Italian. Three are based on the phoneme dictionary. One is using a neural network trained entirely on English for feature extraction, one is using a neural network that was pre-trained on English but fine-tuned on Italian and the last one is using a feature front-end with just lMEL features. A fourth system is based on a grapheme dictionary and uses a network that was trained entirely on English.

Our primary submission is a confusion network combination (CNC) using all three phoneme-based systems. The first contrastive system uses the phoneme dictionary and the

Text corpus	# Words
TED	3m
News + News-commentary + -crawl	4,478m
Euronews	780k
Commoncrawl	185m
GIGA	2323m
Europarl + UN + multi-UN	829m
Google Books	(1b n-grams)

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.*

network that was trained using only English data. The second contrastive system is based on graphemes and is using the same neural network. Our third contrastive system is a ROVER of the two phoneme-based systems using a neural network and the grapheme-based system using the network trained on English entirely.

For our English submission we trained 5 different DBNF GMM acoustic models in total by combining different feature front-ends (M2 and lMEL) and different phoneme sets (CMU and BEEP). In addition to these systems, we trained 2 DBNF DNN hybrid systems, one for each phoneme set. For our primary submission, we combined all 7 systems in a

Text corpus	# Words
TED	3,050k
ECI	480k
Euronews	725k
Google Books	(437m n-grams)

Table 4: *Italian language modeling data after cleaning and data selection. The total number of words was 4.3 million, not counting Google Books.*

System	Dev
lMel+FFV+Pitch EN-NN	38.4
lMel+FFV+Pitch EN-NN Grapheme	38.7
lMel+FFV+Pitch EN-NN IT-ft	40.7
lMel	40.8
ROVER	37.4
CNC	37.1

Table 5: *Italian language results on development data (dev2014)*

CNC. The 5 DBNF GMM systems were adapted in an unsupervised manner on the combination of the first stage outputs using VTLN, MLLR, and cMLLR. A second CNC was computed using the adapted systems and the 2 unadapted hybrid systems. The final submission consists of a ROVER of both CNCs, the 5 adapted DBNF GMM systems and the 2 hybrid systems.

The German setup consisted of 9 separate subsystems 5 with discriminativly trained GMM acoustic models (**bmmie**) and 4 using DNN acoustic models (**hyb**). A confusion network combination is performed on the output of these 9 systems which is then used to adapt the 5 GMM based acoustic models for which a 2nd pass speaker adaped pass is then performed. In the 2nd confusion network combination the 2nd pass systems replace the orginal GMM systems. A ROVER of the hybrid systems, the 2nd pass GMM system and both CNCs results in the final output.

8. Results

Our German evaluation setup has improved noticeably since last year from 18.3% to 17.6% (see Table 7). The best first pass system now has a WER of 19.2%, an improvement of 0.8% abs. over last year's best first pass system. The best 2nd pass system has improved by 1.0% abs.

We evaluated our Italian system on the 2014 dev set (dev2014). Tabel 5 shows the results for different single systems and ROVER and CNC combinations.

The English system has been evaluated on the test sets "dev2012". The results are listed in Table6.

9. Conclusions

In this paper we presented our Italian, English and German LVCSR systems, with which we participated in the 2014 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.

In German we were able to considerably improve the system over last year's system. For Italian we created for the first time a large scale Italian speech recognition system for

System	dev2012
M2+T-CMU	15.7
lMEL+T-CMU	15.5
M2+T-16ms-CMU	15.9
M2+T-BEEP	16.0
lMEL+T-BEEP	16.2
lMEL+T-hyb-CMU	15.9
lMEL+T-hyb-BEEP	16.7
CNC-BEEP-01	13.4
M2+T-CMU	14.3
lMEL+T-CMU	14.4
M2+T-16ms-CMU	14.8
M2+T-BEEP	14.6
lMEL+T-BEEP	14.5
CNC-BEEP-02	13.5
ROVER	13.4

Table 6: *Results for English on development test sets.*

evaluation purposes.

10. Acknowledgements

The authors which to thank Roberto Gretter for providing an Italian pronunciation dictionary for us. The work leading to these results has received funding from the European Union under grant agreement $n \circ 287658$.

11. References

[1] 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.

[2] Christian Saam, Christian Mohr, Kevin Kilgour, Michael Heck, Matthias Sperber, Keigo Kubo, Sebastian Stüker, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura, and lex Waibel, "The 2012 KIT and KIT-NAIST English ASR Systems for the IWSLT Evaluation," in *International Workshop on Spoken Language Translation (IWSLT)*, Dec. 2012.

[3] 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.

[4] 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 *Procedings of the 10th Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

System	Dev2012
lMEL-all-hyb-P	19.4
lMEL-nl-hyb-P	19.2
M2+T-G-bmmie	21.0
M2-hyb-P	20.4
lMEL+T-P-bmmie	20.2
lMEL-hyb-P	19.3
M2-G-bmmie	22.2
M2-P-bmmie	20.3
M2+T-P-bmmie	20.0
CNC1	17.9
M2+T-G-bmmie	19.5
lMEL+T-P-bmmie	19.0
M2-G-bmmie	20.9
M2+T-P-bmmie	18.7
M2-P-bmmie	19.3
CNC2	17.6
ROVER	17.6
2013 setup	18.3
best 2013 1. pass	20.0
best 2013 2. pass	19.7

Table 7: *Results for German language on development data. Systems designated with **M2** use MFCC+MVDR features, **lMEL** systems use log Mel feature and **+T** means that the system also uses tonal features. Hybrid systems are marked with **hyb** with **bmmie** corresponding to systems using bmmie trained GMM acoustic models. Some systems are phoneme based **P** while others are grapheme based **G**.*

[5] 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.

[6] S. Stüker, C. Fügen, F. Kraft, and M. Wölfel, "The ISL 2007 English Speech Transcription System for European Parliament Speeches," in *Proceedings of the 10th European Conference on Speech Communication and Technology (INTERSPEECH 2007)*, Antwerp, Belgium, August 2007, pp. 2609–2612.

[7] 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.

[8] H. Yu, Y.-C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz, "The ISL RT04 Mandarin Broadcast News Evaluation System," in *EARS Rich Transcription Workshop*, 2004.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Commun.*, vol. 50, no. 5, pp. 434–451, May 2008. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2008.01.002

[18] 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.

[19] 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.

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

[21] R. C. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," in *Proceedings of ACL*, 2010.

[22] 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.

[23] 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.