The 2013 KIT IWSLT Speech-to-Text Systems for German and English

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

Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany

{kevin.kilgour|christian.mohr|heck|quoc.nguyen|van.nguyen|eugene.sheen}@kit.edu
{igor.tseyzer|jonas.gehring|m.mueller|matthias.sperber|sebastian.stueker|waibel}@kit.edu

Abstract

This paper describes our English *Speech-to-Text* (STT) systems for the 2013 IWSLT TED ASR track. The systems consist of multiple subsystems that are combinations of different front-ends, e.g. MVDR-MFCC based and lMel based ones, GMM and NN acoustic models and different phone sets. The outputs of the subsystems are combined via confusion network combination. Decoding is done in two stages, where the systems of the second stage are adapted in an unsupervised manner on the combination of the first stage outputs using VTLN, MLLR, and cMLLR.

Index Terms: speech recognition, IWSLT, TED talks, evaluation system, system development

1. Introduction

[1] The *International Workshop on Spoken Language Translation* (IWSLT) offers a comprehensive evaluation campaign on spoken language translation. One part of the campaign focuses on the translation of 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) [2]. In order to evaluate different aspects of this task IWSLT organizes several evaluation tracks on this data covering the aspects of automatic speech recognition (ASR), machine translation (MT), and the full-fledged combination of the two of them into speech translation systems.

The goal of the TED ASR track is the automatic transcription of TED lectures on a given segmentation, in order to interface with the machine translation components in the speech-translation track. The quality of the resulting transcriptions are measured in word error rate (WER).

In this paper we describe our English ASR systems with which we participated in the TED ASR track of the 2013 IWSLT evaluation campaign. This year, our system is a further development of our last year's evaluation system [3] and makes use of system combination and cross-adaptation, by utilising both GMM and Neural Network acoustic models which are trained with different acoustic front-ends and employ different phoneme sets. We also included TED talks available via TED's website by training on them in a slightly supervised manner.

We submitted primary systems for both the German and English evaluations.

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 two 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 and our decoding strategy and results are presented in sections 7 and 8.

2. Data Resources

2.1. Training Data

For acoustic model training we used the following English data sources:

- 200 hours of Quaero training data from 2010 to 2012.
- 18 hours of various noise data, such as snippets of applause and music.
- 158 hours of data downloaded from the TED talks website that was released before the cut-off date of 31 December 2010, including the corresponding subtitles provided by the TED conferences archive.

and the following German data sources:

- 179 hours of Quaero training data from 2010 to 2012.
- 24 hours of broadcast news data

These training set or subsets hereof are also used for the training of the automatic segmenters, that are applied to the evaluation data before decoding.

For English language model training and vocabulary selection, we used the subtitles of TED talks and text data from

Text corpus	**# Words**
TED	3M
News + News commentary	2,114M
GIGA parallel	523M
Gigaword 4	1,800M
UN + Europarl	376M
Google Books Ngrams (subset)	(1000M ngrams)

Table 1: English language modeling data after cleaning and data selection. The total number of words was 4.8 billion, not counting Google Books.

Text corpus	**# Words**
TED (translated)	2,259k
Callhome	150k
Europarl	47,306k
HUB5	19k
MultiUN	5,849k
News+News Commentary	284,415k
ECI	12,652k
Euro Language Newspaper	86,785k
German Political Speeches	5,514k
Common Crawl	47,046k
Google Web Ngrams	1.3T

Table 2: German language modeling data after cleaning and data selection. In total, we used 492 million words, not counting Google Ngrams.

various sources (see Table 1) and for the German language model training and vocabulary selection, we used translated subtitles of TED talks and text data from various sources (see Table 2).

2.2. Test Data

Table 3 describes three test sets ("tst2011", "tst2012" and "tst2013") used for this year's English evaluation campaign, as well as our development set for system development and parameter optimization ("dev2012"). "tst2011" is comprised of TED talks newer than December 2010 and serves as progress test set to measure the improvement in systems from 2011 onwards. "tst2012" is last year's evaluation set, and "tst2013" is a collection of some of the most recent recordings made available by TED. All test sets were used with the original pre-segmentation provided by the IWSLT organizers, except for this year's evaluation set ("tst2013") which has been segmented automatically before decoding. For the German system on a single test set "dev2013" was available.

3. Feature Extraction

Our systems are built using several different front ends that use various inputs for computing deep bottle neck features.

Set	#talks	#utt	dur	dur/utt
dev2012	10	1144	1.7h	5.4s
tst2011	8	818	1.1h	4.9s
tst2012	11	1124	1.7h	5.6s
tst2013	28	1438	4.2h	10.5s

Table 3: Statistics of the development set ("dev2012") and the test sets ("tst2011", "tst2012" and "tst2013"), 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" has been segmented automatically.

The two main input variants, each using a frame shift of 10ms and a frame size of 32ms, are the MFCC+MVDR (M2) features that have been shown to be very effective when used in BNFs [4] and standard lMEL features which generally outperform MFCCs as DBNF inputs. These standard features are often augmented by tonal features. In [**?**] we demonstrate, that the addition of tonal (T) features not only greatly reduces the WER on tonal languages like Vietnamese and Cantonese but also results in small gains on non-tonal languages like English.

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. The first 4-5 hidden layers are pretrained layer-wise as denoising autoencoders after which the network the finetuned as a whole [5]. As can be seen in figure 1 the layers after the bottlenet are discareded and 13 (+-6) bottleneck frames are stacked and reduced back down to a 42 dimensional input feature using LDA.

4. Automatic Segmentation

For this year's ASR track, the evaluation set was provided without manual sentence segmentation, thus automatic segmentation of the target data was mandatory. We evaluated the effectiveness of three different approaches to automatic segmentation of audio data, which are:

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 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 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 frame-

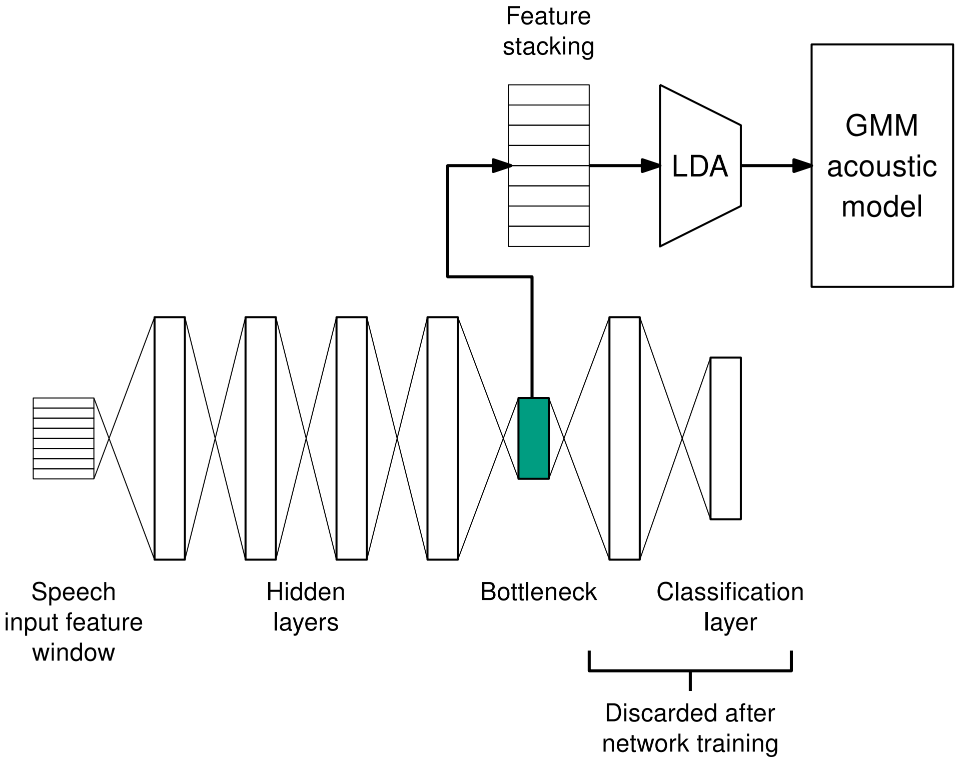

Figure 1: *Overview of our standard DBNF setup.*

work introduced in [7]. The pre-processing makes use of an LDA transformation on 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 segmentation generation.

Table 4 shows the decoding performance of a confusion network combination of hypotheses generated by five development systems after a first pass decoding on the "dev2012" set, for each preliminary application of the various techniques for segmentation.

Segmentation	WER	#utt	dur	dur/utt
Manual	13.2%	1144	1.71h	5.4s
Decoder based	13.8%	594	1.83h	11.1s
SVM based	13.9%	431	1.78h	14.9s
GMM based	14.3%	695	1.77h	9.2s

Table 4: Decoding performance on and statistics of the development set ("dev2012") after automatic segmentation, including the word error rate (*WER*), the total number of utterances (*#utt*), the overall speech duration (*dur*), and average speech duration per utterance (*dur/utt*).

On the English development set the decoder based approach resulted in the best performance in terms of WER, so we decided in favor of the latter for application on the evaluation set. For the German system we used the SVM based segmenters since it performed best on the German development set.

5. Acoustic Modeling

We trained several different acoustic models for each language.

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 to discriminate speech and non-speech and a forced alignment given the subtitles was done where only the relevant speech parts detected by the decoding were used. The procedure is the same that has been applied in [10].

5.2. AM training Setup

The models of all systems are context-dependent quinphones with three states per phoneme, using a left-to-right HMM topology without skip states. All English acoustic models initially use 8,000 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 were 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 one global transformation matrix, and finally refined by one iteration of Viterbi training. All models further use vocal tract length normalization (VTLN).

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

We experimented with using neural network acoustic models. Using the same techniques described in the deep bottleneck layer section we trained neural networks on various input features and with different topologies. Our best setups used deep bottleneck features stacked 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. On the eval2010 test set this system had a WER of 14.61%, which is 0.5% better than this best non hybrid single pass system.

5.4. Pronunciation Dictionary

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 years 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 [13], while for the BEEP dictionary we used Sequitur [14] instead. Both grapheme to phoneme converters were trained on subsets of the respective dictionaries.

5.5. Grapheme System

We built grapheme-based recognizer for both English and German. In order to built the Englsih grapheme-based dictionary, we used a data-driven approach to cluster the most common combinations of letters in order to better reflect the specifics of the English language. These clusters contain for instance combinations such as sch, sh or th. We added these in addition to all the letters of the English alphabet to the set of phones.

Using this dictionary, be trained a system using flatstart training on the training data of the 2011 training set. After doing the context-independent flatstart training, we built a context-dependent system on top of that.

As our best result, we archived to get a WER of 31.8% using a clustertree with 6000 states. Since this WER is quite high compared to the WER of our other systems, we decided not to include this system either in our system-combination or the submission.

The German grapheme system on the other hand performed only slighty worse than our phoneme based system and resulted in overall gains when included in the final system combination.

5.6. BMMIE training

In order to improve the performance of acoustic model, the Boosted Maximum Mutual Information Estimation training (BMMIE) [15] is applied, it is a modified form of the Maximum Mutual Information (MMI) [16]. We wrote lattices for discriminative training using a small unigram language model as in [17]. After lattices generating, the BMMIE training is applied for three iterations with boosting factor b=0.5. This approach resulted in about 0.6% WER improvement for 1st-pass sytems and about 0.4% WER for 2nd-pass systems.

6. Lanuage Models and Search Vocabulary

Language modeling was performed by building separate language models for all (sub-)corpora using the SRILM toolkit [18] with modified Kneser-Ney smoothing. These were then linearly interpolated, with interpolation weights tuned using held-out data from the TED corpus.

6.1. Subword Language Model for German

In order to select a sub-word vocabulary we first perform compound splitting on all the text corpora and tag the split compounds. Linking morphemes are attached to the proceeding word. *Wirtschaftsdelegationsmitglieder* is, for example, split into *Wirtschafts+ Delegations+ Mitglieder (eng: members of the economic delegation)*.

Our compound splitting algorithm requires a set of valid sub-words and selects the best split from all possible splits by maximizing the sum of the squares of all sub-word lengths [19]. For the word *Konsumentenumfrage* this heuristic would correctly choose *Konsumenten Umfrag* over *Konsum Enten Umfrage*.

As a set of valid sub-words we selected the top k words from a ranked word-list generated in the same mannar as our English vocabulary. After applying coumpound splitting to all our text corpora the same maximum likelihood vocabulary selection method is used again to select the best vocabulary from this split corpora resulting in a ranked vocabulary containing both full words and sub-words tagged with a "+".

Pronunciations missing from the initial dictionary are created with both Festival and Mary [20]. The sub-word language model is trained on the split corpora and tuning text analogous to the English language model.

[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

System	Dev2012	Eval2011	Eval2012
M2+T-CMU	15.9	11.6	11.7
lMEL+T-CMU	16.1	11.4	11.4
M2+T-DLabel-CMU	15.8	11.2	11.5
M2+T-BEEP	16.2	12.0	12.6
lMEL+T-BEEP	16.1	12.2	12.6
M2+T-hyb-CMU	16.5	11.9	11.6
M2+T-hyb-BEEP	16.9	12.4	12.4
CNC-BEEP-01	13.7	9.8	9.5
M2+T-CMU	14.7	10.3	10.3
lMEL+T-CMU	15.0	10.2	10.1
M2+T-DLabel-CMU	14.5	10.3	10.1
M2+T-BEEP	14.7	10.8	10.5
lMEL+T-BEEP	14.4	10.6	10.6
CNC-BEEP-02	13.3	9.3	9.2
ROVER	13.3	9.2	9.0

Table 5: Results for English language on development data and evaluation data.

System	Dev	Eval
M2-P-bmmie-i3	21.00	29.40
M2+T-P-bmmie-i4	20.80	30.80
M2+T-G-bmmie-i3	21.70	29.80
M2-hyb-P	21.40	30.50
lMEL+T-P-bmmie-i3	21.10	29.70
lMEL-hyb-P	20.20	29.20
M2-G-bmmie-vit	22.90	30.70
CNC-01	18.60	26.70
M2-P-bmmie-i3-SAT	19.90	27.90
M2+T-P-bmmie-i4-SAT	19.60	27.80
M2+T-G-bmmie-i3-SAT	20.50	27.90
lMEL+T-P-bmmie-i3-SAT	20.10	27.80
M2-G-bmmie-vit-SAT	21.70	29.00
CNC-02	18.30	26.40
ROVER	18.30	26.30

Table 6: Results for German language on development data und evaluation data.

7. Decoding Setup

The decoding was performed with the *Janus Recognition Tool-kit* (JRTk) developed at Karlsruhe Institute of Technology and Carnegie Mellon University [21]. Our decoding strategy is based on the principle of system combination and cross-system adaptation. System combination works on the principle that different systems commit different errors that cancel each other out. Cross-system adaptation profits from the fact that the unsupervised acoustic model adaptation works better when performed on output that was created with a different system that works approximately equally well [22]. The final step in our system decoding set-up is the ROVER combination of several outputs [23].

8. Results

We evaluated our systems on the IWSLT test sets 2011 (tst2011), 2012 (tst2012) and the 2012 dev set. We used the dev2012 set as development set and for parameter optimization and the eval 2012 set to compare our system with last years evaluation results (see table 5). Last year our best system had a WER of 12% on the eval 2012 set which we were able to reduce to 9% with this year's evaluation system.

9. Conclusions

In this paper we presented our English and German LVCSR systems, with which we participated in the 2013 IWSLT evaluation.

10. Acknowledgements

'*Research Group 3-01*' received financial support by the '*Concept for the Future*' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. The work leading to these results has received funding from the European Union under grant agreement $n \circ 287658$. This work was partly realized within the Quaero Programme, funded by OSEO, French State agency for innovation.

11. References

[1] S. Stüker, K. Kilgour, and F. Kraft, "Quaero 2010 speech-to-text evaluation systems," in *High Performance Computing in Science and Engineering '11*, W. E. Nagel, D. B. Kröner, and M. M. Resch, Eds. Springer Berlin Heidelberg, 2012, pp. 607–618.

[2] S. Stüker, F. Kraft, C. Mohr, T. Herrmann, E. Cho, and A. Waibel, "The KIT lecture corpus for speech translation," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, to appear.

[3] 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.

[4] 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.

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

[14] 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

[15] 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.

[16] 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.

[17] 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.

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

[19] T. Marek, "Analysis of german compounds using weighted finite state transducers," *Bachelor thesis, University of Tübingen*, 2006.

[20] 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.

[21] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 214–217.

[22] Sebastian Stüker, Christian Fügen, Susanne Burger, and Matthias Wölfel, "Cross-System Adaptation and Combination for Continuous Speech Recognition: The Influence of Phoneme Set and Acoustic Front-End," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006, ICSLP)*, Pittsburgh, PA, USA: ISCA, Nov. 2006, pp. 521–524.

[23] 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.