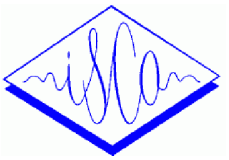
ISCA Archive
http://www.isca-speech.org/archive



International Workshop on
Spoken Language Translation 2011
San Francisco, CA, USA
December 8-9, 2011

The 2011 KIT English ASR System for the IWSLT Evaluation

Sebastian Stüker[1,2], *Kevin Kilgour*[1,2], *Christian Saam*[1,2], *Alex Waibel*[1],

[1]Institute for Anthropomatics
[2]Research Group 3-01 'Multilingual Speech Recognition'
Karlsruhe Institute of Technology
Karlsruhe, Germany

{kevin.kilgour|christian.saam|sebastian.stueker|alex.waibel}@kit.edu

Abstract

This paper describes our English *Speech-to-Text* (STT) system for the 2011 IWSLT ASR track. The system consists of 2 subsystems with different front-ends—one MVDR based, one MFCC based—which are combined using confusion network combination to provide a base for a second pass speaker adapted MVDR system. We demonstrate that this set-up produces competitive results on the IWSLT 2010 dev and test sets.

1. Introduction

In this paper we describe our English *Speech-to-Text* (STT) system with which we participated in the 2011 IWSLT STT evaluation [1]. Our system makes use of system combination and cross-adaptation, by utilising acoustic models which are trained with different acoustic front-ends.

The system has been derived from our 2010 English Quaero ASR evaluation system, by taking acoustic models out of that system and combining them with a language model that has been specifically tailored to the IWSLT lecture task.

1.1. IWSLT

The goal of the *International Workshop on Spoken Language Translation* (IWSLT) evaluation campaign is 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). In order to evaluate different aspects of spoken language translation IWSLT offers 4 evaluation tracks, the ASR and MT tracks are traditional evaluations measuring the word error rate (WER) of ASR systems and the quality (in BLEU) of the MT systems when translating the transcripts. In the SLT track the performance of MT systems on ASR output is evaluated and the SC track evaluates the performance of system combination techniques.

The rest of this paper is structured as follows. Section 2 provides a description of two acoustic front-ends used in our system. An overview of the techniques and data used to build our acoustic models is given in Section 3. We describe the language model used for this evaluation in Section 4 and our decoding strategy is explained in Section 5.

2. Front-ends

We trained systems for two different kinds of acoustic front-ends. One is based on the widely used *Mel-frequency Cepstral Coefficients* (MFCC) obtained from a discrete Fourier transform and the other on the *warped minimum variance distortionless response* (MVDR). The second front-end replaces the Fourier transformation by a warped MVDR spectral envelope [2], which is a time domain technique to estimate an all-pole model using a warped short time frequency axis such as the Mel scale. The use of the MVDR eliminates the overemphasis of harmonic peaks typically seen in medium and high pitched voiced speech when spectral estimation is based on linear prediction.

For training, both front-ends have provided features every 10 ms. During decoding this was changed to 8ms after the first stage. In training and decoding, the features were obtained either by the Fourier transformation followed by a Mel-filterbank or the warped MVDR spectral envelope.

For the MVDR front-end we used a model order of 22 without any filter-bank since the warped MVDR already provides the properties of the Mel-filterbank, namely warping to the Mel-frequency and smoothing. The advantage of this approach over the use of a higher model order and a linear-filterbank for dimensionality reduction is an increase in resolution in low frequency regions which cannot be attained with traditionally used Mel-filterbanks. Furthermore, with the MVDR we apply an unequal modelling of spectral peaks and valleys that improves noise robustness, due to the fact that noise is mainly present in low energy regions.

Both frond ends apply vocal tract length normalization (VTLN) [3]. For MFCC this is done in the linear domain, for MVDR in the warped frequency domain. The MFCC front-end uses 13 cepstral coefficients, the MVDR front-end uses 15. The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. For both front-ends 15 adjacent frames were combined into one single feature vector. The resulting feature vectors were then reduced to 42 dimensions using *linear discriminant analysis* (LDA).

3. Acoustic Modeling

For a given front-end our standard method of training an acoustic model requires first performing LDA to reduce the input dimension. All models are context dependent quinphone systems with three states per phoneme, and a left-to-right topology without skip states. All models use 6,000 distributions and codebooks. The models were trained using incremental splitting of Gaussians training (MAS), followed by *Semi-Tied Covariance* (STC) [4] training using one global matrix, and 2 iterations of Viterbi training. All models use *vocal tract length normalization* (VTLN). In addition to that *feature space constraint MLLR* (cMLLR) speaker adaptive training (SAT) [5] was applied on top.

We improved the initial acoustic models further with the help of Maximum Mutual Information Estimation (MMIE) training [6]. We applied MMIE training firstly to the models after the 2 Viterbi iterations, and secondly to the models after the FSA-SAT training, taking the adaptation matrices from the last iteration of the maximum likelihood FSA-training and keeping them unchanged during the MMIE training.

3.1. Training Data

For acoustic model training we used a mix of data of several types and from different sources:

- 80h of manually trancribed English European Parliament Plenary Session (EPPS) data provided by RWTH Aachen within the TC-STAR project [7]
- 167h of unsupervised EPPS training material that had been collected within TC-STAR by RWTH Aachen but had not been manually transcribed
- 9.8h of data from the Translingual English Data database [8]
- 140h of BroadCast News data from the HUB-4 corpus
- 50h of Quaero data

4. Language Modeling

A 4-gram case sensitive language model with modified Kneser-Ney smoothing was built for each of the text sources listed in Table 1. This was done using the SRI Language Modelling Toolkit [9]. The transcripts of the IWSLT training data were cleaned and split into a 3,000k word training set and a 593k word tuning set. The aforementioned language models built from the text sources in Table 1 were interpolated using interpolation weights estimated on the tuning set resulting in a language model with 47,554k 2-grams, 277,442k 3-grams and 788,400k 4-grams. Even compressed in an easy to load binary format our language model required about 7.4 Gbytes of RAM. Our ASR system deals with this by loading the language model into a region of shared memory and allows multiple decoder instances running on different cores to access it. On a fully utilized 16 core compute node for example the language model will only require about 0.5 GByte per instance.

Text corpus	Word Count	sources
IWSLT training data transcripts	3 million	2
News (+news commentary)	2114 million	4
Parallel Giga Corpus	523 million	1
UN + Europarl documents	376 million	1
google Book Ngrams	1.12 bln ngrams	1
total	3016 million	9

Table 1: *Language Model training data word count per corpus and number of text sources included in corpus. The total word count does not include the google Book Ngrams.*

4.1. Vocabulary Selection

To select the vocabulary the development data text was split into a tuning set and a test set with each containing approximately half the text of every *show*. For each of our text sources (see Table 1) we built a Witten-Bell smoothed unigram language model using the union of the text sources' vocabulary as the language models' vocabulary (global vocabulary). With the help of the maximum likelihood count estimation method described in [10] we found the best mixture weights for representing the tuning set's vocabulary as a weighted mixture of the sources' word counts thereby giving us a ranking of all the words in global vocabulary by their relevance to the tuning set. The top 150k words were selected as our vocabulary. Missing pronunciations were generated using Festival [11].

5. Decoding Strategy

Our decoding strategy is based on the principal 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 systems that works approximately equally well [12]. The set-up used for our evaluation system consists of two stages. In the first stage two systems are being run and in the second stage only one. The two systems' outputs of the first stage is combined with the help of *confusion network comination* (CNC) [13]. On this output the acoustic model of the second stage is then adapted using *Vocal Tract Length Normalization* (VTLN) [3], *Maximum Likelihood Linear Regression* (MLLR) [14], and *feature space constrained MLLR* (fMLLR) [15].

The segmentation of the individual shows into sentence like units was already given by the evaluators. For the sake of simplicity we only assumed one speaker per lecture and did not perform any automatic speaker clustering.

Table 2 shows the word error rates of the different stages

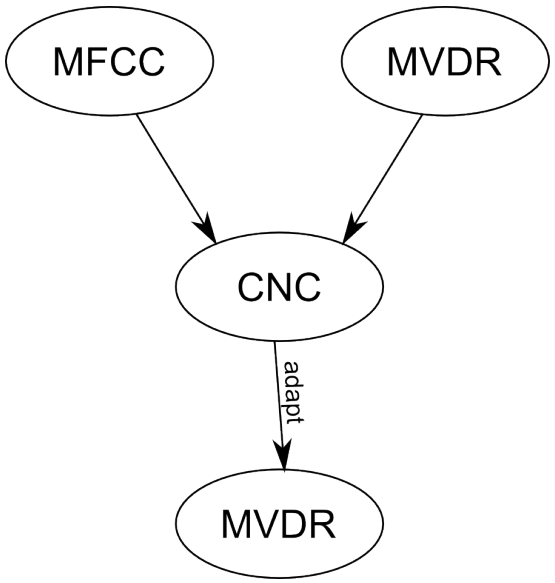


Figure 1: *Decoding Strategy*

system	dev2010	test2010
MVDR	24.6%	22.8%
MFCC	25.0%	23.1%
CNC	24.9%	22.0%
MVDR 2nd pass	21.2%	19.7%

Table 2: *Results on the 2010 test and dev set.*

on the IWSLT 2010 dev and test set for the lecture task.

6. Conclusion

In this paper we described our English speech-to-text system with which we participated in the 2011 IWSLT evaluation on the lecture task. While the acoustic model was unchanged from last year's system, we retrained the language model in order to fit the constraints for this year's evaluation. Our system utilizes a multi-pass strategy with system combination. On the 2010 development set for the IWSLT lecture task our system achieves a WER of 21.2%, and a WER of 19.7% on the 2010 test set.

7. Acknowledgements

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. '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.

8. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the iwslt 2011 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2011*, San Francisco, CA, USA, December 8-9 2011.

[2] M. Wölfel and J. McDonough, "Minimum variance distortionless response spectralestimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, September 2005.

[3] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *ICASSP*, Munich, Germany, April 1997.

[4] M. Gales, "Semi-tied covariance matrices for hidden markov models," Cambridge University, Engineering Department, Tech. Rep., February 1998.

[5] ——, "Maximum likelihood linear transformations for hmm-based speech recognition," Cambridge University, Engineering Department, Tech. Rep., May 1997.

[6] D. Povey and P. Woodland, "Improved discriminative training techniques for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, May 2001.

[7] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross domain automatic transcription on the tc-star epps corpus," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, PA, USA, March 2005.

[8] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the ted corpus lectures," in *International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, March 2003.

[9] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *ICSLP*, 2002.

[10] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," *Arxiv preprint cs/0306022*, 2003.

[11] A. Black and P. Taylor, "The festival speech synthesis system: System docmunation," Human Communication Research Centre, University of Edingburgh, Edingburgh, Scotland, United Kingdom, Tech. Rep., 1997.

[12] S. Stüker, C. Fügen, S. Burger, and M. 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, September 2006, pp. 521–524.

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

[14] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[15] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 357–366, 1995.