The UEDIN ASR Systems for the IWSLT 2014 Evaluation

Peter Bell, Pawel Swietojanski, Joris Driesen,
Mark Sinclair, Fergus McInnes, Steve Renals

School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK

{peter.bell,p.swietojanski, fergus.mcinnes,s.renals}@ed.ac.uk,
jdriesen@staffmail.ed.ac.uk, m.sinclair-7@sms.ed.ac.uk

Abstract

This paper describes the University of Edinburgh (UEDIN) ASR systems for the 2014 IWSLT Evaluation. Notable features of the English system include deep neural network acoustic models in both tandem and hybrid configuration with the use of multi-level adaptive networks, LHUC adaptation and Maxout units. The German system includes lightly supervised training and a new method for dictionary generation. Our voice activity detection system now uses a semi-Markov model to incorporate a prior on utterance lengths. There are improvements of up to 30% relative WER on the `tst2013` English test set.

1. Introduction

This paper describes our system for automatic speech recognition (ASR) of TED talks, used in the 2014 evaluation campaign of the International Workshop on Spoken Language Translation. We describe both our English and German systems, although the development of the two was carried out separately.

This is the third year we have participated in the English ASR task. Our 2012 system [1] used tandem-GMM acoustic models, using deep neural networks (DNNs) to derive bottleneck features, incorporating out-of-domain data from multiparty meetings using the multi-level adaptive networks (MLAN) scheme [2]. In 2013 [3], we combined DNN systems in both tandem and hybrid configurations, again using the MLAN scheme. We also made extensive improvements to our language models, devoting substantial efforts to text normalisation, and data selection using the cross-entropy difference score proposed by [4]. These improvements led to a WER reduction from 12.4% to 10.2% on the `tst2011` progress test set.

This year, our final system features a system combination of several complementary systems built using the HTK and Kaldi toolkits. On the language modelling side, other than using a larger 4-gram model for final rescoring, there are very few changes from last year. This year's system does not employ recurrent neural network language models, as we were unable to obtain gains with the size of models used. On the acoustic modelling side, there are a number of new features: improved speaker adaptation for the DNNs with our recently proposed Learning Hidden Unit Contributions (LHUC) scheme [5]; the use of Maxout [6] and rectified linear units for the DNNs [7]; sequence training of some neural networks [8]; and the use of mixed-band training data. These features of the system are described in more detail in Section 3.

The German system is described separately in Section 4. For German, our major challenge is the lack of reliably-transcribed in-domain acoustic training data, and a good quality dictionary, neither of which we have access to. Like last year, we rely in bootstrapping a system from the German portion of the GlobalPhone corpus, using a biased language model method. We also use a new technique for dictionary expansion [9].

In the 2013 evaluation, ASR systems were required for the first time to operate without manually-supplied segmentation of the test data into utterances. We therefore used an automatic voice activity detection (VAD) based segmenter on the `tst2013` set as input to the ASR. We have since identified a number of problems with the baseline VAD system used in 2013, including a mismatch to the acoustic conditions, and a tendency to segment too tightly, leading to word deletions at sentence boundaries: we describe our work to address these problems in Section 2.

2. Voice activity detection

There were substantial changes to this year's VAD system, used for both English and German systems. After comparing the ASR performance with VAD-based segmentation on manually-segmented development data, we observed a reduction in performance compared to when the manual segmentations were used directly, even when local speech/silence decisions were generally correct. We hypothesised that this is because the utterances are often semantically segmented by human annotators, making them better-suited to language models trained on complete sentences. Additionally, in our system we observed an unfortunate trade-off between an over-sensitive segmenter which

This work was supported by the European Union under the FP7 projects inEvent (grant agreement 287872) and EU-Bridge (grant agreement 287658), and by EPSRC Programme Grant grant EP/I031022/1, *Natural Speech Technology*.

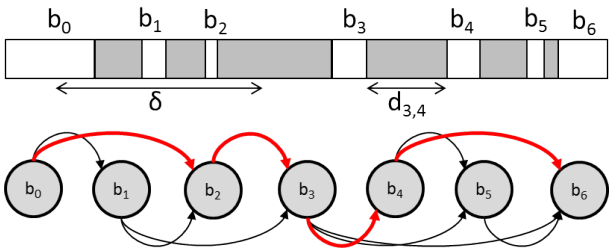

Figure 1: *An example of a candidate break sequence and associated state topology. The transitions highlighted in red show an example optimal break sequence* $B^* = \{b_0, b_2, b_3, b_4, b_6\}$

results in lots of short utterances, and an under-sensitive one, which can lead to excessively long segments or insertions and deletions at utterances boundaries.

As a solution to both problems, in [10], we proposed a novel technique based using a semi-Markov model with an prior on the duration of an utterance designed to yield segments more closely matching the distribution found in training data. For the prior, we used a log-normal distribution with parameters estimated on manually segmented training data. We found that the log-normal distribution generally provides a good fit to the distribution of utterance durations in the training data.

As input to the the semi-Markov decoder, we use a highly sensitive segmentation with small minimum duration constraint of 100ms. This produces many break points that would normally be detrimental to ASR if used directly. We decode this sequence of breaks using a semi-Markov decoder, to find the globally optimal sequence of breaks. The method is illustrated in Figure 1. Further details may be found in [10].

The initial segmentation is produced with an GMM-HMM based model. Speech and non-speech are modelled with diagonal covariance GMMs with 12 and 5 mixture components respectively. Features are calculated every 10ms from a 30ms analysis window and have a dimensionality of 14 (13 PLPs and energy). Models were trained on 70 hours of scenario meetings data from the AMI corpus using the provided manual segmentations as a reference.

3. English systems

3.1. Language modelling

Our language modelling setup is largely unchanged from last year, but we summarise it here for completeness. We trained standard Kneser-Ney smoothed n-gram language models on a combination of TED talk transcriptions as in-domain data, and out-of-domain data sources specified by the IWSLT rules. Table 1 shows the text data available, to which we applied substantial pre-processing and normalisation.

Following [4], we used all the available in-domain data,

Corpus	Total	Selected
TED	2.4M	2.4M
Europarl	53.1M	6.3M
News Commentary	4.4M	0.7M
News Crawl	693.5M	72.9M
Gigaword	2915.6M	232.9M
OOD total	3666.6M	312.8M

Table 1: Numbers of words in LM training sets.

Language model	Perplexity
TED 3-gram	183.2
OOD (312MW / 751MW) 3-gram	133.5 / 138.3
TED+OOD (312MW / 751MW) 3-gram	125.1 / 124.9
TED 4-gram	179.9
OOD (312MW / 751MW) 4-gram	123.9 / 126.4
TED+OOD (312MW / 751MW) 4-gram	114.9 / 113.4

Table 2: Perplexities of N-gram language models on TED development set.

and selected a subset of out-of-domain (OOD) data, D_s to minimise the cross-entropy difference:

$$D_S = \{s | H_I(s) - H_O(s) < \tau\} \quad (1)$$

where $H_I(s)$ is a cross-entropy of a sentence with a LM trained on in-domain data, $H_O(s)$ is a cross-entropy of a sentence with a LM trained on a random subset of the OOD data of similar size to the TED corpus, and τ is a threshold to control the size of D_S. Interpolation parameters were tuned on the `dev2010` and `tst2010` sets.

Table 2 shows perplexities of the in-domain, OOD and final interpolated LMs. In both Kaldi and HTK decoding pipelines the smaller 3-gram model was used for the primary decoding passes; when Kaldi's WFST-based decoder was used, the 3-gram was pruned to reduce memory requirements. In both cases, lattices were finally rescored using an unpruned 4-gram LM. Compared to 2013, when only models trained on 312MW set were used, this year we used the substantially larger 4-gram model trained on 715M words for the final pass. Due to the limitations of HDecode, we again limited the vocabulary to below 64k words based on occurrence count. This limit was also applied in the Kaldi systems, a restriction we plan to remove in future.

We also investigated the use of RNN models, which were interpolated with the 4-gram model, and used to rescore the 3-gram lattices. However, we did not use these models in the system, as we were unable to observe any performance improvements over the large 4-gram model on its own. This is probably due to the fact that the RNNs available at the time of submission were trained on much smaller quantities of text.

Corpus	Quantity (hrs)
TED talks	143
Switchboard	285
AMI meetings (a)	127
AMI meetings (b)	78

Table 3: Training data quantities

3.2. Acoustic modelling

3.2.1. Training data

For in domain training data, as in previous years, we used 813 TED talks recorded prior to the end of 2010, which were aligned to the transcriptions available online using an efficient lightly-supervised technique [11]. We also used two sources of out-of-domain data: the Switchboard 1 corpus of conversational telephone speech, and the AMI corpus of multi-party meetings[1]. The quantities of speech data are summarised in Table 3.

As can be seen from the table, we use the AMI meetings corpus in two configurations. Previously, we have assumed that the AMI corpus is not well-matched to the TED domain, and used it purely as a means of generating bottleneck features for the MLAN scheme described in Section 3.2.2. In this case, we use a setup (a) described in [12]. Following last year's evaluation, however, we observed that with the passing of time, the changing format and expanding scope of TED talks has led to the pre-2010 data no longer being the best match for future test sets. This year, therefore, we decided to train one set of acoustic models on a combination of the TED and AMI data. In this case, we used a more recently-defined training setup (b) that aims to be reproducible by other sites and forms the basis of a Kaldi recipe. This is described in detail in [13].

3.2.2. Tandem MLAN systems

The multi-level adaptive networks (MLAN) scheme [14] aims to make optimal use of mismatched OOD data in training a system for which limited data is available for the target domain. Taking advantage of the fact that features derived from neural networks are known to be portable across domains, OOD DNNs with a bottleneck layer [15] are used to generate features for the in-domain data. In the MLAN scheme, a second-level network is trained on these features, augmented with the original acoustic features, to ensure robustness when the input bottleneck features are poorly-matched to the new domain, and – since each DNN incorporates several frames of acoustic context – allowing wider acoustic context to be incorporated without additional parameters.

The MLAN scheme has a particular advantage when used with the Switchboard telephone data, as it allows us to make good use of narrowband data without the need for upsampling, which may cause performance degradation. To do this, the first level nets are trained on the 8khz Switchboard data. To generate features for the TED data, we can simply downsample this data in to match the telephone data. The bottleneck features are then augmented with standard acoustic features derived *without* the need for any change in sample rate.

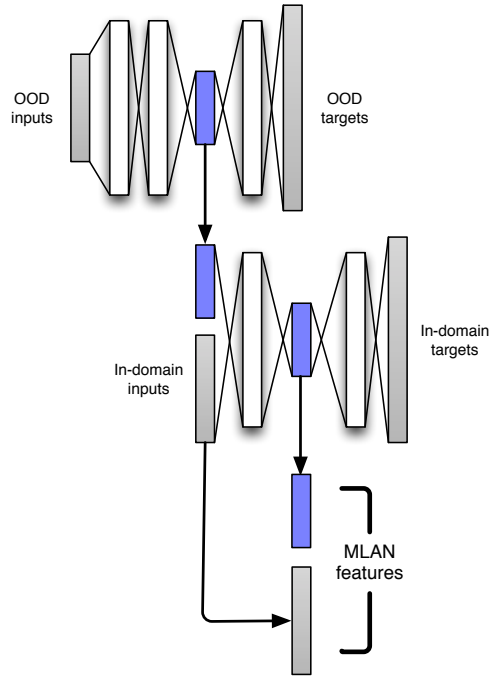

Figure 2: Tandem MLAN feature generation

In this year's system, we used MLAN purely in a tandem configuration [16], whereby the final bottleneck features are augmented with the original acoustic features and used to train a GMM. The complete feature generation process is illustrated in Figure 2. The advantage of this configuration is that it allows us to take advantage of the large quantity of training data available for each test speaker in the TED task by estimating multiple CMLLR adaptation transforms per speaker with a regression class tree.

All tandem networks use 6 wide layers with 2048 hidden units per layer; the bottleneck layers have 30 units. The nets are trained with the standard cross-entropy criterion using approximately 6,000 context-dependent triphone targets derived from a baseline GMM. Input acoustic features are PLPs with first and second derivatives – 39 features in total. Both first- and second-level networks use 9 frames of acoustic context. The final GMMs have MPE training applied. All tandem systems use HTK, as we were unable to achieve comparable performance with Kaldi on these features.

3.2.3. Hybrid LHUC systems

We have previously used DNNs in a hybrid configuration, whereby the nets are used to generate posterior probabili-

[1] http://corpus.amiproject.org/

ties over tied-state triphones for direct use in the decoder. We have noted that speaker adaptation, using a global fMLLR transform per speaker, is essential for competitive performance on the TED task. This year, in addition, we experimented with the use of our recently-proposed technique [5] for creating speaker-dependent DNNs by adapting each hidden layer on a per-speaker basis, which we term Learning Hidden Unit Contributions (LHUC). We briefly summarise it here. Consider the l-th hidden layer of a DNN, given by

$$\mathbf{h}^l = \phi^l \left(\mathbf{W}^{l\top} \mathbf{h}^{l-l} \right). \quad (2)$$

where $\mathbf{W}^{l\top}$ are the weights and ϕ^l is the nonlinear transfer function at the l-th hidden layer. We modify a standard speaker independent (SI) DNN by defining a set of speaker-dependent (SD) parameters for speaker m, $\theta^m = \{\mathbf{r}_m^1, \ldots, \mathbf{r}_m^L\}$, where $\mathbf{r}_m^l \in \mathbb{R}^{M^l}$ is the vector of SD parameters for the lth hidden layer. If $a(\mathbf{r}_m^l)$ is element-wise function that constrains the range of $\mathbf{r}_m^l$, then we can modify (2) to define a hidden layer output that is specific to speaker m:

$$\mathbf{h}_m^l = a(\mathbf{r}_m^l) \circ \phi^l \left(\mathbf{W}^{l\top} \mathbf{h}_m^{l-l} \right), \quad (3)$$

where $\circ$ is an element-wise multiplication. The SD term can be viewed as applying different weights to the contributions from each the hidden units on a per-speaker basis. We define $a(\cdot)$ as a sigmoid with amplitude 2, $a(c) = 2/(1+\exp(-c))$, so that each speaker-dependent weighting is strictly positive and centered at one. This re-parametrisation is for optimisation purposes only; at runtime $a(\cdot)$ can be evaluated once for a given set of θ^m and directly used as a scaling factor. The SD parameters are optimised with respect to the negative log posterior probability $\mathcal{F}(\theta^m)$ over T^m adaptation data-points of the m-th speaker, similar to the SI case:

$$\mathcal{F}(\theta^m) = -\sum_t^{T^m} \log P(s_t | \mathbf{x}_t^m; \theta^m). \quad (4)$$

given speech samples $\mathbf{x}_t$ and tied state labels s_t

We investigated the use of LHUC with three different non-linearities ϕ^l: in addition to the standard sigmoid, we use rectifying linear units [7] and Maxout units [17] which we proposed for ASR in [6]. Rather than applying any explicit function, the maxout network groups linear activations, and passes forward the maximum value in each group:

$$h_i^l = \max_{k=0}^{K-1} (z_{j+k}^l), \quad j = i.K \quad (5)$$

where the z_i^l are the linear outputs of the l-th layer.

Our hybrid DNNs again use 2048 hidden units per layer, but with 12,000 tied-state outputs. The input features are again PLPs with first and second derivatives, and 9 frames of context in total. For the maxout non-linearity we set the number of hidden maxout units to 1500, with a group size of two. All models had fMLLR applied to the input feature space. The LHUC nets were trained only on the 143 hours of TED data. All adaptation on the test set was performed on a per-talk basis using the output from a first-pass decode.

We also trained a single DNN system on a combination of the TED data and the AMI corpus setup (b), with sequence training following the recipe of [8]. As we will show in the results section, the use of the AMI corpus appears to particularly benefit performance on `tst2013`, perhaps due to its poorer match to the pre-2010 TED data.

3.3. Results

We present development results on `tst2011` generated with manual segmentations. Table 4 compares performance of tandem MLAN systems with a baseline trained purely on in-domain features. Consistent with previous results, it may be seen that the use of OOD data gives significant performance improvements: it is interesting to see that the use of entirely mismatched narrowband telephone speech from Switchboard still leads to a 13.5% relative WER reduction with the 3gram LM. The results of the Hybrid LHUC systems are shown in Table 5 (these results are not fully comparable with the results from the previous table as a weaker LM is used). The LHUC technique leads to gains with all three types of non-linearity investigated, and appears to be complementary to the use of fMLLR transforms on the input space. Both the ReLU and Maxout non-linearities appear to derive greater benefit from LHUC.

Model	3gram	4gram
Baseline tandem	12.6	-
SWB MLAN	10.9	10.3
AMI MLAN	11.2	9.8
ROVER	-	9.3

Table 4: Tandem MLAN DNN development results on `tst2011`. All systems are trained with MPE.

Model	WER (%)
DNN	15.2
+LHUC	13.7 (-9.9)
+fMLLR	13.9 (-8.5)
+LHUC	12.9 (-15.1)
ReLU	15.2
+LHUC	13.5 (-11.2)
+fMLLR	13.6 (-10.5)
+LHUC	12.7 (-16.4)
Maxout	14.3
+LHUC	12.8 (-10.4)
+fMLLR	12.5 (-12.6)
+LHUC	11.9 (-16.8)

Table 5: Hybrid DNN development results on `tst2011` using weak 3gram LM. Relative improvements are given in parentheses w.r.t. the corresponding SI model.

Model	WER (%)
2013 systems	
AMI MLAN	22.9
Final submission	21.5
HTK tandem systems	
AMI MLAN	18.1
SWB MLAN	17.2
Kaldi hybrid systems	
ReLU + LHUC	18.4
MaxOut + LHUC	18.7
TED+AMI Seq	15.3
ROVER combinations	
Tandem MLAN	16.6
All Hybrid	15.3
All systems	**14.4**

Table 6: Final systems with automatic segmentation on `tst2013`

Finally, we present results on `tst2013` with automatic segmentation in Table 6. All these results use lattice rescoring with the 751MW 4gram model. The system comination weights for ROVER were tuned on the development sets `dev2010`, `tst2010` and `tst2011`. Note that our scoring is not entirely consistent with that performed in the 2013 evaluation: we obtain hypothesis-to-reference alignments over the entire talk, rather than on a per utterance basis. We believe this approach is fairer as it makes the scoring more robust to slight discrepancies in segment timings between the human reference and the automatic system, which can lead to single words being counted as a deletion error in one segment and an insertion error in the adjoining segment. For comparison, our final 2013 scores 21.5% with full-talk scoring, compared to 22.1% by the official method.

From the table, we see that the new VAD system gives an absolute WER reduction of 3.8% on the AMI MLAN system, which is otherwise unchanged from 2013. Again, the two tandem MLAN systems are highly complementary when used in combination; the sequence-trained DNN trained with both TED and AMI data seems to perform particularly well on the `tst2013`, perhaps reflecting the more diverse range of accents in this test set. Finally, the tandem and hybrid systems are seen to be complementary, resulting in a further reduction in WER to 14.4%. On the `tst2014` test set, this final system has an official score of 12.7%. However, as noted above, this result includes a number of erroneous insertions and deletions at utterance boundaries. Scoring on a per-talk basis against the same reference transcription yields a WER of 10.7%.

4. German system

A major hurdle in achieving high-quality recognition lies in the collection of appropriate training data, both for acoustic modelling and language modelling. For acoustic modelling, participants in this year's ASR evaluation track were provided with German data from the Euronews corpus, a speech corpus that contains news broadcasts in a multitude of languages [18, 19]. The permitted training data was not limited to Euronews, however. Any speech recording made before a certain cut-off date (17/07/2012) could be included. We have chosen to include recordings of plenary sessions of the European parliament, made between January 2007 and December 2010. These recordings are publicly available online, along with their approximate transcriptions [20, 21]. Both text and audio are available in German making this data readily usable for acoustic model training. We will henceforth refer to this set of data as *Europarl*. Lastly, we have included the GlobalPhone corpus in the training data [22].

For LM training, we used the same method that was described in [20] and used in the ASR track of IWSLT 2013. Briefly, it consists of selecting 30% of the training data according to maximum cross-entropy with the target domain [23]. Then, a 3-gram language model is trained on this selected data using Kneser-Ney smoothing and a vocabulary is determined by selecting the top 1-grams in this model, ranked according to decreasing smoothed 1-gram probability. Finally, 4-gram LM training is performed on the same data selection, in which the words are restricted to those in the chosen vocabulary. RNN language model were trained using the RNNLM toolkit [24]. During evaluation, these RNN models were used to rescore 100-best lists, i.e. the 100 most likely utterance recognition hypotheses, that were generated with the 4-gram LM.

4.1. Language Modelling

German Language models were trained on all the German monolingual text corpora provided in the ACL statistical machine translation workshop 2014 [25], and the in-domain text data provided by the organisers of IWSLT 2014. They are listed in table 7. The text in each of these corpora was tokenised as follows: first, all the punctuation is removed. Then all numbers in the text are expanded, as are the most common units, e.g. currency, distance, volume, weight, etc. Any word that is completely capitalised, or in which the letters are separated by full stops, is treated as an abbreviation, and its letters are spelled out. For further details, see [20].

Full-sized 4-gram LMs are trained on each of these text corpora, after which they are interpolated. The interpolation weights are optimised, so as to reduce the perplexity of the resulting LM on an in-domain text corpus, here the text of `dev2012`. Since the list of words contained in this LM is prohibitively large for ASR, it has to be limited to the top words in the ranked list described above. Choosing the size of the vocabulary is a trade-off between model perplexity and OOV-rate, as is shown in Table 8. We have opted for a vocabulary of size 300k. This list of words is turned into a lexicon for ASR, as discussed below, in section 4.2. We will refer to this lexicon as $dict_1$. Since the final 4-gram LM is too large to use in ASR directly, we prune it with a threshold of

corpus	10^6 words
Europarl-v7	47.4
News Commentary	4.5
News Crawl 2007	31.5
News Crawl 2008	107.9
News Crawl 2009	101.6
News Crawl 2010	45.9
News Crawl 2011	252.8
News Crawl 2012	319.7
News Crawl 2013	543.0
IWSLT	2.8
Total	1455.0

Table 7: The different training corpora used for German language modelling, and their sizes

#words	ppl	oov-rate (%)
$1 \cdot 10^5$	235.45	4.22
$2 \cdot 10^5$	261.49	2.85
$3 \cdot 10^5$	274.33	2.36
$4 \cdot 10^5$	280.29	2.14

Table 8: Perplexities on `dev2012`, along with the OOV-rate of the resulting 4-gram LMs, limited to different vocabulary sizes.

10^{-7}. The resulting reduced LM will be referred to below as LM_1. For RNN training, the vocabulary was further reduced to 50k, for computational reasons. We train it on a random selection of 10M lines from the corpora listed in table 8. The hidden layer of the network contains 30 nodes.

4.2. Acoustic Modelling

As discussed above, data sources available for German acoustic model training are Euronews, GlobalPhone, and Europarl. Since Europarl has only approximate transcriptions, we have to apply some form of light supervision on it, in order to obtain a subset in which the transcriptions are accurate. We do this using the same method as in [26]. We use an initial acoustic model, GMM_0, and a biased language model, LM_0, to perform recognition on the entire data, and define a new training set which contains only the segments where the recognition matches the approximate transcriptions. Although a new model trained on this set can in principle be used to repeat the procedure iteratively, there are no guarantees that models from such subsequent iterations will be significantly superior. On the contrary, one even runs the risk of degrading the model by applying this technique iteratively [27]. We have therefore only run a single iteration of data selection on Europarl. The biased Language model, LM_0, was obtained by interpolating the LM provided with GlobalPhone with a language model trained on the annotations of the Europarl speech data. The initial acoustic model, GMM_0, was trained on a combination of Euronews and GlobalPhone. The

corpus	GP	EN	EP	total
#hours	14.85	57.35	79.90	152.10

Table 9: The size of all the different data sources for acoustic model training.

data	WER (%)
GP	49.64
+ EN	44.05
+ EP	41.38

Table 10: Word Error Rates on `dev2012` using different acoustic models

acoustic features were extracted in frames of 25 ms, with a shift 10 ms. 13 MFCC coefficients in each frame were stacked within context windows of 9 frames, and the resulting 117-dimensional representations were projected down to 40 dimensions using LDA/MLLT [28]. GMM_0 has 3000 context dependent states, with a total of 48000 Gaussians. No adaptation was performed. From an initial estimated total of 733 hours of Europarl data, this model allows us to select about 80 hours. This number may seem small, but the total data is likely an overestimate due to overlapping speech segments. Moreover, the majority of the data consists of non-German segments, the speech and transcriptions of which are translated into German separately. The disagreement between text and audio is therefore very large. The amount of useful data from each corpus is listed in table 9, where GP stands for GlobalPhone, EN for Euronews, and EP for Europarl.

To demonstrate the benefits of adding each of these data sets, we have trained simple acoustic models on GlobalPhone (GP), on a combination of GlobalPhone and Euronews (GP+EN), and on all data combined (GP+EN+EP). The dictionary used in this training, which we will call $dict_0$ comprises the GlobalPhone dictionary, augmented with all the OOV words from the three training sets, altogether about 140000 words. The transcription of new words is generated with Sequitur G2P [29], trained on the 40000 words of GlobalPhone. The performance of the resulting models was evaluated on `dev2012`. The WERs are shown in table 10. We can see that, even though the domains of the different training sets are quite far apart, and none close to that of the development set, they all contribute to some extent in improving the results. We will therefore use a combination of these three sets for all acoustic model training that follows. The error rates shown in table 10 are rather high because little effort was taken to tune these evaluations to the target domain. $dict_0$ is a relatively small dictionary (for German), and the language model LM_0 is biased towards Europarl, not TED.

Using all available training data, i.e. GP+EP+EN, we perform speaker-adaptive training in order to obtain speaker dependent GMM-HMM models. The number of context dependent states in this new model was set at 9000, and

the number of Gaussians to 100000. We call this model GMM_1. Repeating the evaluation above with this model yields a WER of 37.65%, an absolute improvement of almost 4%. When we use the same acoustic model in conjunction with the LM_1, the pruned LM trained in section 4.1 and its associated dictionary, $dict_1$, the WER decreases further to 35.88%. This improvement is quite modest, considering the complexity of this LM and the fact that is specifically optimised for the TED domain. A likely reason is that the dictionary only contains about 40000 pronunciations that were manually transcribed. All the others have been generated using a grapheme-to-phoneme (G2P) conversion. All errors made during this process are propagated further through the ASR evaluation. To reduce this problem, we have performed dictionary expansion as proposed in [9]. In practice, we used G2P to generate the 10 most likely pronunciations for every word in dictionary $dict_1$, including the 40000 from the original GlobalPhone lexicon. For the latter, if none of the 10 generated pronunciations matched the original phonetic transcription, it was added as an 11th pronunciation. Initially, all pronunciations of a word are assigned a uniform probability. An alignment of the training data using model GMM_1 is then made, where the different pronunciations of each word of the transcription are set in parallel. The resulting alignments show the pronunciation of each word that best fits its acoustic realisation. Counting the occurrences of each pronunciation then allows an update of their probability in the dictionary, and a re-alignment. This is an iterative process in which the dictionary is refined in each iteration. Every few iterations, the acoustic model can be retrained as well. Here, we have chosen to do just 2 iterations, in each of which the acoustic model is retrained. We will refer to the resulting acoustic model as GMM_2. The resulting dictionary, $dict_2$, is an improvement over $dict_1$, not only because it contains pronunciation probabilities, but also because it lists pronunciations that make sense acoustically, rather than enforcing G2P's best guess. We ran an evaluation on `dev2012` with this pronunciation lexicon, using GMM_1 and the pruned LM, LM_1. The resulting WER was 29.86%, an absolute improvement of almost 6% compared to the original $dict_1$. When replacing the acoustic model GMM_1 for GMM_2, the WER becomes slightly higher: 30.91%. A possible explanation is that the degrees of freedom introduced by pronunciation variation allow the model to over-train.

A DNN is then trained up in a hybrid configuration with model GMM_2. This DNN consists of 6 hidden layers, with 2048 nodes each, connecting through a logistic sigmoid non-linearity. The output layer performs a softmax operation. At the input of the network are the MLLT-transformed speaker-adapted MFCC features we described above, stacked within a context window of 11 frames, which results in a 440-dimensional representation per frame (40× 11). The output is a vector of posterior probabilities over the context-dependent states of the GMM, converted into scaled likelihoods using prior probabilities obtained from training data [30]. The network is pre-trained with layer-wise RBM training, and finetuned by optimising a negative log-likelihood cost function. Evaluating this hybrid DNN setup on `dev2012` gives a WER of 27.83%. Note that all results thus far have either been obtained with the Europarl LM, or with a heavily pruned LM optimised for TED. The full TED-specific model has not been used due to computational limitations. We can, however, rescore the results with this larger LM, obtaining further reductions in WER. Similarly, all of the previous results can be rescored using the RNNLM. All results on `dev2012` are summarised in table 11.The system has an official score of 35.7% on the `tst2014` test set.

GMM_1 + $dict_0$ + LM_0	37.65
GMM_1 + $dict_1$ + LM_1	35.88
GMM_1 + $dict_2$ + LM_1	29.86
GMM_2 + $dict_2$ + LM_1	30.91
+ LM rescore	28.07
+ RNNLM rescore	27.59
GMM_2 + DNN + $dict_2$ + LM_1	27.83
+ LM rescore	25.33
+ RNNLM rescore	**24.90**

Table 11: The results of the German system on `dev2012`

5. Conclusions

We have described our ASR systems for the English and German 2014 IWSLT evaluation. Improvements to our English system, most particularly the use of AMI data, and the deployment of hybrid DNNs with LHUC and sequence training, result in a relative WER reduction of around 30% on the challenging `tst2013` evaluation set compared to our 2013 system. We intend to carry over these benefits to our German system, where a lack of suitable training data remains a challenge.

In the future, we plan to further investigate methods for robust DNN training and adaptation when the training data is limited or poorly-transcribed, something which should enable us to develop systems in new languages more rapidly. We also plan to work on removing the dependence on a dictionary completely, perhaps by adapting grapheme-based models. We also aim to re-incorporate RNN language models in our most competitive English system.

6. References

[1] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN systems for the IWSLT 2012 evaluation," in *Proc. IWSLT*, 2012.

[2] P. Bell, M. J. F. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland, "Transcription of multi-genre media archives using out-of-

domain data," in *Proc. IEEE Workshop on Spoken Language Technology*, Dec. 2012.

[3] P. Bell, F. McInnes, S. R. Gangireddy, M. Sinclair, A. Birch, and S. Renals, "The UEDIN english ASR system for the IWSLT 2013 evaluation," in *Proc. IWSLT*, 2013.

[4] H. Yamamoto, Y. Wu, C. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR system for IWSLT 2012," in *Proc. International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012.

[5] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. SLT*, Lake Tahoe, USA, December 2014.

[6] P. Swietojanski, J. Li, and J.-T. Huang, "Investigation of maxout networks for speech recognition," in *Proc ICASSP*, 2014.

[7] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010.

[8] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, Lyon, France, August 2013.

[9] L. Lu, A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Proc. ASRU*, 2013.

[10] M. Sinclair, P. Bell, A. Birch, and F. McInnes, "A semi-markov model for speech segmentation with an utterance-break prior," in *Proc. Interspeech*, 2014.

[11] A. Stan, P. Bell, and S. King, "A grapheme-based method for automatic alignment of speech and text data," in *Proc. SLT*, Miama, Florida, USA, Dec. 2012.

[12] 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.

[13] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. ASRU*, 2013.

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

[15] F. Grézl, M. Karafiát, S. Kontar, and J. Černokcý, "Probabilistic and bottleneck features for LVCSR of meetings," in *Proc. ICASSP*, 2007.

[16] H. Hermanksy, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1630.

[17] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv:1302.4389*, 2013.

[18] R. Gretter, "Euronews: A multilingual speech corpus for ASR," in *Proc LREC*, Reykjavik, Iceland, May 2014.

[19] ——, "Euronews: A multilingual benchmark for ASR and LID," in *Proc Interspeech*, Singapore, September 2014.

[20] J. Driesen, P. Bell, M. Sinclair, and S. Renals, "Description of the UEDIN system for German ASR," in *Proc IWSLT*, Heidelberg, Germany, December 2013.

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

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

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

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

[25] "The website of the ACL statistical machine translation workshop," 2014. [Online]. Available: www.statmt.org/wmt14

[26] 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.

[27] R. Zhang and A. Rudnickey, "A new data selection approach for semi-supervised acoustic modelling," in *Proc ICASSP*, Toulouse, France, May 2006.

[28] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

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

[30] H. Bourlard and N. Morgan. Kluwer Academic Publishers, 1994.