The I^{2}R ASR System for IWSLT 2015

Tran Huy Dat, Jonathan William Dennis, Ng Wen Zheng Terence

Human Language Technology Department
Institute for Infocomm Research, A*STAR, Singapore
{hdtran,jonathan-dennis,wztng}@i2r.a-star.edu.sg

Abstract

In this paper, we introduce the system developed at the Institute for Infocomm Research (I^{2}R) for the English ASR task within the IWSLT 2015 evaluation campaign. The front-end module of our system includes a harmonic modelling based automatic segmentation and the conventional MFCC feature extraction. The back-end module consists of an auxiliary GMM-HMM training to provide the speaker adaptive transform (SAT) and the initial forced alignment, followed by a discriminative training DNN acoustic modelling. Multi-stage decoding strategy is employed with a semi-supervised DNN adaptation which uses weighted labels generated by the previous-pass decoding output to update the trained DNN models. Finally, Recurrent Neural network (RNN) is used to train and rescore the language modelling to further improve the performances. Our system achieved 8.4 % WER on the tst2013 development set, which is better than the official results on the same set reported from the previous evaluation. For this year's tst2015 test set, we obtained 7.7% WER.

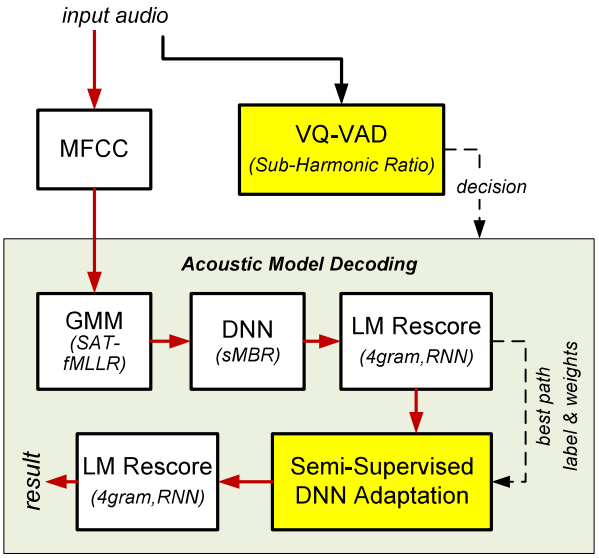

Figure 1: Overview of the I^{2}R ASR system for IWSLT 2015.

1. Introduction

The goal of the Automatic Speech Recognition (ASR) track for IWSLT 2015 is to transcribe TED talks and TEDx talks [1]. The speech in English TED talks are lectures related to Technology, Entertainment and Design (TED) in spontaneous speaking style. Despite that the speech in the TED talks is in general planned, well articulated, and recorded in high quality, the task is challenging due to the large variability of topics, the presence of non-speech events, the ascents of non-native speakers, and the informal speaking style. In this paper, we introduce our system for English TED ASR track of the 2015 IWSLT evaluation campaign. We choose to focus on developing a single system rather than a fusion of multiple platforms. The overview of our ASR system is illustrated in Fig.1. Since the TEDs' audio samples, during the test phase, are provided without class labels and timing information, automatic segmentation is necessary to split audio file into speech sentences to input the ASR system. In this work, we develop a voice activity detection (VAD) method based on harmonic modelling of speech signals and build the automatic segmentation on top of that. As the TEDs audio is normally recorded in relatively high quality, no noise compensation method is needed and we just apply the conventional MFCC features as the input to the ASR system. The training is started with an auxiliary GMM-HMM training to provide the speaker adaptive transform (SAT) and the initial alignment. Then the DNN acoustic modelling is carried out on top of SAT features with a fixed size concatenating window. The hidden layer weights are initialised using layer-wise restricted Boltzmann machine (RBM) pre-training, using 100 hours of randomly selected utterances from the training materials. Multi-stage decoding strategy is employed with semi-supervised DNN model adaptation using weighted lattices generated by the previous-pass decoding output. Finally, Recurrent Neural network (RNN) is used to train and re-score the language modelling to further improve the performances. Our system obtained WER of 8.4% on the development set (tst2013) and 7.7% on the test set (tst2015), respectively. The organisation of the rest of the paper is as follows. Secs.2 introduces the automatic segmentation. Secs.3 and 4 describes the acoustic modelling and language modelling, respectively. Secs.5 reports the experimental results and analyzes the role of each module into the ASR performances. Finally, Secs 6 concludes the paper.

2. Automatic Segmentation

The VAD module detects the speech segments based on the harmonic to sub-harmonic ratio, and uses an adaptive threshold to reject regions of noise and other non-speech and a post-processing to smooth the result.

Our approach uses a vector quantisation (VQ) system as the basis for voice activity detection (VAD), with frame selection based on both energy and the harmonic to sub-harmonic ratio (SHR) [2, 3], which is a feature for voiced speech detection. Three acoustic categories are targeted in this knowledge-based approach:

Speech - voiced speech is characterised by having both a high SHR and high energy, due to the strong harmonic structure produced during speech vocalisation.

Background Noise - for the task of lecture-style speech, where the signal-to-noise ratio (SNR) is high, the noise will typically have a much lower energy than the speech signal.

Clapping - impulsive noise has a high energy but a low SHR, which is due to the physical nature of the way the sounds are generated.

To compute the SHR within each short-time windowed frame, using a frame length of 32 ms, the amplitude spectrum $E(f)$ is first computed. For voiced segments of speech, $E(f)$ has strong peaks at the harmonics of the fundamental frequency $F0$. From this spectrum, the summation of harmonic amplitude (SHA) and summation of sub-harmonic amplitude (SSA) is computed for each frequency in the range $[F0_{min}, F0_{max}]$ as follows:

$$SHA(f) = \sum_{k=1}^{N_{harm}} \sum_{a=-\Delta}^{\Delta} E(k.f + a) \quad (1)$$

$$SSA(f) = \sum_{k=1}^{N_{harm}} \sum_{a=-\Delta}^{\Delta} E((k - \frac{1}{2}).f + a) \quad (2)$$

where only the first N_{harm} harmonics are taken into account in the summation, and a window of $\Delta = 1$ neighbouring bins are included in the summation to account for inharmonicity. Finally, the harmonic to sub-harmonic ratio (SHR) is the ratio of the two, as follows:

$$SHR(f) = \frac{SHA(f)}{SSA(f)} \quad (3)$$

where the maximum value $\max_f (SHR(f))$ is taken as the value of the feature for each frame, $SHR[t]$.

The VQ process is applied on each TED talk independently, and uses basic Mel-frequency cepstral coefficient (MFCC) as the underlying features. Our approach is to use k-Means clustering to build a set of representative vectors for each of the three categories. The top 10% of the available frames, ranked according to the above-mentioned frame-selection criteria, are used for both the speech and noise categories, while only the top 2% of frames are used for the clapping category in anticipation that less data is available.

To allow a threshold to be set for the VAD, the VQ distances are compared using the following formula:

$$VQR = \min(D_{noise}, D_{clapping}) - min(D_{speech}) \quad (4)$$

where the distances D for each category are calculated as the minimum Euclidean distance of the quantised vectors for that category. We used a threshold set at $thresh = 0$ such that speech frames are those with $VQR > thresh$.

Note that the frame-level output decision is first smoothed to join together segments separated by a gap of less than 500 milliseconds, with an additional hangover of length 500 milliseconds then applied to ensure that unvoiced speech at the start and end of the segments are not missed.

3. Acoustic Modelling

This section describes the acoustic modelling used in the I^{2}R ASR system, as shown in Figure 1. The following three aspects are detailed: (1) training data selection, (2) feature extraction and auxiliary GMM-HMM, and (3) DNN acoustic modelling.

3.1. Training Data

Following the success of the NICT system for IWSLT 2014 [4], we use a similar set of training data based on the following three corpora:

Wall Street Journal - this comprises of 81.1 hours of read speech, available from the Linguistic Data Consortium (LDC), from LDC93S6B and LDC94S13B.

HUB4 English Broadcast news - unlike [4] we use the full 201 hours of broadcast news data from LDC97S44 and LDC98S71.

TEDLIUM version 2 - this corpus contains 204 hours of lecture-style TED speech [5] consisting of 1481 talks after the removal of non-permissible talks.

Further experiments were conducted with an additional 44 hours of data extracted from the Euronews corpus [6], provided by the organisers. However, this was found to degrade the WER results by approximately 4% relative so in the final system we did not include it in the training.

3.2. Feature Extraction and Auxiliary GMM-HMM

The acoustic models (both GMM-HMM and DNN) are trained on 13-dimensional MFCCs, without energy, which are mean normalised over the speech segments extracted from each conversation for the speaker. Later, these features are spliced by ± 3 frames adjacent to the central frame and

projected down to 40 dimensions using linear discriminant analysis (LDA).

Prior to DNN training, an auxiliary GMM-HMM is first trained to provide speaker adaptive transforms (SAT) and the initial alignments for training the subsequent DNN system by forced alignment, which inherits the same tied-state structure. To train the GMM-HMM, a monophone system is first trained using the shortest twenty thousand utterances, to make the initial alignments based on a flat-start approach easier. Next, triphone and LDA GMM-HMM systems are trained with 2500 and 4000 tied states respectively, followed by SAT training to give a final SAT GMM-HMM system with 6353 tied triphone states and 150k Gaussians. The SAT approach uses feature-space maximum likelihood linear regression (fMLLR) transforms, with speech segments extracted from each conversation assumed to come from the same speaker. For training, the fMLLR transforms are computed from forced alignments, while for testing, the fMLLR transforms are computed from lattices by using 2 passes of decoding.

3.3. DNN Acoustic Modelling

The DNN acoustic model is trained on top of SAT features that are spliced ± 5 frames and rescaled to have zero mean and unit variance. The DNN has 5 hidden layers, where each hidden layer has 2048 sigmoid neurons, and a 6353 dimensional softmax output layer. The hidden layer weights are initialised using layer-wise restricted Boltzmann machine (RBM) pretraining, using 100 hours of randomly selected utterances from the TEDLIUM corpus [5]. After pretraining, fine-tuning is performed to minimize the per-frame cross-entropy between the labels and network output. The first stage of fine-tuning was performed using the same 100 hour subset as for pretraining with a learning rate of 0.008 and halving beginning when the network improvement slows. This then generated alignments for a full training set to perform a second stage of fine-tuning. Finally, the DNN is retrained by sequence-discriminative training to optimise the state minimum Bayes risk (sMBR) objective. Two iterations are performed with a fixed learning rate of 1e-5. The Kaldi toolkit is used for all experiments [7].

3.4. Semi-supervised DNN adaptation

During decoding, semi-supervised DNN adaptation is utilised on a per-talk basis to reduce any mismatch between training and testing conditions and to provide speaker adaptation of the acoustic model [8, 9]. Additional iterations of fine-training of the DNN requires a frame-level label, and potentially also a confidence measure, and these are generated based on the initial output of the system, as shown in Figure 1.

The frame-level confidence c_{frame_i} is extracted from the lattice posteriors $\gamma(i, s)$, which express the probability of being in state s at time i. The decoding output gives us the best path state sequence, $s_{i,1best}$, and the confidence values are the posteriors under this sequence, as follows [9]:

$$c_{frame_i} = \gamma\left(i, s_{i,1best}\right) \quad (5)$$

The best path state sequence and confidence measures are then used as the target labels and weightings respectively for additional iterations of DNN fine-tuning, with weights less than $c = 0.7$ set to zero. In our experiments, all weights in the network are updated, as our experiments suggested this performed better than adapting only the first layer of the DNN. The learning rate is 0.0008, with halving performed each iteration until no improvement is observed.

Category	Corpus	Sentences selected	Pct% of Original
In-domain	TED Talks	92k	-
Out-of-domain	CommonCrawl	770k	9%
	Europarl	140k	6%
	Gigaword FR-EN	0.9M	4%
	NewsCommentary	47k	19%
	News	12.3M	18%
	Yandex	310k	31%

Table 1: Training data for the language models.

4. Language Modelling

This section describes the language modelling and rescoring approaches used in the I^{2}R ASR system. The following three aspects are detailed: (1) training data selection, (2) n-gram language model training, and (3) RNN language modelling and rescoring.

4.1. Training Data and N-gram Language Model

Table 1 shows the data used for training the language models in the I^{2}R ASR system. The out-of-domain data is provided as part of the enhanced TEDLIUM version 2 corpus [5], and consists of text selected from corpus from the WMT 2013 evaluation campaign. The selection is based on the XenC tool [10], which is a filtering framework that trains both in-domain and out-of-domain language models and uses the difference in the computed scores on the out-of-domain text as an estimation of the closeness of those sentences to the in-domain subject. Text from each corpus is concatenated together to form a single large set that is used for training each of the subsequent language models.

Two n-gram language models are trained using the data selected from the available corpus as described above. The first is a 3-gram model, trained using the "Kaldi LM" package [7], which is used for DNN-based lattice generation during the first pass of decoding. The second is a 4-gram model, which is trained in an identical fashion to the one above, and is used for rescoring of the word lattice to provide a consistent improvement in WER performance.

Processing Step	WER tst2013	
	Ground Truth	Segmentation
SAT GMM	21.3%	22.4%
DNN sMBR	12.3%	11.6%
+ LM Rescore	10.8%	10.1%
+ DNN Adapt 1	9.5%	8.7%
+ DNN Adapt 2	9.4%	8.5%
+ DNN Adapt 3	9.1%	8.4%

Table 2: Detailed experimental results on the tst2013 development set showing the performance at each stage of the decoding system. Note that the DNN semi-supervised adaptation step includes a final round of language model rescoring.

4.2. RNN Language Model Training and Rescoring

A recurrent neural network (RNN) language model is trained and used for n-best list rescoring to further enhance the WER performance. The RNNLM package version 0.3e [11] was used, with 30k words in the vocabulary, 480 hidden units, 300 classes, and 2000M direct connections. Back-propagation Through Time (BPTT), with truncated time order 5, was used for RNN training, which performs joint training with a maximum entropy model to reduce the hidden layer size. The training data for the RNN was the same as above, although to enable a faster training time a random subset of 2M sentences (14% of the filtered corpora) were selected for training.

The RNN language model has a perplexity of approximately 60, and is used to rescore the output decoding lattice, with interpolation weight of 0.3 instead of using the 4-gram LM. With lower perplexity, the RNN language model can be beneficial in reducing the WER, since final ASR performance is quite dependent on a strong language model. Note that the CMU pronouncing dictionary [12] was used, limited to the words that appear in the language training databases.

5. Experimental Results

In this paper, we opt to use a single system without any combination using ROVER [13] or other techniques. At the decoding stage, we first decode the whole test set from the trained DNN acoustic models and 3-gram LM. Then the 4-gram LM rescoring is carried out, following by another RNN rescoring, described above. Next, the semi-supervised adaptation is applied for each TED test file. Each round of semi-supervised adaptation includes DNN models lattice outputting, 4-gram LM rescoring, RNN LM rescoring and DNN model adaptation. After 3-rounds of semi-supervised adaptation of the DNN acoustic model, there was no further improvement in WER on the devleopment sets, hence we applied the same number during final testing. For this year's tst2015 test set, we obtained 7.7% WER.

Processing Step	WER Gain (tst2013)
DNN sMBR	9%
+ LM Rescoring	1.5%
+ Semi-supervised DNN	1.7%

Table 3: Comparison of the approximate WER improvements given by the key components of the system, compared to the SAT-GMM result.

5.1. Results and Discussions

Table 2 reports detailed experimental results on the tst2013 development set showing the performance at each stage of the training and decoding with ground truth segmentation and the proposed automatic segmentation. We can see that the performance of the proposed segmentation is comparable to the ground truths at the baseline SAT-GMM models and and even outperformed the latter at the more comprehensive training models. The best result from tst2013 development set is 8.4% WER and it was obtained with multi-stage semi-supervised adaptation with rescoring of LM. This result is better than the official result of 10.6% WER on the same tst2013 set from last evaluation. The DNN with sMBR discriminative training yields a reasonable result of 11.6% WER and that system is fast enough to be real-time and hence recommended for the live engines.

5.2. Analysis of Word Error Rate Improvements

A summary of the contribution of each processing step to the final WER result is shown in Table 3. It can be seen that the DNN with sMBR discriminative training gives the most significant improvement in performance over the baseline SAT-GMM. In addition, the DNN decoding strategy gives a total of around 2-3% improvement, with the biggest contribution coming from the semi-supervised DNN speaker adaptation, combined with a consistent improvement achieved through language model rescoring. The semi-supervised DNN adaptation is suitable for TED and TEDx talks since it involves a single speaker and long enough to be effective. However, a big jump of performance is normally seen in the first round of adaptation while it is very time consuming. Hence, in practical situations, using one round of adaptation is recommended.

6. Conclusions

In this paper, we described our English ASR system for IWSLT 2015 evaluation campaign. This is a single system consisting of harmonic modelling voice activity detection (VAD) for automatic segmentation, speaker adaptive training (SAT) GMM-HMM initial forced alignment, DNN acoustic modelling with sMBR discriminative training, RNN language modelling and rescoring, and semi-supervised DNN adaptation in decoding. We obtained good performances on both the development and test sets. Among the system, the

harmonic modelling VAD, the DNN acoustic modelling with discriminative training, the semi-supervised DNN adaptation have found to be the key components which contributed to the ASR improvements compared to the baseline systems.

7. References

[1] "Ted," https://www.ted.com/talks.

[2] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–333.

[3] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics." in *Interspeech*, 2011, pp. 1973–1976.

[4] P. Shen, X. Lu, X. Hu, N. Kanda, M. Saiko, and C. Hori, "The NICT ASR system for IWSLT 2014," in *Proceedings of IWSLT 2014*, 2014, pp. 113–118.

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

[6] R. Gretter, ""Euronews: a multilingual speech corpus for ASR," in *Proceedings of LREC*, 2012, pp. 4161–4164.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, Ondej Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2011.

[8] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[9] K. Vesely, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 267–272.

[10] A. Rousseau, "Xenc: An open-source tool for data selection in natural language processing," *The Prague Bulletin of Mathematical Linguistics*, vol. 100, pp. 73–82, 2013.

[11] T. Mikolov, "Statistical language models based on neural networks," 2012.

[12] C. M. University, "The carnegie mellon university pronouncing dictionary v07a," in *[Online] http://www.speech.cs.cmu.edu/cgi-bin/cmudict*, 2015.

[13] J. Fiscus, "A post-processing system to yield reduced error rates: recognizer output voting error reduction (ROVER)," in *Proceedings of LREC*, 1997, p. 347354.