A lecture transcription system combining neural network acoustic and language models

Peter Bell[1], *Hitoshi Yamamoto*[2], *Pawel Swietojanski*[1],
Youzheng Wu[2], *Fergus McInnes*[1], *Chiori Hori*[2] *and Steve Renals*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
[2]Spoken Language Communication Laboratory,
National Institute of Information and Communications Technology, Kyoto, Japan

peter.bell@ed.ac.uk, hitoshi.yamamoto@nict.go.jp

Abstract

This paper presents a new system for automatic transcription of lectures. The system combines a number of novel features, including deep neural network acoustic models using multi-level adaptive networks to incorporate out-of-domain information, and factored recurrent neural network language models. We demonstrate that the system achieves large improvements on the TED lecture transcription task from the 2012 IWSLT evaluation – our results are currently the best reported on this task, showing an relative WER reduction of more than 16% compared to the closest competing system from the evaluation.

Index Terms: large vocabulary speech recognition, lecture transcription, deep neural networks, MLAN, factored RNN language model

1. Introduction

Since 1984, TED (Technology, Entertainment, Design; `http://www.ted.com`) has organised "riveting talks by remarkable people, free to the world" and now organises lecture series around the globe. The short TED talks in diverse disciplines are made freely available online under a Creative Commons license. There is now an abundance of crowd-sourced transcriptions and translations of the lecture material available online, making the transcription and translation of the TED talks an ideal evaluation task for automatic speech recognition (ASR) and machine translation (MT) systems. The audio quality of the recorded TED talks is very good (speakers typically use a head-mounted microphone) compared to domains such as conversational telephone speech or meetings. However, the large vocabulary and diversity of topics presents a significant speech recognition and machine translation challenge. The International Workshop on Spoken Language Translation (IWSLT; `http://iwslt2012.org`) now uses the TED talks as part of its evaluation campaigns in speech transcription, machine translation, and speech translation.

Until recently, typical state-of-the-art systems for speech recognition [1, 2, 3] were based on the HMM-GMM, discriminatively trained with an objective such as MMI or MPE, unsupervised speaker adaptation using VTLN and MLLR/CMLLR, perhaps using tandem or bottleneck features obtained from neural networks. Decoding was performed with a 3-gram or 4-gram language model. However, since 2010, there has been a resurgence in the use of neural network-based models previously explored in the 1990s [4, 5, 6], with work such as [7, 8, 9] showing that hybrid neural network / HMM systems can offer substantial improvements in accuracy over state-of-the-art HMM-GMM systems. The improvements can be explained by a number of factors: the use of deep neural networks (DNNs) with many hidden layers; modelling context-dependent phone states, resulting in a much larger number of output classes; and the use of unsupervised pre-training. In language modelling, too, neural network based approaches have seen success, including both feed-forward [10, 11] and recurrent [12] neural network language models.

In this paper we investigate the improvements on accuracy resulting from these new techniques on the IWSLT TED talks task. We present results combining elements from two previous entries to the English ASR track of the IWSLT 2012 evaluation campaign [13], from the University of Edinburgh, UK (UEDIN) [14] and the National Institute of Information and Communications Technology, Japan (NICT) [15]. Our final system, which follows the IWSLT rules regarding training data, uses acoustic models from the UEDIN system and language models from the NICT system. Together, the system contains several novel features: for acoustic models, we use DNNs incorporating out-of-domain information using the recently-proposed Multi-level Adaptive Networks (MLAN) scheme, in both tandem and hybrid configuration [16]. For language modelling, we employ factored recurrent neural networks (fRNN) [17] alongside standard n-grams. The considerable advantage of combining the strengths of both original systems is evident in our final results, where on our main test sets we achieve a 16% relative reduction in WER compared to the best competing system from the 2012 evaluation, equivalent to a 31% relative reduction over the best system from the evaluation of 2011 [18, 19].

2. Acoustic modelling

2.1. Training data

Our in-domain acoustic model training data comprised 813 TED talks recorded before the end of 2010 (to avoid overlap with the development and test sets). These recordings were automatically segmented using SHOUT [20][1], giving 153 hours of speech. Crowd-sourced transcripts of all talks are available online; these do not include filled pauses and non-vocal noise,

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*.

[1] `http://shout-toolkit.sourceforge.net`

nor are the time-alignments reliable. Therefore we used an efficient lightly supervised technique to select the correct portion of text for each audio segment [21], in which a finite-state “skip network” is constructed from the transcript, from which the text segment is produced for each segment using Viterbi alignment with a previous acoustic model. This resulted in 143 hours of labelled training data. Untranscribed filled pauses that may be present in the speech segments are in effect modelled by states in our generic silence model. We trained baseline HMM-GMMs on this training set, using 12th order (+C0) PLP features with first, second, and third derivatives, projected to a 39-dimension feature vector using HLDA. The models were standard three-state left-to-right HMMs modelling clustered cross-word triphone states. The models used approximately 5000 tied states with 16 Gaussians per state.

2.2. Deep neural networks

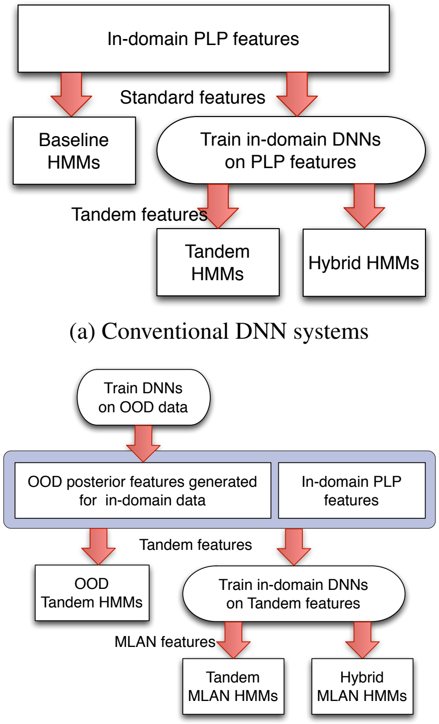
Our acoustic models use deep neural networks in both tandem and hybrid configurations. For tandem systems [22, 23], the net is used to generate log posterior probabilities over monophones. These probabilities are decorrelated and projected to 30 dimensions using PCA and augmented with the original acoustic features, giving a total feature vector size of 69 dimensions. The concatenated feature vectors are used to train new GMMs: we perform the full training procedure, including triphone clustering, from scratch.

For hybrid systems, the networks are used to generate posterior probabilities over tied-state triphones [7], using the state clustering obtained from the matching tandem system. These probabilities are transformed to scaled likelihoods by dividing by the state priors estimated from the training data [4].

In both cases, the neural networks consist of several hidden layers connected via a logistic sigmoid nonlinearity. The output layer uses a softmax function. Nets are pre-trained using stacked RBMs [24], and then fine-tuned to optimise a frame-wise cross entropy criterion. The structure of the nets and training procedure was optimised using held-out validation data from the training set using frame error rate as the optimisation target.

Due to the large amount of data from single known speakers on the TED task, it is beneficial to perform unsupervised speaker adaptation on the test set. For tandem systems, it possible to use the standard CMLLR scheme [25] to adapt the parameters of each Gaussian with a linear transform: we use a maximum of 32 transforms per speaker using a regression class tree. We also apply speaker-adaptive training (SAT) using CMLLR transforms on the training set. The problem of how best to adapt hybrid models in an unsupervised manner is still unresolved [8], but we find that adaption of some kind is essential for the hybrid systems to achieve competitive performance with the adapted tandem systems (see Section 5.1). For our system, we apply CMLLR adaption, using a single transform per speaker to adapt the acoustic features prior to input to the DNN. These transforms are estimated relative to the baseline GMM trained on the same features. It is essential to perform the adaptation on the training set also, so that the DNN is operates in a speaker-normalised space.

For input, the nets used the same 39-dimension PLP feature vectors used to train the baseline GMMs, with 9 frames of temporal context. For the tandem systems, the final nets used had four hidden layers with 1024 hidden units per layer; the hybrid systems used six hidden layers with 2048 hidden units per layer. The nets were trained using our own tool based on the Theano library [26] on NVIDIA GeForce GTX 690 GPUs.



(a) Conventional DNN systems

(b) MLAN DNN systems

Figure 1: Standard and MLAN DNN systems

2.3. Multi-level adaptive networks

Neural network features are known to have good cross-domain portability. We recently proposed an effective method for incorporating information out-of-domain (OOD) data [16], without causing a deterioration in performance when the domain is mismatched to the target domain. The technique, referred to as Multi-level Adaptive Networks (MLAN), involves training a first-level neural network on OOD data. Features are generated from this network for the in-domain training data and augmented with the standard acoustic features. A second-level network is then trained on these features; this network is in effect able to select the most useful parts of the OOD features for phonetic discrimination in the target domain. The second-level networks may use either the tandem or hybrid configurations. A comparison of MLAN with standard DNN systems is shown in Figure 1. In this work we used out-of-domain data from the AMI corpus of multiparty meetings totalling 127 hours. The corpus contains natural spontaneous speech, not particularly well-matched to the TED domain, but contains a diverse range of speakers and accents.

3. Language modelling

3.1. Training data

For language model training, we used three of the English text corpora allowed by the IWSLT evaluation campaign, shown in Table 1. The TED talk transcriptions is a relatively small corpus of in-domain data, whilst the other texts constitute a relatively large out-of-domain corpus. We applied pre-processing

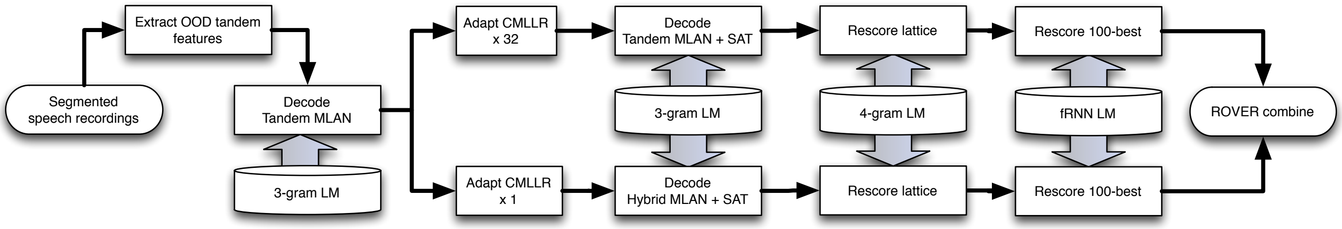


Figure 2: The full decoder architecture

Table 1: Training data of language models (tokens are counted after the pre-processing).

	Corpus	Tokens
In-domain	TED Talks	2.4M
Out-of-domain	News Commentary v7	4.5M
	English Gigaword 5th ed.	2.7G

to converted non-standard words (numbers, abbreviation, etc.) to simple words [27] and removed duplicated sentences.

3.2. Domain adapted n-gram LM

The large news corpus is likely to include many sentences that are highly mismatched to the TED domain. Such sentences are probably harmful to the LM. Therefore, we adopted domain adaptation by selecting only a portion of the news corpus. For this purpose, we employed a cross-entropy difference metric [28], which is biased towards sentences that are both similar to the in-domain (target) corpus D_I and unlike the average of the out-of-domain (source) corpus D_O. Here a selected subset D_S is represented as follows,

$$D_S = \{s \mid H_I(s) - H_O(s) < \tau\},\ s \in D_O, \tag{1}$$

where $H_c(s)$ is a cross-entropy score of sentence s according to LM_c trained on D_c ($c \in \{I, O\}$), and τ is a threshold to control the size of D_S. Note that the LM_O can be trained on a subset sentences randomly selected from D_O. We applied the metric to the news corpus, regarding the TED corpus as D_I.

For the TED corpus D_I and the selected news corpus D_S, modified Kneser-Ney smoothed n-gram LMs ($n \in \{3, 4\}$) were constructed using the SRILM toolkit [29][2]. They were linearly interpolated to form our n-gram LMs by optimizing the perplexity of the development set defined in the IWSLT evaluation campaign. Here the D_S contained 30.0M sentences (559M tokens) of the news corpus. The threshold τ in Eq. (1) was empirically set to minimize the perplexity of the development set based on LM_S. Note that the vocabulary used for LM training contained 133K words from the CMU Pronunciation Dictionary[3] and all words from the TED corpus. In the decoding setup presented here, the LMs were further restricted to the top 60k words.

3.3. Factored RNNLM

Recently, neural network based LMs (NNLMs) have become an increasingly popular choice for LVCSR tasks. Among various NNLMs, our system employed an extended version of the RNNLM [12] called factored RNNLM [17] which can exploit additional linguistic information such as morphological, syntactic, or semantic features. Here we used two kinds of features, word surface and part-of-speech tagged by the GENIA tagger[4]. The number of units in the hidden layer and classes in the output layer were 480 and 300, respectively. The training data for the factored RNNLM was the same as that of the n-gram LM described above. However, since it is very time consuming to train the model on a large amount of data, we reduced the size of D_S to 1.1M sentences (30M tokens).

4. Decoder architecture

The IWSLT evaluation provides an utterance-level segmentation of the test set audio to aid machine translation evaluation. Talk labels are also provided. Each defined test set contains 8–11 talks. For the purposes of adaptation, we assumed a single speaker per talk. For each segment, the system generated OOD tandem features using the AMI networks, as shown in Figure 1b. A first decoding pass was performed with a trigram LM using the tandem MLAN models — the best available acoustic models without no speaker adaptation. The one-best hypotheses from the first pass is used to estimate 32 CMLLR transforms on the tandem MLAN models for each speaker, and also, for each speaker, a single feature-space transform of a set of HMM-GMM models trained on the OOD tandem features. The feature space transforms are used to generate speaker-normalised features for input to the SAT hybrid MLAN models for a second-pass decoding. An additional second-pass decode is performed using SAT tandem MLAN models with standard CMLLR adaptation. Both second-pass decoders use the same trigram LMs. For decoding, we use HTK HDecode[5] with modifications to allow the use of the scaled likelihoods generated using the hybrid models.

In the later stages of the process, we apply the more powerful LMs. Given an output lattice for each utterance, we first rescore with a 4-gram LM. Since the fRNN LMs are not finite-state, to further rescore with the these models, we generate an n-best list from the lattice, rescoring on a whole-utterance basis, in combination with the original 4-gram model. Finally, we perform system combination of the two rescored n-best lists generated from the tandem and hybrid systems using the implementation of ROVER from the SRILM toolkit. The complete process is illustrated in Figure 2.

[2] http://www.speech.sri.com/projects/srilm/

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

[4] http://www.nactem.ac.uk/tsujii/software.html

[5] http://htk.eng.cam.ac.uk

Table 2: *Acoustic model development results (WER%). All results use a lightweight trigam LM.*

System	dev2010	tst2010
PLP	30.4	31.7
+ SAT	26.6	25.3
+ MPE	21.0	20.3
Baseline tandem	22.9	23.3
+ SAT	21.1	19.7
+ MPE	19.4	17.9
Baseline hybrid	21.0	20.3
+ SAT	18.6	17.6
Tandem MLAN	21.6	20.6
+ SAT	20.0	18.1
+ MPE	18.5	**16.4**
Hybrid MLAN	19.2	17.8
+ SAT	**17.8**	**16.4**

5. Results

5.1. System development

To illustrate the performance gains from various components of our system, we present development results on two past development/test sets defined by IWSLT, `dev2010` and `tst2010`. Our first experiments comparing different acoustic models used a lightweight trigram LM which was not trained on the full set of allowed training data [14]. Table 2 shows the results of our acoustic model development. The top three sections of the table show system which do not use OOD training data. Here, it may be noted that: a baseline PLP system obtains large benefits from speaker adaptation and MPE discriminative training; the tandem DNN models are substantially better than the baseline and also derive benefits from adaptation and MPE training; although the baseline hybrid DNN system outperforms its tandem equivalent, speaker-adaptive training is required for it to be competitive with the best tandem system.

The second section of table 2 shows the performance of MLAN systems, incorporating OOD features from the AMI nets. This reduces WER still further in all configurations. The relative performance of two best systems is not consistent across the two development sets. We have observed that a single speaker in `dev2010` has a much higher WER than all the others, which may result in a bias in results on that set toward systems that perform particularly well on that single speaker.

We then conducted language model experiments with both the best performing tandem MLAN and hybrid MLAN systems. Results are shown in Table 3. Replacing the lightweight trigram LM with the full trigram model described in section 3 yielded substantial benefits, leading to a consistent reduction in WER of around 2% absolute. The use of the fRNN LM lead to further performance gains. Further to this, ROVER system combination gave gains of up to 1% absolute, suggesting that the tandem and hybrid systems are complementary, particularly on `tst2010`, where the two systems are closest in performance. The final line of the table shows a combined system with additional optimisation of the relative system and LM weights. To our knowledge, these results are the best reported on these sets. As a comparison, the best-performing system at the 2011 evaluation, from MITLL [19], reported WER of 17.8% and 15.8% on `dev2010` and `tst2010` respectively; in 2012, the system from KIT–NAIST [18] reported a WER of 14.0% on `tst2010`.

Table 3: *Language model development results (WER%)*

System	dev2010	tst2010
Tandem MLAN	16.4	14.4
+ 4gram	15.6	13.8
+ fRNN	14.5	12.8
Hybrid MLAN	15.8	14.4
+ 4gram	15.1	13.5
+ fRNN	14.0	12.7
ROVER combination	15.3	13.4
+ 4gram	14.6	12.7
+ fRNN	13.7	11.9
+ tuning	**13.5**	**11.7**

Table 4: *Results on the* tst2011 *and* tst2012 *test sets (WER%)*

System	tst2011	tst2012
FBK	15.4	16.8
RWTH	13.4	13.6
UEDIN	12.4	14.4
KIT-NAIST	12.0	12.4
MITLL	11.1	13.3
NICT	10.9	12.1

(a) Results from entries to the IWSLT 2012 evaluation.

System	tst2011	tst2012
Tandem MLAN + fRNN	10.2	11.4
Hybrid MLAN + fRNN	10.3	11.3
ROVER combination	**9.3**	**10.3**

(b) Results from the new UEDIN-NICT system

5.2. Final system

We present results of the final system on the `tst2011` and `tst2012` tests sets which we use under the conditions of a formal evaluation, in that we do not tune results to this set. Table 4 compares our results to those of a number of other systems, taken from [13] including the results of independent UEDIN and NICT entries to the 2012 evaluation. On `tst2011`, our final system achieves a WER of 9.3%, a relative reduction of 15% over the 2012 NICT system, 16% over the best competing system, and 31% over the best system from 2011. The trend is similar on `tst2012`. We consider that the substantial reduction we achieve over the 2012 results demonstrates the gains to be had from the novel components of our joint system.

6. Discussion

We have presented our system for automatic transcription of TED lectures, introducing novel features including the use of combinations of speaker-adapted multi-layer adaptive networks in tandem and hybrid configurations, and the use of factored recurrent neural network language models. We demonstrated that, combined in a single system, these innovations are able to achieve large improvements over state-of-the-art systems from several other research labs.

In future, we will explore several directions for further improvements. On the acoustic modelling side, we will investigate the use of other forms of out-of-domain features, from other domains and with varying dimensionality, for example, wider bottleneck features. We plan to investigate more powerful methods for speaker adaptation of the hybrid DNN systems.

7. References

[1] S. F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.

[2] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1513–1525, 2006.

[3] T. Hain, L. Burget, J. Dines, P. N. 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.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[5] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[6] T. Robinson, "The application of recurrent nets to phone probability estimation," *IEEE Trans Neural Networks*, vol. 5, pp. 298–305, 1994.

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[8] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, 2011.

[9] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[11] H. Schwenk, "Continuous space language models," *Computer Speech and Language*, vol. 21, pp. 492–518, 2007.

[12] T. Mikolov, M. Karafiát, L. Burget, J. Černokcý, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010.

[13] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the 9th International Workshop on Spoken Language Translation*, 2012.

[14] 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.

[15] H. Yamamoto, Y. Wu, C.-L. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR system for the IWSLT2012," in *Proc. IWSLT*, 2012.

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

[17] Y. Wu, X. Lu, H. Yamamoto, S. Matsuda, C. Hori, and H. Kashioka, "Factored language model based on recurrent neural network," in *Proc. COLING*, 2012.

[18] C. Saam, C. Mohr, K. Kilgour, M. Heck, M. Sperber, K. Kubo, S. Stüker, S. Sakti, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, "The 2012 KIT and KIT-NAIST english ASR systems for the IWSLT evaluation," in *Proc. IWSLT*, 2012.

[19] A. R. Aminzadeh, T. Anderson, R. Slyh, B. Ore, E. Hansen, W. Shen, J. Drexler, and T. Gleason, "The MIT-LL/AFRL IWSLT-2011 MT system," in *Proc. IWSLT*, 2011.

[20] M. Huijbregts, C. Wooters, and R. Ordelman, "Filtering the unknown: Speech activity detection in heterogeneous video collections," in *Proc. Interspeech*, 2007.

[21] A. Stan, P. Bell, and S. King, "A grapheme-based method for automatic alignment of speech and text data," in *Proc. IEEE Workshop on Spoken Language Technology*, Miama, Florida, USA, Dec. 2012.

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

[23] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," in *Proc. Interspeech*, 2004.

[24] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[25] M. Gales, "Maximum likelihood linear transforms for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 75-98, 1998.

[26] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. SciPy*, Jun. 2010.

[27] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer Speech and Language*, vol. 15, pp. 287–333, 2001.

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

[29] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.