Cross-lingual adaptation with multi-task adaptive networks

Peter Bell, Joris Driesen, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

{peter.bell, s.renals}@ed.ac.uk, jdriesen@staffmail.ed.ac.uk

Abstract

Posterior-based or bottleneck features derived from neural networks trained on out-of-domain data may be successfully applied to improve speech recognition performance when data is scarce for the target domain or language. In this paper we combine this approach with the use of a hierarchical deep neural network (DNN) network structure – which we term a multi-level adaptive network (MLAN) – and the use of multitask learning. We have applied the technique to cross-lingual speech recognition experiments on recordings of TED talks and European Parliament sessions in English (source language) and German (target language). We demonstrate that the proposed method can lead to improvements over standard methods, even when the quantity of training data for the target language is relatively high. When the complete method is applied, we achieve relative WER reductions of around 13% compared to a monolingual hybrid DNN baseline.

Index Terms: deep neural network, multilevel adaptive networks, cross-lingual speech recognition, TED talks, European Parliament

10.21437/Interspeech.2014-5

1. Introduction

In cross-lingual automatic speech recognition (ASR), models applied to a target language are enhanced using data from a different source language. In this scenario, the target language is typically low-resourced: transcribed acoustic training data for the target language may be difficult or expensive to acquire. The cross-lingual approach is motivated by the fact that the source language data, despite being mismatched to the target, may capture common properties of the acoustics of speech which are shared across languages, improving the generalisation of the final models to unseen speakers and conditions.

Cross-lingual ASR may be viewed as a form of adaptation. In contrast to domain or speaker adaptation, the major problem with cross-lingual adaptation arises from the differences in phone sets between the source and target languages. Even when a universal phone set is used, it has been found that realisation of what is ostensibly the same phone still differs across languages [1]. In this paper, we focus on approaches where source and target languages are assumed not to share a phone set, which is probably a valid assumption when a small number of source languages are used, which are unlikely to provide complete phone coverage for an arbitrary target language.

Arguably the simplest approach to the problem of cross-lingual phoneset mismatch is to define a deterministic mapping between source and target phone sets [2] which may be estimated in a data-driven fashion [3]. However, this hard mapping leads to a loss of information from the target language acoustics that cannot be represented by a single source language phone. An alternative is to learn a probabilistic mapping, in which the distribution of target phonemes is expressed over a feature space comprising source language phone posterior probability estimates, which may be formulated as a product-of-experts model [4] or as a KL-HMM [5]. Here, the source languages may be viewed as defining a low-dimensional subspace in which to estimate target language models. This is the motivation behind the work of [6], where a subspace GMM (SGMM) is used, in which the source languages define a subspace of full covariance Gaussians.

Neural networks have been used extensively for cross-lingual ASR. Broadly, the approaches may be categorised as follows:

- *hierarchical approaches*, where source and target language neural networks are combined into a deep-structured model, where the top level net trained on the target acts as a 'merger MLP', incorporating outputs from other nets [7, 8, 9];
- *feature space approaches*, cross-lingual variants of the standard tandem [10] or bottleneck [11] techniques, where neural networks trained on source languages are used to obtain posterior or bottleneck features, on which target-language models – typically GMMs – are trained [12, 13, 14, 15], a technique first used for monolingual domain adaptation [16]. These approaches may be motivated as a variant of the subspace approaches defined above, since the aim is to find a feature space where target language models may be more readily trained, than if the raw input features were used. One advantage of neural networks for this purpose is that the feature space is implicitly discriminative, unlike the SGMM case;
- *regularisation approaches*, where target language neural network training is improved by the use of source language data, for example, by better initialisation of the nets or by weight-sharing.

These approaches have all been shown to be highly effective, and they are not mutually exclusive. There has been intense interest in regularisation approaches in recent years, with a focus on deep neural network (DNN) acoustic models used in a hybrid configuration [17, 18], directly modelling tied context-dependent states in the target language. [19] proposed to use restricted Boltzmann machine (RBM) pre-training on source languages to improve the initialisation of target-language deep neural networks (DNNs). [20] trained hybrid DNNs on a sequence of target languages, progressively swapping the output layer with each new language, whilst in [21, 22], samples from all languages are presented in an interleaved fashion during training, with the output layer swapped according to the target language being presented, following earlier work [23] where

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology), and the European Union under FP7 project grant agreement 287658 (EU-Bridge).

context-independent targets were used. The effect is to regularise the networks by sharing lower hidden layers, whilst the use of entirely different tied-state targets across languages removes the need for phone set mapping. Following [24], this technique is often called *multitask learning*, and (to our knowledge) was first applied to ASR in [25].

These approaches differ in the number of layers that are shared between languages. Often, all but the final hidden layer are shared; in other approaches, only the output layer may vary between languages.

A number of recent papers [26, 15, 27] have used multitask learning on a bottleneck network to generate language-independent features for training a conventional tandem-GMM system for the target language, thus combining the feature space and regularisation approaches, demonstrating that language-independent bottleneck features show consistent improvements over purely in-language bottleneck features.

We have recently proposed multi-level adaptive networks (MLAN) for domain adaptation [28, 29]. In the MLAN scheme, feed-forward DNNs are trained on out-of-domain data and used to generate features (either bottlenecks or decorrelated posterior features) for in-domain data. Second-level in-domain nets are trained on these features augmented with the original acoustic features. A similar somewhat similar architecture, using shallow nets, was independently proposed by [14] for spoken term detection. We showed that this architecture leads to consistent performance improvements even when the source domain is poorly matched to the target, as the second-level network implicitly selects which source features are relevant for discrimination in the target domain.

In this work, we investigate a variant of the MLAN scheme for cross-lingual ASR, where we compare out-of-language and multi-lingual bottleneck features as inputs to the second-level network. The proposed method combines all three of the approaches listed above, with the use of multitask learning, out-of-domain features and a hierarchical structure. We investigate to what extent the approaches are complementary. Unlike much other work in this area, we investigate the extent to which a cross-lingual approach is merited when the amount of target language training data is relatively large: 50 hours of speech.

2. Cross-lingual adaptive networks

The baseline deep neural network systems use a standard feed-forward architecture. The net may be viewed as a cascade of feature extractors, followed by a classification layer: this property motivates the use of intermediate layers for cross-lingual adaptation. As usual, we trained the DNNs to model frame posterior probabilities over context-dependent tied states – unique for each language – using the cross-entropy criterion. Each frame of training data is assigned a tied-state target with alignment by a language-dependent HMM-GMM system, from which the state-clustering is obtained.

The basic structure of the DNNs was held constant across all experiments: all had 6 wide hidden layers with 2048 hidden units per layer. The hidden layers use logistic sigmoid nonlinearities; the output layers use a softmax function. Where bottleneck features were required, an additional layer of 30 hidden units was placed before the final hidden layer (this placing follows experiments in [27]). The inputs to the nets use 11 frames of acoustic context. Training was performed using the Theano library [30] on NVIDIA GeForce GTX 690 GPUs.

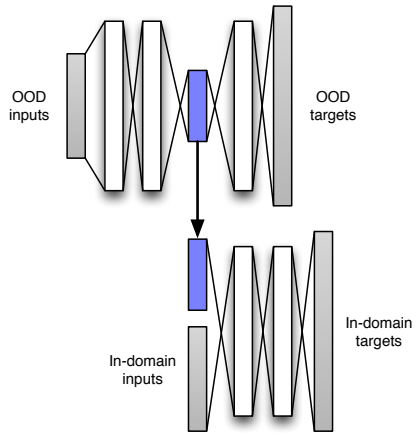

Figure 1: *Standard MLAN for domain adaptation. Input and output neurons are shaded grey.*

2.1. Multitask training

Multitask learning [24] is the method of training a classifier to operate on two related tasks using a shared representation, with the aim of improving generalisation. As implemented in a DNN framework, this involves sharing lower hidden layers of the network, whilst the output layer (and possibly the higher hidden layers) are swapped according to different sets of targets that are presented during learning. The advantages of this approach are that it effectively increases the amount of training data for each network; it may be easier for a net that would otherwise learn poorly from hidden inputs, to have more relevant inputs selected by the other net; and the other net may improve generalisation by acting as a source of noise.

As we discussed in the introduction, multitask learning was first applied to cross-lingual ASR by [23] and later used by a number of authors in various configuration [26, 27, 21, 22, 20]. The phoneset mismatch problem is solved by swapping the higher layers according to the language of the training sample presented to the network. From the perspective of a low-resource target language, the method should allow a shared, language-independent speech representation to be more robustly learned in the lower layers due to the increased training data presented. Also, compared to simply using out-of-domain neural network features as in the feature-space approaches discussed, including the target language in the neural network training ensures that the network retrains the ability to make phone discriminations not present in the source languages.

2.2. Multi-level adaptive networks

We have previously used the multi-level adaptive networks scheme (MLAN) for domain adaptation [28, 29]. This combines features derived from an out-of-domain (OOD) neural network with standard in-domain acoustic features, but then trains a new in-domain DNN on these inputs. Both levels of network use 11 frames of context. One advantage of this setup is that the second-level network is able to perform discriminative feature selection on the out-of-domain inputs. Additionally, the scheme may be viewed as a single network with parameter sharing in the lower layers, due to the fact that an identical first-level network is repeated across all frames of context input to the second, which may improve generalisation. Figure 1 illustrates the scheme. For simplicity, the diagram does not show the

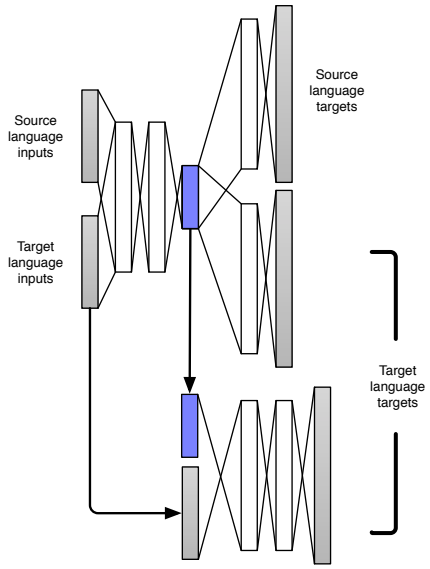

Figure 2: *Multitask MLAN in its standard implementation. Input and output neurons are shaded grey.*

use of frame stacking, and the number of hidden layers shown is purely illustrative.

We now propose to combine the MLAN scheme with multitask learning. In this case, both sets of inputs are used in training the first-level network, presented in alternate mini-batches. The second network, used to generate state posteriors for the decoder, uses only the target language acoustic features, combined with bottleneck features from the first input, a feature vector size of 69 per frame. The entire procedure is illustrated in Figure 2. Again, the use of frame stacking is not illustrated and more hidden layers are used in practice.

2.3. Refinements

We investigate a number of refinements to the basic scheme described above. First, we apply speaker adaptive training (SAT) to the second-level DNNs: as is now standard, the input features for both training and evaluation data are transformed using a global CMLLR transform per speaker, estimated using a GMM trained on the same features. We have previously found [31] that this technique reduces the word error rate of hybrid DNN systems significantly when there is plenty of data for each test speaker.

Second, our standard acoustic features comprise 13 perceptual linear prediction (PLP) features with first and second order deltas, normalised to zero mean and unit variance on a per-speaker basis. However, it is widely known that alternatives such as filterbank coefficients, not used by GMM systems due to the feature correlations, are effective as DNN inputs [32]. We therefore investigated the use of DNNs trained on filterbanks, using 23 coefficients without dynamic features. Where this was done, we continued to use PLP features as inputs to the second-level networks, as they are required for the operation of SAT. The use of complementary features at different levels may also lead to gains in its own right.

We also investigate the use of a new state-clustering to obtain targets for training the second level networks, derived from the input features to this network, rather than from the original acoustic features. Intuitively, we would expect this to lead to improvements due to the DNN being able to concentrate on modelling variation in the new feature space, rather than variation that may not be important when the BN features are included.

Finally, it is of course possible to apply multitask learning to the second-level network too. To perform this, we repeat an identical feature-generation process for the source language, and use the output as alternating inputs to the second-level nets, where again, the targets are the source language tied states.

3. Experiments

3.1. ASR task

We were primarily interested to investigate the performance of the proposed cross-lingual technique when relatively large amounts of target language data are available. We therefore chose German as the target language, owing to the ready availability of training data. For a similar reason, we used English as the source language.

Our primary experiments were carried out on a test set of German-language TED talks[1] defined by the 2013 International Workshop on Spoken Language Translation (IWSLT), where a German evaluation was introduced for the first time. This consists of a series of 10 minute lectures, with typically only one speaker per lecture. Details of the task can be found in [33].

Our English corpus matched the domain of the target speech – we used speech data obtained from TED talks in English available from the TED website. The transcriptions are crowd-sourced; a lightly supervised technique was used to match them to the speech. 50 hours of speech data was selected for training.

For target language training data, it was not possible to use German TED talks to the unavailability of transcriptions. Instead, we used a collection of European Parliament plenary sessions ("Europarl") available from the Parliament's website[2], for which approximate transcriptions are available. Again, a lightly supervised method was used for alignment – for more details see [34]. We again selected 50 hours of speech for training (note that this is a smaller quantity than in the previous reference to give more rapid experiment turnarounds. Due to the domain mismatch between training data and test set giving the potential for misleading results, we additionally created an in-domain German test set, consisting of a further 2 hours of held-out Europarl data.

3.2. Experimental setup

All systems used 3-state cross-word triphone HMMs. Our baseline system was trained on 50 hours of German Europarl, using PLP features with first, second and third temporal derivatives, projected down to 39 dimensions using an HLDA transform. The system had 2,800 tied states with 16 Gaussians per state. For the baseline, we used a two-pass system with SAT, using up to 32 CMLLR transforms per speaker.

All experiments used a trigram language model trained on approximately 300M words of news crawl data with a 60k-word vocabulary. Our lexicon was derived from the lexicon supplied with the GlobalPhone corpus [35], with missing pronunciations predicted automatically using the Sequitur G2P tool [36].

[1] `http://www.ted.com`
[2] `http://europarl.europa.eu`

3.3. Results

We present the results of the various DNN systems on the TED task in Table 1 and on the matched-domain Europarl task in Table 2. Frame error rates (FER) for the hybrid DNNs are shown for interest, but are not discussed. The trends are broadly similar for both tasks. Firstly, on the PLP-based systems, it may be observed that the baseline German (de) hybrid DNN system outperforms the baseline GMM/SAT system, although no speaker adaptation is used in the DNN system. The use of multitask training is effective here, leading to relative WER reductions of 5% and 7% on the respective tasks over the in-domain baseline. The use of the standard MLAN technique, whereby English (en) tandem features are used as inputs to a second-level network leads to reductions over the use of purely in-domain features, in the case of TED, but a deterioration in performance on Europarl. This may be due to the fact that the English features are better matched to the TED German domain, as discussed above, whilst they are trained on no more data than the German features. However, when the full multitask MLAN is used (final row of tables 1 and 2) there are gains over multitask baseline on both test sets; and larger gains over the standard MLAN method. Again, the improvement is more pronounced on TED, where the use of multitask MLAN leads to a 1.5% relative improvement over purely multitask learning, and 5.5% over the standard MLAN technique.

Table 1: *Frame and Word Error Rates on TED dev2012 (%)*

	PLP		FBANK	
System	FER	WER	FER	WER
Baseline de GMM/SAT	-	36.6	-	-
Baseline de DNN	45.6	36.1	53.7	35.4
multi DNN	44.9	34.3	52.5	33.1
de MLAN	-	-	40.9	33.9
en MLAN	43.3	35.8	42.9	34.3
multi MLAN	38.7	33.8	39.9	31.8

Table 2: *Frame and Word Error Rates on Europarl test set (%)*

	PLP		FBANK	
System	FER	WER	FER	WER
Baseline de GMM/SAT	-	16.6	-	-
Baseline de DNN	45.6	15.7	53.7	15.8
multi DNN	44.9	14.6	52.5	14.6
de MLAN	-	-	40.9	14.6
en MLAN	43.3	16.6	42.9	15.0
multi MLAN	38.7	14.5	39.9	13.9

Second, we discuss the results of hybrid systems trained on filterbank ("FBANK") features. Here, the use of multitask learning again leads to improvements over the monolingual hybrid system. The answer as to whether filterbank or PLP features are better as the input to the DNN is not consistent between tasks. The MLAN systems now incorporate filterbank-based bottleneck features into the second-layer hybrid DNN. This leads to improvements on both tasks: the German (de) features are more useful, which is no surprise considering that the training sets are the same size. However, it is notable that the English-MLAN systems outperform the purely German-trained hybrid system on both test sets. Again, the multitask MLAN system has the lowest performance on both sets. Overall, the use of multitask training, the MLAN structure, and the use of complementary filterbank features gives total relative reductions of more than 11% on both tasks, over the standard in-language DNN baseline.

Table 3: *Word Error Rates after system refinements*

	TED dev2012	Europarl
MLAN multi	31.8	13.9
+ SAT	31.1	13.7
+ states	31.0	13.1
+ multi	30.5	12.8

Finally, we show the results of additional refinements to the system in Table 3, applied to the best-performing system from the above results. We show the affect of applying SAT to the MLAN feautures used in the second-level DNN training, using global transforms estimated for each training and test speaker. Note that we used block diagonal transforms, independently transforming the bottleneck and PLP components: we have found this necessary for the transforms to be well-conditioned. As might be expected due to the larger quantity of adaptation data per test speaker for the TED task, it benefits from adaptation more than Europarl. There is additional gain from moving to a new state tying obtained from the new features. This appears to give greater benefit to Europarl, which may be because it gives a closer fit to the training data which therefore gives less benefit to the mismatched TED domain. Interestingly, there is further benefit to applying multitask training a second time to the second-level network: this gives relative improvements of around 2% on both tasks. Stripping out the effect of SAT and the change in state tying, which we did not apply to the original hybrid networks, the final system gives WERs of 31.2% and 13.7%, a net reduction over the hybrid DNNs trained on PLP features of around 13% relative on both tasks.

4. Conclusions

We have demonstrated that multitask learning may be effectively combined with a hierarchical network structure based on the MLAN scheme, for cross-lingual ASR. Use of the technique has enabled us to obtain considerable performance improvements on two German ASR tasks through the use of English training data, even when 50 hours of German training data is used.

In future we will investigate whether the first-level bottleneck features used as inputs to the second-level network can be improved further, for example, by experimenting with alternative acoustic features, and by applying speaker-adaptation on the input feature space. We also intend to investigates variants of this technique in a monolingual setting, for speaker adaptation.

5. References

[1] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. ASRU*. IEEE, 2013, pp. 138–143.

[2] V.-B. Le and L. Besacier, "First steps in fast acoustic modelling for a new target language: application to Vietnamese," *Proc. ICASSP*, 2005.

[3] K. C. Sim and H. Li, "Stream-based context-sensitive phone mapping for cross-lingual speech recognition," in *Proc. Interspeech*, 2009.

[4] K. C. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *Proc. ASRU*, 2009.

[5] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence to improve ASR for under-resourced languages," in *Proc. ICASSP*, 2012.

[6] L. Lu, A. Ghoshal, and S. Renals, "Cross-lingual SGMMs for low resource speech recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 1, 2013.

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource LVCSR systems," in *Proc. Interspeech*, 2010.

[8] J. Pinto, M. Magimai.-Doss, and H. Bourlard, "Hierachical tandem features for ASR in Mandarin," in *Proc. Interspeech*, 2011.

[9] V.-B. Le, L. Lamel, and J.-L. Gauvain, "Multi-style MLP features for BN transcription," in *Proc. ICASSP*, 2010, pp. 4866–4869.

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

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

[12] A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006.

[13] O. Çetin, M. Magimai.-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and F. J., "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc. ASRU*, 2007.

[14] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in *Proc. Interspeech*, 2012.

[15] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. ICASSP*, 2012.

[16] S. Sivadas and H. Hermansky, "On use of task independent training data in tandem feature extraction," in *Proc. ICASSP*, 2004.

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

[18] 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.

[19] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. IEEE Workshop on Spoken Language Technology*, 2012.

[20] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," *Proc. ICASSP*, 2013.

[21] J.-T. Huang, J. Li, D. Yu, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013.

[22] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multi-lingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, 2013.

[23] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the use of a multilingual front-end," in *Proc. Interspeech*, 2008.

[24] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[25] S. Parveen and P. Green, "Multitask learning in connectionist robust ASR using recurrent neural networks," in *Proc. Interspeech*, 2003.

[26] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU*, 2011.

[27] Z. Tüske, R. Schlüter, and H. Ney, "Multi-lingual hierarchical MRASTA features for ASR," in *Proc. Interspeech*, 2013.

[28] 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.

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

[30] 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.

[31] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals, "A lecture transcription system combining neural network acoustic and language models," in *Proc. Interspeech*, Aug. 2013.

[32] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. IEEE Workshop on Spoken Language Technology*, 2012.

[33] J. Cettolo, Niehues, S. Stüker, L. Bentivogli, and M. Frederico, "Report on the 10th IWSLT evaluation campaign," in *Proc. International Workshop on Spoken Language Translation*, 2013.

[34] J. Dreisen, P. Bell, M. Sinclair, and S. Renals, "Description of the UEDIN system for German ASR," in *Proc. International Workshop on Spoken Language Translation*, 2013.

[35] T. Schultz, "GlobalPhone: A multilingual speech and text corpus developed at Karlsruhe University," in *Proc. Interspeech*, 2002.

[36] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, 2008.