

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Transcription of multi-genre media archives using out-of-domain data

Citation for published version:
Bell, PJ, Gales, MJF, Lanchantin, P, Liu, X, Long, Y, Renals, S, Swietojanski, P & Woodland, PC 2012, Transcription of multi-genre media archives using out-of-domain data. in *Spoken Language Technology Workshop (SLT), 2012 IEEE.* Institute of Electrical and Electronics Engineers (IEEE), pp. 324-329. https://doi.org/10.1109/SLT.2012.6424244

Digital Object Identifier (DOI):
10.1109/SLT.2012.6424244

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Spoken Language Technology Workshop (SLT), 2012 IEEE

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Download date: 29. Mar. 2024

Bell, P. J., Gales, M. J. F., Lanchantin, P., Liu, X., Long, Y., Renals, S., Swietojanski, P., & Woodland, P. C. (2012). Transcription of multi-genre media archives using out-of-domain data. In Spoken Language Technology Workshop (SLT), 2012 IEEE. (pp. 324-329). IEEE. doi: 10.1109/SLT.2012.6424244

TRANSCRIPTION OF MULTI-GENRE MEDIA ARCHIVES USING OUT-OF-DOMAIN DATA

PJ Bell[1]*, MJF Gales*[2]*, P Lanchantin*[2]*, X Liu*[2]*, Y Long*[2]*, S Renals*[1]*, P Swietojanski*[1]*, PC Woodland*[2]

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

{peter.bell,s.renals}@ed.ac.uk, p.swietojanski@sms.ed.ac.uk

(2) Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

{mjfg,pkl27,xl207,yl467,pcw}@eng.cam.ac.uk

ABSTRACT

We describe our work on developing a speech recognition system for multi-genre media archives. The high diversity of the data makes this a challenging recognition task, which may benefit from systems trained on a combination of in-domain and out-of-domain data. Working with tandem HMMs, we present Multi-level Adaptive Networks (MLAN), a novel technique for incorporating information from out-of-domain posterior features using deep neural networks. We show that it provides a substantial reduction in WER over other systems, with relative WER reductions of 15% over a PLP baseline, 9% over in-domain tandem features and 8% over the best out-of-domain tandem features.

Index Terms— speech recognition, tandem, cross-domain adaptation, media archives

1. INTRODUCTION

Technologies for delivering broadcast content online have matured considerably, but automatic transcription, metadata extraction, and indexing are still underdeveloped. The situation is particularly serious for archive material, for which human-generated metadata is often sparse or non-existent. Automatic processing of such archives—which in some cases are extremely large, dating back many decades—would "unlock" them, indexing historic content, and enabling search based on transcriptions, speaker identity, and other extracted metadata.

The British Broadcasting Corporation (BBC) has a stated aim to open its broadcast archive to the public by 2022. In a collaboration with BBC Research and Development, we have begun to investigate the automatic transcription of broadcast material across the full range of genres. This is still an immature research area. Although there has been a focus on the transcription of broadcast content since the mid-1990s, this has focused strongly on broadcast news and related content. Automatic transcription of arbitrary, multi-genre media content is a much more challenging task, since the material to recognise includes "on location" broadcasts in diverse environments (for instance, sport or documentary features) and drama with highly-emotional speech, overlaid background music and sound effects. It is also the case that the rights issues are much simpler for broadcast news, compared with content with substantial creative input, an important issue when considering resources for research and evaluation.

Recent work which has focused on the automatic transcription or indexing of multi-genre broadcast data has included work on the automatic transcription of podcasts and other web audio [1], automatic transcription of YouTube [2, 3], the MediaEval rich speech retrieval evaluation which used blip.tv semi-professional user created content [4], and the automatic tagging of a large radio archive [5].

This paper concerns the automatic transcription of multi-genre content from the BBC archive. We used a transcribed corpus containing 18 hours of talk-radio output (recorded from one station over a 24-hour period) and several episodes of a television drama (11 hours in total). The latter part of the corpus is rich with sound effects, emotional speech, and background music. We split the corpus into training and test sets, and it is described in more detail in Section 4. As discussed in Section 5.1, state-of-the-art transcription systems built for domains such as conversational telephone speech (CTS), meetings, and (North American) broadcast news (BN) perform with low accuracy on the multi-genre BBC data due to the high mismatch in environment, speaker, speech style, and accent, as well as the broadly challenging nature of the content. In-domain HMM-GMM (Gaussian mixture model - hidden Markov model) systems trained on this corpus outperform these out-of-domain (OOD) systems, despite the fact that there is an order of magnitude less in-domain training data (Section 5.2).

The focus of our work is development of methods which can effectively combine in-domain and OOD training data, using neural networks in the tandem framework [6], whereby context-dependent HMMs with GMM output distributions are trained on standard acoustic features concatenated with fea-

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology). Thanks to Andrew McParland and Sam Davies of BBC R&D.

tures derived from neural networks. We refer to systems using either posterior features derived from the network outputs, or using features derived from the hidden units (e.g. bottleneck features) as tandem. Our use of neural networks is motivated by prior work (discussed in Section 2) showing they are able to provide good cross-domain portability, in addition to their inherent strengths in phone discrimination and ability to model wide acoustic context.

Although neural networks have been used for some time in acoustic modelling both in the tandem framework and in hybrid systems (in which neural networks are trained to directly estimate posterior probabilities which can then be used as scaled likelihood estimates for HMM states), there has been intensive research recently on deep neural networks (DNNs), with extremely promising results [7, 8, 9]. We have used deep neural networks (DNNs) with generative pre-training to obtain posterior features used in the tandem framework. Using DNNs to produce discriminative features for GMMs in a tandem framework is attractive for cross-domain modelling, since it allows the GMM and DNN parameters to be adapted independently.

This paper presents a novel technique for posterior feature combination in a cross-domain setting, that we refer to as Multi-Level Adaptive Networks (MLAN) (Section 3). We have investigated this technique using the multi-genre broadcast corpus, in terms of cross-domain speech recognition using different acoustic training data sources across different target genres (Section 5). We evaluate the new technique in terms of a discriminatively-trained speaker-adaptive speech recognition system, comparing in-domain and out-of-domain posterior features with the features obtained using MLAN (Section 6).

2. CROSS-DOMAIN ADAPTATION WITH NEURAL NETWORKS

Posterior features derived from neural networks have been successfully used in cross-domain and cross-lingual settings [10, 11, 12, 13]. In [10], Sivadas and Hermansky obtained reductions in word error rate (WER) on a small-vocabulary task by retraining HMMs using neural network features trained entirely on task-independent data, whilst Stolcke et al [11] reported that neural networks trained on CTS data could be used directly in a meeting recognition task. Their results were supported by Le et al [12] who used CTS data and accented speech data to improve performance on a Broadcast News task.

There are a number of design choices when building systems using OOD posterior features. Typically, nets trained on OOD data are used to generate posterior features for in-domain data but, as investigated in [11], it is also possible to adapt the nets to the new domain by performing additional training iterations. Similarly, the HMMs used may simply be those trained on the OOD data; or may be adapted to the new domain using MAP adaptation [10, 11]; or retrained from scratch. Additionally, outputs from nets trained on different domains may be combined using a merger MLP, as in [12, 13]. Reductions in WER are reported from all these approaches.

3. METHODOLOGY

3.1. Neural network posterior features

We trained DNNs to model frame posterior probabilities over monophones. We used unsupervised restricted Boltzmann machine pretraining [14], which may be viewed as a form of regularizer, enabling DNNs to be robustly trained with limited quantities of training data [15]. The networks were fine-tuned (further discriminatively trained using stochastic gradient descent) to minimise the negative log-posterior probability of the true class labels, which were fixed by Viterbi alignment of the training data with a baseline HMM.

The structure of the DNNs was fixed following analysis of the frame error rate on held-out validation data. We finally used nets with four hidden layers, nine frames of acoustic context and 1024 units in each hidden layer. All DNN training was performed using GPU machines. To obtain posterior features, monophone log-posterior probabilities output from the nets were decorrelated using a single PCA transform, with dimensionality reduced to 30 [6]. These posterior features were concatenated with the original acoustic features.

For our experiments comparing posterior features obtained from different nets, we trained context-dependent HMMs with maximum likelihood (ML) training. Since discriminative training has been shown to yield additional benefits for tandem HMMs [16], our final system used HMMs trained with the minimum phone error (MPE) criterion [17].

3.2. Adaptation with multi-level networks

The use of an initial out-of-domain DNN, adapted to a new domain can be viewed as imposing a form of regularization on the resulting net. However, we have observed relatively small benefits from this approach when deep architectures are used and the quantity of in-domain data is fairly large, as the generative pre-training itself acts as a regularizer—so when the domain mismatch is high, purely in-domain tandem features may be more effective than OOD features. We therefore propose an alternative adaption procedure which we call Multi-Level Adaptive Networks (MLAN). In the first level of this scheme, networks trained on OOD acoustic data are used to process in-domain acoustic data to generate posterior features, which are concatenated with the original in-domain acoustic features, as in the tandem framework. We would expect the OOD posterior features to enhance the discriminative abilities of the simple in-domain acoustic features. In the second level, we train *additional* DNNs, using the first level

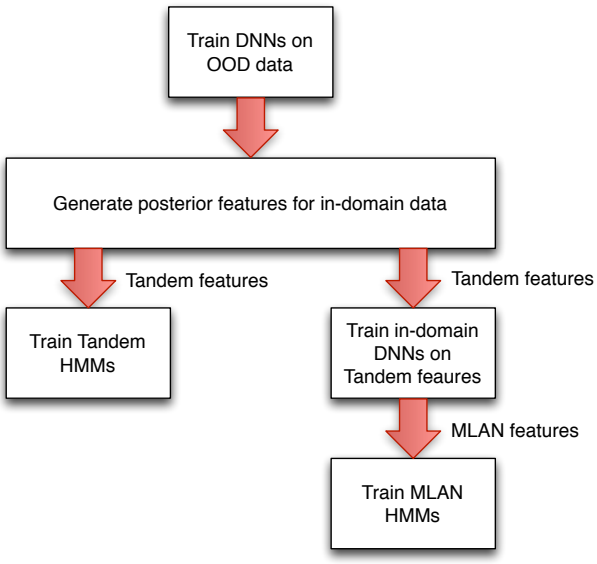


Fig. 1. Multi-Level Adaptive Network (MLAN) architecture

tandem features as input, to minimise an *in-domain* objective function of log-posterior phone probabilities. The outputs from these DNNs are used to generate the final tandem features for HMM training. By expanding the input tandem feature vector used at the second level, output from multiple networks, trained on different domains, may be included with no modification to the architecture. This scheme is similar to that recently published by Thomas et al [18], with the differences in that work being: (i) two-layer MLPs were used; (ii) the MLPs in both levels are trained on OOD data (and on data from a different language at the first level); (iii) a separate merger MLP was used to merge multiple sets of posterior features at the first level, whereas in our case the second level DNN can take multiple sets of first level features.

In all the experiments reported in this paper, HMMs are trained independently for each new feature set. The process is outlined in Figure 1. The motivation for the MLAN scheme is that the new DNNs, trained discriminatively, are able to learn which elements of the OOD posterior features are useful for discrimination in the new domain; whilst the direct inclusion of in-domain acoustic features in the input means that the resulting frame error rates ought never to be worse than DNNs trained purely in-domain. The additional generative pre-training carried out ensures than the new DNN does not over-fit to the in-domain data.

4. RECOGNITION OF MEDIA ARCHIVES

4.1. BBC dataset

The multi-genre broadcast corpus, kindly made available by BBC Research and Development, contains speech that is mostly British English, with a range of regional accents. It is composed of 2 datasets, with audio encoded in stereo MPEG 1.0 Audio Layer III with a bitrate of 128 kbit/s and sample frequency of 48kHz. The first comprises 36 talk-radio shows broadcast on the same radio channel over 24 hours in February 2009. Different genres are represented: news, weather reports, book readings, documentaries, panel games, debates, and dramas. The show durations range from 2 minutes (weather reports) to 3 hours (morning news / current affairs programme) to give a total duration of 18 hours. The transcriptions were done manually and included time-stamps (quantised to 1s), speaker names, and additional metadata such as indications of music or sound effects. The second comprises 14 episodes of a TV drama series broadcast in 2010. Episode durations range from 40–75 minutes, to give a total duration of 11 hours. The transcriptions were derived from subtitles for hearing impaired and included time-stamps and other metadata such as indications of music and sound effects, or indications of the way the text has been pronounced. Most of the shows include several speakers. Speaker identities were indicated by the use of four different text colours.

For both datasets, the audio content covers a broad range of genres, environment and speaking style. For purposes of analysis, we divided the data into three categories by broad genre:

- `studio`: speech in controlled, studio conditions, and news reports, sometimes including telephone speech from reporters or contributors;
- `location`: material produced on location: in this data, this includes parliamentary proceedings and a visit to a farm;
- `drama`: the TV drama series, containing dramatic, fast emotional speech, and high background noise levels, making ASR particularly challenging.

The time stamps were found to be unreliable due to quantisation effects in the talk-radio shows and to time lags that occur in the TV drama subtitles (presumably arising from the respeaking process for subtitle creation). We refined the transcriptions using a light supervision approach based on decoding with a biased language model [19]. Each show was first segmented and clustered by speaker using the CU RT-04 diarisation system [20]. Each speech segment was decoded in two passes using speaker adaptation, with the decoding employing a biased language model trained on the raw transcription. The decoder output was compared with the raw transcription to identify matching sequences. Non-matching word sequences from the raw transcription were force aligned to the remaining speech segments. Once realigned, the position of time stamps in the transcription could be corrected. After refinement, there was around 23 hours of transcribed and aligned speech in total (from an initial 29 hours of raw audio). We divided the data at the show level into a training set of 20.7 hours (`bbc.train`) and a test set of 2.3 hours (`bbc.eval`), each containing roughly the same balance across genres. For the drama series, some speakers ap-

Domain	Training data (hrs)
AMI	126.8
CTS	276.0
BBC (in domain)	20.8

Table 1. Amount of in-domain and OOD training data

System	Studio	Location	Drama	All
AMI	28.4	50.1	76.6	51.8
CTS	43.4	66.2	83.2	64.1

Table 2. Development system results (WER/%) on `bbc.eval` for unadapted OOD systems (HMM-GMM trained on PLP coefficients)

peared in both the training and test set.

4.2. Out-of-domain data

For the cross-domain experiments, two diverse sets of out-of-domain data were used. Firstly, we used 277 hours of US-English conversational telephone speech (CTS) taken from the Switchboard I, Switchboard II and CallHome corpora. Secondly, we used recordings of multi-party meetings from the AMI corpus (AMI), using the training setup described in [21]. Recordings made with multiple distant microphones are combined into a single channel using a beamformer and overlapping speech was removed. Table 1 compares the quantities of data available from each domain.

4.3. System descriptions

Development experiments were performed using a simple one-pass system which allowed rapid experimental turnaround for investigation of the different feature sets. Cross-word triphone HMM-GMM acoustic models were trained on the 20.7 hours of BBC data, `bbc.train`, using ML estimation. Models were tied with phonetic decision trees to give a total of about 3,000 tied states with an average of 16 mixture components per state.

The baseline acoustic features were perceptual linear prediction (PLP) coefficients with first, second and third temporal derivatives, projected to 39 dimensions with an HLDA transform. Posterior features were generally computed from these same projected features, except in the case of the AMI posteriors, where a stacked bottleneck architecture with a filterbank input was used [21]. In all experiments, the posterior features were projected to 30 dimensions[1], giving a total augmented feature vector dimension of 69. Feature vectors were normalised for mean and variance at the speaker level. To compute the OOD posterior features the BBC data was coded

[1]The AMI bottleneck features were obtained from a 30-dimension bottleneck and no further projection was performed.

Feature set	Studio	Location	Drama	All
PLP	17.0	32.5	67.3	39.4
BBC tandem	14.4	27.5	59.2	34.1
AMI tandem	14.3	26.6	59.2	33.8
CTS tandem	14.3	28.6	62.3	35.5

Table 3. Development system results (WER/%) on `bbc.eval` without posterior features (PLP) and using in-domain (BBC) and OOD (AMI, CTS) tandem features

Feature set	Studio	Location	Drama	All
AMI MLAN	13.5	25.0	56.1	32.0
CTS MLAN	12.5	25.5	56.6	31.9
AMI+CTS MLAN	12.5	24.3	54.9	**31.0**

Table 4. Development system results (WER/%) on `bbc.eval` using MLAN features

and normalised to match the data used to train the nets, which in the case of CTS included downsampling to 8khz.

A one-pass decoder architecture was used. A 50,000 word vocabulary trigram language model [21] was linearly interpolated with a model trained on 0.23M words of in-domain BBC data and the vocabulary supplemented to 52,000 words.

The final evaluation system differs primarily from the system used in the development experiments through the use of MPE discriminative training [17], and a two-pass decoding architecture, in which a first pass generated initial transcriptions using unadapted models, and the second pass used speaker adapted models to generate the final transcription. The other differences were the use of a global block diagonal (39x39 and 30x30) semi-tied covariance [22] transform to further remove correlation in the tandem feature space, an average of 12 mixture components per state, a 61,000 word vocabulary, and a language model which interpolated component trigram models estimated from 0.23M words of in-domain BBC data (weight 0.55) and 1,400M words of North American Broadcast News data (weight 0.45).

5. DEVELOPMENT EXPERIMENTS

5.1. Baseline unadapted systems

As a preliminary experiment, we performed recognition of the `bbc.eval` data using two out-of-domain acoustic models trained on PLP features from the AMI and CTS training sets. BBC PLP features were mean and variance normalised on a per-speaker basis to match each OOD training set, and in the case of CTS, the data was downsampled to 8khz.

The results, shown in Table 2, clearly demonstrate the large acoustic mismatch between these domains and the BBC domain, with the CTS models performing particularly poorly. We briefly experimented with the use of MAP adaptation of these models to the BBC training data, but did not improve

Feature set	1-pass (unadapted)				2-pass (adapted)			
	Studio	Location	Drama	All	Studio	Location	Drama	All
PLP	12.0	25.9	58.8	32.7	11.5	23.6	58.9	31.8
BBC tandem	11.7	23.3	54.9	30.4	11.3	22.3	54.4	29.8
AMI tandem	11.3	22.6	55.0	30.1	11.1	21.5	54.2	29.4
AMI+CTS MLAN	10.2	20.9	50.5	27.6	9.8	20.0	50.2	27.1

Table 5. Final MPE system results (WER/%) on `bbc.eval` using PLP, tandem, and MLAN features.

upon the results of HMMs trained on purely in-domain data.

5.2. Cross-domain tandem systems

We next investigated the performance of tandem features. Table 3 compares models trained purely on the `bbc.train` dataset with models trained on tandem features obtained using OOD nets. Firstly, it may be noted that even a simple system trained on in-domain data outperforms any of the OOD systems shown in Section 5.1, despite the much smaller quantities of training data available, again illustrating the large domain mismatch. Secondly, even when trained solely on `bbc.train`, DNN tandem features are shown to give large gains over standard PLP features in this setup.

Comparing the out-of-domain tandem features from AMI and CTS, with the simple PLP results, it is seen that both out-of-domain posterior features improved performance for all genres, with the overall WER reduced by 5.6% absolute and 3.9% absolute using AMI and CTS features respectively. This supports earlier work suggesting that posterior features are portable across domains. Comparing with the BBC tandem results note that the well-trained OOD posteriors are often better than equivalent in-domain posteriors: for example, CTS and AMI are both better for Studio speech; AMI is best for Location speech; whilst the AMI and in-domain features are equally matched for Drama, the genre most mismatched to the out-of-domain acoustic models.

5.3. MLAN systems

Table 4 shows the performance of the MLAN technique. MLAN provides substantial additional gains over standard tandem features, for both domains. The CTS posteriors, which are seen in Tables 2 and 3 to be worst-matched to the BBC domain, gain the most benefit from MLAN with a 3.6% absolute WER reduction overall. Finally, the combination of both OOD posterior features with MLAN reduces WER still further, suggesting the second-level DNN is successfully able to exploit complementary information between AMI and CTS. On CTS, we compared the use of MLAN with a shallow softmax adaptation to the BBC data, which gave a significantly higher overall WER of 34.3%. Other baselines will be explored in future work.

6. FINAL SYSTEM EVALUATION

Based on the results of Section 5, we selected the best-performing in-domain and out-of-domain tandem features, and the best MLAN features, for use in training a more competitive final system. Table 5 shows the final system results on `bbc.eval` with and without speaker adaptation. The HMMs were trained with MPE only on `bbc.train` using STC-projected PLP features and the relevant posterior features.

It can be seen that all of the new features outperform the baseline PLP features in both the unadapted and speaker-adapted MPE systems. This supports the findings from the development system, and indicates that the posterior features can bring complementary information to the PLP features even when the HMMs are discriminatively trained.

As we found in the earlier results, the out-of-domain tandem features trained on AMI perform better on both the Studio and Location subsets than the tandem features trained only on the small amount of in-domain BBC data, but have almost an equal acoustic and linguistic match to the Drama genre. When the best MLAN features, additionally incorporating the CTS posteriors, are used, the performance improves still further. Overall, the improvement over the baseline PLP features, in both the unadapted speaker-adapted systems, is dramatic, with absolute WER reductions of 5.1% and 4.7% respectively.

Table 5 indicates that speaker adaptation is effective in reducing the WER for all three posterior feature sets, compared with the baseline PLP features which only offers gains for the Location and Studio subsets, although for these two subsets, the gains from adaptation are larger than for the posterior features. We hypothesise that the posterior features are better able to capture speaker-invariant information in the these subsets, whilst in the noisy Drama subset, are able to model speaker-dependent structures more effectively than PLPs.

7. CONCLUSIONS

We have presented a method for recognition of multi-genre media archives with neural network posterior features, successfully using out-of-domain data to improve performance. Our results consistently show that our Multi-Level Adaptive Networks scheme results in substantial gains over both in-

domain or out-of-domain nets used in a tandem setup, with relative WER reductions of 9% and 8% respectively on our final system – and a 15% relative reduction over the PLP baseline.

In future work we will investigate the technique in an HMM-GMM system that also incorporates speaker-adaptive training and fMPE transforms. We will also adapt the method for use in a hybrid DNN system.

8. REFERENCES

[1] J. Ogata and M. Goto, "PodCastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription," in *Proc. Interspeech*, 2009.

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proc. ICASSP*, 2009, pp. 4873 – 4876.

[3] R. C. van Dalen, J. Yang, and M. J. F. Gales, "Generative kernels and score-spaces for classification of speech: Progress report," Tech. Rep. cued/f-infeng/tr.676, Cambridge University Engineering Department, 2012.

[4] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. J. F. Jones, "Overview of MediaEval 2011 rich speech retrieval task and genre tagging task," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, 2011.

[5] Y. Raimond, C. Lowis, R. Hodgson, and J. Tweed, "Automated semantic tagging of speech audio," in *Proc. WWW 2012*, 2012.

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

[7] G.E. 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] A. Mohammed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[9] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. ICASSP*, 2012.

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

[11] 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.

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

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

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

[15] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, and P. Vincent, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[16] J. Zheng, O. Cetin, M.-Y. Hwang, X. Lei, A. Stolcke, and N. Morgan, "Combining discriminative feature, transform, and model training for large vocabulary speech recognition," in *Proc. ICASSP*, 2007.

[17] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*. IEEE, 2002, vol. I, pp. 105–108.

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

[19] N. Braunschweiler, M.J.F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, 2010, pp. 2222–2225.

[20] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland, "The development of the Cambridge University RT-04 diarisation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, nov 2004.

[21] T. Hain, L. Burget, J. Dines, P.N. Garner, F. Grezl, A.E. 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.

[22] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 272–281, May 1999.