INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Comparative Evaluation of GMM-Free State Tying Methods for ASR

Tamás Grósz[1], *Gábor Gosztolya*[1,2], *László Tóth*[2]

[1]University of Szeged, Institute of Informatics, Szeged, Hungary
[2]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
{ groszt, ggabor, tothl } @ inf.u-szeged.hu

Abstract

Deep neural network (DNN) based speech recognizers have recently replaced Gaussian mixture (GMM) based systems as the state-of-the-art. While some of the modeling techniques developed for the GMM based framework may directly be applied to HMM/DNN systems, others may be inappropriate. One such example is the creation of context-dependent tied states, for which an efficient decision tree state tying method exists. The tied states used to train DNNs are usually obtained using the same tying algorithm, even though it is based on likelihoods of Gaussians, hence it is more appropriate for HMM/GMMs. Recently, however, several refinements have been published which seek to adapt the state tying algorithm to the HMM/DNN hybrid architecture. Unfortunately, these studies reported results on different (and sometimes very small) datasets, which does not allow their direct comparison. Here, we tested four of these methods on the same LVCSR task, and compared their performance under the same circumstances. We found that, besides changing the input of the context-dependent state tying algorithm, it is worth adjusting the tying criterion as well. The methods which utilized a decision criterion designed directly for neural networks consistently, and significantly, outperformed those which employed the standard Gaussian-based algorithm.

Index Terms: deep neural networks, context-dependent state tying, Kullback-Leibler divergence, entropy

1. Introduction

Deep neural network (DNN) based hybrid speech recognizers are nowadays regarded as the state-of-the-art and have replaced conventional Gaussian mixture modeling (GMM) based hidden Markov models (HMMs). Since the introduction of HMMs, the speech community have developed many methods to optimize the process of the training of GMM-based acoustic models. HMM/DNN hybrid systems have inherited most of these techniques, even though some of these may be suboptimal for them. Two such examples are the flat start training scheme and the creation of context-dependent (CD) phone models, which are vital components of standard HMM/GMM systems.

Conventionally, HMM/GMM systems are trained by an iterative re-estimation and re-alignment of the models, known as ”flat start” training. This procedure is quite straightforward for HMM/GMMs, but performing the same for HMM/DNNs is not so obvious. In the past few years, however, it was shown by several researchers that, if done with proper caution, flat start training can also be performed with DNNs [1, 2, 3, 4].

While hybrid models applied only context-independent (CI) phone models for a long time [5], there is now common agreement that HMM/DNN systems also greatly benefit from using context-dependent tied states [6, 7]. Thus, it is necessary to find an approach for efficiently creating context-dependent tied states for systems built on DNNs. Currently, the dominant solution is the decision tree-based state tying method of Young et al. [8]. This technique fits Gaussians on the distribution of the states, and uses the likelihood gain to govern a decision tree-based state-splitting process. Thanks to the Gaussian assumption and the decision tree representation, this approach is computationally very efficient. However, as we have already mentioned, it may be inappropriate to just impose the common HMM/GMM-based techniques on the HMM/DNN training procedure, and this may hold for this state tying algorithm as well.

GMM-based methods assume that the Gaussian components have diagonal covariance matrices, and thus require decorrelated features like cepstral coefficients (MFCCs). However, HMM/DNN hybrids tend to work better on more primitive features like mel filter bank energies [9]. Since conventional HMM/GMM systems cannot be efficiently trained on these features, the usual approach is to build a HMM/GMM system on a standard feature set like MFCCs, create the tied state inventory and alignment, and then throw away the feature set and the whole model. Furthermore, intuitively, the state clustering algorithm should split those states where the splitting would be beneficial for the respective classifier. Since the objective functions during GMM and DNN training are different, measuring how a Gaussian fits a given class may be unrelated to the difficulty of modeling that class by a DNN. This suggests that if we perform the CD state tying by following the standard approach, we do it on a mismatched feature set and using a mismatched similarity metric.

Quite recently, a number of articles were published about CD state tying for HMM/DNNs. The issue of the ‘inappropriate feature set’ can be handled by performing the state clustering process on the output of a DNN instead of the raw features. This idea was investigated in a couple of studies (e.g. [1, 2, 10, 11]). In those studies, however, only the input of the clustering algorithm was modified, while the whole state tying algorithm remained intact. Other studies proposed novel decision criteria for the standard state tying method, which suit neural networks better. Gosztolya et al. [12] proposed applying the Kullback-Leibler divergence-based decision criterion originally developed for KL-HMMs by Imseng et al. [13]. Zhu et al. [14] constructed a criterion that relied on entropy. Lastly, Wang et al. [15] trained a special network that optimized for Deep Canonical Correlation Analysis, and clustered the output of this network via k-means.

All these studies experienced a drop in the word error rate (WER) compared to the baseline that uses the standard Gaussian likelihood-based state tying method with the MFCC vectors. However, none of these studies compared its results with other neural network-based state tying approaches, which makes these methods quite hard to compare. Furthermore, the datasets used differed to a huge extent as well: Gosztolya et al. used a Hungarian database, Zhu et al. used a German one, while Wang et al. used the quite small TIMIT corpus, where

http://dx.doi.org/10.21437/Interspeech.2017-899

only phoneme error rates can be reported. In this study we compare four such approaches on the same LVCSR task, where the same context-independent neural network will provide the input vectors for the state clustering. Note that, since we obtain the frame-level CI labels by the MMI DNN flat-start method proposed in [3], the CI models have no inherent GMM dependency. Therefore those state tying methods which have a decision criterion designed for DNNs (i.e. the ones proposed by Gosztolya et al. [12] and Zhu et al. [14]) are completely GMM-free.

2. Decision Tree-Based State Tying

The decision tree-based state tying algorithm was introduced by Young et al. [8], and evolved into a vital component of training large vocabulary speech recognizers. The main idea is to pool all context variants of a state, and then build a decision tree by successively splitting this set into two. For each step, the algorithm chooses one of the pre-defined questions in such a way that the resulting two non-overlapping subsets of the original state set $\mathcal{S}$ differ maximally. The algorithm measures this difference by using a likelihood-based decision criterion. Although minor improvements to the algorithm like the automatic generation of the questions via clustering were proposed [16], the main scheme of the method proved to be so successful that it has remained unaltered ever since.

2.1. Likelihood based decision criterion

Suppose that we have a set of states $\mathcal{S}$ that we need to tie, using the decision tree-based method of Young et al. At each node, we have a set of questions, and each question can split $\mathcal{S}$ into two non-overlapping sub-sets depending on the answer to the question. Odell formulated a maximum likelihood-based decision criterion [17] and proposed a computationally efficient algorithm by approximating the splitting criterion as

$$L(\mathcal{S}) \simeq -\frac{1}{2}\big(\log[(2\pi)^K|\Sigma(\mathcal{S})|] + K\big)\sum_{s\in\mathcal{S}} N(s), \qquad (1)$$

where $s \in \mathcal{S}$ are the individual states, $\Sigma(\mathcal{S})$ is the variance of data in $\mathcal{S}$, and $N(s)$ is the number of examples (frames) in the training data which belong to state s. Using this formula, we should choose the question q which maximizes the likelihood difference $\Delta L(q|\mathcal{S})$

$$\Delta L(q|\mathcal{S}) = \big(L(\mathcal{S}_y(q)) + L(\mathcal{S}_n(q))\big) - L(\mathcal{S}), \qquad (2)$$

where $\mathcal{S}_y(q)$ and $\mathcal{S}_n(q)$ are the two subsets of $\mathcal{S}$ formed based on the answer to the question q. Notice that the likelihood values do not depend on the training observations themselves, but only on the variance over training samples corresponding to the states, and the total number of frames belonging to each state.

3. Neural Network-Based CD State Tying

Using the state tying method of Young et al. with the decision criterion of Odell, we rely on the variance of the feature vectors assigned to each state. While this assumption fits well with a system employing GMMs, using the above procedure to create tied states for a CD HMM/DNN system is flawed in two distinct ways, because we perform state tying using the wrong feature set and the wrong state similarity criterion.

A solution proposed by several studies is that we first train a context-independent neural network (either DNN or a shallow ANN). This CI NN (referred to as the *auxiliary* neural network from now on) can be trained using the frame-level targets provided by a HMM/GMM system, but we can also utilize some of the DNN flat-start techniques introduced recently (e.g. [2, 3, 4]). Then, for the next step, we create the CD states based on this auxiliary neural network. This approach has the advantage that this network can be trained on the same feature set that is used for the final acoustic model training. However, the optimal way of performing state clustering on the output of this CI DNN is not clear. Next, we will describe four such approaches.

3.1. Clustering the CI DNN output

This approach, proposed by Senior et al. [10], is quite straightforward. They simply use the frame-level outputs of the auxiliary neural network as input for the state tying procedure. The whole clustering process remains the same in every other respect. Senior et al. reported a slight improvement in the WER, and, naturally, with this approach they were able to avoid the feature set mismatch among CD DNN training and the CD state tying process. However, since they used the original state tying method of Odell [17], which relies of likelihoods of Gaussians, in our opinion their method can hardly be regarded as completely GMM-free.

3.2. Clustering the DNN hidden activations

In a parallel study Bacchiani and Rybach [11] proposed performing the clustering on the activations of the last hidden layer of the auxiliary CI NN. Although one cannot expect the activation vectors to be decorrelated (or to follow any predefined distribution), Bacchiani and Rybach were able to use them as inputs for the CD state tying method of Young et al. The WERs they got were reported to be lower for smaller CD state sizes than by using the standard approach, but for larger state counts it was the other way around. They explained this by recalling that the frame-level CI labels were obtained by HMM/GMMs, hence there was a mismatch in the frame-level targets.

3.3. Kullback-Leibler divergence based decision criterion

The DNN-based CD state tying approaches described so far all leave the decision criterion intact. The first study published which utilized a criterion designed for DNN outputs in a standard HMM/DNN framework was that of Gosztolya et al. [12], who used the criterion of Imseng et al., originally developed for their Kullback-Leibler HMM framework [13]. Since in the study of Gosztolya et al. neither the feature set nor the decision criterion relied on Gaussians, this was the first study in which the state tying procedure was in fact completely GMM-free. Next, we give a brief description of this algorithm, based on articles [12], [18] and [19].

The Kullback-Leibler divergence between two posterior vectors (the observed (z_t) and the state prototype (y_s)) is defined as

$$D_{KL}(y_s||z_t) = \sum_{k=1}^{K} y_s(k)\log\frac{y_s(k)}{z_t(k)}, \qquad (3)$$

where $k \in \{1, \ldots, K\}$ is the dimensionality index of the posterior distribution vector [20]. Instead of maximizing the likelihood, we will minimize the KL-divergence

$$D_{KL}(\mathcal{S}) = \sum_{s\in S}\sum_{f\in F(s)}\sum_{k=1}^{K} y_{\mathcal{S}}(k)\log\frac{y_{\mathcal{S}}(k)}{z_f(k)}, \qquad (4)$$

where $\mathcal{S}$ is a set of states s, and $F(s)$ is the set of training vectors corresponding to state s. The posterior vector associated with the set $\mathcal{S}$ (i.e. $y_{\mathcal{S}}$) can be calculated as the element-wise geometrical mean of the example vectors belonging to the elements of $\mathcal{S}$, normalized to add up to one, but Imseng noted that the arithmetic mean can also be used [21]. After expanding and simplifying, we get

$$D_{KL}(\mathcal{S}) = -\sum_{s \in \mathcal{S}} N(s) \log \sum_{k=1}^{K} y_{\mathcal{S}}(k), \tag{5}$$

so the KL divergence of a set of states $\mathcal{S}$ can be calculated based on the statistics y_s and $N(s)$ of the individual states [18]. For the splitting of a set of states $\mathcal{S}$, the straightforward option is to choose the question that maximizes the KL-divergence difference $\Delta D_{KL}(q|\mathcal{S})$:

$$\Delta D_{KL}(q|\mathcal{S}) = D_{KL}(\mathcal{S}) - \big(D_{KL}(\mathcal{S}_y(q)) + D_{KL}(\mathcal{S}_n(q))\big). \tag{6}$$

3.4. Entropy-based decision criterion

The fourth approach we tested was proposed by Zhu et al. [14]. They also replaced the decision criterion of Eq. (1) with another formula that has no implicit GMM dependency. The key idea was to measure the inter-similarity of each merged cluster by calculating the entropy of the examples belonging to it. The entropy of a K-dimensional probability distribution can be calculated as

$$H(p) = \sum_{i=1}^{K} p(i) \log p(i). \tag{7}$$

The probability distributions associated with each initial state (i.e. the y_s vectors) were estimated via the mean of the DNN outputs for all the frames associated with a given state. Then, for a set of states $\mathcal{S}$, the prototype probability vector ($y_{\mathcal{S}}$) was calculated as the arithmetic mean of the prototype (y_s) of the member states, weighted by the number of state occurrences ($N(s)$); from these values, the decision criterion used during state tying can be calculated by using the entropy function, i.e.

$$D_E(\mathcal{S}) = \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log y_{\mathcal{S}}(k). \tag{8}$$

4. DNN Flat Start

CD state tying requires aligned frame-level CI labels. Furthermore, all of the CD state tying methods described in Section 3 require a trained neural network that provides the probability estimates of CI outputs. The alignments can be got by training a HMM/GMM in flat start mode, and the alignments provided by this model can be used to train a context-independent DNN; this solution, however, besides being quite clumsy, has other weaknesses as well. For example, Bacchiani and Rybach [11] experienced that a HMM/GMM system may produce suboptimal alignments for a HMM/DNN framework.

For these reasons, we opted for a more elegant solution, which relies only on neural networks. We trained a CI DNN in flat-start mode, following the adaptation of the Maximum Mutual Information (MMI) algorithm published in [3]. In order to speed up training and increase robustness, Gosztolya et al. proposed a number of small modifications over the original MMI training scheme. First of all, frame-level training targets were determined by using a forward-backward search instead of using the crude binary targets. Secondly, decoding was not done at the word level (which is the common approach in DNN-based sequence-discriminative training; see e.g. [22, 23]), but at the phoneme level instead. As a further simplification, the alignment process was performed by omitting state priors and the language model as well. This allowed the online utterance-by-utterance re-alignment of the phonetic transcriptions, which led to a quicker convergence of the training process. With these small modifications Gosztolya et al. were able to not only speed up DNN training significantly compared with other DNN flat-start methods used such as that of Zhang et al. [2], but they also reported an improvement in the word error rate with the final, context-dependent HMM/DNN system.

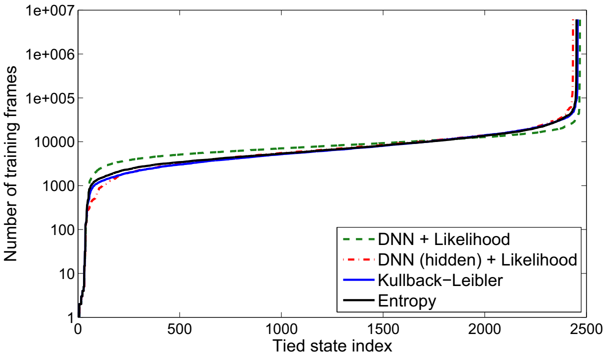

Figure 1: *The number of training frames for the different state tying methods for the case of cca. 2400 CD states.*

5. Experimental Setup

The DNN acoustic models were trained on the 81-hour long Wall Street Journal (WSJ) read speech corpus [24] (specifically the `si-284` set). The recognizers were evaluated on the `eval92` and `eval93` test sets in the “open-vocabulary” (60K word vocabulary) test condition, using a pruned version of the standard trigram language model. We used the `eval93` set as our development set; i.e. we tuned the language model weight and the insertion penalty on it, and also chose the number of tied states for each state tying method based on the WER achieved on this set. Then, in the last step, we evaluated the models using the optimal meta-parameters on the `eval92` set as the test set.

We used our custom DNN implementation, which achieved the best accuracy published so far on the TIMIT database with a phonetic error rate of 16.5% on the core test set [25]. In the actual tests, 40 mel filter bank energies were used along with their first and second order derivatives as input features. The DNNs were trained on 15 neighbouring feature vectors. Both the auxiliary CI and the final acoustic DNNs had five hidden layers, each containing 1000 rectified linear neurons [26, 27], and the softmax function was employed in the output layer. Decoding and evaluation was performed by a modified version of HTK [28]. For all four clustering algorithms, we varied the state tying threshold to get roughly 1800, 2400, 3000 and 3600 tied states.

Fig. 1 shows the distribution of the training classes for the case of roughly 2400 CD states. Besides noticing that the distribution produced by the different state tying methods is quite similar, we should also note that using the original decision criterion with the DNN outputs as input (proposed by Senior et al.) resulted in the best balanced class distribution of tied states.

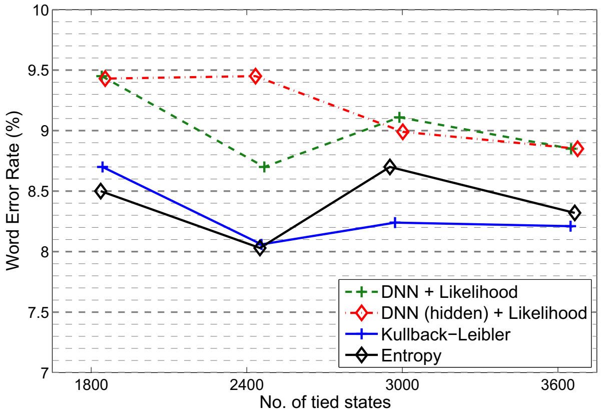

Figure 2: *WER for the different state tying approaches on the development set.*

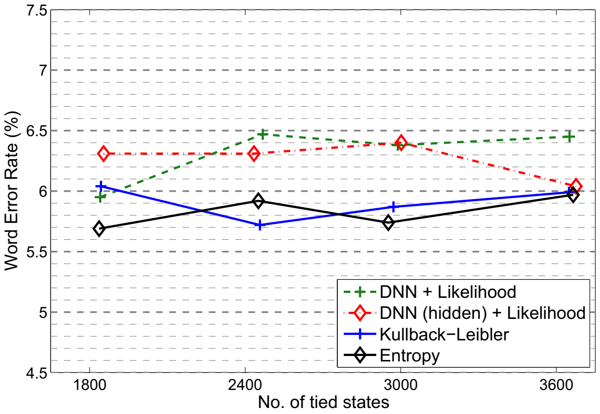

Figure 3: *WER for the different state tying approaches on the test set.*

Table 1: *WER values on the development and test sets got by using the different CD state tying methods*

Input layer	State tying decision criterion	WER (%) Dev.	WER (%) Test
Output	Likelihood (Senior, [10])	8.70%	6.47%
Hidden	Likelihood (Bacchiani, [11])	8.85%	6.04%
Output	KL-divergence (Gosztolya, [12])	8.06%	5.72%
	Entropy (Zhu, [14])	8.03%	5.92%

6. Results and Discussion

Table 1 sums up the best WER scores got on the development set and the corresponding WER values obtained on the test set for the four CD state tying approaches tested. It can be clearly seen that the most basic approach worked the worst of all: using the CI DNN outputs with the standard state tying decision criterion led to an 8.7% WER on the development set and a 6.47% WER on the `eval92` set, used as our test set. Although using the outputs of the last hidden layer with the original state tying method (proposed by Bacchiani and Rybach) led to slightly worse scores on the development set, it significantly outperformed the first approach on the test set. This approach is also justified by the fact that the activation vectors of a DNN are widely used as features in several tasks such as speaker identification [29] and various image processing applications [30].

The remaining two methods utilized some novel decision criteria instead of the Gaussian-based, standard one, and this fact is clearly reflected in their performance. On the development set they achieved practically identical WER scores (8.06% vs. 8.03% for the Kullback-Leibler and the entropy-based decision criteria, respectively); they differed somewhat on the test set, but the difference is statistically not significant. Overall, by relying on the Kullback-Leibler based decision criterion the WER scores were reduced by 0.8% compared to the basic approach of Senior et al., meaning a 12% improvement in terms of relative error reduction.

Figures 2 and 3 show the WER scores attained as a function of the number of CD states for the four state tying approaches tested. We can observe that the two solutions that used the original state tying algorithm, and the two which utilized a decision criterion designed for DNN outputs, are clearly separated, with the latter group producing consistently lower WER scores for both sets regardless of the number of tied states. (Notice that on the development set the highest WER is around 9.5%, while it is 6.5% for the test set, so the same relative WER improvement corresponds to a smaller absolute improvement for the latter set.) This, in our opinion, supports our hypothesis that besides changing the input of the CD state tying algorithm, its behaviour should also be adjusted to better suit DNNs, and so achieve an optimal performance.

Increasing the number of CD states helps those approaches which use the original, likelihood-based criterion; for the other two methods, however, optimality is achieved by having about 2400 states. On the test set, all four approaches seem to be quite insensitive to the number of tied states. Note that these inventory sizes appear to be smaller than those commonly used on the WSJ corpus, which, due to the lower computational requirements, is an improvement by itself.

7. Conclusions

In this study we compared the performance of four state clustering approaches proposed earlier to create context-dependent tied states for DNN acoustic models. What was common in the four approaches is that they utilized the output of a context-independent neural network as their input. We found that besides this, replacing the decision criterion used during state clustering is also beneficial: the two methods that employed this step consistently and significantly outperformed the two variants that used the original formula on the Wall Street Journal corpus, often used as reference. As we employed the MMI DNN flat-start method proposed by Gosztolya et al., in our tests neither the frame-level training targets nor the feature set used during CD state tying had any implicit GMM dependency. This means that, in our tests, the models that used the Kullback-Leibler divergence-based and the entropy-based decision criteria were 100% GMM-free.

8. Acknowledgements

The TitanX video card used for this study was donated by the NVidia Corporation. Tamás Grósz was supported by the ÚNKP-16-3 New National Excellence Programme of the Ministry of Human Capacities.

9. References

[1] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "GMM-free DNN training," in *Proceedings of ICASSP*, 2014.

[2] C. Zhang and P. Woodland, "Standalone training of context-dependent Deep Neural Network acoustic models," in *Proceedings of ICASSP*, 2014, pp. 5597–5601.

[3] G. Gosztolya, T. Grósz, and L. Tóth, "GMM-free flat start sequence-discriminative DNN training," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 3409–3413.

[4] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2751–2755.

[5] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic, 1994.

[6] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DNN-HMMs for real-world speech recognition," in *Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained Deep Neural Networks for large vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[8] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of HLT*, 1994, pp. 307–312.

[9] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using Deep Belief Networks," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 14–22, 2012.

[10] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "GMM-free DNN acoustic model training," in *Proceedings of ICASSP*, 2014, pp. 5639–5643.

[11] M. Bacchiani and D. Rybach, "Context dependent state tying for speech recognition using deep neural network acoustic models," in *Proceedings of ICASSP*, 2014, pp. 230–234.

[12] G. Gosztolya, T. Grósz, L. Tóth, and D. Imseng, "Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying," in *Proceedings of ICASSP*, Brisbane, Australia, Apr 2015, pp. 4570–4574.

[13] M. Razavi, R. Rasipuram, and M. Magimai-Doss, "On modeling context-dependent clustered states: Comparing HMM/GMM, hybrid HMM/ANN and KL-HMM approaches," in *Proceedings of ICASSP*, 2014.

[14] L. Zhu, K. Kilgour, S. Stüker, and A. Waibel, "Gaussian free cluster tree construction using Deep Neural Network," in *Proceedings of Interspeech*, Dresden, Germany, Sep 2015, pp. 3254–3258.

[15] W. Wang, H. Tang, and K. Livescu, "Triphone state-tying via Deep Canonical Correlation Analysis," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 3444–3448.

[16] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," in *Proceedings of ICASSP*, 1998, pp. 805–808.

[17] J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 1995.

[18] D. Imseng and J. Dines, "Decision tree clustering for KL-HMM," Idiap Research Institute, Tech. Rep. Idiap-Com-01-2012, 2012.

[19] D. Imseng, J. Dines, P. Motlicek, P. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proceedings of Interspeech*, 2012.

[20] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[21] D. Imseng, "Multilingual speech recognition A posterior based approach," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2013.

[22] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proceedings of ICASSP*, 2009, pp. 3761–3764.

[23] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech*, 2013, pp. 2345–2349.

[24] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of HLT*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 357–362.

[25] L. Tóth, "Phone recognition with hierarchical Convolutional Deep Maxout Networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 25, pp. 1–13, 2015.

[26] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of AISTATS*, 2011, pp. 315–323.

[27] L. Tóth, "Phone recognition with deep sparse rectifier neural networks," in *Proceedings of ICASSP*, 2013, pp. 6985–6989.

[28] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[29] P. Matějka, O. Glembek, O. Novotný, O. Plchot, F. Grézl, L. Burget, and J. H. Černocký, "Analysis of DNN approaches to speaker identification," in *Proceedings of ICASSP*, 2016, pp. 5100–5104.

[30] Y. Bar, N. Levy, and W. L., "Classification of artistic styles using binarized features derived from a Deep Neural Network," in *Proceedings of ECCV*, 2015, pp. 71–84.