# Recent Improvements to Neural Network based Acoustic Modeling in the EML Transcription Platform

Volker Fischer

*EML European Media Laboratory GmbH, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany*
*Email: fischer@eml.org*

## Introduction

In recent years, automatic speech recognition has enjoyed tremendous improvements from the use of (deep) neural networks (DNNs) for both acoustic modeling and stochastic language modeling [1, 2]. Powerful hardware, in particular graphics processing units (GPUs), and sophisticated training algorithms enable the use of deeper and deeper networks that reduce word error rates achieved with conventional Gaussian Mixture Models (GMM/HMM) by up to 30 percent [3]. However, comparisons of latency and real time factor for conventional and DNN based speech recognizers are only seldom published.

Acoustic modeling for the *EML Transcription Platform* [4, 5] has also adopted the deep learning paradigm in order to deliver improved accuracy for commercial applications ranging from server based media or contact center analytics to real time command and control tasks running on onboard units in cars. At the core of the platform is a state of the art large vocabulary continuous speech recognizer [6], which has been highly optimized for this purpose without giving up the flexibility of a research decoder.

Both tight real time conditions and the limited availability of GPUs in many of our target scenarios make the use of (very) deep neural networks prohibitive and let us seek for improvements in acoustic modeling with rather shallow networks. In the remainder of this paper we first give an overview of our baseline approach to the problem. After discussing some practical aspects of TANDEM and DNN-HMM hybrid models we describe some recently introduced modifications to the training procedure, and finally provide some evaluation on a publicly available database.

## Acoustic Modeling

### Feature Extraction

Feature extraction for GMM based acoustic models computes 16 MFCC features, a degree of voicing feature [7], and — recently introduced for tonal languages — a median smoothed pitch feature every 10 milliseconds. Temporal dynamics are captured by the concatenation of nine consecutive frames and an LDA transformation is used to finally reduce the feature vector dimension to 45. Vocal tract length normalization (VTLN) estimates a speaker specific filterbank warping factor $\alpha$, $\alpha \in \{0.80, 0.82, \ldots, 1.18, 1.20\}$, by maximizing the likelihood of a speaker's data given a speaker independent acoustic model. Training data yielding the same warping factor is modelled by a GMM that serves as a *text independent* warping factor recognizer during recognition [8]; for that purpose a 33 dimensional feature vector (16 MFCC, 16 delta coefficients, 1 acceleration coefficient) is used. Finally, speaker adaptive training (SAT) with coarse target models [9] is applied on top of VTLN in order to support online speaker adaptation for applications that permit multi-pass decoding.

### Neural Network Training

Neural network acoustic modeling is based on the RASR/NN toolkit introduced in [10], and many of our training recipes for *TANDEM acoustic models* borrow ideas from [11]. However, considering our customers' demand for state-of-the-art accuracy *without increased resource consumption*[1] we make use of rather shallow networks only. Within the chosen *hierarchical bottleneck architecture* neural network feature extraction proceeds as follows: First, 20 log-energy features (*CRBE*s) are extracted every 10 msec from the warped filterbank and 101 feature vectors are spliced together. Smoothing by two-dimensional bandpass filters yields a modulation spectrum (MRASTA filtering, [12]) whose fast modulation part is concatenated with the warped CRBEs and input into the first neural network. Warped CRBEs, the remaining (slow modulation) part of the spectrum, and PCA transformed features obtained from the bottleneck of the first network are input to the second neural network. The number of bottleneck features is chosen in order to capture roughly 90 percent of the total variance, and a context window of 9 frames is applied. Finally, 30 features are obtained from the second network's bottleneck and concatenated with the 45 MFCC base features. The so created 75 dimensional feature vectors are then subjected to Gaussian mixture modeling with up to 256 densities per Hidden Markov Model state.

In our standard setup both neural networks have three hidden layers, i.e. two layers with 2000 units each, and a bottleneck layer with 60 units in between. The output layer has 6000 - 8000 units that represent context dependent triphone HMM states. All hidden layers use a sigmoid activation function, whereas the output layer uses a softmax function. Input features are subject to global mean and variance normalization. We use supervised discriminative pre-training [3], stochastic gradient descent with L2 regularization, and a frame-wise cross-entropy training criterion. For most languages our neural networks are trained with 100 - 150 hours of data.[2,3] Usually 10 percent of the training data is put aside for cross validation, whereas the remaining part is used for back-propagation training. For that purpose, it can be further divided into a number of (overlapping) subsets, each used to run training on a copy of the network.

We have only recently started to explore *DNN-HMM hybrid acoustic models* more systematically. In this approach the

[1] Resources comprise memory and — first and foremost — runtime.
[2] EML's language portfolio currently includes English (US and UK), German, French, Spanish, NA Spanish, Italian, and Mandarin Chinese.
[3] Of course, the subsequent GMM training makes use of much more data.

base feature vector consists of the above mentioned LDA-transformed 16 warped MFCCs, degree of voicing, and pitch; optionally, 16 warped PLP coefficients can be added. A context window of 5 to 11 frames is applied which outputs 225 to 495 features that serve as input to a 7 hidden layer DNN with 2000 units per layer.

### Practical Considerations

For both, TANDEM and DNN-HMM hybrid acoustic modeling we have achieved state-of-the-art word error rates (WER) reductions of about 15 – 20 percent (relative) across languages for tasks ranging from grammar based decoding of small vocabularies (approx. 150 words) to large vocabulary speech recognition with approx. 1.5 million words and 4-gram language models.

One reason to prefer TANDEM models over DNN-HMM hybrid models is our customers' desire to adapt acoustic models to their particular needs. Here, the TANDEM approach can rely on established techniques (like, for example, MAP [13]), whereas the adaptation of neural networks still is an active research field that has to battle problems such as "*catastrophic forgetting*" [14]. However, the TANDEM approach requires significant higher decoding time because of:

- additional signal processing steps for the computation of the MRASTA features (this is the least expensive additional effort),
- additional neural network forwarding, which depends on the number of layers and units per layer, and
- increased efforts for feature scoring (or *labeling*), which are due to the higher feature vector dimension.

While – on the other hand – we see less time spent in the speech recognizer's back end search, our need to match the decoding time of the plain GMM-HMM acoustic models requires TANDEM acoustic models with smaller sized mixture sets. Both, the use of a Bayesian Information Criterion [15] for the selection of a proper acoustic model resolution and the computation of neural network features at a reduced frame rate of 15 msec help to fulfill real-time requirements, but lower the gains in WER to 10 – 15 percent relative.

## Recent Enhancements

The practical considerations described above make us focus on the use of TANDEM models, but some of our recent enhancements also apply to DNN-HMM hybrid models. The use of *gender information* is common to several of the methods, but not always available for all our training data. Therefore, we start with the description of a simple, but efficient approach to gender classification.

### Gender classification

Text independent VTLN (see above) provides a simple but quite accurate gender classification method as a byproduct. VTLN computes the warping factor $\alpha_{ML}$ for an utterance with feature vectors X=$\{x_1, ..., x_T\}$ according to

$$\alpha_{ML} = argmax_{\alpha\in A}\left\{\sum_{x\in X} p(x|(\omega,\mu,\Gamma)_\alpha)\right\}.$$

Here, A = {0.80, …, 1.20} is the set of all warping factors under consideration, and $(\omega, \mu, \Gamma)_\alpha$ are the GMM parameters for class α.

Gender classification utilizes the fact that a warping factor $\alpha_{ML}$ <= 1 is usually selected for female speakers, while for male speakers usually $\alpha_{ML}$ >= 1.0 is true. It computes a gender label *g(X)* for an incoming utterance *X* according to

$$l(X,Z) = \sum_{\alpha\in Z}\sum_{x\in X} -\log(p(x|(\omega,\mu,\Gamma)_\alpha))$$

$$g(X) = \begin{cases}\text{female, if } l(X,A_f) < l(X,A_m)\\ \text{male, else}\end{cases}$$

with $A_f$ = {0.80, …, 1.00} and $A_m$ = {1.00, …, 1.20} being the sets of warping factors for female and male speakers.

### Vocal Tract Length Perturbation

Vocal tract length *perturbation* (VTLP, 16) seeks to increase a neural network's generalization ability by introducing small random variations to the training data. VTLP modifies the training data by using a random warping factor α for the computation of neural network input features, e.g. warped CRBEs or MFCCs.

Our implementation of VTLP randomly chooses $\alpha_f \in A_f$ (see above) for utterances from female training speakers, and $\alpha_m \in A_M$ for utterances from male training speakers in each epoch of neural network training. We apply VTLP in both TANDEM and DNN-HMM hybrid acoustic modeling.

### Gender targets

For TANDEM acoustic models we extract features from neural networks with gender depend output units. For that purpose, Viterbi state alignments, which provide the training targets, are tagged with a gender label. Only states for silence and few other non-speech events (noise, music, etc.) are shared between genders, resulting in output layers with almost twice as many units for both networks of the hierarchical bottleneck approach. The sizes of all other layers and the number of features extracted from the bottleneck layers remain the same for both networks.

### Gender adaptation

For DNN-HMM hybrid modeling we compute gender dependent input feature vector transformations. For that purpose all network layers are fixed after training and the network is augmented with an additional input layer with size of the feature vectors (i.e. 225 – 495 units). Two instances of this adaptation layer are trained with data from speakers classified either as male or female, and by using the same network training settings (stochastic gradient training, L2-regularization, framewise cross-entropy). During runtime we use the gender classification GMM for the selection of one of the two transformations and feed the gender transformed feature vectors into the neural network feature scorer.

### DNN state prior

For DNN-HMM hybrid models we experiment with a different way of computing the HMM state priors. Given the trained neural network (with softmax output layer) and a feature vector $x$, the activation of the $i$-th output unit is an estimate for the state posterior $p(s_i|x)$ of state $s_i$. HMM based speech recognition requires the likelihood

$$p(x|s_i) = p(s_i|x) \cdot \frac{p(x)}{p(s_i)},$$

where $p(x)$ is constant,

$$p(s_i) = \frac{N_i}{\sum_j N_j}$$

is the state prior for state $s_i$, and $N_i$ is the number of training frames labeled with state $s_i$. In contrast, the *DNN state prior* $p_{DNN}(s_i)$ is computed from the $i$-th network output $net_i()$ by feeding all $N$ training feature vectors into the network and taking the average network output:

$$p_{DNN}(s_i) = \frac{1}{N}\sum_x net_i(x)$$

## Experimental Results

### Dataset

The freely available CMU Census Database serves as a test-bed for our developments. It consists of roughly 30 minutes of 16kHz training material (948 utterances from 74 speakers, 21 of them females), and 6 minutes of test data (130 utterances from 10 speakers, 3 of them females) and allows the creation of a speech recognizer with a vocabulary size of approx. 100 words and a small trigram language model build from the training transcripts.

The gender information provided in the database is neither used in training nor is it used in the recognition experiments. Instead, training and test use the gender labels provided by the GMM for text independent vocal tract length normalization, which is correct for 96 percent of the 74 *training* speakers (21 of 21 females; 50 of 54 males).

### Experiments

While the standard training procedure outlined above is the same for all languages under development for the EML Transcription Platform, in the experiments reported here we reduced the HMM inventory to only 200 single state triphone models for the 33 phonemes in the data base, plus a single context independent HMM for silence.

All results are obtained in a *streaming single pass mode*, i.e. recognition starts while audio data is still received by the recognizer, and no online adaptation is performed. Decoder parameters, like e.g. Viterbi beam and pruning thresholds, are the same in all experiments and chosen to fulfill customer requirements regarding runtime and latency.

For TANDEM modeling we used both (gender dependent) phoneme and state targets in neural network training, and created models with 3 different acoustic resolutions, i.e. 16, 32, or 64 densities per HMM state. In all cases we added only 15 bottleneck features to the 45 standard MFCC features. The baseline results are obtained with gender independent neural network training targets and VTL*N*. Tables 1 and 2 show that gender targets and VTL*P* improve almost all acoustic models; only in case of low resolution Gaussians gender dependent phoneme training targets showed a degradation.

**Table 1:** Impact of vocal tract length perturbation and gender dependent phoneme targets for TANDEM models with different acoustic resolution. There are 33+1=34 phoneme targets and $2 \times 33 + 1 = 67$ gender dependent phoneme targets.

| TANDEM (phoneme targets) | 64 | 32 | 16 |
|---|---|---|---|
| Base | 6.3 | 6.7 | 7.5 |
| gender target | 6.0 | 6.5 | 6.7 |
| VTLP | 6.1 | 6.6 | 7.1 |
| gender target | 5.8 | 6.1 | 7.4 |

**Table 2:** Impact of vocal tract length perturbation and gender dependent state targets for TANDEM models with different acoustic resolution. There are 201 state targets and $2 \times 200 + 1 = 401$ gender dependent phoneme targets.

| TANDEM (state targets) | 64 | 32 | 16 |
|---|---|---|---|
| Base | 7.9 | 8.0 | 8.8 |
| gender target | 5.8 | 7.0 | 8.5 |
| VTLP | 6.7 | 6.5 | 9.1 |
| gender target | 6.2 | 6.5 | 7.5 |

In the DNN-HMM hybrid approach we fixed the dimension of the input feature vector to 225 (45 x 5), reduced the size of all hidden layers to 512, and used only 5 instead of 7 hidden layers. In order to obtain results that are not biased by the random initialization we trained network with three different random seeds.

**Table 3**: Impact of VTLP and DNN prior on DNN-HMM hybrid models

.

| DNN-HMM | 1st | 2nd | 3rd | avg. |
|---|---|---|---|---|
| base | 10.3 | 8.8 | 10.5 | **9.9** |
| +DNN prior | 8.8 | 7.8 | 9.2 | **8.6** |
| VTLP | 9.1 | 8.4 | 10.2 | **9.2** |
| +DNN prior | 7.9 | 8.3 | 9.4 | **8.5** |

Table 3 shows an average gain of approx. 7 percent relative when using VTLP, and 15 percent (rel.) improvement when using the DNN prior with the baseline neural network without VTLP. The average gain obtained with DNN priors is smaller for the VTLP models, which matches the intuition that recognition needs to rely less on the class prior, if the neural networks classification error rate is low.

**Table 4:** Initial results for unsupervised gender adaptation of DNN-HMM hybrid models with an additional input layer.

| DNN-HMM | 1st | 2nd | 3rd | avg. |
|---|---|---|---|---|
| base + DNN prior | 8.8 | 7.8 | 9.2 | **8.6** |
| + gender layer | 8.0 | 8.2 | 9.0 | **8.4** |

Finally, Table 4 shows results for unsupervised gender adaptation with an additional input layer. Whereas the DNN prior and VTLP yield consistent improvement, we have to observe degradation in one of the three adaptation experiments. Nevertheless, the small average improvement suggests further work in this area, e.g. by a refinement of the GMM classifier.

## Summary and future work

In this paper we presented an overview of neural network based acoustic modeling in the EML Transcription Platform. We discussed practical considerations that let us prefer TANDEM acoustic models of moderate depth and presented two recent enhancements made to the training procedure, namely gender dependent training labels and vocal tract length perturbation. We extended the experimental evaluation of VTLP to DNN-HMM hybrid models and showed that additional gains can be obtained from the use of state priors retrieved from the network output. Finally we sketched first steps towards the unsupervised adaptation of neural networks based on gender clustering of training data. Future work will deal with more experiments in this direction and will extend the approach to speaker clusters.

## References

[1] G. E. Hinton, L. Deng, D. Yu et al.: *Deep Neural Networks for acoustic modeling in speech recognition. The shared views of four research groups.* In: *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82 -97, 2012.

[2] D. Yu, L. Deng: *Automatic Speech Recognition: A Deep Learning Approach*. Springer, London, 2015.

[3] F. Seide, G. Li G., and D. Yu: *Conversational Speech Transcription Using Context-Dependent Deep Neural Networks.* In: *Proc. of Interspeech 2011.* Florence, Italy, 2011.

[4] V. Fischer and S. Kunzmann: *The EML Transcription Platform --- A Flexible Transcription Environment for Robust Speech Recognition.* In: *Proc. der Jahrestagung der Deutschen Arbeitsgemeinschaft für Akustik (DAGA 2013).* Merano, Italy, 2013.

[5] EML European Media Laboratory GmbH: *EML Transcription Platform. Architecture, Technical Information, and Services, Version 7.0*. Heidelberg, 2015.

[6] D. Nolden, H. Ney, and R. Schlüter: *Time Conditioned Search in Automatic Speech Recognition Reconsidered.* In: *Proc. of Interspeech 2010.* Makuhari, Chiba, Japan, 2010, 234 – 237.

[7] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney: *The RWTH Aachen University Open Source Speech Recognition System.* In: *Proc. of Interspeech 2009.* Brighton, UK, 2009, 2111–2114.

[8] Welling, L., Kanthak, S., Ney, H.: *Improved Methods for Vocal Tract Normalization.* In: *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1999).* Phoenix, AZ., 1999, 761–764.

[9] Stemmer, G., Brugnara, F., Giuliani, D.: *Adaptive Training Using Simple Target Models.* In: *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005).* Philadelphia, PA., 2005, 997–1000.

[10] S. Wiesler, A. Richard, P. Golik, R. Schlüter, and H. Ney: *RASR/NN: The RWTH neural network toolkit for speech recognition.* In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014).* Florence, Italy, 2014, 3313-3317.

[11] C. Plahl, R. Schlüter, and H. Ney: *Hierarchical Bottle Neck Features for LVCSR.* In: *Proc. of Interspeech 2010*, Makuhari, Japan, 2010, 1197–1200.

[12] H. Hermansky and P. Fousek: *Multi-resolution RASTA filtering for TANDEM-based ASR.* In: *Proc. of Interspeech 2005.*, Lisbon, Portugal, 2005.

[13] J. Gauvain and C. Lee: *Maximum a Posteriori Estimation of Multivariate Gaussian Mixture Observations of Markov Chains.* In: *IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2*, 1994, 291 – 298.

[14] D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio: *Adaptation of Artificial Neural Networks Avoiding Catastrophic Forgetting.* In: *Proc. of the Int. Joint Conf. on Neural Networks 2006.* Vancouver, Canada, 2006, 1554 – 1561.

[15] S. Chen and P. Gopalakrishnan: *Clustering via the Bayesian Information Criterion with Applications in Speech Recognition.* In: *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1998).* Seattle, OR., 1998, 645 – 648.

[16] N. Jaitly and G.Hinton: *Vocal Tract Length Perturbation (VTLP) improves Speech Recognition.* In: *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, Georgia, 2013.