Improved Acoustic Feature Combination for LVCSR by Neural Networks

Christian Plahl, Ralf Schlüter and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Aachen, Germany
{plahl,schlueter,ney}@cs.rwth-aachen.de

Abstract

This paper investigates the combination of different acoustic features. Several methods to combine these features such as concatenation or LDA are well known. Even though LDA improves the system, feature combination by LDA has been shown to be suboptimal. We introduce a new method based on neural networks. The posterior estimates derived from the NN lead to a significant improvement and achieve a 6% relative better word error rate (WER).

Results are also compared to system combination. While system combination has been reported to outperform all other combination techniques, in this work the proposed NN-based combination outperforms system combination. We achieve a 2% relative better WER, resulting in an improvement of 7% relative to the baseline system.

In addition to giving better recognition performance w.r.t. WER, NN-based combination reduces both, training and testing complexity. Overall, we use a single set of acoustic models, together with the training of the NN.

Index Terms: feature extraction, multi-layer neural network, speech recognition

1. Introduction

Within different projects, several state-of-the-art LVCSR speech recognizers have been set up at RWTH. In order to take advantage of system combination, multiple complementary subsystems based on different acoustic front-ends [1, 2] are built. Therefore, a complete GMM/HMM based training and decoding have to be performed independently for each acoustic front-end followed by a system combination step, resulting in high computational costs.

In order to reduce and optimize the resources available, several approaches for combining acoustic features have been proposed in the last years. For example, in [3] the combination is done explicitly on the feature level by linear discriminant analysis (LDA), though LDA has been shown to be suboptimal [4]. Furthermore, the combination in [4] is done in an acoustic rescoring framework. Even though both approaches achieve reasonable improvements, system combination seems to be superior [5] and can be performed on different levels. Implemented in the adaptation step of the system it is referred to as *cross adaptation* and proves to give considerable improvements [6]. Alternatively, lattice or N-best-list based system combination is applied to the final output of the individual systems [7].

In this paper, we focus on the question, how multiple acoustic features are combined in an acoustic front-end, resulting in a reduction of the overall training and decoding effort, one acoustic model only, while achieving competitive results compared to system combination. Specifically, we propose the combination of several acoustic features by neural networks (NN) and explore the feature combination method of several acoustic front-ends, continuing the work started in [3].

First, we compare LDA and the proposed neural network feature combination method using one, two or three input feature streams. Finally, we show that our NN combination technique achieves competitive results with state-of-the-art system combination techniques on a Spanish broadcast news task. We almost achieve the same recognition result compared to system combination even if one feature stream is used for NN-based combination. Moreover, we outperform system combination by the proposed NN combination technique when two or more feature streams are fed into the neural network. Now, we can optimize the overall process. Finally, instead of training multiple system, training of one system containing all features is sufficient and the overall training and decoding process could be simplified.

The paper is structured as follows:
In Section 2 we describe the proposed NN combination method based on multiple feature streams. Acoustic modeling and the different acoustic features used are described in Section 3. The experimental setup is explained in Section 4, followed by the experimental results for LDA, NN combination and system combination in Section 5. Finally, we end up with a summary of the paper and the conclusions in Section 6.

2. Neural Network Feature Combination

In the last years phoneme posterior estimates derived from a NN have recently become a major component of state-of-the-art automatic speech recognition (ASR) systems [1, 2]. The structure of the neural networks as well as the input features for the neural network have been under investigation in the last years. In [8] the hierarchical processing of NNs introduced in [9] and the bottle neck topology in [10] are combined, resulting in further improvements. Moreover, the resulting hierarchical bottle neck structure benefits from both concepts. In [11] and [12] long-term features such as features based on long temporal pattern (TRAP) or multi-resolution RASTA features have been introduced. These features use a large temporal context of up to one second and therefore provide additional information to the conventional short-term features with a temporal context of 25ms [1, 12].

These previous experiments report that NN features can provide complementary information to the final ASR system. The most important advantage of the neural network is the nonlinear transformation —most of the time the sigmoid function is used— of the input features by the neural network. We want to benefit from the nonlinearity of the NN to improve the combination of several feature streams.

In order to analyze the NN combination techniques the structure of the neural network is kept as simple as possible. We

Copyright © 2011 ISCA

use the well known TANDEM approach first mentioned in [13]. There, a neural network is trained on short-term features. Posterior estimates are derived from the NN and are transformed further by logarithm and PCA. Finally, the log/PCA transformed posterior estimates are used as input to train a HMM/GMM based recognition system.

2.1. Input Features

We feed 9 consecutive frames of the classical short-term features as input to the neural network. The short-term features used are MFCC, and PLP, and Gammatone (GT) features. The features are augmented with first order temporal derivatives and the second order temporal derivatives of the first dimension, resulting in a 33 dimensional feature vector for MFCCs and PLPs and 31 components for Gammatone features. A short description of the Gammatone features is given in Section 3.2. Finally, mean and variance normalization are performed. In order to focus on the different combination methods, we have skipped any further transformation method of the input features for the NN training. Overall, the input feature vector consist of $33 \times 9 = 297$ elements, and $(33 + 33) \times 9 = 594$ elements for single feature stream and two feature stream combination respectively —(31×9) and $(33 + 31) \times 9$ when GTs are used. When all short-term features are combined we end up with a $(33 + 33 + 31) \times 9 = 873$-dimensional feature vector.

2.2. Training

The training of the network is performed using a single feed-forward NN consisting of three layers. While the first and last layer are taken for I/O, we have enlarged the middle layer to provide the necessary model power. We have taken 4000 nodes in the hidden layer and 37 nodes in the output layer corresponding to the 37 phonetic targets. These targets are derived from a forced alignment of a previously trained ASR system based on MFCC features only. On a cross set —containing 8% of the training data— we achieve a frame accuracy of 71% for single stream combination and 74% for multiple stream combination. In the training, up to 75% of the frames are classified correctly. Finally, the TANDEM phoneme posterior estimates derived from the NN are transformed by logarithm and reduced by PCA. The final 15-dimensional feature vector keeps 95% of the variability of the phoneme posteriors. In the acoustic front-end the posterior features are augmented with the LDA reduced MFCC features.

3. Acoustic Modeling

As in [1, 2], several systems that differ only in their acoustic front-ends have been set up. The front-ends used are built by different feature combination techniques —feature combination done by LDA or a NN— and are based on the Spanish RWTH speech recognition system developed for the Quaero 2010 Evaluation Campaign. The feature combination techniques differ in the number of feature streams combined, the type of base features used and the combination method applied. As shown in [1, 2] our ASR systems are competitive to other ASR systems within the project. In the experiments some system uses NN posterior estimates, while other systems, for comparison, are based on classical features only. Nevertheless, the acoustic training is performed independently for each of the systems. The training of the neural network posterior estimates, described in Section 2, as well as the training of the acoustic models are performed on the whole speech corpus of 60 hours. The acoustic models are trained using the RWTH Speech Recognition system [14].

3.1. Acoustic Features

The acoustic front-ends of the systems consist of different base features. The features used are MFCCs, and PLP, and Gammatone features, which are described in detail in [15]. We will give a short description of the Gammatone features in Section 3.2 as well.

The base features, augmented with a voiceness feature, are normalized by segment-wise mean and variance normalization. All features within a sliding window of length nine are concatenated and are projected to a 45-dimensional feature space by LDA. As proposed in [1, 3] we use a common LDA for the whole feature stream. The HMM/GMM system trained on these feature sets are used as baseline systems for all other experiments.

For all experiments concerning NN estimates as described in Section 2 we augment the 45-dimensional LDA transformed MFCC features and the NN features. Overall, the final MFCC + NN feature stream contains of 60 components. For comparison, all two short-term feature combination approaches by LDA consist of 60 components also.

3.2. Gammatone Features Extraction

As introduced in [15], the Gammatone (GT) features are extracted by auditory filter banks realized by Gammatone filters. These filters are defined in the time domain by the following impulse response:

$$h(t) = k \cdot t^{n-1} \exp(-2\pi \cdot B \cdot t) \cdot \cos(2\pi \cdot f_c \cdot t + \phi).$$

Here, k defines the output gain, B defines the bandwidth, n is the order of the filter, f_c is the filter's center frequency, and ϕ is the phase. In this work, 4th order Gammatone filters are used. A similar filter bank has been used by [16] for robust speaker identification.

The center frequencies of the filters are distributed over the frequency range according to the Greenwood function with human parameters [17]:

$$\rho(x) = A \cdot (10^{a \cdot x} - k)\text{Hz}$$

where $A = 165.4$, and $a = 0.88$, and $k = 2.1$.

The 68 Gammatone filter outputs are temporally integrated followed by a spectral integration and a 10th root compression. Finally, after cepstral decorrelation the 15 coefficients are mean and variance normalized.

In [15] we have shown that Gammatone features are competitive to MFCC features and that they are robust against noise in the speech signal. Therefore, GT features could provide contrary information to the classical short-term MFCC features.

3.3. Acoustic Training

The acoustic models for all systems are based on triphones with cross-word context, modeled by a 6-state left-to-right Hidden Markov Model (HMM). A decision tree based state tying is applied resulting in a total of 4500 generalized triphone states. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix.

In order to compensate for speaker variations we use constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR). In addition, during recognition, MLLR is applied to the means of the acoustic models. For computational reason we have not performed a full training including discriminative training.

4. Experimental Setup

Approximately 60 hours of Spanish Broadcast News (BN) and speech data collected from the web are used both for training the neural network posterior estimates and for training the acoustic models. The whole training data are provided within the Quaero project.

We evaluate the performance of the different systems built on the Quaero task also. While the system parameters are tuned on the development corpus of 2010 (dev10) —marked by * in Table 1— the evaluation corpora of 2010 (eval10) and 2009 (eval09) are used for testing only. The development and evaluation corpora consist of a mix of different speech sources. The data contain broadcast news, speeches of the European Parliament, several pod cast shows and other speech data collected from the web. Table 1 shows some statistics of the training and testing corpora used.

Table 1: Acoustic data of the training and testing corpora. The corpus marked by * is used for tuning the parameters of the individual systems.

	Training and testing data			
	train	dev10*	eval10	eval09
total data	60h	2.8h	3.3h	3.2h
# segments	19519	1016	1267	924

In recognition a 4-gram language model (LM) is used which consists of 60k words. The LM is trained on the final text editions and verbatim transcriptions of the European Parliament Plenary Sessions, and on data from the Spanish Parliament and Spanish Congress, provided within the TC-STAR project. Language model data provided within the Quaero project and the training transcription have been included as well.

5. Multiple Feature Combination

5.1. Feature Combination by LDA

In the first experiments we continue the work started in [3]. We use a single LDA matrix to combine several short-term acoustic feature streams. Results, similar to the results of [3], are shown in Table 2.

Table 2: Recognition results of a GMM/HMM trained system using linear discriminant analysis to combine single (a) or multiple feature streams (b). The input features used are MFCCs, PLPs or Gammatone features.

	Method	Features	Test Corpora (WER[%])		
			dev10	eval10	eval09
(a)		MFCC	22.0	18.5	17.0
		PLP	23.0	19.6	17.8
		GT	22.3	19.0	17.3
(b)	LDA	MFCC + PLP	22.2	18.7	17.0
		MFCC + GT	**21.7**	**18.4**	**16.9**

While the first lines of Table 2, marked by (a), show LDA combination results on a single stream, the last lines, marked by (b), show results of combining two feature streams by LDA. The best result for the single stream combination is achieved by the MFCC system. The best multiple stream combination result is obtained by adding GT features. Here, the system can benefit from the Gammatone features, which are robust against noise in the speech data. The single system is improved by 0.3% absolute for the development set, but only 0.1% absolute on the evaluation corpora. Overall, the improvement by LDA for multi feature combination is slight. As shown in Table 2 the LDA combined MFCC+PLP system even degrades in word error rate. This is due to the effect of numerical problems in the LDA estimation covered by [3].

5.2. Neural Network Feature Combination

Next, we have trained several neural networks based on different number of input feature streams. The final posterior probabilities are concatenated with the LDA reduced MFCC features to train a GMM/HMM-based ASR system. Results are shown in Table 3. As expected, adding neural network posterior estimates improve the baseline system trained on LDA reduced MFCCs. We gain up to 1.4% absolute on the development corpus and up to 1% absolute on the evaluation corpora. Moreover, the results show that NN posteriors estimates provide complementary information even if they are trained on the same input features (here: 0.7-0.9% absolute improvements for MFCC-based NN features, marked by (b) in Table 3).

Table 3: Recognition results of a GMM/HMM trained system using MFCC features, marked by (a), and, in addition, neural network posterior estimates trained on a single, marked by (b), or multiple input feature streams, marked by (c) and (d).

	Method	Features	Test Corpora (WER[%])		
			dev10	eval10	eval09
(a)		MFCC	22.0	18.5	17.0
(b)	NN	MFCC	21.1	17.7	16.3
		PLP	21.2	17.8	16.5
		GT	20.6	17.5	16.2
(c)	NN	MFCC + PLP	20.6	17.4	16.2
		MFCC + GT	**20.4**	**16.9**	**15.9**
		PLP + GT	20.5	17.3	15.9
(d)	NN	**MFCC+PLP+GT**	**20.4**	**17.0**	**15.7**

As shown in Table 3 the combination of two or more feature streams outperforms the single NN feature approach. The best single NN stream approach could be further improved by 0.2% absolute or more over all corpora by using all input feature streams. When we combine all short-term acoustic features, MFCCs, PLPs, and GTs, we improve in the non-speaker adapted case only (results not reported). After speaker adaptation no improvement between the best two stream and the three stream combination experiments is observable.

Nevertheless, the best results are achieved when several feature streams are combined. Compared to the multiple feature stream combination by LDA of Section 5.1 we improve the best system by more than 1% absolute or up to 6% relative.

5.3. System Combination

Finally, we have performed system combination of the baseline systems, line (a) in Table 2. Results are shown in Table 4 marked by SC. While ROVER is the easiest method, we have tested lattice based combinations as well as N-best-list combinations. We achieve the best system combination results by the min.fWER method [18]. As mentioned in [5], system combination seems to be superior to other combination techniques. As shown in Table 4, this is no longer the case and neural network feature combinations outperform the system combination of the baseline systems.

Even though system combination achieves improvements of about 5% relative, the results for one stream NN combination are slightly worse only. The best NN combination method using two or more input streams even beat system combination. The difference is around 2% relative for all corpora.

These results are very impressive, because we have used a simple topology and short-term features for NN feature combination only.

Table 4: Recognition results for a GMM/HMM trained system using MFCCs and posterior estimates of a neural network trained on multiple input features, marked by (NN). Lines marked by (SC) are results for system combination of the baseline MFCC, PLP and GT system. System combination has been performed by the min.fWER approach.

Method	Features	Test Corpora (WER[%])		
		dev10	eval10	eval09
SC	MFCC +PLP	21.3	17.7	16.3
	MFCC +GT	20.9	17.5	16.0
	PLP +GT	21.4	18.1	16.6
	MFCC +PLP +GT	**20.9**	**17.5**	**16.0**
NN	MFCC +GT	20.4	16.9	15.9
	MFCC +PLP +GT	**20.4**	**17.0**	**15.7**

6. Summary and Conclusions

In order to simplify the training and decoding process, the aim of this paper was to make feature combination competitive to system combination. Therefore, we compared different feature combination techniques such as LDA, system combination, and our proposed neural network processing. We showed, that the results for system combination on the development corpus could be almost reached by the NN-based feature combination when one feature stream is used. The NN-based combination approach using all three feature streams outperformed system combination of the baseline system by 2.4-2.9% relative on the development and evaluation sets of 2010 respectively.

Instead of setting up several subsystems based on different acoustic front-ends and combining these systems afterwards by system combination, training of one system with all features combined by the neural network approach is sufficient. The NN feature combination method simplifies the system development circle and optimizes the training and decoding. Instead of three systems, we use a single set of acoustic models, together with the training of the NN.

In addition, we verified that the LDA approach is suboptimal for feature combination. The LDA approach does not perform as well as the NN based posterior estimates trained on one feature stream. By combining all three features streams the system improved further up to 6% relative compared to LDA and up to 7% relative to the baseline MFCC system.

Though the NN feature combination achieved significant improvements, questions still remaining for optimal combination results. These questions concern the topology of the neural network, the combination of NN probabilistic features, the training of several NN, and finally the best combination of the short-term and long-term features for the NN training. In the future, we will investigate this area to find the best feature combination approach.

7. Acknowledgements

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

[1] C. Plahl, B. Hoffmeister, G. Heigold, J. Lööf, R. Schlüter, and H. Ney, "Development of the GALE 2008 Mandarin LVCSR system," in *Interspeech*, Brighton, U.K., Sep. 2009, pp. 2107–2110.

[2] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 Quaero ASR evaluation system for English, French, and German," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2212–2215.

[3] R. Schlüter, A. Zolnay, and H. Ney, "Feature combination using linear discriminant analysis and its pitfalls," in *Interspeech*, Pittsburgh, PA, USA, Sep. 2006, pp. 345–348.

[4] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Philadelphia, PA, USA, Mar. 2005, pp. 457–460.

[5] A. Zolnay, "Acoustic feature combination for speech recognition," Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, Aug. 2006.

[6] D. Guiliani and F. Brugnara, "Acoustic model adaptation with multiple supervisions," in *Proc.* TC-STAR *Workshop on Speech-to-Speech Translation*, Barcelona, Spain, Jun. 2006, pp. 151–154.

[7] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, College Park, MD, 2000.

[8] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical bottle neck features for LVCSR," in *Interspeech*, Makuhari, Japan, Sep. 2010, pp. 1197–1200.

[9] F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, and R. Schlüter, "Hierarchical neural networks feature extraction for LVCSR system," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 42–45.

[10] F. Grézl, M. Karafiat, S. Kontar, and J. Cernock, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Honolulu, HI, USA, Apr. 2007, pp. 757–760.

[11] H. Hermansky and S. Sharma, "TRAPs - classifiers of temporal patterns," in *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, Dec. 1998.

[12] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 361–364.

[13] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2000, pp. 1635–1638.

[14] 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 *Interspeech*, Brighton, U.K., Sep. 2009, pp. 2111–2114.

[15] R. Schlüter et al., "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Honolulu, HI, USA, Apr. 2007, pp. 649–652.

[16] Q. P. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, Taxas, USA, Mar. 2010, pp. 4514–4517.

[17] D. D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *Acoustical Society of America Journal*, vol. 87, pp. 2592–2605, Jun. 1990.

[18] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *Interspeech*, Pittsburgh, PA, USA, Sep. 2006.