COMPARISON AND COMBINATION OF DIFFERENT CRBE BASED MLP FEATURES FOR LVCSR

Zoltán Tüske, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany

{tuske,schluter,ney}@cs.rwth-aachen.de

ABSTRACT

Multi Layer Perceptron (MLP) features extracted from different types of critical band energies (CRBE) —derived from MFCC, GT, and PLP pipeline— are compared on French broadcast news and conversational speech recognition task. Though the MLP structure is kept fixed, ROVER combination of different CRBE based systems leads to 4% relative improvement. Furthermore, aiming at the combination of state-of-the-art features based on various signal analysis methods into one single stream, posterior feature space based combination technique is proposed. The speaker normalized features originated from different CRBEs are merged after additional MLP training by Dempster-Shafer rule. The performance of these posterior features unifying the different CRBE based features is superior to the best single CRBE based posterior features by 6% relative. Further results reveal that the concatenated cepstral and unified posterior features perform nearly as well as the ROVER combination of the different CRBE based systems.

Index Terms— MFCC, GT, PLP, CRBE, MRASTA, Dempster-Shafer, LVCSR

1. INTRODUCTION

Recently numerous MLP based feature extractions have been used in state-of-the art large vocabulary continuous speech recognition (LVCSR) systems. Besides the investigation of (more and more complex) MLP topologies, several different short and long-term representations of speech are suggested in the literature to train the MLPs. As short-time input features to MLP, the different cepstral coefficients (MFCC or PLP) are directly used [1, 2]. Considering the long-term features, different CRBE processing structures have been shown to improve the recognition performance substantially, like MRASTA [3], TRAPS/HATS [4]. A fair comparison of the different MLP structures on the same CRBE has been done for LVCSR only recently [1]. Although the original MRASTA and TRAPS/HATS were introduced on Bark-scaled CRBE, the long-term features could be extracted using Mel-triangular filterbank, as well [5].

The different cepstral features have been shown to be complementary and could be efficiently combined [2] e.g. through ROVER [6], nevertheless they are based on the same concepts. However, the effect of different signal analysis methods on the MLP based posterior features has not been fully investigated, in particular not on the long-term posterior features.

Therefore, keeping the MLP and the long-term feature extraction structures fixed, the posterior features are investigated from the viewpoint of signal analysis. In this paper we systematically compare the different CRBE based short-term (cepstral) and long-term (MRASTA based) posterior features. We extend our study not only to MFCC and PLP but also to the GT [7] based CRBE. Furthermore, we also examine the performance of Dempster-Shafer (DS) [8] combined (multi-stream) short and long-term posteriors of the different CRBEs. Our goal is not only to find the best CRBE which we think should be task dependent, as shown in the past for cepstral features [2, 7], but also to find an effective way to combine those different but similarly performing state-of-the-art features into a single feature stream. Based on the results of our previous work where we showed that MLP training could benefit from the features normalized for Speaker Adaptive Training (SAT) [9], we demonstrate through experiments that the different CRBE based features can be efficiently combined – after speaker normalization and additional MLP training – into one single stream by the use of DS rule.

The paper is organized as follows: Section 2 shortly summarizes the different cepstral (and CRBE) feature extraction steps. Section 3 gives the details of the corpus used in our experiments. We describe the experimental setups in Section 4 followed by results (Section 5). The paper closes with conclusion (Section 6).

2. FEATURE EXTRACTION

This section gives a brief overview of the three cepstral features and the corresponding CRBEs and points out their differences.

2.1. Mel-Frequency Cepstral Coefficients — MFCC

The feature extraction is based on the Short-Time Fourier Transform (STFT) of the pre-emphasized speech signal [10]. After integration of the amplitude spectrum with triangular filterbank, where the filters are equally spaced on Mel-scale, the logarithm is taken. The MFCC features are extracted by applying Discrete Cosine Transformation (DCT) on the output of the previous step. Finally, segment-wise mean and variance normalization is applied. As MFCC-CRBE we refer to the logarithmized triangular filter energies.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

2.2. GammaTone features — GT

Instead of the STFT based analysis, the features are extracted by an audiologically motivated filterbank realized by time-domain gammatone filters [7]. The auditory filters are equally spaced on Greenwood-scale. After spectral and temporal integration the $10th$ root is taken instead of the logarithm. The DCT decorrelation is followed by mean and variance normalization. In the rest of the paper GT-CRBE means the spectro-temporal smoothed and root compressed filterbank energies.

2.3. Perceptual Linear Predictive coefficients — PLP

These features were proposed in [11] and they are based again on the STFT of speech. Simulating the critical band masking, the amplitude spectrum is integrated with trapezoid filters equally spaced on Bark-scale. The filterbank output is pre-emphasized according to equal-loudness curve. To simulate the relation between the intensity and perceived loudness of sound, cubic root amplitude compression is performed followed by all-pole model parameter estimation (linear predictive (LP) analysis). The autoregressive coefficients are directly transformed to cepstral coefficients which are mean and variance normalized. In the followings the PLP-CRBE denotes the logarithmized critical band energies reconstructed from the LP coefficients.

2.4. Summary

One of the main differences between the features refers to the shape of the critical band filters: *triangular*, *gammatone*, or *trapezoid*. Further difference concerns the distinction how the decreasing frequency resolution of the human ear is modeled with higher frequencies: *Mel*, *Greenwood*, or *Bark-scales*. Moreover, all-pole model fitting and Hamming-windowing are applied in PLP and GT pipeline, respectively, to *smooth* the CRBE, whereas there is no additional smoothing in the MFCC extraction. The cepstral features also differ w.r.t loudness-intensity compression: MFCC uses *logarithm*, whereas PLP and GT apply *root* function. However, the PLP-CRBE is compressed by logarithm as in case of MFCC-CRBE.

3. CORPUS DESCRIPTION

The QUAERO project is a large vocabulary speech recognition task focusing on transcription of web data. The data include different speech types: Broadcast News (BN) and Broadcast Conversation (BC) like comedy, cooking sessions, interviews, and talk-shows. Recognition on the data is challenging due to the huge variability in the acoustic conditions and to the relatively large portion of spontaneous speech. Within the QUAERO project about 250 hours of transcribed French speech data are collected and used for training the acoustic models and neural networks (Train). While the system parameters are tuned on development corpus (Dev10) of 2010, the evaluation set (Eval10) of 2010 is used for measuring the final recognition performance. According to the aim of the project the fraction of the BC is increased every year. The evaluation set of the present year consists of 50% BN and 50% BC. Table 1 summarizes the corpus statistics of training and testing data.

Table 1. *Training and testing corpora*

	Train	Dev10	Eval10
total data [h]	257	3.7	2.9
# running words	9,800,000	41,000	36,000

4. EXPERIMENTAL SETUPS

4.1. Feature extraction

We follow the feature extraction method as described in our earlier work [9], however, the cepstral feature and CRBE calculation is changed according to the Section 2. Based on our previous result, vocal tract length normalized CRBEs and cepstral features are extracted. The dimensions of the different cepstral features and CRBEs are showed in Table 2.

Table 2. *Dimension of the Cepstral features and CRBEs*

Dimension	Feature extraction		
	MFCC	GT	PLP
Cepstral	16	15	16
CRBE	20	15	20

First, the cepstral features are extracted from the audio signal. Then two MLPs are trained in parallel and their phoneme posterior outputs are combined by Dempster-Shafer [8] rule. The cepstral features are fed to the first MLP (short-term), while the second (long-term) posterior estimates are based on hierarchical processing [12] of two MLPs, where the corresponding inputs are the fast and slow modulation frequencies of multi-resolution rasta filtered (MRASTA) [3] CRBE with temporal context of one second.

Thus, in this study the following five types of features from three different CRBEs are calculated for the experiments: cepstral coefficients, short and long-term posteriors, their DS-combination, and the concatenated cepstral and DS features. In order to combine the three different CRBE based features in posterior space by applying again the DS rule, three MLPs are trained on the corresponding speaker normalized features in addition. The feature combination experiments are done with concatenated cepstral and DS combined short and long-term posterior features only (Fig. 1). The reason for the choice of DS rule is that it has been experimentally proven to be one of the most efficient method to combine MLP classifiers [8].

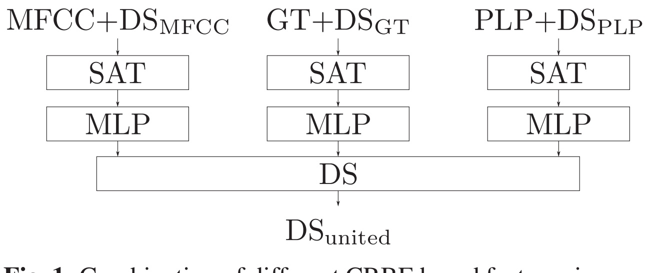

Fig. 1. Combination of different CRBE based features in posterior feature space after speaker normalization

All MLPs consist of three layers and are trained to approximate class posterior probabilities of 44 phonemes. Furthermore, the number of nodes in the hidden layer is fixed to 7000. According to the TANDEM approach [13] the extracted posteriors are logarithmized and PCA transformed.

4.2. Acoustic modeling

Triphone-level acoustic models (AM) with cross-word context are modeled by 3-state left-to-right Hidden Markov Models, where the number of different triphone states are reduced to 4,500 with phonetic decision tree based state-tying.

The Gaussian Mixtures Models (GMM) parameters of the AM are trained by Maximum Likelihood estimation resulting in 2M Gaussians. Instead of training the AMs from scratch, an initial alignment is generated by the previous best system [14], and used to estimate the decision tree of the state-tying, the LDA matrix for cepstral features, and the mixture parameters in the first iteration steps.

To mitigate the effect of speaker variation, the Constrained Maximum Likelihood Linear Regression (CMLLR) speaker normalization technique is applied in training and recognition, where the affine transformation matrix w.r.t speaker is estimated using simple target model approach [15]. Compared to [14] we skip the cross adaption and discriminative training steps due to computational reasons.

4.3. Language modeling

Based on the available data in the QUAERO project, 4-gram language model was estimated and smoothed by the modified Kneser-Ney method. The vocabulary contains 200K words and the out-of-vocabulary ratio is about 0.5%. More details are available in [14].

5. EXPERIMENTAL RESULTS

In the following, all of the results reported below are obtained with CMLLR based speaker adaptive trained AM in two recognition passes. In contrast to [14] the LM rescoring with the unpruned LM is not performed.

5.1. Performance of cepstral features based system

The results of the first experiment, where the performance of the cepstral features is compared, are shown in Table 3. All the cepstral features give comparable results, however, the best performance is achieved on this task by MFCC on both the development and evaluation sets. Though cepstral features follow similar concepts and provide similar representation of the speech signal, the ROVER [6] combination of these systems leads to more than 4% relative Word Error Rate (WER) improvement compared to the best single system. This result indicates the complementarity between the three cepstral features.

Table 3. *Word Error Rate (WER) for different Cepstral Coefficients (CC)*

WER [%]	CC			ROVER
	MFCC	GT	PLP	
Dev10	23.8	24.3	24.3	22.8
Eval10	25.1	25.2	25.8	24.0

5.2. Comparison of the different CRBE based posterior features

In the second experiment we measure the recognition performance of the short and long-term posterior features extracted from different Cepstral Coefficients (CC) and CRBEs respectively. Table 4 shows that the short-term posteriors give similar results; however, the MRASTA based long-term posterior features show larger differences w.r.t CRBE. Moreover, compared to the best single stream posterior results, the DS-combination of the short and long-term features improves the recognition performance by more than 7% relative in case of MFCC-CRBE and GT-CRBE. Surprisingly, PLP-CRBE based DS posteriors achieve modest results, the relative gain in WER is 4%. The lowest WER is achieved by the MFCC-CRBE based DS$_{\text{MFCC}}$ posterior features.

Table 4. *Recognition performance of posterior features*

WER [%]	Posterior features					
	CC (short-term)		MRASTA (long-term)		DS$_{crbe}$ (combined)	
CRBE	Dev10	Eval10	Dev10	Eval10	Dev10	Eval10
MFCC	25.3	27.2	25.6	26.8	23.4	24.3
GT	25.4	26.9	26.2	27.3	23.5	24.8
PLP	25.2	27.0	26.1	27.2	24.2	25.9

5.3. Comparison of different CRBE based concatenated cepstral and posterior features

In state-of-the-art GMM/HMM speech recognition systems the concatenated cepstral and posterior features are widely used. Therefore, Table 5 shows the WER for the different CRBE based cepstral features augmented with the corresponding DS features. There is less than 4% rel. WER differences between the distinct CRBE based systems, nonetheless the MFCC based system gives the best results. The ROVER combination of them leads to 4% improvement relative to the best system as in case of pure cepstral systems. The results underline that the different speech signal representations could be exploited even with MLP based features.

Table 5. *Recognition results for the concatenated cepstral and DS-combined short and long-term posterior features*

WER [%]	CC+DS$_{crbe}$			ROVER
CC/CRBE	MFCC	GT	PLP	
Dev10	21.5	21.8	21.8	20.7
Eval10	22.5	22.8	23.3	21.6

5.4. Combination of different CRBE based features into single stream

Based on our previous work [9] we transform the speaker normalized concatenated cepstral and DS features into posterior feature space with an additional MLP training. The recognition results of these posterior features (without concatenation to cepstral features!) are shown in Table 6. Comparing to the DS results in Table 4, we can confirm our former statement, that the MLP could profit from the speaker normalized (SAT) features. Furthermore, by DS combination of the posteriors extracted from the different CRBE based speaker normalized features, we are able to improve the recognition performance further. The posterior features unifying the different CRBE based speech representation are denoted as $\text{DS}_{\text{united}}$ and outperform the DS_{MFCC} features from Table 4 by relative 6%.

Table 6. *Recognition performance in WER [%] of the speaker normalized (SAT) and MLP transformed concatenated cepstral and DS combined short and long-term posterior features*

WER [%]	MLP(SAT(CC+DS_{crbe})			$\text{DS}_{\text{united}}$
CC/CRBE	MFCC	GT	PLP	
Dev10	22.8	23.4	23.2	22.4
Eval10	23.4	23.9	24.4	22.8

Table 7. *Recognition performance of the concatenated cepstral and the different CRBEs unifying posterior features*

WER [%]	CC+$\text{DS}_{\text{united}}$		
CC	MFCC	GT	PLP
Dev10	21.1	21.1	21.2
Eval10	21.9	21.8	22.0

Since the posterior and cepstral features tend to make complementary errors, the performance of the concatenated cepstral and the $\text{DS}_{\text{united}}$ features is also investigated (Table 7). All of the cepstral features benefit from the concatenation of the united posterior features resulting in very similar performance which is comparable to the ROVER based system combination results in 5. Comparing to the corresponding concatenated cepstral and DS features based system in Table 5, the relatie improvement is between 3% (MFCC) and 5% (PLP).

6. CONCLUSION AND FUTURE DIRECTIONS

Different types of critical band energy based cepstral and posterior features were evaluated on a French LVCSR task. Besides, experimentally selecting the CRBE giving the best results (MFCC-CRBE), an effective way was presented to combine the different state-of-the-art features based on different signal analysis methods. Reconfirming our previous statement that MLP could benefit from the SAT features, the proposed combination method operates in posterior feature space to merge the different CRBE based speech signal representations. The unified posterior features in concatenation with MFCC performed comparable to ROVER based system combination. In this study three layer perceptrons and hierarchical MRASTA based long-term feature extraction were applied as part of the experiments. As direction of future work, we intend to repeat the experiments with other long-term feature representations (TRAPS, HATS, DCT/wLP-TRAPS), which could serve as further basis for a final system combination. Increasing the complexity of the MLP – more hidden-layer and more phonetic targets in the output layer – should be also part of the further investigation.

7. ACKNOWLEDGEMENT

This work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. [213850]. 11, Speech Communication with Adaptive Learning - SCALE. Furthermore, the work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. REFERENCES

[1] F. Valente *et al.*, "Analysis and Comparison of Recent MLP Features for LVCSR Systems," in *Proc. of Interspeech*, 2011, pp. 1245–1248.

[2] C. Plahl *et al.*, "Improved Acoustic Feature Combination for LVCSR by Neural Networks," in *Proc. of Interspeech*, 2011, pp. 1237–1240.

[3] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. of Interspeech*, 2005, pp. 361–364.

[4] B. Chen *et al.*, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. of Int. Conf. on Spoken Language Processing*, 2004, pp. 612–615.

[5] F. Grézl and M. Karafiát, "Hierarchical Neural Net Architectures for Feature Extraction in ASR," in *Interspeech*, 2010, pp. 1201–1204.

[6] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 347–354.

[7] R. Schlüter *et al.*, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 649–652.

[8] F. Valente, "Multi-stream speech recognition based on Dempster-Shafer combination rule," *Speech Communication*, vol. 52, no. 3, pp. 213–222, Mar. 2010.

[9] Z. Tüske *et al.*, "A study on speaker normalized MLP features in LVCSR," in *Proc. of Intererspeech*, 2011, pp. 1089–1092.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[12] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 4165–4168.

[13] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2000, pp. 1635–1638.

[14] M. Sundermeyer *et al.*, "The RWTH 2010 QUAERO ASR Evaluation System for English, French, and German," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 2212–2215.

[15] G. Stemmer *et al.*, "Adaptive training using simple target models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 997–1000.