Normalization of spectro-temporal Gabor filter bank features for improved robust automatic speech recognition systems

Marc René Schädler[1] *and Birger Kollmeier*

[1]Medical Physics, Carl-von-Ossietzky University Oldenburg / D-26111 Oldenburg / Germany

marc.r.schaedler@uni-oldenburg.de

Abstract

Physiologically motivated feature extraction methods based on 2D-Gabor filters have already been used successfully in robust automatic speech recognition (ASR) systems. Recently it was shown that a Mel Frequency Cepstral Coefficients (MFCC) baseline can be improved with physiologically motivated features extracted by a 2D-Gabor filter bank (GBFB). Besides physiologically inspired approaches to improve ASR systems technical ones, such as mean and variance normalization (MVN) or histogram equalization (HEQ), exist which aim to reduce undesired information from the speech representation by normalization. In this study we combine the physiologically inspired GBFB features with MVN and HEQ in comparison to MFCC features. Additionaly, MVN is applied at different stages of MFCC feature extraction in order to evaluate its effect to spectral, temporal or spectro-temporal patterns. We find that MVN/HEQ dramatically improve the robustness of MFCC and GBFB features on the Aurora 2 ASR task. While normalized MFCCs perform best with clean condition training, normalized GBFBs improve the ETSI MFCCs features with multi-condition training by 48%, outperforming the ETSI advanced front-end (AFE). The MVN, which may be interpreted as a normalization of modulation depth works best when applied to spectro-temporal patterns. HEQ was not found to perform better than MVN.

Index Terms: robust ASR, physiological Gabor filter bank features, modulation depth, normalization

1. Introduction

After decades of research in the area of automatic speech recognition (ASR) still no system exists that would equal humans ability to recognize speech. Especially in acoustically adverse conditions (background noise, spectral coloring, reverberation) there is a big gap in performance of about 15 dB between humans and machines. Tackling the long-term goal to improve the robustness of ASR systems to the level of humans, several approaches exist. One approach is to mimic the signal processing of the human auditory system or rather, to integrate its principles in terms of effective models into ASR systems. This proved to work for the well known part of the auditory system as today many robust ASR systems employ features based on a logarithmically scaled Mel-spectrogram like the one depicted in Fig. 1. This representation of speech roughly reflects the frequency selectivity and the compressive loudness perception of the human ear. Beyond the log Mel-Spectrogram there were several successful attempts to integrate single auditory principles, like the extraction of physiologically motivated [1] spectro-temporal patterns, into an ASR system to improve its robustness [2]. The early spectro-temporal features used additional processing with neural nets to improve a MFCC baseline. Recently, a filter bank of spectro-temporal filters which extracts features that can be used directly with GMM/HMM recognizers and improved a MFCC baseline was presented [3]. But generally, the use of the most detailed models of the auditory system does not result in the most robust ASR systems. One reason for this might be that the use of GMM/HMM based back-ends entrains certain restrictions on the feature characteristics. A different approach is therefore the use of statistical methods to better match the requirements of state-of-the-art GMM/HMM based back-ends. Normalization techniques like MVN [4] or HEQ [5] have shown to improve the robustness of systems based on traditional MFCC features. In this study both approaches are combined and normalization methods are applied to the physiologically motivated spectro-temporal Gabor filter bank (GBFB) features in comparison to traditional MFCC features. Further, the effect of MVN/HEQ is interpreted as a normalization of modulation depth and its

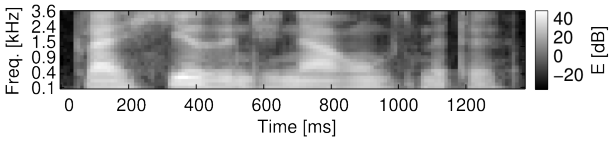

Figure 1: *Logarithmically scaled Mel-spectrogram of speech. Light areas denote high energy. The representation of speech through a log Mel-spectrogram is an element of many feature extraction algorithms for robust ASR systems.*

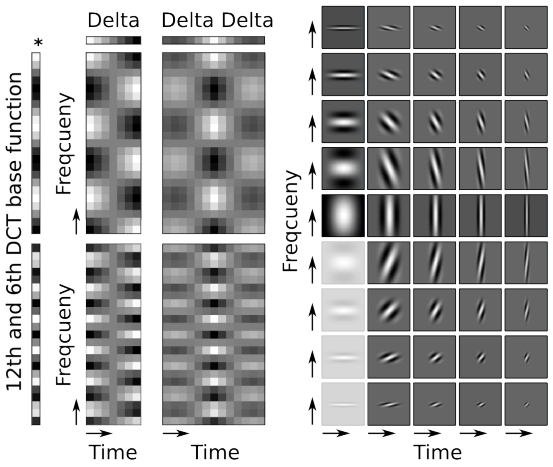

Figure 2: *Left panel: Effective spectro-temporal patterns of combined traditional spectral DCT and temporal $\Delta\&\Delta\Delta$ processing. Right panel: The 41 2D-Gabor filters that are used for feature extraction with the Gabor filter bank. The patterns are scaled and their real spectral extension is the same as of the MFCC-DD patterns in the left panel.*

effect on temporal, spectral, and spectro-temporal patterns is investigated.

2. Methods

2.1. Gabor filter bank features

The Gabor filter bank (GBFB))features are based on a log Mel-spectrogram with 23 Mel-bands between 64 Hz and 4 kHz, 10 ms window shift, and 25 ms window length. An examplary log Mel-spectrogram is depicted in Fig. 1. While for the extraction of MFCCs with $\Delta\&\Delta\Delta$ this spectro-temporal representation is processed spectrally with a DCT and temporally with slope-filters, GBFB features are extracted with 2D-Gabor filters that perform a simultaneous spectral and temporal processing. Fig. 2 depicts the relation of the spectro-temporal 2D-Gabor filters and the effective MFCC-DD spectro-temporal patterns. The outer product of a DCT base function and a Delta base function gives the effective spectro-temporal pattern that the corresponding MFCC-DD dimension encodes. The GBFB feature extraction is illustrated in Fig. 3. First, spectro-temporal patterns are extracted by 2D-convolving the 2D-Gabor filter functions with the log Mel-spectrogram. A subsequent selection of representative channels by critically sampling the filtered log Mel-spectrograms limits the systematical correlation of the feature dimensions. Each 2D-Gabor filter extracts patterns of a pair of a spectral and a temporal modulation frequency. These features were shown to improve the robustness of a MFCC baseline system when fed directly into an GMM/HMM recognizer [3]. The range of modulation frequencies covered is about 6 to 25 Hz and 0.03 to 0.25 $\frac{\text{cycles}}{\text{Mel-band}}$. Some properties of the GBFB features are compared with those of MFCC features in Tab. 1. The MFCC-DD processing can be described by separate spectral and temporal operations, while the GBFB processing cannot.

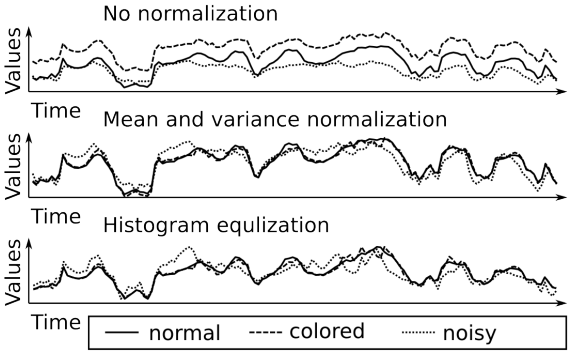

Figure 4: *Illustration of mean and variance normalization and histogram equalization of the first MFCC values for a speech signal in different acoustic contexts.*

Table 1: *Properties of MFCCs and GBFBs compared*

Feature	spectral	temporal	separable	dim.
MFCC-DD	DCT	$\Delta\&\Delta\Delta$	yes	39
GBFB	Gabor	Gabor	no	311

2.2. Normalization of feature value statistics

It has been shown that the robustness of an ASR system with MFCC features can be increased by removing the mean value and normalizing the variance of each feature dimension [4]. This processing is called mean and variance normalization (MVN) and normalizes the first and the second moments of the feature value distributions. An extension to MVN is mapping the feature values to a specific reference distribution [5]. This processing is called histogram equalization (HEQ) and normalizes all moments of the feature value distributions. The effect of MVN and HEQ on the first (not zeroth) MFCC is illustrated in Fig. 4. A spectral coloring (eg. preemphasis) of a speech signal leads to a systematic changes in the log Mel-spectrogram and consequently to a change of the derived features (cf. offset/mean value in Fig. 4 *colored*). Likewise, additive noise or reverberation result in a reduction of the dynamic range by filling up the "valleys" of the log Mel-spectrogram, which may be interpreted as a reduction of modulation depth (cf. scale/variance in Fig. 4 *noisy*). Applying MVN/HEQ to MFCC/GBFB features counteracts the influence of the most common sources of variability in noisy speech by normalizing the modulation depth, because the feature values scale lin-

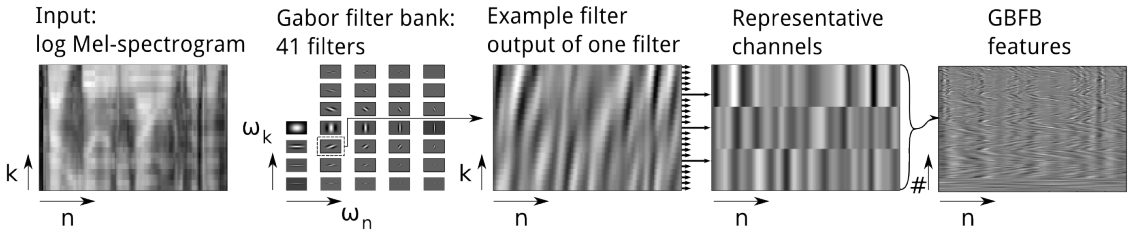

Figure 3: *Illustration of the Gabor filter bank feature extraction. n: temporal index; k: spectral index; ω: modulation frequencies. The input log Mel-spectrogram is filtered with each of the 41 filters of the Gabor filter bank. Representative channels of the filter outputs are selected and concatenated. The 311-dimensional output is used as feature vector.*

early with it. The recognition performance of GBFB and MFCC features is evaluated with and without MVN and HEQ.

2.3. Recognition experiment and baseline

The effect of the different front-ends on the robustness of an ASR system is evaluated within the Aurora 2 framework [6]. The task is the recognition of English connected digits which are contaminated with eight different everyday background noises from 20 dB to -5 dB. The framework provides speech data for training and testing as well as a GMM/HMM classifier and trainings rules. A reference setup defines whole-word left-to-right HMMs with 16 states, 3 mixtures per state, and without skips over states. The back-end is not modified and used with the same parameters as in the reference. Two different training conditions exist. For *clean* training only utterances *without* added noise are used, while for *multi* training utterances *with and without* added noise are used. Although only four noise types that occur in the testing data are also included in *multi* training data, it allows the recognizer to learn the reliability of feature patterns in noise. As reference features the first 13 MFCCs with first and second order discrete temporal derivative ($\Delta\&\Delta\Delta$) are used, resulting in 39-dimensional MFCC-DD features. Additionally the baseline results for ETSI MFCC [7] and ETSI Advanced Front-End (AFE) [8] features are reported. The word recognition accuracies are compared at signal-to-noise ratios (SNR) from 20 to -5 dB.

2.4. Spectral and temporal contribution

With the aim of evaluating the effect of normalizing only spectral, only temporal, or spectro-temporal patterns, the separability of spectral and the temporal processing with MFCC-DD features is exploited The normalization (N) is applied at the following stages of MFCC-DD feature calculation: MFCC-N-DD, MFCC-DD-N, DD-N-MFCC. With MFCC-N-DD features, spectral patterns are integrated by the DCT before normalization. With DD-N-MFCC features, short term temporal patterns are integrated by the $\Delta\&\Delta\Delta$ processing before normalization. And with MFCC-DD-N features, spectral and short term

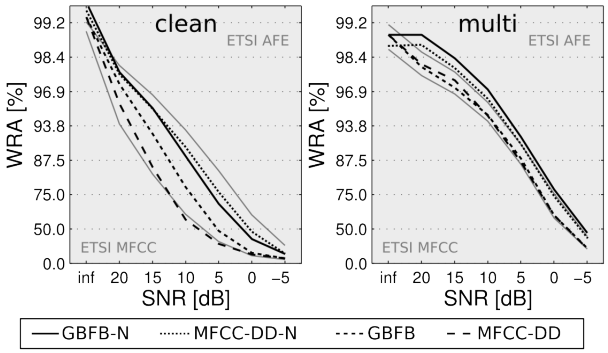

Figure 5: *Word recognition accuracies for GBFB and MFCC-DD feature with and without mean and variance normalization (N) at different test signal to noise ratios and training styles.*

temporal patterns are integrated before normalization. The recognition performance of the differently normalized features is evaluated.

3. Results

3.1. Normalized GBFB features

Average word recognition accuracies (WRA) for GBFB and MFCC features with and without MVN are reported in Fig. 5. With ***clean*** condition training MVN dramatically improves the robustness of MFCCs by 5-7 dB over a wide range of WRAs (50% to 95%). The improvements for GBFBs with 2-3 dB are smaller, but they perform about 3 dB better without MVN. Thus, MFCCs perform about 1 dB better than GBFBs at low SNRs, but cannot improve the highly optimized ETSI AFE baseline. However, GBFB features outperform all features when testing on clean data. In terms of average relative improvement over SNRs from 20 dB to 0 dB, MFCCs with MVN improve the WRA of the ETSI MFCC baseline by 58% on, while GBFBs with MVN improve the baseline by 54%. With ***multi*** condition training MVN improves the performance of MFCCs almost independently of the SNR by about 2-3 dB. For GBFB features the improvements are with 2.5 dB at low SNRs and up to 6 dB at high SNRs

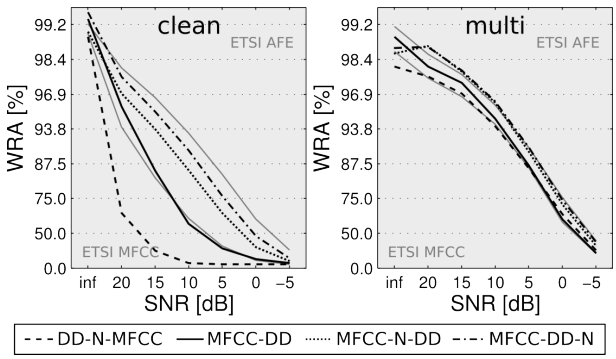

Figure 6: *Word recognition accuracies for MFCC-DD features with and without mean and variance normalization of spectral (MFCC-N-DD), temporal(DD-N-MFCC), and spectro-temporal (MFCC-DD-N) patterns at different signal to noise ratios for clean and multi style training.*

more pronounced. In terms of average relative improvement over SNRs from 20 dB to 0 dB, MFCCs with MVN improve the WRA of the ETSI MFCC baseline by 37%, while GBFBs with MVN improve the baseline by 48%. GBFB features outperform all other features, including ETSI AFE, in every noisy testing condition. The the improvements with HEQ were found to be similar to the improvements with MVN within a range of ± 1 dB and are therefore omitted. The very high recognition scores for clean testing data with *clean* condition training, as well as for high SNRs with *multi* condition training (which contains speech data at $\{\infty, 20, 15, 10, 5\}$ dB SNR) indicate a certain sensitivity of GBFB features to mismatched SNR conditions. Possibly, the 311-dimensional GBFB features encode more precise information about the speech signal than the 39-dimensional MFCC features which results in a higher sensitivity to the SNR. This finding puts the one-model-for-all-SNRs approach into question, as speech at 0 dB SNR and speech at 20 dB SNR have quite different characteristics. If the hypothesis holds, than GBFB features with MVN should perform even better in context-dependent models, which should be evaluated in future experiments.

3.2. Spectral vs. temporal normalization

Average word recognition accuracies (WRA) for MFCC features with and without MVN of spectral, temporal, and spectro-temporal patterns are depicted in Fig. 6. Normalizing the output after the temporal processing and before the spectral processing results in worse performance than without normalization, with an exception at very low SNRs with *multi* condition training. The MVN effectively normalizes all Mel-bands to have the same energy and the same modulation depth which seems to accompanied by a loss of information that is relevant for robust ASR. Normalizing the output after the spectral processing and before the temporal processing results in important improvements, but the best performance is achieved by normalizing after the spectral and temporal processing. This indicates that spectro-temporal patterns are best extracted from an unprocessed spectro-temporal representation and normalization is best performed after spectral and temporal integration.

4. Conclusions

The most important findings of this work can be summarized as follows:

- Normalization increases the robustness of physiologically motivated spectro-temporal Gabor filter bank features by 2.5-5 dB SNR on a digit recognition task, outperforming ETSI AFE features with multi-style training.
- Normalization of separable spectro-temporal patterns was found to be best applied after spectral and temporal integration.
- Normalized Gabor filter bank features seem work well in matched signal to noise ratio conditions, which should be further investigated with SNR-dependend models.

5. Acknowledgements

This work is funded by DFG SFB/TRR 31 "The active auditory system".

6. References

[1] A. Qiu, C. Schreiner, and M. Escabi, "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition," *J. Neurophysiology*, vol. 90, no. 1, pp. 456–476, 2003.

[2] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. of Interspeech 2002*, 2002, pp. 25–28.

[3] M. Schädler, B. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Am.*, vol. accepted, 2012.

[4] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[5] A. De La Torre, A. Peinado, J. Segura, J. Pérez-Córdoba, M. Benítez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 355–366, 2005.

[6] D. Pearce and H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ICSLP 2000*, vol. 4, 2000, pp. 29–32.

[7] ETSI standards document, "201 108 v. 1.1.3, speech processing transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," *European Telecommunications Standards Institute*, 2003.

[8] ——, "202 050 v. 1.1.5, speech processing transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *European Telecommunications Standards Institute*, 2007.