Evaluation of modulation-depth normalizing methods for the improvement of robust ASR systems with spectro-temporal features

Marc René Schädler and Birger Kollmeier
Medizinische Physik - Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Deutschland
Email: marc.r.schaedler@uni-oldenburg.de

Introduction

After decades of research in the area of automatic speech recognition (ASR) still no system exists that would match humans robustness. Especially under acoustically adverse conditions (background noise, spectral coloring, reverberation) there is a big gap in performance of about 15 dB between humans and machines. One approach to improve ASR recognition performance 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

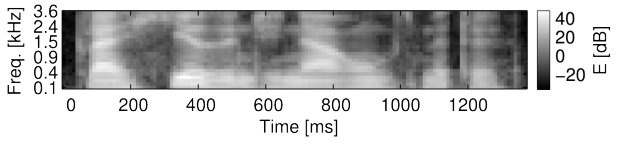

Figure 1: *Logarithmically scaled Mel-spectrogram of speech. Light areas denote high energy. The representation of speech through a log Mel-spectrogram is an element of many feature extraction algorithms for robust ASR systems.*

there were several successful attempts to integrate further auditory principles. In one of these a filter bank of physiologically motivated 2D-Gabor filters was used to extract spectro-temporal patterns [2]. But the use of more detailed models of the auditory system for robust speech feature extraction does not convince yet in terms of robustness. 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 [3] have shown to improve the robustness of systems based on traditional MFCC features. In this study both approaches are combined and MVN is applied to the physiologically motivated spectro-temporal GBFB features in comparison to traditional MFCC features.

Gabor filter bank features

The Gabor filter bank features (GBFB) 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 example is depicted in Fig. 1. While for the extraction

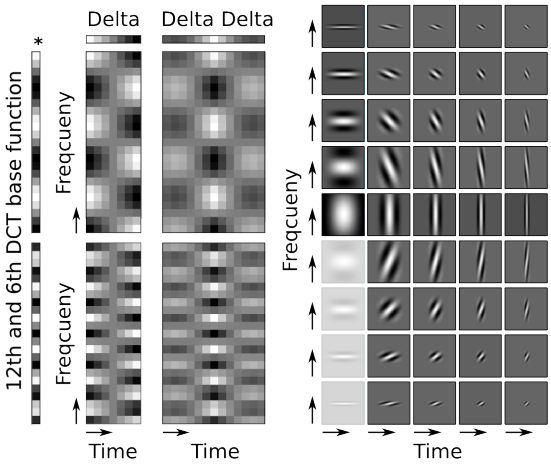

Figure 2: *Left panel: Effective spectro-temporal patterns of combined spectral DCT and temporal Δ&ΔΔ processing. Right panel: The 41 2D-Gabor filters that are used for feature extraction with the Gabor filter bank. The patterns are scaled. Their spectral extension is the same as of the MFCC-DD patterns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

of MFCCs 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 are extracted as follows. 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 of the resulting filtered log Mel-spectrograms limits the systematical correlation of the GBFB feature dimensions. Each 2D-Gabor filter extracts patterns of a certain spectral and temporal modulation frequency. The filter bank parameters used in this study are taken from the study that introduces the GBFB features [2]. The range of modulation frequencies covered by the 311-dimensional GBFB features is about 6 to 25 Hz and 0.03 to 0.25 $\frac{\text{cycles}}{\text{Mel-band}}$.

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 [3]. This processing is called mean and variance normalization (MVN) and its effect on the first MFCC is illustrated in Fig. 3. Applying MVN to

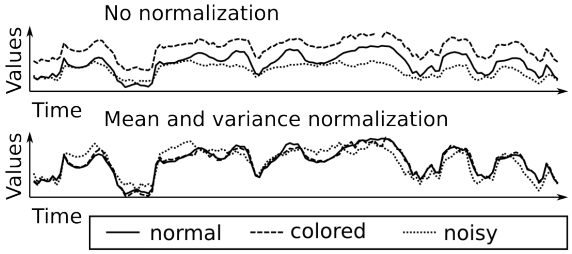

Figure 3: *Illustration of mean and variance normalization of the first MFCC values for a speech signal in different acoustic contexts.*

MFCC or GBFB features alleviates the influence of the most common variability in noisy speech. MFCC and GBFB feature values scale linearly with the modulation depth of the analyzed signal. In fact, the 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. 3 *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. 3 *noisy*). The recognition performance of GBFB and MFCC features is evaluated with and without MVN.

Recognition experiment and baseline

The effect of MVN on the robustness of an ASR system is evaluated within the Aurora 2 framework [1]. 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. 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. 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 and ETSI Advanced Front-End (AFE) features are reported. The word recognition accuracies are compared at signal-to-noise ratios (SNR) from 20 to -5 dB.

Results and discussion

Average word recognition accuracies (WRA) for GBFB and MFCC features with and without MVN are reported in Fig. 4. With ***clean*** condition training MVN dramatically improves the robustness of MFCCs by 5-7 dB over

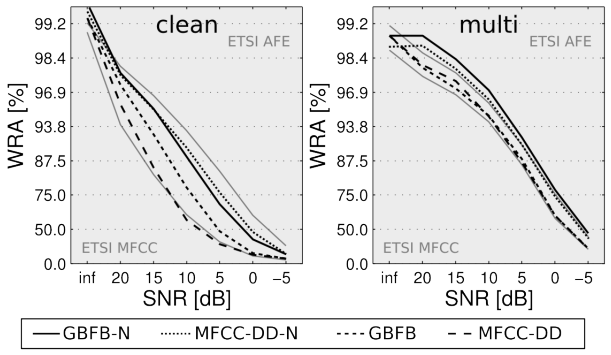

Figure 4: *Word recognition accuracies for GBFB and MFCC-DD feature with and without mean and variance normalization (N) at different test signal to noise ratios and for different training styles.*

a wide range of WRAs (50% to 95%). The improvements for GBFBs with about 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. 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 more pronounced. GBFB features with MVN outperform all other features, including ETSI AFE features, in every noisy testing condition.

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 should perform even better context-sensitive models, which should be evaluated in future experiments.

References

[1] D. Pearce and H.G. Hirsch. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. of ICSLP 2000*, volume 4, pages 29–32, 2000.

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

[3] O. Viikki and K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3):133–147, 1998.