Noise Robust Automatic Speech Recognition with Adaptive Quantile Based Noise Estimation and Speech Band Emphasizing Filter Bank

Casper Stork Bonde, Carina Graversen, Andreas Gregers Gregersen,
Kim Hoang Ngo, Kim Nørmark, Mikkel Purup,
Thomas Thorsen and Børge Lindberg.

Department of Communication Technology, Aalborg University, Niels Jernes Vej 12,
DK-9220 Aalborg Ø, e-mail lindberg@kom.aau.dk

Abstract. An important topic in Automatic Speech Recognition (ASR) is to reduce the effect of noise, in particular when mismatch exists between the training and application conditions.

Many noise robutness schemes within the feature processing domain use as a prerequisite a noise estimate prior to the appearance of the speech signal which require noise robust voice activity detection and assumptions of stationary noise. However, both of these requirements are often not met and it is therefore of particular interest to investigate methods like the Quantile Based Noise Estimation (QBNE) mehtod which estimates the noise during speech and non-speech sections without the use of a voice activity detector. While the standard QBNE-method uses a fixed pre-defined quantile accross all frequency bands, this paper suggests adaptive QBNE (AQBNE) which adapts the quantile individually to each frequency band.

Furthermore the paper investigates an alternative to the standard mel frequency cepstral coefficient filter bank (MFCC), an empirically chosen Speech Band Emphasizing filter bank (SBE), which improves the resolution in the speech band.

The combinations of AQBNE and SBE are tested on the Danish SpeechDat-Car database and compared to the performance achieved by the standards presented by the Aurora consortium (Aurora Baseline and Aurora Advanced Fronted). For the High Mismatch (HM) condition, the AQBNE achieves significantly better performance compared to the Aurora Baseline, both when combined with SBE and standard MFCC. AQBNE also outperforms the Aurora Baseline for the Medium Mismatch (MM) and Well Matched (WM) conditions. Though for all three conditions, the Aurora Advanced Frontend achieves superior performance, the AQBNE is still a relevant method to consider for small foot print applications.

1 Introduction

Car equipment control and the rapidly growing use of mobile phones in car environments have developed a strong need for noise robust automatic speech recognition systems. However, ASR in car environments is still far from achieving sufficient performance in the presence of noisy conditions.

While significant efforts are needed both in the acoustic modelling domain and in the feature processing domain of speech research, the present paper focusses on the latter in the context of standard Hidden Markov modelling of acoustic units.

One feature processing method proposed by Stahl et al. is Quantile Based Noise Estimation (QBNE) which estimates the noise during speech and non-speech sections without the explicit use of a voice activity detector [1].

In QBNE the noise estimate is based on a predefined quantile (q-value) being constant for all frequencies and independent of the characteristics of the noise. In an attempt to adapt to the noise characteristics of the data, this paper suggests adaptive QBNE (AQBNE) in which the q-value is determined independently for each frequency according to the characteristics of the data for that particular frequency. This gives a non-linear noise estimate compared to QBNE.

While the standard MFCC is motivated mainly from perceptual considerations, we suggest in this paper to consider a Speech Band Emphasizing (SBE) filter bank which has the purpose to better focus on the speech information available in the signal.

The paper is organised as follows. In section 2 QBNE, spectral subtraction, AQBNE and the SBE filter bank are described in further details. In section 3, the experimental framework, based on the Danish SpeechDat-Car database, is described. In section 4 the experimental results are presented and compared to the performance achieved by the Aurora consortium. In section 5 the conclusions are drawn.

2 Methods

2.1 Quantile Based Noise Estimation and Spectral Subtraction

QBNE, as proposed in [1], is a method based on the assumption that each frequency band contains only noise at least the q'th fraction of time, even during speech sections. This assumption is used to estimate a noise spectrum $N(\omega, i)$ from the spectrum of the observed signal $X(\omega, i)$ by taking the maximum value of the q-quantile in every frequency band. For every frequency band ω the frames of the entire utterance $X(\omega, i)$, $i = 0, \cdots, I$, are sorted so that $|X(\omega, i_0)| \leq |X(\omega, i_1)| \leq \cdots \leq |X(\omega, i_I)|$, where i denotes the frame number. This means that for each frequency band the data is sorted by amplitude in ascending order. From this the q-quantile noise estimate is defined as:

$$\hat{N}(\omega) = |X(\omega, i_{\lfloor qI \rfloor})| \tag{1}$$

Fig. 1 shows a plot of the sorted data for a sample utterance. The curves in the figure corresponds to different frequencies. The chosen q-value can be seen as a vertical dashed line in the figure. The intersection between the vertical line and each frequency curve is the noise estimate for the corresponding frequency.

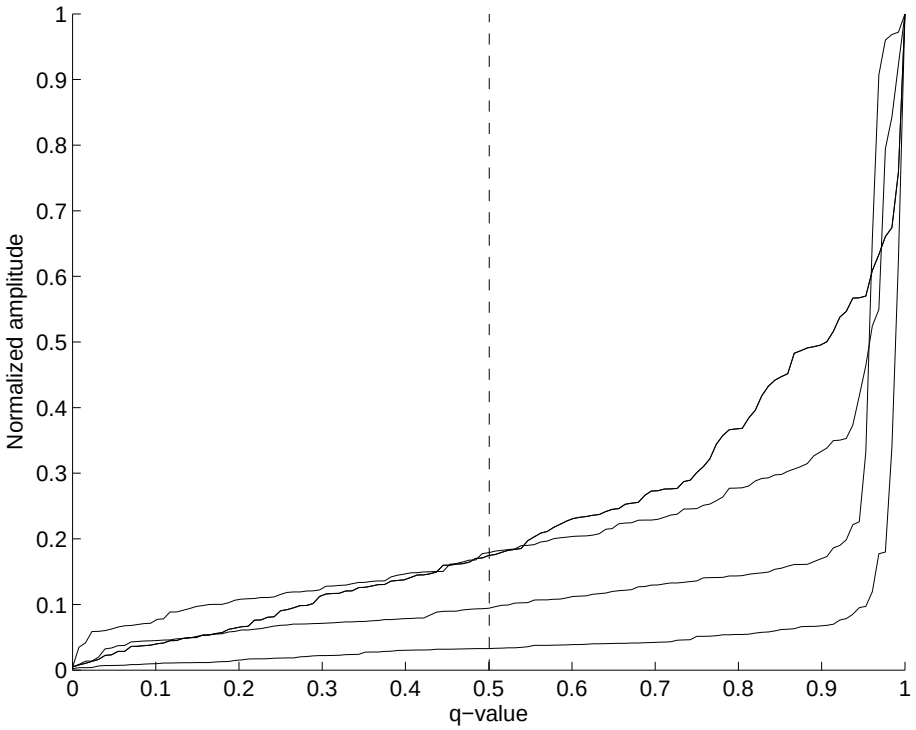

Fig. 1. Data sorted by amplitude for different frequency bands ($q = 0.5$). Only a few frequencies are plotted.

Ideally, the noise can be eliminated from the observed signal by applying amplitude spectral subtraction in which the observed signal $X(\omega)$ is modeled as a speech signal $S(\omega)$ to which uncorrelated noise $N(\omega)$ is added.

The amplitude spectrum of the speech is estimated by subtracting a weigthed amplitude spectrum of the estimated noise from the amplitude spectrum of the observed signal:

$$|\hat{S}(\omega)| = |X(\omega)| - \eta|\hat{N}(\omega)| \tag{2}$$

where η, the weighting constant, in the present experiments has been choosen to $\eta = 2.5$. This value was found to be an optimal value in [1]. To avoid negative amplitude, the implemented spectral subtraction in its final form is as follows:

$$|\hat{S}(\omega)| = \max\left(|X(\omega)| - \eta|\hat{N}(\omega)|\ ,\ \gamma|\hat{N}(\omega)|\right) \tag{3}$$

The value of γ, the fraction of the estimated noise, is set to 0.04, which has been found to be a reasonable value in [1].

2.2 Adaptive QBNE

The rationale behind AQBNE is to adapt to the utterances and noise levels by adjusting the quantile individually for each frequency band. The primary purpose of the method is to be able to train with low noise utterances (lab. recordings) and test with high noise utterances (application recordings) without the typical associated performance degradation.

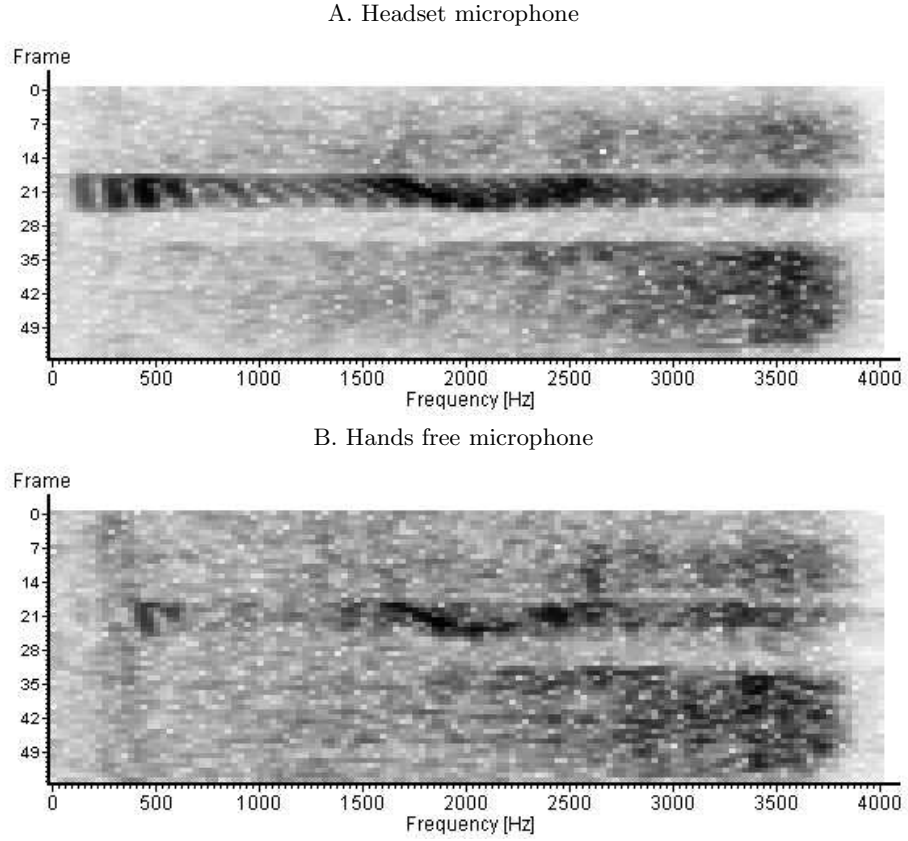

Fig. 2. Two spectograms of the same utterance recorded with two different microphones. A. with a headset microphone. B. with a hands free microphone at the rear view mirror.

When training with low noise data the recognizer models will contain a much more detailed description of the utterance compared to what is possible to obtain by eliminating the noise from a high noise test utterance. Fig. 2 shows two spectograms of an utterance synchroneously recorded with a headset microphone (A) and a hands free microphone placed at the rear view mirror (B). In Fig. 2B a larger part of the information in the speech is below the noise floor compared to Fig. 2A. These two signals represent typical training and test signals in a high mismatch scenario. The purpose of introducing AQBNE is two-fold. First, the purpose is to provide a better noise-estimate in the high-noise signals. Second, the purpose is to let the signal (A) be more equal to the signal (B) during training, resulting in a model that does not describe the utterance at a level of detail that is unobtainable under test. For the second purpose, however, experiments have then to reveal to what extent the increased performance under mismatched conditions is at the expense of decreased performance under well matched conditions.

AQBNE is developed by examining quantile plots as shown in Fig. 3. First observe that frequency bands with high noise contains more energy in the majority of the utterance than low noise frequency bands. This leads to the assumption that a smaller q-value is desired for the high noise frequency bands and a higher q-value is desired for the low noise frequency bands. Statistically a low fixed q-value corresponds to a low noise estimate for all frequency bands and a higher fixed q-value to a higher noise estimate. By adapting the q-value to the noise level of the frequency band the low and high noise utterances will converge to similar representations when the noise is eliminated, which subsequently should lead to better ASR performance.

In contrast to the fixed q-value method in QBNE (represented by a vertical line in Fig. 1 in subsection 2.1), the q-value in AQBNE is determined by a q-estimation curve as shown in Fig. 3. The intersection between this curve and each frequency curve is the noise estimate for the corresponding frequency. The vertical dashed line in Fig. 3 refers to the minimum desired q-value.

The q-estimation curve is defined as follows:

$$f(q) = e^{(q_{\min} - q)\tau} \tag{4}$$

where $q_{\min}$ is the minimum allowed q-value and τ is the slope of the curve.

Define $\tilde{q}$ as the q-value associated with the intersection of the q-estimation curve and the frequency curve, then $\tilde{q}$ is the solution of:

$$f(\tilde{q}) = |X(\omega, i_{\lfloor \tilde{q} I \rfloor})| \tag{5}$$

The noise estimate is then defined by

$$\hat{N}(\omega) = |X(\omega, i_{\lfloor \tilde{q} I \rfloor})| \tag{6}$$

and subsequently used as the noise estimate during spectral subtraction.

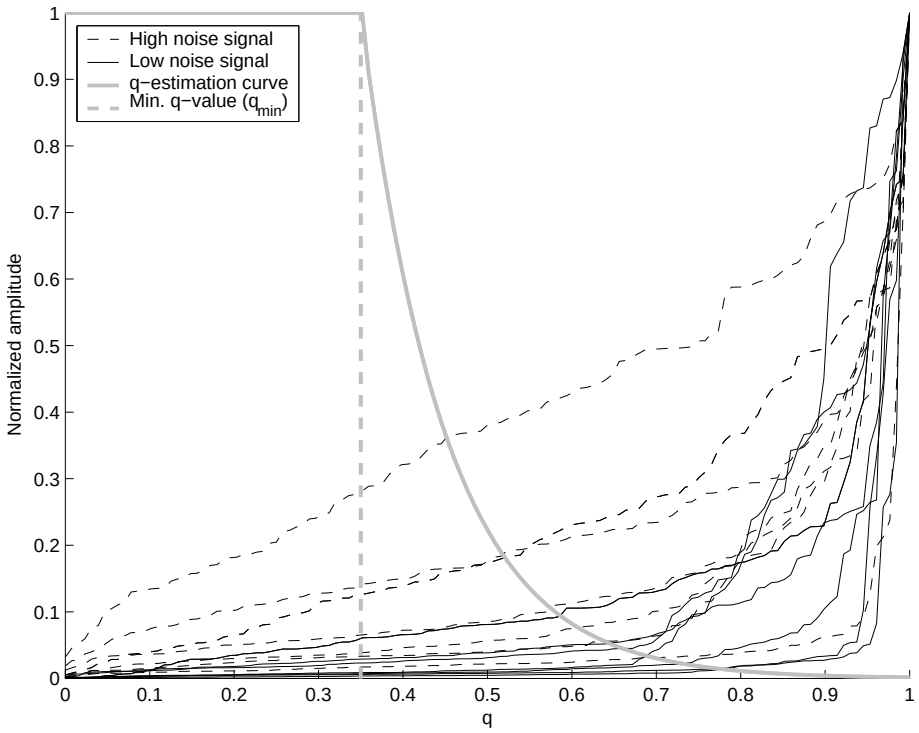

Fig. 3. Data sorted by amplitude for different frequency bands for a low noise (solid black) and high noise (dashed black) speech signal. The solid gray curve is the q-estimation curve, and the dashed gray line is the minimum allowed q-value.

2.3 Speech Band Emphasizing Filter Bank

The predominant parametric representation of features in speech is Mel Frequency Cepstrum Coefficients (MFCC), introduced in 1980 by Davis and Mermelstein as an improvement to Linear Frequency Cepstrum Coefficients (LFCC) [2].

The method compresses the spectral information by applying the psycho acoustic theory of critical bands combined with a mel scale warping of the frequency axis, in order to more closely resemble human perception.

The method is implemented by applying a mel filter bank to the speech signal consisting of half overlapping triangular filters linearly distributed on the mel scale:

$$\text{Hertz2Mel}\{f\} = 2595 \log\left(1 + \frac{f}{700}\right) \tag{7}$$

The output from each filter is integrated to produce the reduced and warped spectrum which is then transformed into cepstrum coefficients. Fig. 4 (top) shows

the mel filter bank with 23 triangular filters as specified in the ETSI ES 201 108 standard [3].

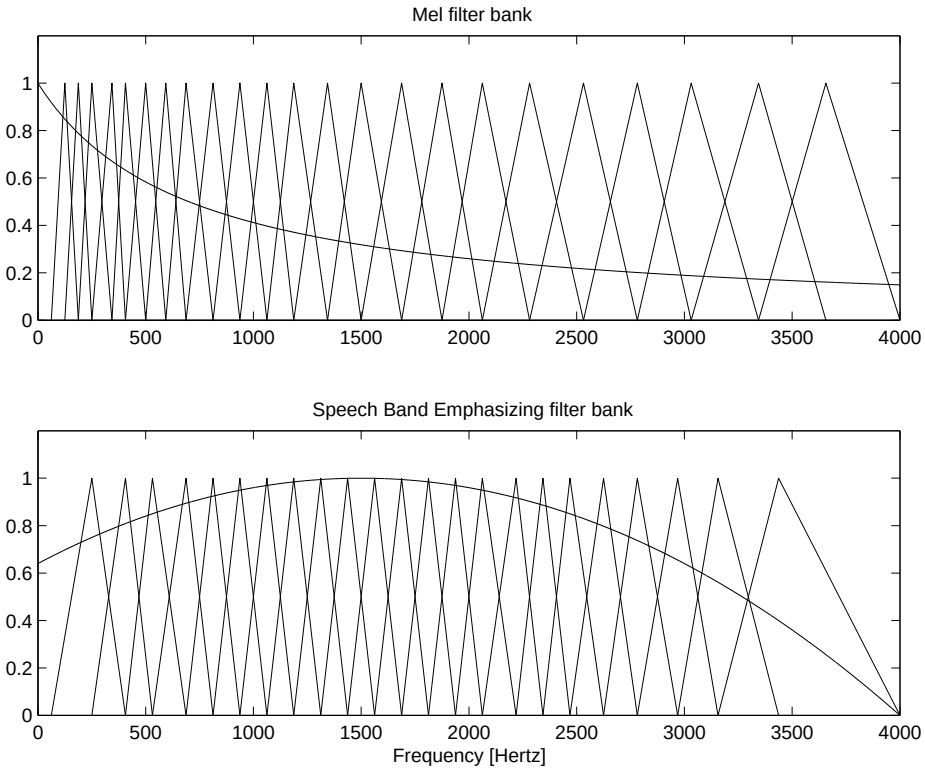

Fig. 4. The mel filter bank compared to the Speech Band Emphasizing filter bank (SBE). The continuous lines are the normalized derivatives of the mapping functions of Eq. 7 and 8indicating the concentration of triangular filters (importance functions).

While both the mel-scale and critical band assumptions are well founded theoretically as well as experimentally they do not necessarily translate to robust ASR. As Fig. 4 (top) reveals the filters in the mel filter bank are concentrated with maximum resolution at low frequencies.

With the purpose to better focus on the speech information avaliable in the signal, we will investigate a speech band emphasizing (SBE) filter bank, where the filter concentration is empirically chosen to be highest at 1500 Hz and to decrease with higher and lower frequencies. To define the distribution using the existing framework the mapping function between Hertz and mel is replaced by a new mapping function between Hertz and the imaginary unit SBE:

$$\text{Hertz2SBE}\{f\} = 12\,000\,000 f + 4500 f^2 - f^3 \tag{8}$$

Instead of distributing the triangular filters linearly on the mel scale the filters are now distributed linearly on the SBE scale, which is defined by Eq. 8. The resulting SBE filter bank is illustrated in Fig. 4 (bottom).

The mapping function in Eq. 8 is obtained by first defining an *importance function*. This function attains values between 0 and 1, where 1 indicates "most important" and 0 indicates "least important". The importance function of the SBE scale is defined by the polynomial (depicted as a continuos line in Fig. 4 (bottom)):

$$i(f) = \frac{-(f - 1500\text{Hz})^2}{(4000\text{Hz} - 1500\text{Hz})^2} + 1 \tag{9}$$

The mapping function is the indefinite integral of the importance function with the arbitrary constant discarded (or set to a convenient value):

$$Hertz2SBE\{f\} = I(f) = \int i(f)df \tag{10}$$

Because the unit SBE has no physical meaning or interpretation, any scale factor can be applied to simplify the expression. In particular Eq. 8 is scaled by 18 750 000 to make the highest order coefficient unity.

3 Experimental Framework

To evaluate the methodology an experimental framework based on the Danish SpeechDat-Car Digits Database is used, which is part of the Aurora-3 database[4]. The corpus comprises 2457 utterances recorded in a car under different noise conditions, 265 utterances for quiet (motor idling, car stopped), 1513 utterances for low noise (town traffic or low speed on rough road) and 679 utterances for high noise (high speed on rough road). Each utterance is recorded synchroneously by two microphones Close Talking (CT) and Hands Free (HF) using a sample rate of 16 kHz and a 16 bit quantization. This results in a total of 4914 speech recordings.

Training and test definitions proposed by the Aurora consortium is used which divide the test in three parts: Well Match (WM), Medium Mismatch (MM) and High Mismatch (HM). In WM training and test are performed under all conditions for both microphones. Because of the matched conditions under training and test the WM scenario does not reveal the frontends ability to adapt to unexpected noise during test. In MM training is performed at low noise level and tested at high noise level, both with the HF microphone. MM is primarily a test of the frontends ability to suppress additive noise. In HM training is performed at all noise levels with the CT microphone and tested at high noise level with the HF microphone. HM is primarily a test of independence of microphone type, placement and distance relative to the speaker.

The speech recognizer used in Aurora is based on whole word HMMs. The structure of each HMM is a simple left-to-right model 16-state 3-mixture per state for each digit. The silence model is a 3-state 6-mixture and the short pause

Table 1. Results with mel filter bank.

Frontend	**WM**	**MM**	**HM**	**Average**	**AI**
Aurora Baseline	80.13	51.56	33.61	58.50	-
Aurora Adv. Frontend	93.37	81.49	79.59	85.77	-
QBNE					
q=0.40	**84.60**	66.80	21.72	62.65	3.73
q=0.45	84.19	**68.10**	35.95	**66.50**	**14.99**
q=0.50	83.77	65.23	38.05	65.85	14.40
q=0.55	81.58	63.80	38.63	64.62	12.77
q=0.60	80.46	63.15	**40.68**	64.46	13.29
Adaptive QBNE					
q_{min}=0.30, τ=10	83.77	64.06	**53.68**	**69.35**	**25.23**
q_{min}=0.30, τ=15	**85.12**	**67.45**	36.32	66.74	15.29
q_{min}=0.35, τ=10	82.60	65.36	48.46	68.03	21.65
q_{min}=0.35, τ=15	83.10	66.67	41.22	66.88	17.40
q_{min}=0.40, τ=10	81.58	64.97	51.95	68.36	23.47
q_{min}=0.40, τ=15	82.24	65.49	48.54	67.95	21.61
q_{min}=0.45, τ=10	80.33	65.23	18.84	59.67	-1.61
q_{min}=0.45, τ=15	81.05	66.02	49.81	67.98	22.32

Table 2. Results with SBE filter bank.

Frontend	**WM**	**MM**	**HM**	**Average**	**AI**
SBE baseline	82.04	54.43	31.59	59.76	1.40
SBE + Adv. Frontend	N/A	N/A	N/A	-	-
QBNE					
q=0.40	**84.94**	67.97	12.88	60.99	-1.88
q=0.45	83.49	**71.74**	32.21	66.56	14.33
q=0.50	83.75	70.83	**46.15**	**69.83**	**24.22**
q=0.55	81.22	68.49	23.20	62.26	4.29
q=0.60	80.78	65.76	36.82	64.53	12.35
Adaptive QBNE					
q_{min}=0.30, τ=10	81.82	69.14	45.33	68.26	21.49
q_{min}=0.30, τ=15	**83.92**	**70.57**	31.47	66.14	13.20
q_{min}=0.35, τ=10	82.00	66.93	48.79	68.42	22.66
q_{min}=0.35, τ=15	81.84	69.66	36.65	66.28	15.40
q_{min}=0.40, τ=10	81.22	62.37	**60.63**	69.48	**27.98**
q_{min}=0.40, τ=15	82.69	68.36	49.81	69.45	24.73
q_{min}=0.45, τ=10	79.27	66.02	42.86	65.53	16.27
q_{min}=0.45, τ=15	81.84	67.45	52.90	**69.57**	25.99

Average is weighted 0.4WM+0.35MM+0.25HM. **AI** is the weighted average of the improvements for each of the three test conditions compared to *baseline*. Best accuracy in each category is in boldface.

model is a 1-state 6-mixture. The HMM parameters are estimated using Viterbi training and Baum-Welch re-estimation procedure. To build and manipulate the HMM's the Hidden Markov Model Toolkit (HTK) is used [5].

4 Experimental results

The baseline reference is a 32 bit floating point precision C implementation of the Aurora standard frontend (WI007) which is supplied with SpeechDat-Car and fully specified in the ETSI ES 201 108 standard [3].

To enable easy implementation of the various methods an extensible clone of this frontend has been written in Java using 64 bit floating point precision. This frontend is the baseline for all relative improvement calculations, and will be denoted *baseline*. This frontend has also been implemented with the SBE filter bank in place of the mel filter bank. It is denoted *SBE baseline*. All tested methods are implemented as extensions to these two frontends.

The results are reported in table 1 for methods implemented on the *baseline* frontend, and in table 2 for methods implemented on the *SBE baseline* frontend. Table 1 also shows the equivalent results obtained using the Aurora Advanced Frontend as reported in [6].

The first three columns are the accuracies obtained for each of the three test conditions. The averages are calculated as a weighting of these three results with the equation 0.4WM+0.35MM+0.25HM. The rightmost column is a weighted average of the improvement in each of the three results compared to *baseline*, using the same weighting. For each method the best accuracies are typeset in boldface.

The SBE filter bank performs slightly better than the mel filter bank. QBNE shows even greater improvements, especially with the SBE filter bank. Adaptive QBNE which sought to improve high mismatch performance succeeds ind this regard, yielding even better results than QBNE, again especially in combination with the SBE filter bank.

Fig. 5 shows a graphical approach of method comparison. For each method and test condition, the best score is selected, and the improvement to *baseline* is calculated. The white columns are methods implemented on *baseline* (mel filter bank) and grey columns are methods implemented on *SBE baseline* (SBE filter bank).

AQBNE is a clear advantage over QBNE in the high mismatch scenario with an 80.39% improvement compared to a 21.04% improvement for QBNE with mel filter bank. It achieves this without significantly compromising the performance in WM and MM, reflected in an average improvement of 27.98% compared to 14.99% for QBNE with mel filter bank. Given the very small computational impact of the method, compared to QBNE, it is a relevant improvement in all ASR applications.

Furthermore it is evident that the SBE filter bank in general improves recognition performance, and especially combines well with both QBNE and AQBNE.

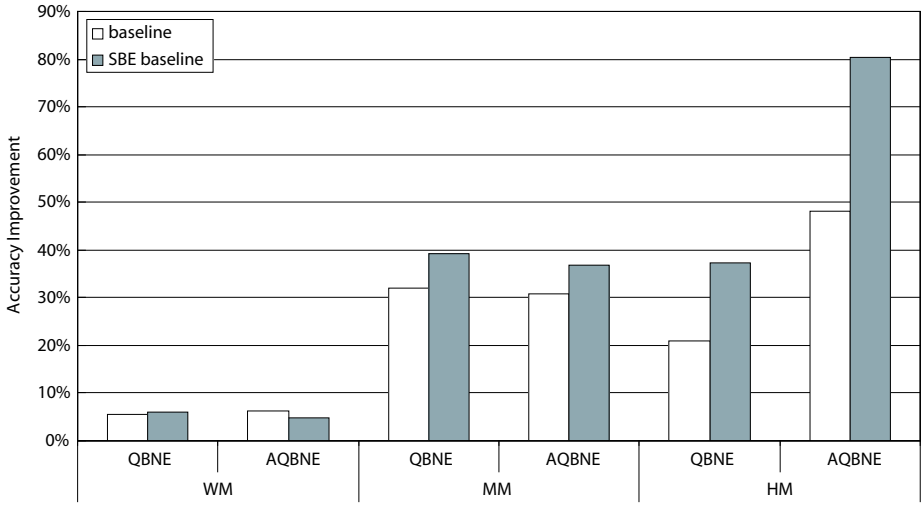

Fig. 5. Comparison of methods by improvement to *baseline*. For each method, the best result for each of the three test conditions is selected.

Finally, it is observed that for MFCC the Aurora Advanced Frontend achieves superior performance compared to AQBNE. So far the SBE frontend has not been integrated with the Aurora Advanced Frontend in order to compare the two feature representations.

5 Conclusion

In this paper the task of eliminating unwanted noise from a speech signal recorded in a noisy car environment has been investigated. Two methods have been considered seperately and in combination.

The first method is Quantile Based Noise Estimation (QBNE), which estimates the noise during both speech and non-speech sections. Adaptive QBNE (AQBNE) is suggested and concluded to improve the standard QBNE. In addition an empirically chosen Speech Band Emphasizing filter bank (SBE) is suggested as an alternative to the standard mel filter bank.

The experiments conducted on the Danish SpeechDat-Car database show that SBE generally show an improvement over the standard mel filter bank (Aurora baseline), achieving significant increases in recognition performance when combined with both QBNE and AQBNE.

Specifically AQBNE with SBE achieved remarkable results under highly mismatched training and test conditions with an 80.39% improvement compared to a 21.04% improvement for QBNE with mel filter bank. The average improvement for SBE with AQBNE was as high as 27.98% compared to 14.99% for QBNE with mel filter bank.

Though for all test conditions, the Aurora Advanced Frontend achieves superior performance compared to AQBNE, the AQBNE is still a relevant method to consider for small foot print applications. In addition, it is relevant to investigate the integration of the AQBNE and SBE methods into the Aurora Advanced Frontend.

References

1. Volker Stahl, Alexander Fischer and Rolf Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering", pp. 1-4, ICSLP 2000.
2. Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", pp. 357-366, IEEE Trans. Acoustics, Speech, Signal Processing, vol. 28(4).
3. European Telecommunications Standards Institute, "ES 201 108 v.1.1.2", http://www.etsi.org/, 2000.
4. Asuncin Moreno, Boerge Lindberg, Christoph Draxler, Gal Richard, Khalid Choukri, Stephan Euler and Jeffrey Allen, "SpeechDat-Car. A Large Speech Database for Automotive Environments", pp. 1-6, LREC 2000.
5. Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev and Phil Woodland, "The HTK Book (for HTK Version 3.2.1)", http://htk.eng.cam.ac.uk, 2002.
6. Dusan Macho, Laurent Mauurary, Bernhard No, Yan Ming Cheng, Doug Ealey, Denis Jouver, Holly Kelleher, David Pearce and Fabien Saadoun, "Evaluation of a Noise-Robust DSR Front-end on Aurora Databases", pp. 17-21, Proc. ICSLP 2002, Denver, Colorado