ADAPTIVE DIFFERENTIAL MICROPHONE ARRAYS USED AS A FRONT-END FOR AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Elmar Messner, Hannes Pessentheiner, Juan A. Morales-Cordovilla, Martin Hagmüller

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

ABSTRACT

For automatic speech recognition (ASR) systems it is important that the input signal mainly contains the desired speech signal. For a compact arrangement, differential microphone arrays (DMAs) are a suitable choice as front-end of ASR systems. The limiting factor of DMAs is the white noise gain, which can be treated by the minimum norm solution (MNS). In this paper, we introduce the first time the MNS to adaptive differential microphone arrays. We compare its effect to the conventional implementation when used as front-end of an ASR system. In experiments we show that the proposed algorithms consistently increase the word accuracy up to 50 % relative to their conventional implementations. For PESQ we achieve an improvement of up to 0.1 points.

Index Terms— beamforming, differential microphone arrays (DMAs), automatic speech recognition (ASR), microelectromechanical systems (MEMS) microphones

1. INTRODUCTION

Voice recording is a simple task that can be achieved by means of a single directional microphone. The use of a uni-directional microphone is not always satisfactory, since every 4 - 5 dB improvement of the SNR may raise the speech intelligibility by 50 % [1]. In realistic scenarios, the captured signal consists of a desired speech signal and other interfering signals, e.g. music, speech, noise, etc. In this work we consider a system that is able to record the target speaker and to simultaneously suppress interfering sources. This can be realized by means of microphone arrays and beamforming algorithms. For a compact arrangement and limited resources, differential microphone arrays (DMAs) can be used.

The usage of adaptive differential microphone arrays (ADMAs) is limited by the so called white noise gain [2], which renders second- and higher-order implementations impractical. The authors of [2] present the minimum-norm solution (MNS) for DMAs, which features a higher robustness against the white noise gain. However, to the best of our knowledge, MNS has never been used in ADMAs, and the effect on ASR is not investigated. In this paper we apply the MNS in ADMAs and compare them with the conventional implementations, used as a front-end for an ASR system. In our experiments we consider close-talking speaker scenarios in a reverberant environment with up to three interferer and SNR values from -6 dB to 12 dB. Not surprising, ADMAs show a clear and consistent improvement over a single omnidirectional microphone in terms of perceptual evaluation of speech quality (PESQ) and word accuracy rates (WAcc). Furthermore, ADMAs with MNS consistently outperform the conventional implementation.

The paper is organized as follows. Sections 2 and 3 present the theory of the algorithms and Section 4 describes their implementation. Section 5 gives an overview on the recordings that were made for the evaluation of the algorithms and Section 6 presents the results. Section 7 concludes the paper.

The authors acknowledge funding by the European project DIRHA FP7-ICT-2011-7-288121 and the K-Project ASD funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria and The Technology Agency of the City of Vienna (ZIT). The programme COMET is conducted by Austrian Research Promotion Agency (FFG).

2. ADAPTIVE DMAS

References [3] and [4] present the realization of a DMA with variable beamformers. These beamformers are suppressing the interfering sources by directly nullforming towards the corresponding directions. The adaptive beamformer combines the output signals of the fixed beamformer to obtain the final beamformer output. Figure 1 shows the schematic implementation.

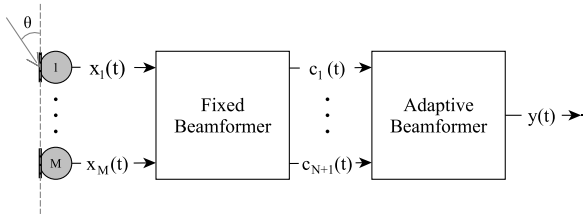


Fig. 1. *Schematic implementation of an ADMA. M ... number of microphones, N ... Order of the DMA. $c_n(t)$ ... output signal of fixed beamformer.*

2.1. First-Order ADMA

The conventional first-order-implementation of the ADMA [3] needs $M = N + 1 = 2$ microphones. The fixed beamformer combines the microphone signals to form its output signals. The frequency and angular dependent responses of the fixed beamformer are

$$C_1(\omega, \theta) = \begin{bmatrix} 1 & e^{-j\omega\tau_0 \cos\theta} \end{bmatrix} \begin{bmatrix} 1 \\ -e^{-j\omega\tau_0} \end{bmatrix} S(\omega) \qquad (1)$$

$$C_2(\omega, \theta) = \begin{bmatrix} 1 & e^{-j\omega\tau_0 \cos\theta} \end{bmatrix} \begin{bmatrix} -e^{-j\omega\tau_0} \\ 1 \end{bmatrix} S(\omega), \qquad (2)$$

where $S(\omega)$ is the spectrum of the signal source, ω is the angular frequency, θ is the azimuthal angle and $\tau_0 = \delta/c$ is the delay with the speed of sound c and the microphone distance δ (cf. Fig. 2(a)). The approximate speed of sound in dry (0% humidity) air is

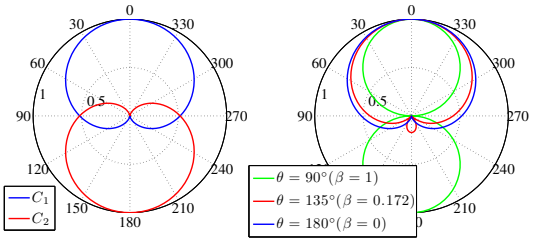
$c = (331.1 + 0.0606\vartheta)$, where ϑ is the temperature in degrees Celsius ($^\circ C$). These signals are adaptively combined to obtain the final beamformer output signal. The beamformer output normalized by the input spectrum $S(\omega)$ is

$$\left|\frac{Y(\omega,\theta)}{S(\omega)}\right| = \left|\left(C_1(\omega,\theta) - \beta C_2(\omega,\theta)\right) H_L(\omega)\right|, \quad (3)$$

where β is a real constant and $H_L(\omega)$ the compensation filter. The resulting beam pattern depends on the value of β, ranging between $0 \leq \beta \leq 1$. The NLMS-algorithm updates the value of β. The update equation written in the time-domain is

$$\beta_{t+1} = \beta_t + \mu \frac{y(t)c_2(t)}{\|c_2(t)^2\| + \Delta}, \quad (4)$$

with the step-size μ and the regularization parameter Δ. Figure 2(b) depicts the beam pattern of the beamformer output for different values of β.

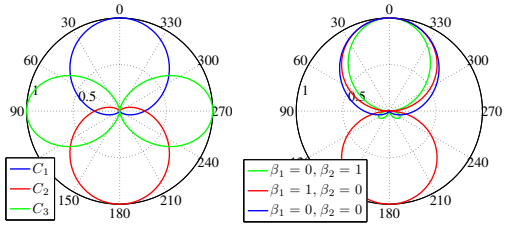


(a) (b)

Fig. 2. *Beam patterns of the first-order ADMA: (a) Fixed beamformer outputs; (b) Beamformer output for different values of β.*

2.2. Second-Order ADMA

The conventional second-order-implementation of the ADMA [4] needs $M = N + 1 = 3$ microphones for the fixed beamformer. The fixed beamformer provides three output signals. These three output signals are adaptively combined to obtain the final beamformer output. Figure 3 depicts the corresponding beam patterns. The second-order ADMA is able to place two distinct zeros in the output beam pattern (the first-order ADMA only one).



(a) (b)

Fig. 3. *Beam patterns of the second-order ADMA: (a) Fixed beamformer outputs; (b) Beamformer output for different values of β.*

3. NOVEL ROBUST ADAPTIVE DMAS

3.1. Robust First-Order ADMA

Due to the compensation of the high-pass characteristics of DMAs (a slope of 6 dB/octave for first-order DMAs) the so-called white noise gain arises [2]. An approach to reduce the white noise gain is the implementation with a microphone number $M > N + 1$. The authors of [2] realize this with the minimum-norm solution. For a more robust implementation of the first-order ADMA we implement the fixed beamformer with this approach. Figure 4 depicts the schematic implementation for this novel fixed beamformer. The closed form solution for the filter elements is

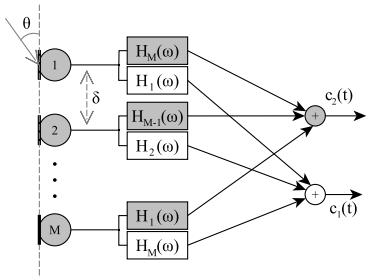


Fig. 4. *Schematic implementation of the novel fixed beamformer of a first-order ADMA with the minimum-norm solution.*

$$\mathbf{h}(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{D}^T(\omega, \boldsymbol{\alpha})[\mathbf{D}(\omega, \boldsymbol{\alpha})\mathbf{D}^T(\omega, \boldsymbol{\alpha})]^{-1}\boldsymbol{\beta}, \quad (5)$$

where $\mathbf{D}^T(\omega, \boldsymbol{\alpha})$ is the constraint matrix of size $M \times (N + 1)$ and the design vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The parameters to design a first-order cardioid are:

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 & -1 \end{bmatrix}^T, \quad (6)$$

$$\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 \end{bmatrix}^T. \quad (7)$$

The constraint matrix for $M = 4$ microphones is

$$\mathbf{D}(\omega, \boldsymbol{\alpha}) = \begin{bmatrix} 1 & e^{-j\omega\tau_0} & e^{-j2\omega\tau_0} & e^{-j3\omega\tau_0} \\ 1 & e^{j\omega\tau_0} & e^{j2\omega\tau_0} & e^{j3\omega\tau_0} \end{bmatrix}. \quad (8)$$

We obtain the solution for the filter vector $\mathbf{h}(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by solving Eq. 5.

3.2. Robust Second-Order ADMA

The second-order DMA ($M = 3$) features a high-pass characteristic with a slope of 12 dB/octave that has to be compensated. This entails a stronger amplification of the white noise compared to the first-order DMA. Figure 5 shows the schematic implementation of the novel fixed beamformer for a second-order ADMA with the minimum-norm solution. In the first stage we apply two first-order ADMA fixed beamformer for $M - 1$ microphones (cf. Fig. 4). In the second stage we consider three conventional first-order DMAs to form the three fixed beamformers' output signals. For further details see [5].

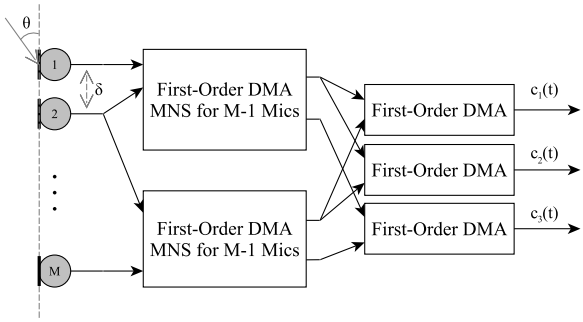


Fig. 5. *Schematic implementation of the novel fixed beamformer of a second-order ADMA with the minimum-norm solution.*

3.3. Robust First/Second-Order Hybrid ADMA

Although the MNS, applied for the second-order ADMA, entails an enhancement regarding the white noise gain, the amplification in the low frequency range is still too high for a real usage. An approach that allows to utilize a second-order ADMA in real applications is a hybrid version in combination with a first-order ADMA [6]. A first-order ADMA (with $M-1$ microphones) operates in the low frequency range and above the transition frequency f_t operates a second-order ADMA.

4. IMPLEMENTATION

We investigated the following implementations of the ADMAs:

- First-order ADMA ($M = 2$)
- Robust first-order ADMA ($M = 4$)
- First/second-order hybrid ADMA: ($M = 3$)
- Robust first/second-order hybrid ADMA ($M = 5$)

The implementation of each algorithm is based on block processing with the overlap-add method and 50% overlapping. The used window-type is Hanning and the sampling frequency $f_s = 48$ kHz. The frame size for the block-processing is 2^8 samples. The value for the step-size is $\mu = 0.6$ and the regularization constant is $\Delta = 10^{-4}$. The compensation filter features an amplification of infinity at $f = 0$ Hz; thus, the first frequency pin for the designed filter is set to zero.

For the first-/second-order hybrid ADMA ($M = 3$) the transition frequency is $f_t = 1850$ Hz, and for the robust first-/second-order hybrid ADMA ($M = 5$) it is $f_t = 1050$ Hz.

5. RECORDINGS

For the design of DMAs the microphone distance has to be very small. No speech-corpus is available for this microphone array setup. Therefore, we designed a small linear microphone array. We investigated the performance of the algorithms in a small conference room. We simulated different realistic scenarios with a target speaker and up to three interfering speakers.

5.1. Recording Environment

The recordings took place in a small conference room ($5.99 \times 5.33 \times 3.13$ m) at the Signal Processing and Speech Communication Laboratory (SPSC Lab) at the TU Graz. The temperature in the room varied during the recordings between $\vartheta = 31°$C and $\vartheta = 33°$C. We placed the microphone array at the center of the room and surrounded it by four loudspeakers, distributed on a circle with a radius of $r = 1$ m (see Fig. 6). The height of the top of the microphone array with respect to the floor is $h_{MA} = 1,25$ m. We mounted the loudspeakers on a height of $h_{LS} = 1,21$ m, measured from their bottom. The first loudspeaker (LS1) is acting as the target speaker and the rest as interfering speakers coming from different directions. As a reference for the sound pressure level we adjusted the loudspeakers to reach an A-weighted equivalent sound level of $L_{Aeq} = 80$ dB by playing back white Gaussian noise.

5.2. Recording Equipment

The playback setup consists of *Yamaha MSP5 Studio Loudspeakers* connected with the audio interface *Focusrite Liquid Saffire 56*. For playback and recording we used the real-time graphical dataflow programming environment PureData.

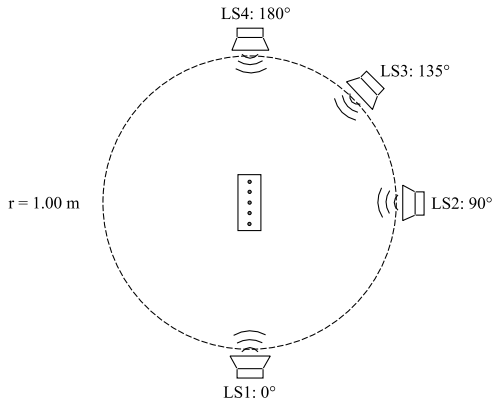


Fig. 6. *Recording setup.*

The MP34DT01 are omnidirectional, digital MEMS microphones with a size of $3 \times 4 \times 1$ mm. They exhibit a frequency range of 20 Hz to 16000 Hz and feature a SNR of 63 dB. Up to eight microphones are operating on the STM32 MEMS microphones application board.

We mounted the microphones on a microphone array grid with the dimensions $9.7 \times 4.8 \times 0.5$ cm. The distance between two adjacent microphones of the linear microphone array is $\delta = 2.14$ cm.

5.3. Playback

We generated the playback signals with MATLAB. For each scenario we generated four 4-channel WAVE files, each with a different SNR (-6dB, 0 dB, 6 dB and 12 dB). The target speaker signal consists of a sequence of German commands from the male speaker 001 of the GRASS corpus [7]. Within one minute we played back 24 commands. The target speaker is present in each scenario with the same level. We played back the interfering speakers [7] from different direction ($90°$, $135°$ and $180°$), whereas the target speaker had a fixed position ($0°$). Also the number of interfering speakers is changing ($\# = 1, 2$ and 3). Each scenario lasts one minute.

6. RESULTS

We evalueted the performance of the ADMAs by means of the PESQ and ASR Word Accuracy Rate (WAcc). For the estimation of the WAcc, a short description of the ASR engine follows.

6.1. Speech Database

The training material consists of a clean training set, i.e. without reverberation. This contains 5046 isolated utterances corresponding to 55 male and female speakers: 19 GRASS [7] speakers (with different commands, keywords, and read sentences than in the test set) and 36 PHONDAT-1 [8] speakers. We mixed two databases to make the recognition more robust to speaker variation. The training sets include the speaker 001 [7].

6.2. ASR Engine

The front-end and the back-end of the ASR Engine are HTK-based recognizers [9, 10]. This recognizer is appropriate for a medium vocabulary size. The front-end takes the enhanced signal and obtains mel frequency cepstrum coefficients (MFCCs) using: 16 kHz sampling frequency, frame shift and length of 10 and 32ms, 1024

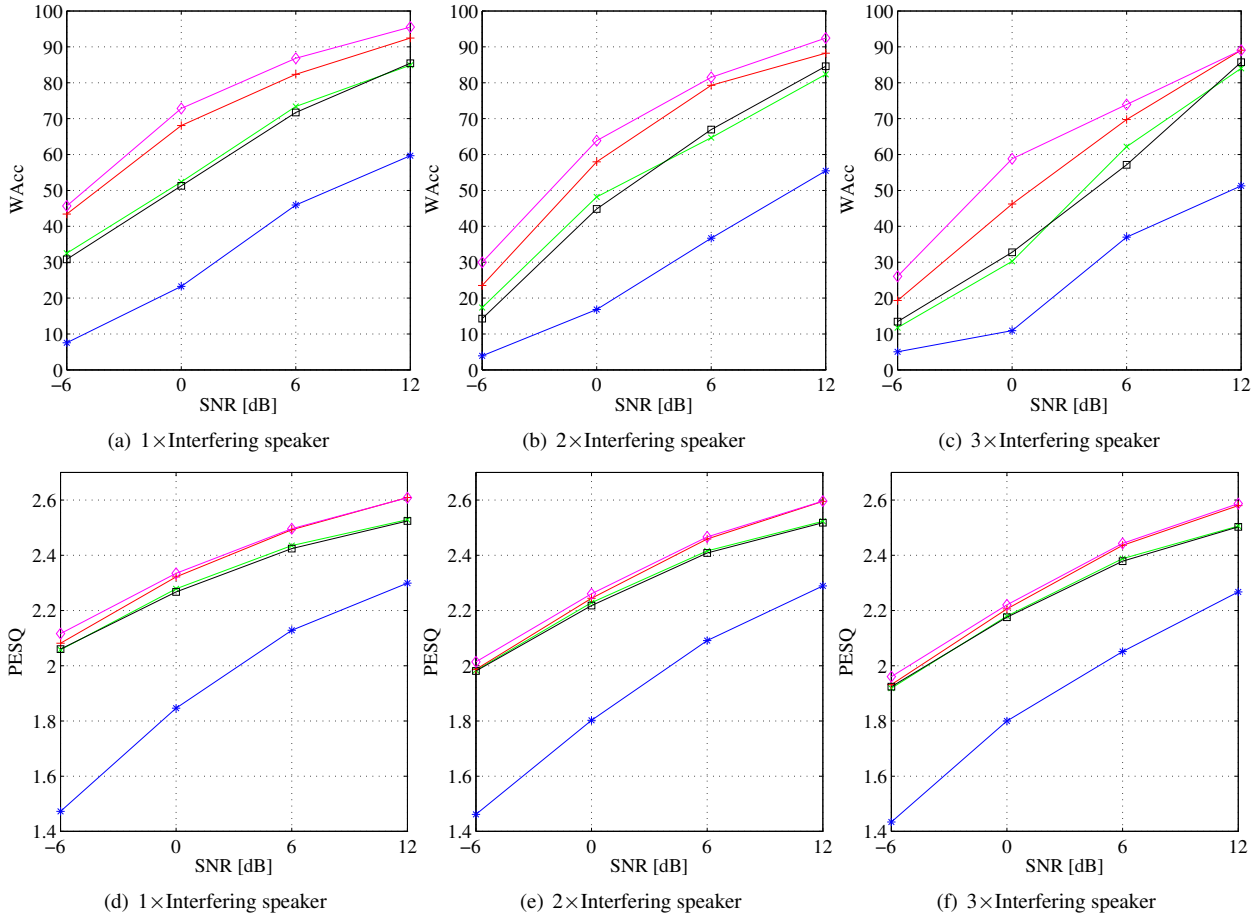


Fig. 7. *Results for different scenarios and SNR values: (a - c) WAcc, (d - f) PESQ.*
Legend: -∗- Single omnidirectional microphone; -×- First-order ADMA ($M = 2$); -+- Robust first-order ADMA (MNS: $M = 4$); -□- First/second-order hybrid ADMA ($M = 3$); -◇- Robust fist/second-order hybrid ADMA ($M = 5$).

frequency bins, 26 mel channels and 13 cepstral coefficients with cepstral mean normalization. We also append delta and delta-delta features, obtaining a final feature vector with 39 components. The back-end employs a transcription of the training corpus based on 34 monophones to train triphone-HMMs. We model each triphone by a HMM of 6 states and 8 Gaussian-mixtures per state. The lexicon is a set of 295 words derived from the German commands of the GRASS corpus [7]. We train a general bigram using these commands. These commands include some of the 24 test utterances. We train the HMMs with the center microphone signal of the training set without any enhancement.

6.3. Evaluation

Figure 7 shows the results for the PESQ and the WAcc. We evaluate the measures for scenarios with up to three interfering speakers and different SNR values.

We see that for every scenario and SNR condition all ADMAs increase the WAcc (cf. Fig. 7(a-c)) compared to a single omnidirectional microphone front-end. With the robust implementations of the ADMAs we achieve an improvement of up to 50% compared to their conventional implementations. In addition to suppressing the interfering signals, the ADMAs dereverberate the target signal and therefore also reduce the miss-match between training and test data. For the evaluation with the PESQ (cf. Fig. 7(d-f)) we observe a similar behaviour as for the WAcc. With the robust ADMAs we achieve an improvement of up to 0.1 points compared to the converntional ADMAs.

Looking at the different ADMA implementations, we see that the robust first/second-order hybrid ADMA ($M = 5$) gives the best results for most scenarios.

7. CONCLUSIONS

DMAs are a suitable front-end for an ASR system in close-talking scenarios. Their compact arrangement makes them an interesting alternative to conventional microphone arrays.

We conclude that for an ASR system with clean training used in a reverberant environment, an ADMA can improve the WAcc for every SNR condition. In this scenario, the novel robust implementations outperform the conventional ones, while the robust first/second-order hybrid ADMA with $M = 5$ microphones yielding the best results. With the used microphone distance of $\delta = 2.14$ cm between two adjacent microphones, for a linear microphone array with up to $M = 5$ microphones, we still achieve a compact arrangement.

As future work, we plan to investigate the effect of retraining the ASR with ADMA processed material and combining noise reduction algorithms with an ADMA.

8. REFERENCES

[1] Wim Soede, Augustinus J Berkhout, and Frans A Bilsen, "Development of a directional hearing instrument based on array technology," *The Journal of the Acoustical Society of America*, vol. 94, pp. 785, 1993.

[2] Jacob Benesty and Jingdong Chen, *Study and Design of Differential Microphone Arrays*, Springer, 2012.

[3] G.W. Elko and A.T.N. Pong, "A simple adaptive first-order differential microphone," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995, pp. 169–172.

[4] G.W. Elko and J. Meyer, "Second-order differential adaptive microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP*, 2009, pp. 73–76.

[5] Elmar Messner, "Differential Microphone Arrays," M.S. thesis, Graz University of Technology, 2013.

[6] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2915–2929, 2005.

[7] B. Schuppler, M. Hagmüller, J. A. Morales Cordavilla, and H. Pessentheiner, "GRASS: the Graz corpus of Read And Spontaneous Speech," LREC 2014.

[8] F. Schiel and A. Baumann, "Phondat 1, corpus v.3.4.," Tech. Rep., Bavarian Archive for Speech Signals (BAS), 2006.

[9] J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, P. Mowlaee, F. Pernkopf and G. Kubin, "A German distant speech recognizer based on 3D beamforming and harmonic missing data mask," 2013, in AIA-DAGA.

[10] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task," Tech. Rep., ETSI STQ-Aurora DSR, 2002.