Evaluation of Microphone Array Front-Ends for ASR – an Extension of the AURORA Framework

Harald Höge[1], Josef G. Bauer[1], Christian Geißler[1], Panji Setiawan[2], Kai Steinert[3]

[1] Siemens AG, Corporate Technology
Munich, Germany
{harald.hoege, josef.bauer, christian.geissler}@siemens.com

[2] Universität der Bundeswehr München
Munich, Germany
panji.setiawan.ext@mch.siemens.de

[3] Technische Universität München
Munich, Germany
kai.steinert@mytum.de

Abstract

The paper is focused on evaluating recognizers for automotive applications using microphone arrays. We propose to extend the framework developed within the AURORA project (Hirsch & Pearce, 2000). Based on measurements of the impulse responses within a given car between the position of the speaker and the microphone array and based on multi-channel in-car noises recorded at the microphone array output we propose to convert the close-talk speech recordings of the AURORA databases to multi-channel databases. The resulting data can be used to evaluate algorithms of microphone array front-ends. To compare the performance of these algorithms we start with the AURORA Advanced Front-End (AFE) as a baseline for mono-channel processing and demonstrate the evaluation with basic multi-channel algorithms. Our actual goal is to propose a certain evaluation procedure and to encourage others to support it.

1 Introduction

Competitive evaluation campaigns as done within the DARPA projects (Garofolo et al., 1997) have been proven to be a very successful organizational approach to push progress in automatic speech recognition (ASR) technology. Another aspect of evaluation is the creation of standards for ASR components. For this purpose high performing components for the given set of applications have to be developed. Recently the AURORA project provided a fruitful framework to standardize acoustic front-ends suited for distributed speech recognition (ETSI, 2000).

Competitive evaluation of commercial recognizers is still an open issue. Companies like Telcos offering ASR supported IVR systems or car manufacturers offering ASR supported navigation systems would like to have a reasonable procedure which proves that a given ASR system is suitable for their application.

Our paper is focused on two issues:

- push robust speech recognition technology using microphone arrays as front-ends,
- support the evaluation of recognizers for automotive applications.

The framework we propose is an extension of the framework built up by the second phase of the AURORA project which was focused on improving the front-end with respect to noise robustness. In a competitive evaluation campaign partners of the AURORA consortium had to present a front-end which showed best performance on speech databases specified by the AURORA consortium. The new front-end called AURORA Advanced Front-End (AFE) achieved a reduction of error rates of about a factor of two compared to the first front-end (Macho et al., 2002). ELDA - an operational unit of ELRA (http://www.elra.info) - has supported AURORA's evaluation campaign by making publicly available AURORA databases as described in the ELRA catalogue of language resources (http://www.elda.fr/rubrique14.html). To summarize the AURORA project showed clearly that the framework set up

- pushed progress in noise robustness substantially
- led to an accepted standard.

In order to extend the AURORA approach to our purpose, i.e. to push ASR technology based on microphone arrays and to support the evaluation of recognizers for automotive applications we propose a procedure for evaluation. This procedure includes the generation of a multi-channel database derived from an existing single-channel car database by signal processing methods and a setup for evaluation with a speech recognizer using as back-end the HTK framework as in the AURORA project. With the setup baseline results for single-channel processing can be generated and results for algorithms with additional multi-channel processing can be compared to them.

The main advantage of this approach lies in the fact that no costly in-car speech recordings from many speakers have to be done for each microphone array and each car. We are convinced that this approach will finally lead to a cost effective and perhaps standardized procedure to evaluate the performance of recognizers using microphone array front-ends.

The paper is organized as follows: First the evaluation scenario is described. Second the procedures for training and test are presented including the single-channel baseline results. Finally examples of multi-channel processing are given and benchmarked with the baseline.

2 Evaluation Scenario

2.1 Evaluation Criterion

The performance of multi-channel front-ends is often characterized by the gain (SNR/SNI improvement) or directivity. But since the target application is ASR the relevant evaluation criterion will be the recognizer's performance improvement in terms of the error rate reduction of multi-channel solutions compared to the "best" single-channel solution. In the proposed framework the latter is assumed to be the AURORA AFE serving as baseline.

2.2 Evaluation Environment

The target environment selected for the evaluation of multi-channel technology is an in-vehicle environment as in the AURORA project. This environment has attracted several companies as can be seen by the consortia SpeechDat-Car (http://www.speechdat.org/SP-CAR) and SPEECON (http://www.speecon.com) where in-vehicle recordings were made for many languages to train recognizers.

2.3 Evaluation Database

2.3.1 General Approach

For evaluating the multi-channel technology appropriate training and test databases are necessary. In the AURORA project already existing databases from the SpeechDat-Car project could be used. For microphone arrays no such real-world car databases are generally available. To avoid the costly effort of generating real car recordings we propose to adapt an already existing mono database to the target environment by signal processing methods resulting in a multi-channel environment adapted database (MEADB). Optimally the initial database is a clean one recorded in an environment that leads to recordings neither containing effects of room acoustics (e.g. reverberation) nor any noise. The clean database is converted into the MEADB in two steps:

- the utterances of the clean database are convolved with the room's impulse responses to add the room acoustics of the target environment and
- multi-channel noises are superposed to the resulting utterances.

Both components – the impulse responses and the noises - will be measured and recorded respectively within the real-world environment.

2.3.2 Generated MEADB

For the noise recordings and measurements of impulse responses a 4-channel microphone array was mounted at the ceiling of a car in front of the driver's seat. For the linear array uni-directional microphone cartridges with a constant interelement spacing of 4 cm were used and oriented in broadside direction towards the assumed speaker. The cartridges used were four samples of the EM145N from Primo Company Ltd.

The measurement of the impulse responses was performed using maximum-length sequences (MLS) (Vanderkooy & Rife, 1989) with a loudspeaker at the assumed driver's mouth position and by simultaneous recordings with all microphones of the array. The multi-channel noise recordings were carried out while driving around in the car. Various driving situations were included to cover a wide range of typical noises and to be able to create a multi-condition database. The sampling rate used was 8 kHz.

A clean database as described in Section 2.3.1 was not available so we had to resort to a close-talk database approximating the required characteristics. We selected the close-talk channel of the AURORA 3 German database (ELRA-AURORA/CD0003-03). The main advantage of using this car data is that it already represents the way people adapt their voice to the specific environment (Lombard effect).

The procedure of artificially creating the MEADB with the described components is illustrated in Figure 1.

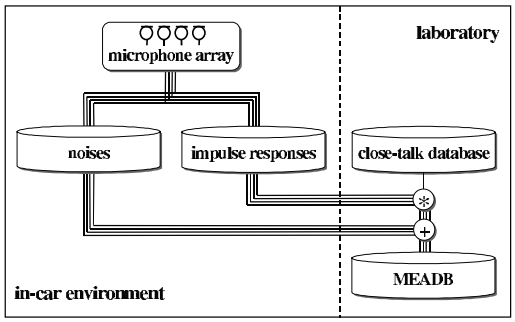

Figure 1: Generation of an environment adapted database. By means of the multi-channel noises and impulse responses the single-channel close-talk input database becomes a multi-channel one, as well.

The figure points up that only a limited portion of the work has to be carried out in the target environment and the remaining work can be performed in the laboratory. So an adaptation to a new environment or microphone array can be done with a limited effort.

2.3.3 Noise Field Characterization

The scaling of the added noise was selected in a way to achieve a pseudo-random SNR within a predefined typical range for car environments. Figure 2 shows the SNR distribution of the original close-talk data compared to one of the corresponding real-world far-talk channels and to one of the channels of the MEADB.

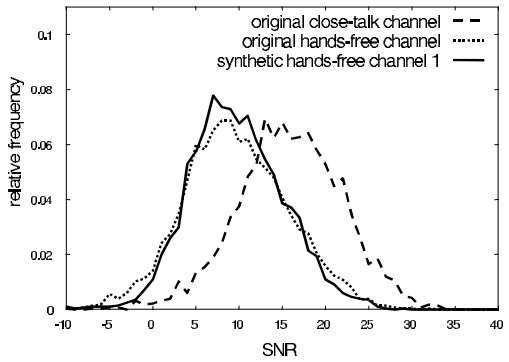

Figure 2: Frequency of occurrences of utterances with certain SNR.

In a multi-channel signal processing context a noise field is often described by the mean coherence of the multi-

channel noise signal. The coherence of the noise-only signals $N_i(f)$ and $N_j(f)$ is given as

$$\Gamma_{N_iN_j}(f) = \frac{P_{N_iN_j}(f)}{\sqrt{P_{N_iN_i}(f)P_{N_jN_j}(f)}}$$

where $P_{N_iN_j}(f)$ is the cross spectral density and $P_{N_iN_i}(f)$ and $P_{N_jN_j}(f)$ are the auto spectral densities of the signals $N_i(f)$ and $N_j(f)$, respectively. Office or car noise can be modeled using a diffuse noise field which is described by the sinc function (Brandstein & Ward, 2001). Figure 3 shows the mean of the real part of all channel combinations' coherence for the diffuse noise field theory and for an estimation from real car noise recordings.

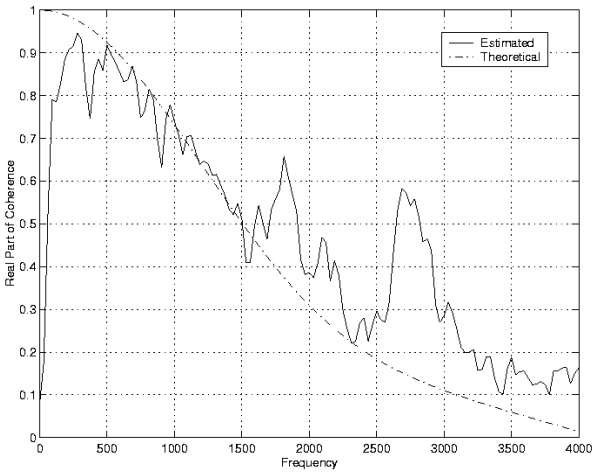

Figure 3: Theoretically calculated and estimated real mean coherence for the diffuse noise field of a car.

2.3.4 Training and Test Data

In the original AURORA 3 subsets the already quite small test lists would have been further reduced by our procedure where only the close-talk utterances could have been used. So we re-partitioned the speech material into a new training and one[1] new test set according to the official SpeechDat-Car speaker lists. But the overall material is still limited to the widespread AURORA 3 German digit database.

The close-talk part of the database consists of 1444 files from 112 speakers, 61 female and 51 male. A total of 1016 files from 75 speakers, 40 female and 35 male, are selected for the training part and 428 files from 37 speakers, 21 female and 16 male, for the test part. The MEADB finally consists of 4064 training utterances and 1712 test utterances corresponding to four different channels.

3 Baselines for Training and Test

To achieve a mono-channel baseline result, error rates are measured on utterances from each of the four channels of the microphone array individually. The mono-channel baseline result is then defined by the averaged error rates of all the channels. Figure 4 shows the baseline mono- and the multi-channel training and test procedures. Note that only one single HMM which has been trained on all four-channel utterances available from the training part was used for both mono- and multi-channel recognition performance measurement.

An example for the multi-channel processing is described in Section 4. Error rates achieved with multi-channel processing are then compared to the mono-channel baseline. With this approach more advanced multi-channel front-ends as reported in McCowan & Bourlard (2003) can be evaluated in order to standardize the "best" multi-channel advanced front-end (MCAFE).

The HTK training steps were done following Hirsch & Pearce (2000) where each digit is modeled as a whole word HMM having 16 states, simple left-to-right model without skips over states, a mixture of 3 Gaussians per state and using only diagonal covariance matrix. Other 3 states and single state HMMs with 6 Gaussians per state are used to model pauses.

The HTK inputs are vectors with 39 coefficients as processed by the AFE, consisting of 12 cepstral coefficients and a combined logarithmic frame energy and zeroth cepstral coefficient plus the delta and acceleration coefficients. The AFE is basically applying a two-stage mel-warped Wiener filter scheme and SNR-dependent waveform processing prior to cepstral calculation followed by a blind equalization scheme.

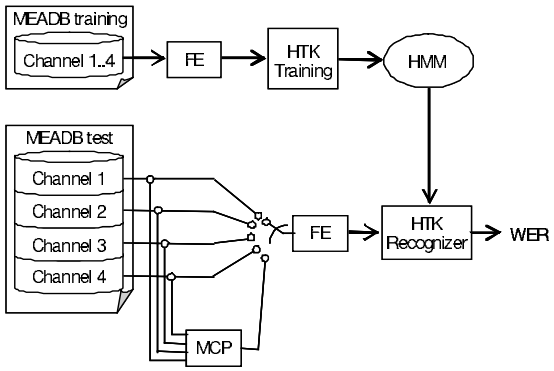

Figure 4: Mono- and multi-channel processing (MCP) procedures.

Note that for the mono-channel baseline the front-end (FE) used in training and test is the AFE representing the "best" front-end. For future multi-channel processing *any* FE can be used, also the AFE. The mono-channel baseline result is shown in Table 1.

mono-channel baseline	*WER*	*INS*	*DEL*
Advanced Front-End	8.4 %	1.1 %	2.4 %

Table 1: Mono-channel baseline result.

4 Examples for Multi-Channel Processing

Multi-channel processing is done using a delay-and-sum beamformer with an additional Wiener post-filter.

Suppose all M input signals (M being the number of channels) are given at the microphone output as

$$x_i(t) = s(t-\tau_i^*) + n_i(t)$$

where $s(t)$ is the desired signal, τ_i^* is the signal propagation time from the desired signal source to microphone i and $n_i(t)$ is the noise in channel i for $i \in \{0,\ldots,M-1\}$. Each of the M input signals is first split

[1] In the original AURORA 3 databases three different test lists for different matching conditions were defined.

into blocks of 25 ms (200 samples). Because a circular convolution should be avoided (Brandstein & Ward, 2001) each block must then be zero-padded to the double length, i.e. 400 samples. For a fast FFT computation a power of two is chosen as block length which is 512 samples in our case. We are using a Hann window and therefore the frame shift has to be half the number of non-trivial information samples, i.e. 100 samples, to enable a perfect reconstruction of the signal.

After applying an FFT at each block the time lag τ_i between the desired signal in each channel is aligned by adding a linear phase term to the noisy signal phase (see Figure 5). Therefore appropriate time delays τ_i have to be chosen. The resulting signals are $V_i(f)$ for all channels i.

Summing the $V_i(f)$ for all channels and dividing by M yields the delay-and-sum output signal with a noise power reduced by the factor M if the noise field is perfectly incoherent, i.e. the noise cross power spectrum between any two different channels is zero (Meyer & Simmer, 1997).

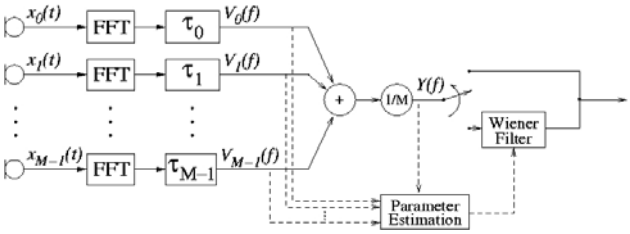

Figure 5: Structure of the delay-and-sum beamformer where the Wiener post-filter can be added optionally.

To further enhance the delay-and-sum output signal a post-filter is used which estimates a Wiener filter given as

$$W(f) = \frac{P_S(f)}{P_S(f) + P_N(f)}$$

where $P_S(f)$ and $P_N(f)$ are the signal and noise power at the output of the beamformer, respectively. Under the assumption of zero cross correlation between the desired signal and noise and the same noise power density spectrum at all channels the post-filter can be estimated using the multi-channel signals as follows

$$\hat{W}(f) = \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} P_S^{(ij)}(f)}{\frac{1}{M} \sum_{i=1}^{M} P_{V_i V_i}(f)}$$

where $P_{V_i V_i}(f)$ is the power density spectrum of signal $V_i(f)$ and $P_S^{(ij)}(f)$ is an estimate for the speech power density spectrum derived from the channel pair (i,j) as follows

$$P_S^{(ij)} = \frac{\Re\{P_{V_i V_j}\} - \frac{1}{2}\Re\{\Gamma_{N_i N_j}\}\left(P_{V_i V_i} + P_{V_j V_j}\right)}{1 - \Re\{\Gamma_{N_i N_j}\}}$$

The frequency dependency is omitted for the sake of brevity. $\Re\{.\}$ is the real part, $P_{V_i V_j}$ the cross spectral density between channel i and j. $\Gamma_{N_i N_j}$ is the complex coherence of the noise field, as discussed in Section 2.3.3.

The critical part is a good estimation of the spectral densities. As for the estimation of the Wiener filter either the theoretical coherences of a diffuse noise field can be used or an appropriate estimation using the multi-channel auto and cross spectral densities.

Table 2 shows the recognition results achieved using the multi-channel setup and theoretical coherences. The slightly worse results when an additional Wiener post-filter is applied to the delay-and-sum beamformer may be due to the concatenation with the already effective noise suppression of the AFE. Another FE may be more appropriate in this case as preliminary results show.

multi-channel processing	*WER*	*INS*	*DEL*
mono-channel baseline	8.4 %	1.1 %	2.4 %
delay-and-sum (D&S)	6.9 %	1.0 %	1.7 %
D&S + Wiener post-filter	7.0 %	1.9 %	1.3 %

Table 2: Multi-channel recognition results compared to the single-channel baseline.

5 Conclusions

We have presented a cost effective procedure to evaluate recognizers using microphone arrays as front-ends. Still it has to be proven that this approach is as good as making new in-car recordings with many speakers for each car and each microphone array. The performance of the front-ends with multi-channel processing tested here turned out to be only slightly superior to the mono-channel baseline using AFE. So there is much room for improvement in the development of microphone array algorithms.

References

Brandstein, M. & Ward, D. (2001). Microphone Arrays. Signal Processing Techniques and Applications. Springer 2001.

ELRA-AURORA/CD0003-03. Subset of SpeechDat-Car German. Distributed by ELDA (Evaluations and Language Resources Distribution Agency). http://www.elda.fr.

ETSI standard doc (2000). Speech Processing, Transmission and Quality Aspects (STQ). Distributed Speech Recognition. Front-end Feature Extraction Algorithm, Compression Algorithms. ETSI ES 201 108 V1.12 (2000-04). Available from http://pda.etsi.org/pda/queryform.asp.

Garofolo, J.S. & Fiscus, J.G. & Fisher, W.M. (1997). Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora. Proc. DARPA Speech Recognition Workshop, Feb. 1997.

Hirsch, H.-G. & Pearce, D. (2000). The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. ISCA ITRW ASR 2000.

Macho, D. et al. (2002). Evaluation of a noise-robust DSR front-end on Aurora Database. ICSLP 2002 (pp. 17—20).

McCowan, I. & Bourlard, H. (2003). Microphone array post-filter based on noise field coherence. IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 6, pp. 709-716, November 2003.

Meyer, J. & Simmer, K.U. (1997). Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. Proc. ICASSP, Vol. 2, pp. 1167-1170, Munich, Germany, 21-24 April 1997.

Vanderkooy, J. & Rife, D.D. (1989). Transfer-function measurement with maximum-length sequences. J. Audio Eng. Soc., 37(6), 419—444.