Acoustic Estimation of the Head Orientation for In-Car Communication Systems

Rasool Al-Mafrachi, Marco Gimm, Gerhard Schmidt
Christian-Albrechts-Universität zu Kiel, 24118 Kiel, E-Mail:[ral,mgi,gus]@tf.uni-kiel.de

Abstract

In order to overcome the communication difficulties among passengers due to a large amount of background noise in a car driven at high or even moderate speed, so-called In-car-communication (ICC) systems are recently used. Such ICC systems improve the signal-to-noise ratio (SNR) within the car compartment by recording, processing, and playing back the desired speech signal of the talking passenger over loudspeakers located close to the listening passengers. However, due to the acoustic directionality of the human head, ICC systems record a distorted speech signal with a degraded intelligibility and quality when the talking passenger turns his/her head. This contribution proposes first investigations towards an acoustic head orientation estimation of the speaking passenger within a noisy car compartment. The estimated head orientation cues would permit the improvement of speech enhancements technologies of the ICC systems such as appropriate equalization and corresponding residual noise adjustments (time- and frequency selective floor within noise suppression schemes). Our algorithm is based on using power ratio between the microphones signals as input feature vector for an artificial neural network (ANN) whose output is the estimated head orientation. The performance has been evaluated at different conditions and reported effective and robust results.

1. Introduction

In automotive environments, usually there is a high amount of background noise when driving at high or even moderate speed, which can make the communication among the passengers very difficult. Head directivity is another problem which can impair the communication inside a car since the talking passengers are facing the windshield and do not face each other as in normal conversations. Therefore, the signal to noise ratio (SNR) inside the car cabinet will be very bad and the communication partners will start to increase their vocal effort (the so-called *Lombard effect*) to compensate this bad SNR situation. Furthermore, the passengers may start to change their head orientation and lean towards each other to reduce the distance between them and hence increasing the conversation intelligibility. However, these actions are uncomfortable for long-time conversations and at the same time has a safety risk if the driver does them. In-car communication (ICC) systems are a solution for this problem [1]. However, the ICC solution has some challenges resulting from the closed acoustic loop operation and the high amount of disturbing noises (wind, engine, tire noise, etc.) which require various signal processing techniques in order to overcome these challenges and ensure system stability (such techniques include echo control [2], feedback cancellation [3], noise reduction [4], beamforming [5] etc.).

Furthermore, the directionality of human head acoustic field creates another challenge for the ICC systems. According to [6], the speech signal is attenuated by $5 - 10$ dB with respect to the head direction of the speaking person and this attenuation is larger at the higher frequencies. For this reason, the passenger dedicated microphone is recording a degraded low quality and less intelligibility speech signal when the talking passenger turns his/her head and this degradation is directly proportional with the angle of the head orientation (assuming 0° is pointing to the best microphone). Therefore, by estimating the head orientation of the speaking passenger it would be possible to equalize the microphone signal proportionally with the estimated angle to compensate for the frequency selective attenuation resulting from the head turning.

A lot of research have been done to estimate the source orientation for applications such as smart room voice commanded devices [7], robots [8], voice-controlled wheelchairs for disabled people [9], and speech acquisition in reverberant environments [10]. However, no attentions have been given to estimate the head orientation of the speaker in automotive environment where there is low SNR scenarios and the speech signal buried in mixture of stationary and non-stationary background noises. Usually, large arrays of hundreds of microphone units have been used in the estimation of source orientation [11, 12]. Using large microphone arrays may make the orientation estimation robust and reliable against undesired additions to signal intensity such as reverberation and microphone directivity characteristics but at the same time makes the solution expensive and not suitable for automotive applications.

Recently, source orientation estimation has been investigated using small microphones arrays [13] but only for smart room's commanded devices where there is no or little background noise. In this contribution, however, we aim to find a practical head orientation estimation approach using few number of microphones (e.g. only 3 channels) for automotive environments. Estimating head orientation measures in an automotive environment is not trivial task especially with low SNR scenarios and the presence of other speaking passenger within the car compartment. The ultimate goal of our approach is to be robust enough to provide precise and continuous real-time angle estimation for the passenger head orientation in a car driven in diverse automotive environment with a mixture of different background noises and with low SNR scenarios (up to -5 dB).

The paper is organized as follows: Sec. 2 describes our model approach, the details of pre-processing, feature extraction and the neural network estimator approach of the head orientation. Sec. 3 presents the details of the experimental setup while the experimental results, discussions and evaluations can be found in Sec. 4. Finally, a short conclusion is shown in Sec. 5.

2. Model Approach

In this section, we describe our algorithm for head orientation estimation. Fig. 1 depicts the complete process of the proposed algorithm.

Figure 1: The proposed system top view, k is the frame index.

2.1 Pre-processing

To overcome the bad conditions inside the car and to achieve an accurate estimation of the head orientation in term of azimuth, the microphones input signals should pass through a set of pre-processing steps to enhance the signals quality before extracting the feature vector. Fig. 2 shows the block diagram of the pre-processing stage.

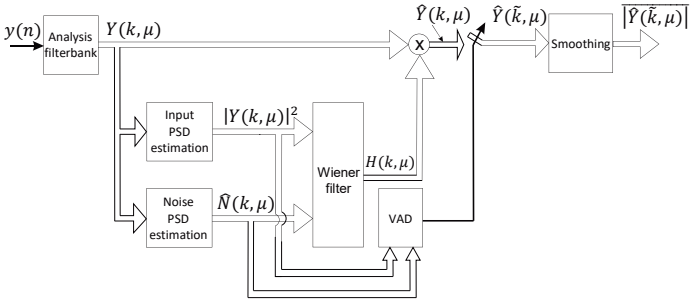

Figure 2: Pre-processing block diagram.

All the pre-processings were performed in the time-frequency domain $Y_i(k,\mu)$ which is generated by applying an analysis filterbank on the discrete input signal $y_i(n)$, where i, k and μ represent the channel index, frame index and the frequency subband index respectively. For ease of notation, the explanation will be for a single channel indicated by dropping the subscript i (it will be used only when necessary). The parameters for the filterbank in this study were: frame length $l = 256$, overlap = 50% , window= Hanning, and the fast Fourier transform length $N_{\mathrm{FFT}} = 256$.

2.1.1 Noise estimation and reduction

A suitable noise estimation is needed in order to efficiently suppress the background noises which can affect the extracted feature in the latter stage and hence the estimation accuracy. A robust extended noise estimation (RENE), proposed and evaluated in [14], was used. The noise estimation $\hat{N}(k,\mu)$ is modeled as the weighted sum of the smoothed input magnitude spectrum $\overline{|Y(k,\mu)|}$ and the slow changing noise pre-estimator $\hat{N}_{pre}(k,\mu)$ as following:

$$\hat{N}(k,\mu) = \big(1 - w(k,\mu)\big) \cdot \hat{N}_{pre}(k,\mu) + w(k,\mu) \cdot \overline{|Y(k,\mu)|}, \quad (1)$$

where $w(k,\mu)$ is the probability weighting for a speech pause. The estimated noise $\hat{N}(k,\mu)$ is employed by computing Wiener filter coefficients $H(k,\mu)$ to reduce the additive car noise signals and produce an enhanced version of the input signal spectrum $\hat{Y}(k,\mu)$ as following:

$$\hat{Y}(k,\mu) = Y(k,\mu) \cdot H(k,\mu) \quad (2)$$

2.1.2 Voice activity detection (VAD)

The VAD has a direct impact on the performance of our head orientation estimator. A simple threshold based voice activity detector was used in our approach to detect the speech presence. The idea behind this detector is to use the estimated noise $\hat{N}(k,\mu)$ as a reference and compare it with the smoothed input magnitude spectrum $\overline{|Y(k,\mu)|}$ multiplied by a certain SNR threshold THR_{SNR}. Whenever the smoothed input magnitude spectrum is larger than the estimated noise reference, the decision flag $\psi(k,\mu)$ is set to 1 otherwise it set to 0. This process is performed framewise for every subband where each flag $\psi(k,\mu)$ represents the decision output at frame k in subband μ.

$$\psi(k,\mu) = \begin{cases} 1, & \text{if } \overline{|Y(k,\mu)|} \cdot THR_{\mathrm{SNR}} > \hat{N}(k,\mu) \\ 0, & \text{else.} \end{cases} \quad (3)$$

The SNR threshold THR_{SNR} has been set to -12 dB. The final decision of the VAD for the frame k is determined by comparing the number of active flags $N_{\psi-\mathrm{act}}(k)$ within the frame k against another threshold VAD_{THR} as following:

$$VAD_{\mathrm{decision}}(k) = \begin{cases} 1, & \text{if } N_{\psi-\mathrm{act}}(k) > VAD_{\mathrm{THR}} \\ 0, & \text{else,} \end{cases} \quad (4)$$

where 1 and 0 represent the presence and absence of the speech respectively, VAD_{THR} has been set to 5 in this study. Only the speech frames are preserved which indicated by $\sim$ superscript above the frame index k.

2.1.3 Smoothing

Smoothing is used to reduce the variance within the magnitude of the enhanced input speech $\overline{|\hat{Y}(\tilde{k},\mu)|}$ and it is performed along the time axis for every subband using a first order IIR filter as following:

$$\overline{|\hat{Y}(\tilde{k},\mu)|} = \alpha|\hat{Y}(\tilde{k},\mu)| + (1-\alpha)\overline{|\hat{Y}(\tilde{k}-1,\mu)|}, \quad (5)$$

where α is the smoothing constant.

2.2 Feature extraction

After the pre-processing stage, the smoothed magnitude of the enhanced input speech signal for each channel i was decomposed into $M = 15$ Mel frequency bands using a Mel filterbank in order to reduce the feature vector dimensionality. Fig. 3 shows a block diagram of the feature extraction stage.

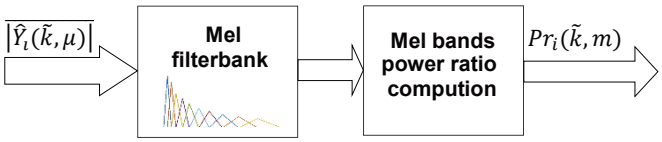

Figure 3: Feature extraction block diagram.

With center frequencies from 300 Hz to 8000 Hz spaced on Mel scale, the power ratio were computed for each Mel band as following:

$$Pr_i(\tilde{k},m) = 10\log_{10}\left[\frac{P_i(\tilde{k},m)}{\overline{P_C(\tilde{k},m)}}\right], \quad (6)$$

where m is the Mel band index and $\overline{P_C(\tilde{k},m)}$ is the mean of the power $P_i(\tilde{k},m)$ over the total number of channels C which is given by:

$$\overline{P_C(\tilde{k},m)} = \frac{1}{C}\sum_{i=1}^{C} P_i(\tilde{k},m) \quad (7)$$

$Pr_i(\tilde{k},m)$ represents the feature vector which is used as the input of the head orientation estimator in the next stage. By using the power ratio related feature, we expect that our head orientation approach can model the radiation pattern of the human head and can estimate its orientation.

2.3 Head orientation estimator

The estimation of the head orientation estimation is performed using an artificial neural network (ANN) which is a massive parallel distributed structure of interconnected elements called "neurons". Each neuron is equipped with a bias and a non-linear activation function while the interconnections between the neurons are equipped with weights. By adjusting these biases and weights in the training phase, the network can learn to estimate the angle of the head orientation correctly if provided with enough training data. In this paper, the ANN was fully connected in a feedforward configuration with two hidden layers (45 and 30 neurons respectively), both equipped with non-linear sigmoid activation function, and one output layer with one neuron corresponding to the angle of the head orientation which equipped with a linear activation function. This is the best structure which was found empirically. The feature data was divided into three data sets: training 70%, validation 15%, testing 15% and the network was trained using Bayesian regularization backpropagation algorithm. Dividing the feature data into three sets allow for early training stopping to avoid network overfitting. Cross-validation was also preformed to ensure that the network does not overfit to the training data and to ensure its ability for generalizing correct angle for fresh data that it has not seen before.

3. Experimental setup

All the experiments were conducted in an acoustically isolated room. An artificial head (HMS II.3 from HEAD Acoustics GmbH) was used to simulate the speaker which placed on a turntable stand with 360° of freedom $1\ m$ above the ground. Three microphones were placed around the head, 1 m above the ground and 60 cm away from the head. The head was rotated at azimuths within the range $[-90, 90]$ in the horizontal plane with steps of 15° (resulting of total 13 angles) and in different positions (the distance between each two successive positions is 20 cm). Fig. 4 shows the experimental setup and a schematic diagram which illustrates all head orientations and positions used in this study. For each head orientation and position, 1 minute clean speech consisting of a set of male and female samples selected from the TIMIT database was played back using the head and the corresponding response was recorded through three distributed microphones around the head. A sampling frequency of 44.1 kHz was used for playback and recording. The recorded signals were down sampled to 16 kHz and car noise was added at different SNRs (10 dB, 5 dB, 0 dB, and −5 dB) to simulate speech signals uttered in a car driven at different speeds. A neural network was trained separately for each position and another single neural network was trained for all the positions.

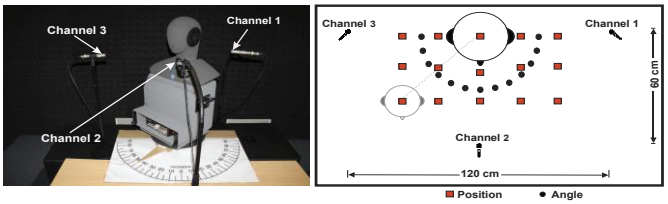

Figure 4: Experimental setup: (A) actual setup, (B) schematic diagram showing the used orientations and positions.

4. Results

Three cost functions were used to evaluate the performance of the ANN head orientation estimator system, mean magnitude error ($\bar{e}$), error variance ($\sigma_{\bar{e}}^2$) and Pearson correlation (r). The first criterion $\bar{e}$ measures the average of the magnitude error between the correct angles θ and their estimated counterpart $\hat{\theta}$. It is computed as:

$$\bar{e} = \frac{1}{N}\sum_{j=1}^{N}\left|(\theta_j - \hat{\theta}_j)\right| \quad (8)$$

The second criterion $\sigma_{\bar{e}}^2$ quantifies how much the estimated angles vary around their mean and it is computed as:

$$\sigma_{\bar{e}}^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(\left|(\theta_j - \hat{\theta}_j)\right| - \bar{e}\right)^2 \quad (9)$$

The last criterion r measures the linear correlation between the estimated and the correct angles, which has a value between $[-1, +1]$ and it is computed as:

$$r = \frac{\sum_{j=1}^{N}(\theta_j - \bar{\theta}).(\hat{\theta}_j - \bar{\hat{\theta}})}{\sigma_\theta \,.\, \sigma_{\hat{\theta}}} \quad (10)$$

where N represents the total number of speech frames, the best head orientation estimator is the one with low mean magnitude error, low variance error and high Pearson correlation. Tab. 1 and Tab. 2 summarize the performance of the ANN head orientation estimator system in terms of these three cost functions for different SNR scenarios.

Table 1: Performance of the ANN in terms of $\bar{e}$, σ_e and r for single position training (middle position).

SNR	Cost functions		
	$\bar{e}$ in °	σ_e in °	r
10 dB	0.15	0.82	1
5 dB	0.50	0.76	0.9999
0 dB	0.76	1.33	0.9996
−5 dB	0.79	1.4	0.9996

Table 2: Performance of the ANN in terms of $\bar{e}$, σ_e and r for all the positions training.

SNR	Cost functions		
	$\bar{e}$ in °	σ_e in °	r
10 dB	0.39	1	0.9998
5 dB	0.61	1.41	0.9997
0 dB	0.85	1.63	0.9995
−5 dB	1.69	2.81	0.9986

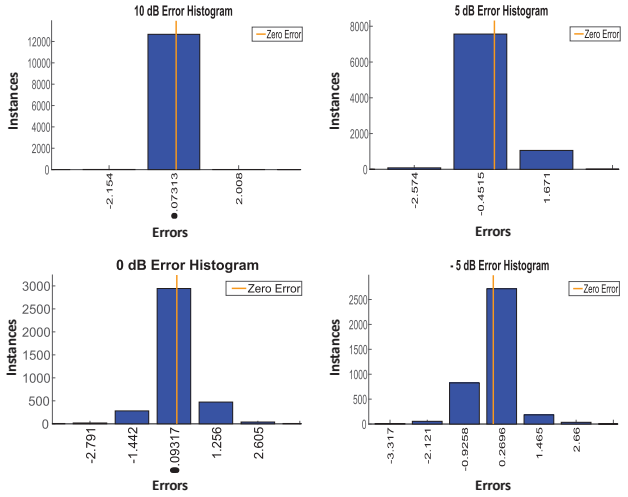

Figure 5: Error histogram for 10, 5, 0 and -5 dB SNR scenarios, single position (middle position).

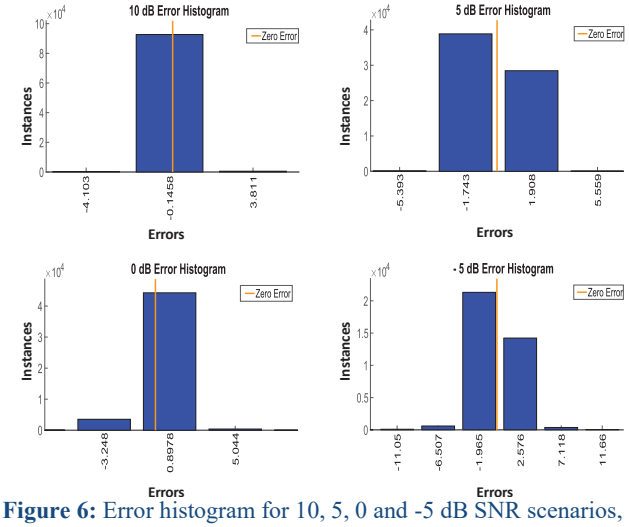

Figure 6: Error histogram for 10, 5, 0 and -5 dB SNR scenarios, all positions.

5. Conclusion

In this work, we proposed pre-studies for an acoustic head orientation estimation method for ICC systems. An ANN was employed to estimate the angle of the head orientation using the extracted feature vector after pre-processing the input signals. The method was tested and evaluated under various SNR conditions and head positions. The results of the analysis were promising showing that the system provides robust performance in estimating the head orientation angle precisely with only less than 2° mean magnitude error at the worst case ($-5\ dB$ scenario). The next step is to deploy and investigate the performance of our method in real car environment with real time processing.

References

[1] T. Haulick, G. Schmidt, Signal Processing for In-Car Communication Systems, Signal Processing, vol. 86(6), pp. 1307–1326, June 2006.

[2] A. Ortega, E. Lleida, E. Masgrau, Acoustic Echo Control and Noise Reduction for Cabin Car Communication, Proc. EUROSPEECH 2001, vol. 3, pp. 1585–1588, 2001.

[3] P. Bulling, K. Linhard, A. Wolf, G. Schmidt: Acoustic Feedback Compensation with Reverb-based Stepsize Control for In-car Communication Systems, ITG Speech, October 2016.

[4] T. Maschmann, M. Gimm, V. Kandade Rajan, and G. Schmidt: Implementation of a new Method for Noise Suppression in Automotive Environments, Proc. DAGA, Kiel, Germany, open access, 2017.

[5] M. Krini, V. K. Rajan, Klaus Rodemer, G. Schmidt: Adaptive Beamforming for Microphone Arrays on Seat Belts, Proc. DAGA 2015, March 16-19, 2015, Nürnberg, Germany.

[6] Brian B. Monsona and Eric J. Hunter: Horizontal directivity of low- and high-frequency energy, J. Acoust. Soc. Am. 132 (1), 2012 Acoustical Society of America.

[7] A.Y. Nakano, S. Nakagawa, K. Yamamoto, Distant speech recognition using a microphone array network, IEICE Trans. Inf. Syst. E93-D (9) (2010) 2451–2462.

[8] S. Hwang, Y. Park, Y. Park, Sound direction estimation using an artificial ear for robots, Robot. Auton. Syst. 59 (3–4) (2011) 208–217.

[9] A. Sasou, Acoustic head orientation estimation applied to powered wheelchair control, in: Second International ICST Conference on Robot Communication and Coordination, 2009, pp. 1–6.

[10] S. T. Shivappa, B.D. Rao, M.M. Trivedi, Role of head pose estimation in speech acquisition from distant microphones, in: Proceedings of ICASSP, 2009, pp. 3557–3560.

[11] J. M. Sachar, H. F. Siverman, A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array, in: Proceedings of ICASSSP, 2004, pp. 65–68.

[12] A. Levi, H. Silverman, A robust method to extract talker azimuth orientation using a large-aperture microphone array, IEEE Trans. Audio Speech Lang. Process. 18 (2) (2010) 277–285.

[13] A. Y. Nakano, P. M. S. Burt, Directional acoustic source orientation using only two microphones. ELSEVIER, Volume 23, Issue 6, December 2013, Pages 1918-1922.

[14] C. Baasch: Verbesserung und Implementierung einer Geräuschschätzung in einem Echtzeitsystem für Anwendungen im Automobilbereich, Bachelor thesis, Kiel University, Faculty of Engineering, 2012. (In German).