2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914770

MODEL-BASED ESTIMATION OF IN-CAR-COMMUNICATION FEEDBACK APPLIED TO SPEECH ZONE DETECTION

Kaspar Müller[1], *Simon Doclo*[2], *Jan Østergaard*[3], *Tobias Wolff*[1]

[1]Cerence GmbH, Acoustic Speech Enhancement, Ulm, Germany, kaspar.mueller@cerence.com
[2]University of Oldenburg, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany
[3]Aalborg University, Department of Electronic Systems, Aalborg, Denmark

ABSTRACT

Modern cars provide versatile tools to enhance speech communication. While an in-car communication (ICC) system aims at enhancing communication between the passengers by playing back desired speech via loudspeakers in the car, these loudspeaker signals may disturb a speech enhancement system required for hands-free telephony and automatic speech recognition. In this paper, we focus on speech zone detection, i.e. detecting which passenger in the car is speaking, which is a crucial component of the speech enhancement system. We propose a model-based feedback estimation method to improve robustness of speech zone detection against ICC feedback. Specifically, since the zone detection system typically does not have access to the ICC loudspeaker signals, the proposed method estimates the feedback signal from the observed microphone signals based on a free-field propagation model between the loudspeakers and the microphones as well as the ICC gain. We propose an efficient recursive implementation in the short-time Fourier transform domain using convolutive transfer functions. A realistic simulation study indicates that the proposed method allows to increase the ICC gain by about 6 dB while still achieving robust speech zone detection results.

Index Terms— feedback suppression, in-car communication, hands-free, speaker activity detection, speech zone detection

1. INTRODUCTION

While multiple built-in loudspeakers for passenger entertainment in cars have been a standard for decades, modern car cabins are increasingly equipped with multiple distributed microphones for applications such as hands-free telephony, automatic speech recognition or in-car communication (ICC). The former applications are designed for communication with external parties and require speech enhancement, e.g. beamforming or noise reduction [1–3]. On the other hand, ICC systems are designed to enhance speech intelligibility between passengers by reinforcing desired speech signals in the car cabin [4–6]. Often, this is achieved by recording the speech signal of a front passenger and reproducing it over loudspeakers at the rear cabin (see Fig. 1) to prevent the driver from turning around or shouting in order to be understood. The main challenge of ICC systems is to stabilize the closed electroacoustic loop resulting from the feedback of the loudspeaker signals to the microphones [4–10]. In practice, hands-free and ICC systems often run on different processors with highly restricted information exchange meaning that the loudspeaker signals used for the ICC system are generally not available for the hands-free system, which rules out a joint processing of both systems as proposed in [11].

This project has received funding from the SOUNDS European Training Network, an European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956369.

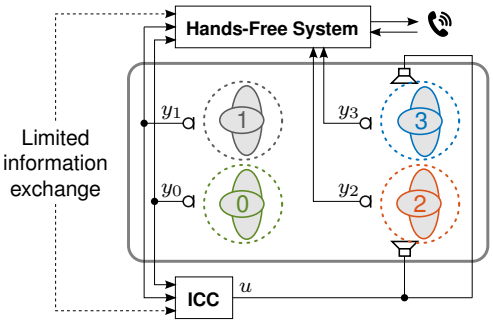


Fig. 1. Exemplary car setup with independent hands-free and in-car communication (ICC) systems. Dashed circles symbolize speech zones.

Depending on the ICC gain, the speech enhancement performance of the hands-free system may be substantially degraded by the ICC system. In this paper, we specifically focus on its influence on speech zone detection: When speech zones are defined in the car (one zone for each seat) to achieve speech enhancement for each zone individually, it is required to distinguish which zone is active, i.e. which passenger is speaking. According to [12], speech zone detection can be achieved by evaluating the maximum signal power ratios of passenger-dedicated microphones, as the speaker-dedicated microphone typically shows the highest signal power. However, this assumption may be violated in combination with an ICC system since the signal power of microphones close to ICC loudspeakers may exceed that of the speaker-dedicated microphone. One might consider classical feedback cancellation techniques [7] to remove the ICC feedback from the microphone signals. This would however require a loudspeaker reference signal, which in practice is not available for speech zone detection.

In this paper, a model-based feedback signal estimation method is proposed. This method estimates the ICC feedback contribution in the observed microphone signals without requiring access to the clean loudspeaker signals and by only considering free-field propagation between the loudspeakers and microphones. We propose an efficient recursive implementation in the short-time Fourier transform (STFT) domain using convolutive transfer functions (CTF) [13]. Finally, we suppress the ICC feedback contribution from the power spectral densities (PSD) of the observed microphone signals, which helps to improve robustness of speech zone detection against ICC feedback.

2. SIGNAL MODEL

We consider a car environment with a single speaker, M passenger-dedicated microphones and an ICC system with L loudspeakers (see Fig. 2). The observed microphone signals $y_m(n)$, with the microphone index $m = \{0, ..., M-1\}$ and the sample time index n, consist of three

components: The direct signals $y_m^{\mathrm{d}}(n)$, i.e. the microphone signals due to in-car speech (including reverberation), the ICC feedback signals $y_m^{\mathrm{fb}}(n)$ induced by the ICC loudspeakers as well as noise $v_m(n)$, i.e.

$$y_m(n) = y_m^{\mathrm{d}}(n) + y_m^{\mathrm{fb}}(n) + v_m(n)\,. \tag{1}$$

This paper will only consider the low-noise case, where $v_m(n)$ can be neglected. The desired direct signal $y_m^{\mathrm{d}}(n)$ is the dry speech signal $s(n)$ filtered with the acoustic impulse response $h_{\mathrm{d},m}(n)$ from the speech source to the m-th microphone. Likewise, the feedback signal $y_m^{\mathrm{fb}}(n)$ can be expressed as

$$y_m^{\mathrm{fb}}(n) = u(n) * h_{L,m}(n) \text{ with } h_{L,m}(n) = \textstyle\sum_{i=0}^{L-1} h_{i,m}(n), \tag{2}$$

where $u(n)$ denotes the ICC loudspeaker signal (same for all loudspeakers), $\{*\}$ denotes convolution and $h_{i,m}(n)$ is the acoustic impulse response from the i-th loudspeaker to the m-th microphone. $m=0$ refers to the ICC reference microphone channel that is reinforced in the car cabin.

The ICC and speech enhancement system including the speech zone detection are implemented in the STFT domain. Accordingly, the signal model needs to be formulated in the same domain. Since the impulse responses $h_{\mathrm{d},m}(n)$ and $h_{L,m}(n)$ are typically longer than the STFT frame size, and especially due to the recursive ICC structure (see Fig. 2), filtering in the STFT domain requires a multi-frame approach such as crossband filtering [14, 15]. To reduce computational complexity, we will consider the convolutive transfer function (CTF) approximation [13], which only considers band-to-band filters, i.e.

$$Y(k,l) = S(k,l) * H(k,l) = \sum_{l'=0}^{N_H-1} S(k,l-l')\,H(k,l')\,, \tag{3}$$

where k denotes the frequency index and l the frame index, $S(k,l)$ and $Y(k,l)$ are the STFT representations of the input and output signal, respectively, and $H(k,l)$ is the CTF consisting of N_H filter coefficients (we only consider causal coefficients). Here, $\{*\}$ denotes the convolution over frames.

In this work, we model the ICC system (see Fig. 2) using a linear model, where it should be realized that this is an approximation of a real ICC system, such as e.g. [4–6]. The forward path incorporates the static ICC CTF $H_{\mathrm{ICC}}(k,l)$, which models the known ICC processing delay τ_{ICC}, as well as a dynamic, real-valued and frequency-dependent ICC gain $\alpha(k,l)$. Since τ_{ICC} is fixed, the ICC gain $\alpha(k,l)$ is the only variable parameter that needs to be transmitted from the ICC to the zone detection at runtime. However, the ICC gain is usually slowly time-varying and thus requires no frame-wise transmission (the frame index will be omitted below). $G_{L,0}(k,l)$ is a linear model of the ICC feedback cancellation filter (see [5, 6, 8–10] for practical examples), which aims at subtracting the feedback signal from the ICC input signal to stabilize the ICC system and only amplify the desired direct signal.

For better readability, the frequency and frame indices k, l are omitted hereinafter in this section. Using this signal model, it can be shown that the loudspeaker signal U and the m-th feedback signal Y_m^{fb} are given by

$$U = \alpha\, H_{\mathrm{ICC}} * (Y_0 - G_{L,0} * U)\,, \tag{4}$$

$$Y_m^{\mathrm{fb}} = H_{L,m} * U\,. \tag{5}$$

By substituting (4) into (5), the feedback signal can be written as

$$\boxed{Y_m^{\mathrm{fb}} = \alpha\, H_{\mathrm{ICC}} * \big(H_{L,m} * Y_0 - G_{L,0} * Y_m^{\mathrm{fb}}\big).} \tag{6}$$

We consider only strictly causal CTFs (exclusively depending on past frames) to ensure that the recursive filters in (4), (6) and in the following equations are realizable.

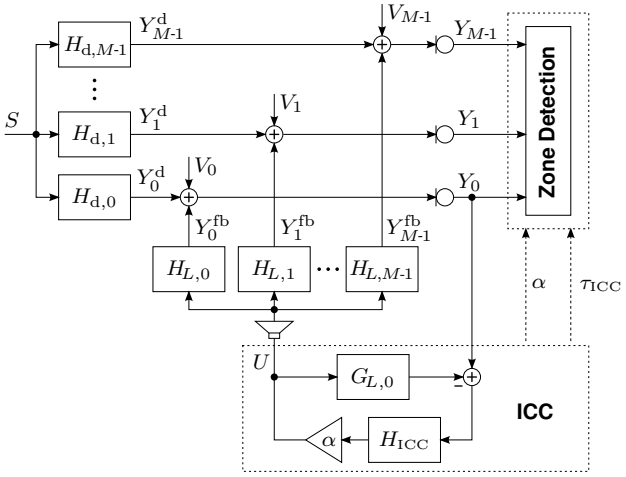


Fig. 2. Signal model with basic in-car communication (ICC) system in the STFT domain (frequency and frame indices are omitted).

To get a deeper understanding of the influence of the feedback cancellation filter $G_{L,0}$, one can define a mismatch

$$\Delta H_{L,0} = H_{L,0} - G_{L,0} \tag{7}$$

between the acoustic transfer function $H_{L,0}$ and $G_{L,0}$. Resolving (7) for $G_{L,0}$ and inserting into (4) yields

$$\begin{aligned} U &= \alpha\, H_{\mathrm{ICC}} * \big(Y_0 - H_{L,0} * U + \Delta H_{L,0} * U\big) \\ &= \alpha\, H_{\mathrm{ICC}} * \big(Y_0 - Y_0^{\mathrm{fb}} + \Delta H_{L,0} * U\big). \end{aligned} \tag{8}$$

By using $Y_0 = Y_0^{\mathrm{d}} + Y_0^{\mathrm{fb}}$ and substituting (8) into (5), the feedback signal can be written as

$$\boxed{Y_m^{\mathrm{fb}} = \alpha\, H_{\mathrm{ICC}} * \big(H_{L,m} * Y_0^{\mathrm{d}} + \Delta H_{L,0} * Y_m^{\mathrm{fb}}\big).} \tag{9}$$

According to this, the mismatch $\Delta H_{L,0}$ can be interpreted as a measure of the amount of feedback reduction: $\Delta H_{L,0} = 0$, i.e. $G_{L,0} = H_{L,0}$, corresponds to perfect feedback cancellation (only the desired direct signal Y_0^{d} would be amplified since there would be no recursion in (9)); $\Delta H_{L,0} = H_{L,0}$, i.e. $G_{L,0} = 0$, represents no feedback cancellation.

3. FEEDBACK SIGNAL ESTIMATION

In practice, the feedback signals have to be estimated based on the observed microphone signals $Y_m(k,l)$ without knowledge of the loudspeaker signal $U(k,l)$. For this purpose, a feedback signal estimator can be derived from (6) as

$$\begin{aligned} \hat{Y}_m^{\mathrm{fb}}(k,l) = {} & \alpha(k)\, \hat{H}_{\mathrm{f},m}(k,l) * Y_0(k,l) - \\ & \alpha_{\mathrm{r}}(k)\, \hat{H}_{\mathrm{r}}(k,l) * \hat{Y}_m^{\mathrm{fb}}(k,l)\,, \end{aligned} \tag{10}$$

where $\hat{H}_{\mathrm{f},m}(k,l) = H_{\mathrm{ICC}}(k,l) * \hat{H}_{L,m}(k,l)$ denotes the feed-forward CTF and $\hat{H}_{\mathrm{r}}(k,l) = H_{\mathrm{ICC}}(k,l) * \hat{G}_{L,0}(k,l)$ the recursive CTF of the feedback signal estimator, respectively (see Fig. 3), and $\{\hat{\cdot}\}$ denotes estimated or modeled quantities. We consider only strictly causal coefficients of the CTFs $\hat{H}_{\mathrm{f},m}(k,l)$ and $\hat{H}_{\mathrm{r}}(k,l)$. Furthermore, we introduced $\alpha_{\mathrm{r}}(k)$ to ensure stability of the estimator, as described below.

According to [13, 14], the CTFs $\hat{H}_{\mathrm{f},m}(k,l)$ and $\hat{H}_{\mathrm{r}}(k,l)$ can be computed from the corresponding time-domain filters

$$\hat{h}_{\mathrm{f},m}(n) = h_{\mathrm{ICC}}(n) * \hat{h}_{L,m}(n),\ \hat{h}_{\mathrm{r}}(n) = h_{\mathrm{ICC}}(n) * \hat{g}_{L,0}(n)\,. \tag{11}$$

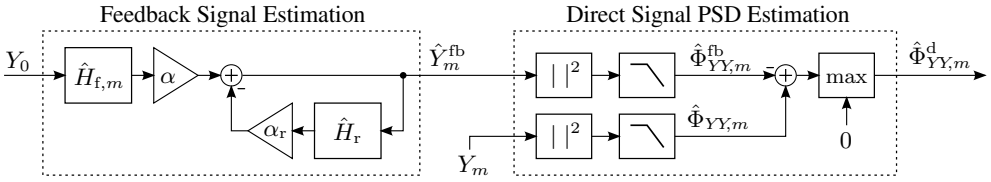


Fig. 3. Signal flow graph of the direct signal PSD estimation in the STFT domain (frequency and frame indices are omitted).

The ICC transfer function $h_{\mathrm{ICC}}(n)$ incorporates the known ICC processing delay τ_{ICC}. The acoustic impulse response $\hat{h}_{L,m}(n)$ can be either measured or modeled. In this work, we consider a simple free-field propagation model. Accordingly, the acoustic impulse response from the i-th loudspeaker to the m-th microphone are approximated by

$$\hat{h}_{i,m}(n) = \frac{\beta_{i,m}}{D_{i,m}}\,\delta(n-\tau_{i,m}) \;\text{ with } \tau_{i,m} = \left\lfloor \frac{f_s \cdot D_{i,m}}{c} \right\rceil, \quad (12)$$

where $D_{i,m}$ is the distance between the i-th loudspeaker and m-th microphone, $\beta_{i,m}$ is a gain factor that incorporates the individual sensitivity of the loudspeakers and microphones[1]. Relative to (7), we propose to model the time-domain feedback cancellation filter as

$$\hat{g}_{L,0}(n) = (1-\lambda)\,\hat{h}_{L,0}(n) \quad (13)$$

with $\hat{h}_{L,0}(n)$ as described above. The mismatch factor $\lambda \in [0,1]$ introduced herein is used to model the amount of feedback reduction, where higher values of λ correspond to less feedback cancellation when assuming that $\hat{h}_{L,0}(n) \approx h_{L,0}(n)$.

In practice, it is crucial to ensure stability of the recursive filter in (10) at runtime, where it should be noted that stability of the ICC system does not guarantee stability of the feedback signal estimator. For this purpose, the gain $\alpha_{\mathrm{r}}(k)$ of the recursive CTF is upper-limited

$$\alpha_{\mathrm{r}}(k) = \min\left\{\alpha(k),\, \alpha_{\mathrm{r,max}}(k) - \delta\right\}, \quad (14)$$

where $\alpha_{\mathrm{r,max}}(k)$ characterizes the stability border of the recursive filter in (10) and δ is a small positive value (we used $\delta = 0.2$ in this work). $\alpha_{\mathrm{r,max}}(k)$ can be determined as the maximum positive, real-valued gain for each subband k, where all poles of the recursive filter

$$\hat{H}_k(z) = 1 \,/\, \left(1 + \textstyle\sum_{l=1}^{L_{\mathrm{H}}} \alpha_{\mathrm{r,max}}(k)\,\hat{H}_{\mathrm{r}}(k,l)\,z^{-l}\right) \quad (15)$$

are strictly within the unit circle [16].

4. ZONE DETECTION ROBUST AGAINST ICC FEEDBACK

In this section, we describe an energy-based speech zone detection approach and moreover propose a method to increase its robustness against ICC feedback using the feedback signal estimates of Section 3.

4.1. Energy-Based Speech Zone Detection

In [12], it was proposed to perform energy-based speech speech zone detection based on the signal power ratios (SPR)[2]

$$\mathrm{SPR}_m(k,l) = 10\,\log_{10} \frac{\max\left\{\hat{\Phi}_{YY,m}(k,l), \epsilon\right\}}{\max\left\{\max\limits_{\substack{m'\in\mathcal{M}\\ m'\neq m}} \left\{\hat{\Phi}_{YY,m'}(k,l)\right\}, \epsilon\right\}} \quad (16)$$

of passenger-dedicated microphones, where $\hat{\Phi}_{YY,m}(k,l)$ is the estimated PSD of the m-th microphone signal, $\mathcal{M}$ is the set of all microphones and ϵ is a small positive number. A broadband SPR value can be determined by averaging over a set of frequencies $k \in \mathcal{K}_{\mathrm{speech}}$ assumed to contain speech energy (we used 100 Hz ... 8 kHz), i.e.

$$\overline{\mathrm{SPR}}_m(l) = \frac{1}{|\mathcal{K}_{\mathrm{speech}}|} \sum_{k\in\mathcal{K}_{\mathrm{speech}}} \mathrm{SPR}_m(k,l)\,. \quad (17)$$

The active zone is finally identified as the passenger-dedicated (= zone-dedicated) microphone with the largest broadband SPR, i.e.

$$\zeta_{\mathrm{active}}(l) = \arg\max_{m\in\mathcal{M}} \overline{\mathrm{SPR}}_m(l)\,, \quad (18)$$

where $\zeta \in \mathcal{M}$ are the possible speech zones.

Robustness of this speech zone detection approach against ICC feedback could be increased by using direct signal PSD estimates $\hat{\Phi}_{YY,m}^{\mathrm{d}}(k,l)$ (no contribution of ICC feedback) instead of the microphone signal PSDs $\hat{\Phi}_{YY,m}(k,l)$ to compute the SPR in (16). Hereinafter, we propose an estimation of the direct signal PSDs using the feedback signal estimates from Section 3.

4.2. Estimation of the Direct Signal PSD

One could consider directly subtracting the estimated feedback signal $\hat{Y}_m^{\mathrm{fb}}(k,l)$ from the observed microphone signal $Y_m(k,l)$ to obtain a direct signal estimate $\hat{Y}_m^{\mathrm{d}}(k,l)$ in a first step. However, this approach is highly sensitive to estimation errors such as phase errors. Therefore we propose to estimate the direct signal PSD in the power domain, which enables a better control of estimation errors. The microphone signal PSD $\Phi_{YY,m}(k,l)$ is defined as

$$\begin{aligned}\mathbb{E}\left\{|Y_m(k,l)|^2\right\} = \mathbb{E}\left\{|Y_m^{\mathrm{d}}(k,l)|^2\right\} + \mathbb{E}\left\{|Y_m^{\mathrm{fb}}(k,l)|^2\right\} + \\ 2\,\Re\left\{\mathbb{E}\left\{Y_m^{\mathrm{d}}(k,l)\,Y_m^{\mathrm{fb}*}(k,l)\right\}\right\},\end{aligned} \quad (19)$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator and $\Re\{\cdot\}$ the real value operator. It can be shown that the cross term $\mathbb{E}\left\{Y_m^{\mathrm{d}}(k,l)\,Y_m^{\mathrm{fb}*}(k,l)\right\}$ is negligible if the STFT frame size $N < \tau_{\mathrm{ICC}} + \min_i\{\tau_{i,m}\}$ (see (12)) and the source signal $S(k,l)$ is a zero-mean random signal. Moreover, this condition implicitly entails that the CTFs $\hat{H}_{\mathrm{r}}(k,l)$ and $\hat{H}_{\mathrm{f},m}(k,l)$ in (10) are strictly causal, which ensures the realizability of the feedback signal estimator.

Real-life scenarios in the car do not fully comply with the conditions stated above as speech signals are short-time stationary and reverberation is ignored. However, assuming low reverberation times in a car and ignoring short-time stationary of speech, we propose to estimate the direct signal PSD (complying with the upper bound of N) as

$$\hat{\Phi}_{YY,m}^{\mathrm{d}}(k,l) = \max\left\{\hat{\Phi}_{YY,m}(k,l) - \hat{\Phi}_{YY,m}^{\mathrm{fb}}(k,l), 0\right\}. \quad (20)$$

Here, the direct signal PSD estimate $\hat{\Phi}_{YY,m}^{\mathrm{d}}(k,l)$ is lower-limited to avoid negative results due to estimation errors. At runtime, the right-hand side PSD estimates in (20) are computed by exponential smoothing of the squared magnitudes of the according spectra $Y_m(k,l)$ and $\hat{Y}_m^{\mathrm{fb}}(k,l)$, respectively (see Fig. 3).

[1] $\beta_{i,m}$ reflects the level difference between the loudspeaker and microphone signal that would be observed in a free field at $D_{i,m} = 1$m. It could also include the directivity of loudspeakers and microphones which is not considered here.

[2] Originally, the authors used the SPR for speaker activity detection.

5. SIMULATION RESULTS

The influence of the proposed direct signal PSD estimation on the described speech zone detection approach was evaluated in simulations.

Microphone signal simulation: The time-domain microphone signals were simulated using the signal model described in Section 2 at a sampling frequency $f_s = 24\,\text{kHz}$ with a clean female speech source sample. The acoustic impulse responses $h_{\text{d},m}(n)$ and $h_{L,m}(n)$ were measured in a car (Audi A6) with reverberation time $T_{60} \approx 80\,\text{ms}$ using $M = 4$ passenger-dedicated microphones as illustrated in Fig. 1. The direct impulse responses $h_{\text{d},m}(n)$ were measured with a GRAS 44AA mouth simulator at zone 0 (driver seat), while $h_{L,m}(n)$ were measured with $L = 2$ built-in rear door loudspeakers. Uncorrelated white noise was added at each microphone ($\text{SNR} = 20\,\text{dB}$ at the reference microphone). To simulate the ICC system, we only considered a stationary broadband ICC gain α and $h_{\text{ICC}}(n)$ consisted of a processing delay $\tau_{\text{ICC}} = 15\,\text{ms}$ and a fixed gain. The latter gain was adjusted so that the ICC system at $\alpha = 0\,\text{dB}$ (typical ICC operation gain) induces the same RMS level at the rear microphones ($m = 2, 3$) as at the co-driver microphone ($m = 1$) due to a signal from the driver seat. For the feedback cancellation filter $g_{L,0}(n)$ in the ICC system, the measured impulse response $h_{L,0}(n)$ was superimposed by a random, white noise mismatch $\Delta h_{L,0}(n)$ (cf. (7)). It was adjusted so that the simulated ICC system became unstable for $\alpha > 4\,\text{dB}$, which matches a realistic system.

Direct signal PSD estimation: The modeled impulse responses $\hat{h}_{L,m}(n)$ and $\hat{g}_{L,0}(n)$ which are required for the feedback signal estimation (Section 3) were defined as follows: $\hat{h}_{L,m}(n)$ was modeled according to (12) (free-field propagation), where the distances $D_{i,m}$ between the ICC loudspeakers and the passenger-dedicated microphones were measured in the car. Furthermore, the same broadband gain $\beta_{i,m}$ was assumed for all impulse responses (same sensitivity for all loudspeakers and microphones is assumed). The modeled feedback cancellation filter in (13) was set to $\hat{g}_{L,0}(n) = \hat{h}_{L,0}(n)$. We intentionally chose this deviation from the simulated feedback cancellation filter $g_{L,0}(n)$ to investigate the robustness against modeling inaccuracies. The feedback signal estimation and direct signal PSD estimation was implemented in the STFT domain with Hann-windowed frames of $N = 256$ samples length (corresponding to 10.7 ms) and 50% overlap.

Results and Discussion

Fig. 4 shows the broadband SPR in (17) of the four passenger-dedicated (= zone-dedicated) microphones for an 8s speech signal from the driver seat (zone 0) at a typical ICC gain $\alpha = 0\,\text{dB}$. The proposed direct signal PSD estimation was activated after 4s. As can be clearly observed, the SPR levels without processing (0s-4s) are largely overlapping whereas a complete separation between the SPR levels of the active zone 0 and the remaining zones is achieved by the proposed method (4s-8s), even though the mean SPR levels are shifted by less than 1 dB.

Fig. 5 shows the zone detection rate of the four speech zones

$$\text{zone detection rate}(\zeta) = \frac{\#\,\text{frames with}\ \zeta_{\text{active}}(l) = \zeta}{\#\,\text{total frames}} \qquad (21)$$

for different ICC gains α, where the same 8s speech signal as before was used. The $\text{zone detection rate}(0)$ (circle markers) thus reveals the rate of correct zone detection. The speech zone detection either directly used the unprocessed microphone signal PSDs (a) or the estimated direct signal PSDs (b). Fig. 5a shows that the correct zone detection rate based on the microphone signal PSDs [12] is degraded to less than 50% due to ICC feedback at a typical ICC gain $\alpha = 0\,\text{dB}$. In contrast, the results in Fig. 5b indicate a substantial improvement due to the proposed direct signal PSD estimation, where the ICC-gain can be increased by about $+6\,\text{dB}$ to obtain similar detection rates. Moreover, good speech zone detection results are obtained until the stability limit of the simulated ICC system at $\alpha = 4\,\text{dB}$.

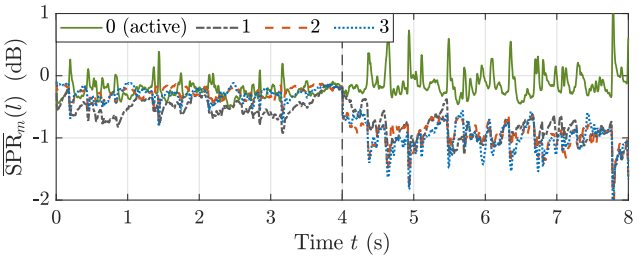
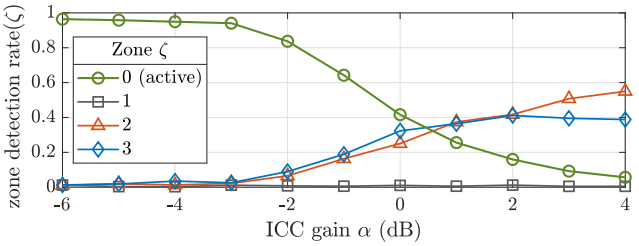
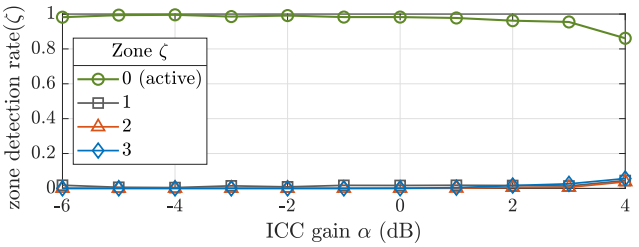


Fig. 4. Broadband SPR levels for an 8s speech signal at ICC gain $\alpha = 0\,\text{dB}$ based on the unprocessed microphone signal PSDs (0s-4s) and on the proposed direct-sound PSD estimates (4s-8s).



(a) Zone detection rates using unprocessed microphone signal PSDs.



(b) Zone detection rates using the proposed direct signal PSDs.

Fig. 5. Zone detection rates of an 8s speech signal in zone $\zeta = 0$ (driver seat) for different ICC gains. $\text{zone detection rate}(0)$ (circle markers) corresponds to the rate of correct detection.

6. CONCLUSIONS

This work described an approach to enhance the robustness of energy-based speech zone detection against an independently operating, interfering in-car communication system. The proposed method was designed in particular to cope with very limited information exchange between the speech zone detection and the ICC system. Specifically, we introduced a model-based ICC feedback signal estimation based on a free-field propagation model between loudspeakers and microphones, which requires no clean loudspeaker reference signal but only the slowly time-varying ICC gain. A computational efficient implementation in the STFT domain was derived consisting of one feed-forward and one recursive CTF. The resulting feedback signal estimates were used to estimate the microphone signal PSDs without ICC feedback. Simulations with measured impulse responses in a car indicated a robustness gain of about 6 dB against ICC feedback.

While this work considered low-noise scenarios with a simulated ICC system and speech with frontal head orientation, future work should also focus noisy environments with a real, more complex ICC system and different head orientations of a speaker.

7. REFERENCES

[1] J. Benesty, J. Chen, and E.A.P. Habets, *Speech Enhancement in the STFT Domain*, Springer Berlin Heidelberg, 2012.

[2] M. Buck, E. Hänsler, M. Krini, G. Schmidt, and T. Wolff, "Acoustic array processing for speech enhancement," in *Handbook on Array Processing and Sensor Networks*, chapter 8, pp. 231–268. John Wiley & Sons, Ltd, 2010.

[3] T. Matheja, M. Buck, and T. Fingscheidt, "A dynamic multichannel speech enhancement system for distributed microphones in a car environment," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, Dec. 2013.

[4] A. Ortega, E. Lleida, E. Masgrau, and F. Gallego, "Cabin car communication system to improve communications inside a car," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, Orlando, USA, May 2002, pp. IV–3836–3839.

[5] G. Schmidt and T. Haulick, "Signal processing for in-car communication systems," *Signal Processing*, vol. 86, no. 6, pp. 1307–1326, June 2006.

[6] C. Lüke, G. Schmidt, A. Theiß, and J. Withopf, "In-car communication," in *Smart Mobile In-Vehicle Systems*, pp. 97–118. Springer New York, Oct 2013.

[7] T. van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: state of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, 2011.

[8] P. Bulling, K. Linhard, A. Wolf, and G. Schmidt, "Acoustic feedback compensation with reverb-based stepsize control for in-car communication systems," in *Proc. ITG Symposium on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 337–341.

[9] P. Bulling, K. Linhard, A. Wolf, and G. Schmidt, "Stepsize control for acoustic feedback cancellation based on the detection of reverberant signal periods and the estimated system distance," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 176–180.

[10] M. Gimm, P. Bulling, and G. Schmidt, "Residual feedback suppression with extended model-based postfilters," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, May 2021.

[11] M. Gimm, A. Namenas, and G. Schmidt, "11 Combination of hands-free and ICC systems," in *Vehicles, Drivers, and Safety*, pp. 165–182. De Gruyter, May 2020.

[12] T. Matheja, M. Buck, and T. Fingscheidt, "Speaker activity detection for distributed microphone systems in cars," in *Vehicle Systems and Driver Modelling*, chapter 10, pp. 145–160. De Gruyter, Aug. 2017.

[13] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, May 2009.

[14] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[15] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 162–173, Jan. 2008.

[16] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-time signal processing*, Prentice Hall, Upper Saddle River, N.J, 1999.