# A Two-Microphone Diversity System and its Application for Hands-Free Car Kits

*Juergen Freudenberger and Klaus Linhard*

DaimlerChrysler AG, Research & Technology, Dialog Systems
Wilhelm-Runge-Str. 11, 89081 Ulm, Germany
phone: +49 731 505 4816, email: juergen.freudenberger@daimlerchrysler.com

## Abstract

In this paper we consider a two-channel diversity technique that combines the processed signals of two separate microphones. For in-car applications, this enables a better compromise for the microphone positions. The advantage of the proposed system is its insensitivity with respect to varying speaker sizes or local noise sources. To achieve this we choose the microphone position in that way that one microphone is optimum for a tall speaker, and the second one is suitable for a small speaker. For local noise sources we may apply a similar design to choose the microphone position in accordance with the location of the noise sources. A corresponding signal combiner has to tasks: compensation of phase shifts and weighting proportional to the signal strength. We propose solutions for both problems and demonstrate the effectiveness of diversity combining.

## 1. INTRODUCTION

Due to the obvious dangers of holding a telephone in one hand, and steering a car with the other, many countries either strongly recommended, or legally enforced hands-free telephone operation in all moving vehicles. Thus for safety and comfort reasons, a hands-free telephone system that provides the same quality of speech as conventional fixed telephones is desirable. A natural bottleneck for the speech quality of a hands-free car kit is the position of the microphone. Obviously, speech has to be picked up as close to the mouth as possible. The important question, where to place the microphone inside the car, is however difficult to answer. The *ideal* position has to consider noise sources like airflow from electric fans or car windows. Furthermore, the distance microphone-driver depends significantly on the position of the driver and therefore on the size of the driver. A practical position is apparently a compromise for different speaker sizes. Good noise robustness of single microphone systems requires the use of single channel noise suppression techniques, most of them derived from spectral subtraction [1]. A disadvantage of such systems is that they introduce considerable speech distortion.

Alternatively, a multi microphone setup promises to overcome some of these difficulties. Microphone array techniques based on beamformer algorithms efficiently improve the signal-to-noise ratio (SNR) with less audible distortion [2, 3]. However, the SNR still depends on the actual speaker distance from the microphones. Another approach, cross-spectral subtraction, is based on two microphones that are positioned separately (e.g. 80cm apart) in order to insure incoherent recording of noise [4, 5]. Here one sensor is used only for filter adaptation and the system output is the filtered signal of a single microphone.

In this paper we consider a two-channel diversity technique that combines the processed signals of two separate microphones. This enables a compromise for the microphone position with respect to different speaker sizes and noise sources. In communication systems diversity combining is a convenient approach to cope with fading due to multipath propagation of the radio signals. Similarly, Flanagan and Lummis [6] and Allen *et al.* [7] considered multi channel signal processing systems to reduce signal distortion due to reverberation. Both concepts consider an essentially noise free environment. As we focus on in-car applications our aim is of course noise robustness. Usually reverberation is of minor concern, because for in-car acoustics the direct path dominates the early reflections. Nevertheless, diversity combining is an effective means to reduce signal distortion due to reverberation and therefore improves the speech intelligibility.

In section 2, we discuss the influence of the microphone positioning on the room impulse response and the achievable signal to noise ratios at different microphone positions. The discussions are supported by acoustic measurement results obtained in a Mercedes S-Class. In the subsequent section, we describe the basic signal processing components required for diversity combining. That is, we consider the design of appropriate noise suppression filters and the estimation of the phase difference of the two microphone signals that is required for coherent signal summation. In section 4, we present some simulation results.

## 2. Microphone positioning

In this section, we present measurement results for speech signal propagation within a car (measured in a Mercedes S-Class). In particular, we consider the transfer characteristics of speech signal propagation between speaker and the microphones. Furthermore, we investigate the signal-to-noise ratio at the sensors for typical background noise.

We measured the transfer characteristic of two cardioid microphones with positions suited for car integration. One microphone (denoted by *mic. 1*) was installed close to the interior rear mirror (overhead console). The second microphone (*mic. 2*) was mounted at the end of the A-pillar. The microphones have a high pass characteristic and attenuate frequencies below 500Hz. The transfer functions depicted in Fig. 1 were measured with aid of an omni directional microphone placed at the microphone reference point of an artificial head. Thus, the transfer functions represent the acoustic path from the speaker's mouth to the car microphone. We considered two speaker sizes: a tall speaker of about 194cm highs and a small speaker of 164cm highs.

One observes from Fig. 1 that although the reverberation

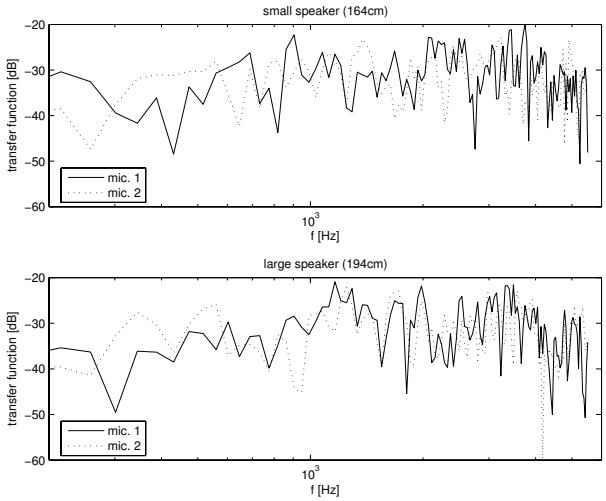


**Fig. 1**: Transfer functions for different microphone positions and speaker sizes.

| | 100km/h | 140km/h | defrost |
|---|---|---|---|
| small speaker | 1.3/1.3 | -0.6/-2.3 | 1.8/-0.5 |
| tall speaker | 0.4/9.0 | -1.5/5.4 | 0.9/7.2 |

Table 1: SNR values [dB] at mic. 1/mic. 2 for different driving situations.

time inside a car is relatively short ($T_{60} \approx 50$ms) it has a strong influence on the transfer function and causes frequency selective fading.

The influence of the speaker size on the transfer characteristic becomes more apparent when we consider signal-to-noise ratios. For this purpose we recorded noise samples for two driving situations 100km/h and 140km/h, respectively, and one example of noise caused from the electric fan (defrost). Speech samples were recorded using an artificial head in two different seat positions. The corresponding SNR values are given in table 1. From these values one observes that a small speaker would prefer the position of microphone 1, while for a tall speaker the second position is superior providing more than 6dB better SNR.

## 3. Basic system structure

A straightforward proposal to adapt the acoustic frond-end to the different situations considered in the previous section would be to select the best microphone input according the actual SNR condition. For communication systems with multiple receiving antennas, such an approach is well known as *selection combining*. However, we know from communication theory that *maximum ratio combining* promise some potential for improvements if we combine both input signals (cf. for example [8]).

It is worthwhile to consider the communication situation. Let $r_l(t) = \alpha_l e^{-j\phi_l} s(t) + n_l(t)$ be the $l$th received baseband signal, where $s(t)$ is the actually sent signal, $n_l(t)$ is an additive noise term, $\alpha_l$ and $\phi_l$ are the attenuation factor and the phase shift of the $l$th channel, respectively. The combiner that achieves the best performance is one in which each signal is multiplied by the corresponding complex conjugate channel gain $\alpha_l e^{j\phi_l}$. The effect of this multiplication is to compensate for the phase shift in the channel and to weight the signal by a factor that is proportional to the signal strength.

Similarly, a signal combiner for our microphone system has to perform the corresponding tasks: compensation of phase shifts and weighting proportional to the signal strength. Because room reverberation results in frequency selective fading, it is more convenient to consider the system in the frequency domain $R_l(f) = \alpha_l(f)e^{-j\phi_l(f)}S(f) + N_l(f)$ where the channel coefficients $\alpha_l(f)e^{-j\phi_l(f)}$ are now also frequency dependent. Then, for our two sensor system, the maximum ratio combiner would output the combined signal $R(f) = \alpha_1(f)e^{j\phi_1(f)}R_1(f) + \alpha_2(f)e^{j\phi_2(f)}R_2(f)$. The problem at hand is that different to the situation in radio communication we have no means to explicitly estimate the coefficients $\alpha_l(f)e^{-j\phi_l(f)}$ (i.e. the room transfer characteristic) for our microphone system which is a prerequisite for maximum ratio combining. Therefore, we are looking for spectral weights that are proportional to the coefficients $\alpha_l(f)$, i.e. $G_1(f) \propto \alpha_1(f)$ and $G_2(f) \propto \alpha_2(f)$. Furthermore, the aim of signal processing for speech signals is usually not to restore the absolute signal phase. Hence, it is sufficient to compensate the phase difference $\phi_\Delta(f) = \phi_1(f) - \phi_2(f)$. This results in the combiner rule $R(f) = G_1(f)R_1(f) + G_2(f)e^{j\phi_\Delta(f)}R_2(f)$. A corresponding processing system is depicted in Fig. 2, where an additional post filter after the combiner is included. In the following we will consider the tasks, spectral weighting and phase compensation, separately.

### 3.1. Spectral weighting

We are looking for spectral weights proportional to the signal strength, or more precisely proportional to the channel coefficients $\alpha_l(f)$. $G_l(f) = \sqrt{SNR_l(f)}$ would therefore be natural choice, because $\sqrt{SNR_l(f)} \propto \alpha_l(f)$ for stationary source and noise processes. However, car-noise is only quasi-stationary and speech signals are non-stationary and the actual signal power is time-varying. A direct weighting with a short term estimate of $\sqrt{SNR_l(f)}$ would typically result in a larger amplification of voiced sounds. To omit this problem we normalize the filter coefficients such that $G_l(f) \in [0, 1]$. An example for such a filter would be

$$G_l(f) = \sqrt{\frac{SNR_l(f)}{1 + SNR_l(f)}} \quad , \tag{1}$$

which corresponds to independent spectral subtraction in each input channel. For two inputs this can be generalized to:

$$G_l(f) = \sqrt{\frac{SNR_l(f)}{1 + a_1 SNR_1(f) + a_2 SNR_2(f)}} \quad . \tag{2}$$

This formula can be considered as a spectral weighting of the individual input signals with the nominator term $\sqrt{SNR_l(f)}$, where the denominator is a normalization to avoid signal distortions due to varying target signal powers. The parameters $a_1$ and $a_2$ enable a trade off between noise suppression and signal distortion as we will see in Section 4. A reasonable constraint for these parameters is $a_1 \geq 0$, $a_2 \geq 0$, and $a_1 + a_2 \leq 2$. The special case of independent spectral subtraction in each input channel corresponds to $a_l = 1$ and $a_1 + a_2 = 1$.

For an implementation of the spectral weighting we have to estimate the SNR, i.e. the power spectral densities (PSD) of the speech signal and the noise components. However, only the noisy speech signals are available. A practical solution is obtained with a time-frequency dependent voice activity detection (VAD) as presented in [9] which stops the estimation of

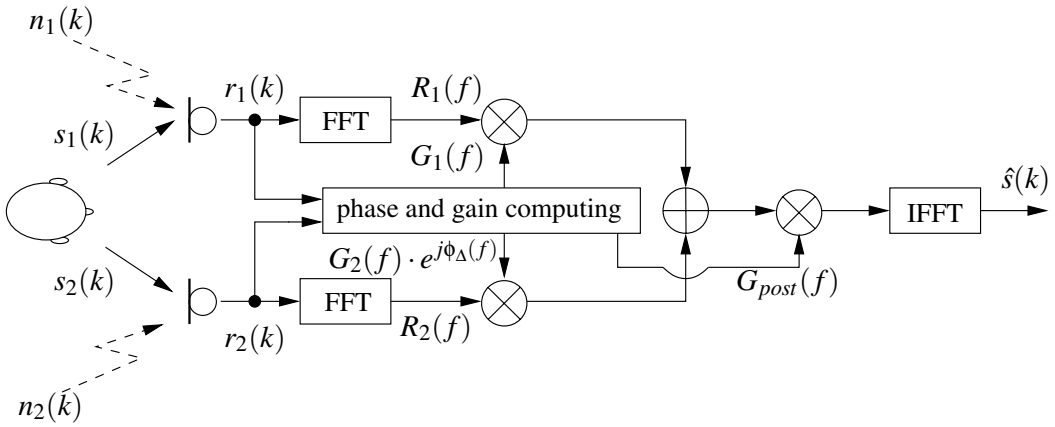


**Fig. 2**: Basic system structure of the two-channel diversity system.

the noise PSD during speech activity. The PSD of the speech signals is then obtained by spectral subtraction.

In addition to the spectral weighting and combining of the individual input channels we have included an post filter after the combiner in Fig. 2. This post filter exploits the noise decorrelation of the two microphone inputs due to the large separation of the installation positions and provides an additional SNR gain. We use the absolute value of the coherence of the two input signals as a post filter ($*$ denotes complex conjugate)

$$G_{post}(f) = \frac{\mid R_1(f)R_2(f)^* \mid}{\sqrt{R_1(f)R_2(f)}} \quad .$$

### 3.2. Phase compensation

Similar to the problem of SNR estimation the phase compensation requires estimates of the phase differences $\phi_\Delta(f)$, where the phase differences can only be reliably estimated during speech activity. Using the current phase difference

$$e^{j\phi_\Delta(f)} \approx \frac{R_1(f)R_2(f)^*}{\mid R_1(f) \mid \mid R_2(f) \mid}$$

leads to unreliable phase values for all time-frequency points without speech activity. Diversity combining using this short-term estimate leads to additional signal distortions.

A coarse estimate of the phase difference can be obtained from the time-shift $\tau$ between the direct path components in both room impulse responses. This time-shift can for example be found by searching for the maximum value of the cross-correlation of the two input signals whenever speech activity is detected. Hence, the phase estimate is $\phi_\Delta(f) \approx 2\pi f\tau$, where $\tau$ is recursively smoothed. Note that a combiner using these phase values would be equivalent to a delay-and-sum beamformer. However, in contrast to an ordinary microphone array, the two room impulse responses (speaker to microphone) of the diversity system are largely decorrelated. Such a beamformer would therefore result in a loss of signal power as signal components resulting from reflections are added in-coherently.

In order to avoid such signal distortions due to phase switching we combine the two mentioned phase estimates. A practical approach is to weight the two estimates using the coefficients of the filters $G_1$ and $G_2$ as a measure of the speech activity in the individual channels, i.e. when speech is detected the current phase is used. Then our final phase term is

$$\begin{aligned} e^{j\phi_\Delta(f)} &= (1 - G_1(f)G_2(f))e^{j2\pi f\tau} + \\ & \quad G_1(f)G_2(f)\frac{R_1(f)R_2(f)^*}{\mid R_1(f) \mid \mid R_2(f) \mid} \quad . \end{aligned} \tag{3}$$

## 4. Experimental results

For our simulations we consider the same microphone setup as in Section 2. We used car-noise recordings and speech samples recorded from an artificial head with a sampling frequency of 11025Hz, an FFT length of 256 and a Hamming window for time windowing.

As an objective measure of speech distortion we calculated the cosh spectral distance (a symmetrical version of the Itakura-Saito distance [10]) between the power spectra of the clean input signal (without reverberation and noise) and the output speech signal (filter coefficients were obtained from noisy data).

Table 2 provides results for single channel noise reduction, where we used spectral subtraction as proposed in [9]. We observe from these results that the position of microphone 1 would be a suitable compromise for both speaker sizes, whereas position 2 would result in up to 5dB better SNR for a tall speaker. The results for the diversity combining scheme are given in Table 3. Here we considered two parameter settings, with $a_1 = a_2 = 0.5$ and with $a_1 = a_2 = 0$. In the first case we observe SNR values that are 1-2dB better than the best value of the single channel noise suppression, where the signal distortion is significantly smaller than in the single channel case. The speech is free of musical tones and sounds more natural compared to ordinary spectral subtraction. The reduction of the signal distortion can be explained by the dereverberation effect of the diversity combining. If we omit the filter normalization, i.e. choose $a_1 = a_2 = 0$ we observe a larger signal distortion, but also a larger improvement of the noise suppression. In this case about 5dB can be gained compared to the best single channel case. As might be expected in this case, some voiced utterances sound un-naturally stressed. Values with $a_1 > 0$ and $a_2 > 0$ enable a trade off between noise suppression and signal distortion.

| | 100km/h | 140km/h | defrost |
|---|---|---|---|
| SNR small speaker | 13.1/10.6 | 10.7/6.8 | 11.6/8.7 |
| SNR tall speaker | 12.3/17.0 | 9.7/13.3 | 10.8/15.6 |
| dist. small speaker | 1.8/2.0 | 1.8/1.9 | 1.2/1.3 |
| dist. tall speaker | 2.7/1.3 | 2.5/1.3 | 1.8/1.2 |

Table 2: Output SNR values [dB] for single channel noise reduction with input signal from mic. 1/mic. 2, respectively.

| | 100km/h | 140km/h | defrost |
|---|---|---|---|
| SNR small speaker | 14.3/18.8 | 12.5/16.6 | 13.6/17.0 |
| SNR tall speaker | 17.6/22.1 | 15.0/19.0 | 16.9/20.4 |
| dist. small speaker | 1.3/1.8 | 1.4/2.0 | 0.8/1.1 |
| dist. tall speaker | 1.2/1.7 | 1.2/1.7 | 0.8/1.3 |

Table 3: Output SNR values [dB] for diversity combining for two parameter sets. First value corresponds to $a_1 = a_2 = 0.5$, the second value to $a_1 = a_2 = 0$.

## 5. CONCLUSIONS

In this work we have presented a two-channel diversity technique that combines the processed signals of two separate microphones, where the aim of our approach was noise robustness for in-car hands-free applications. We have demonstrated that single channel noise suppression methods are sensitive to the microphone location and in particular to the distance between speaker and microphone. The proposed two-channel diversity scheme achieves better SNR values than the better of the two single channel systems and is therefore less sensitive to varying speaker positions. Moreover, diversity combining is an effective means to reduce signal distortions due to reverberation and therefore improves the speech intelligibility.

## 6. References


[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[2] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[3] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control; A Practical Approach*, Wiley, 2004, ISBN 0-471-45346-3.

[4] A. Akbari Azirani, R. Le Bouquin-Jeannes, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech, and Audio Processing*, vol. 5, no. 5, pp. 484–487, 1997.

[5] A. Guerin, R. Le Bouquin-Jeannes, and G. Faucon, "A two-sensor noise reduction system: applications for hands-free car kit," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1125–1134, 2003.

[6] J. L. Flanagan and R. C. Lummis, "Signal processing to reduce multipath distortion in small rooms," *The Journal of the Acoustical Society of America*, vol. 47, no. 6, pp. 1475–1481, June 1970.

[7] J. B. Allen, D. A. Berlkey, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

[8] J. G. Proakis, *Digital Communications*, McGraw-Hill, 1995, ISBN 0-07-113814-5.

[9] H. Puder, "Single channel noise reduction using time-frequency dependent voice activity detection," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), Pocono Manor*, 1999, pp. 68–71.

[10] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electronics and Communications in Japan*, vol. 43, no. A, pp. 36–43, 1970.