ABSTRACT

Distributed microphone systems in cars usually provide dedicated microphones for several speakers where each microphone captures the desired speech signal at the best. The signal quality may differ strongly among the speaker channels depending on the microphone position, the microphone type, the kind of background noise, and the speaker himself. When combining these signals to a weighted mix annoying switching artifacts may result. In this contribution a new dynamic signal mixer is presented that uses spectral preprocessing to compensate both for different speech signal levels and for different background noise levels and colorations. Thus, artifacts are avoided and smooth transitions can be achieved for the various speech level and the background noise spectrum at speaker changes.

Index Terms—Distributed microphones, signal mixer, speech enhancement, teleconferencing

1. INTRODUCTION

In digital signal processing many multi-microphone arrangements have been addressed where two or more microphone signals have to be combined to one single output signal. Compared to beamforming setups (e.g., [1]) that require a small spacing between the microphones and a predefined geometry distributed microphone systems do not have such restrictions. Each microphone should capture the speech of one speaker at the best and is therefore mounted in his near vicinity. To realize a conference system the microphone signals usually are mixed with weighting functions which vary depending on activity of the different speakers.

Especially in car environments the speech signal is corrupted by strong background noise arising from, e.g., the engine or the airflow from open windows or electrical fans. There may be large differences in the speech and background noise level and therewith in the signal quality of the different microphone channel signals. Switching the mentioned weights at speaker changes bears the risk of producing artifacts. Either noise jumps may be noticeable after signal combining or an unnatural sound may occur due to including heavily distorted channels.

There is a wide range of solutions for distributed microphone signal combining. In [2] a concept for “automatic mixing” is described which aims for live sound scenarios. Effects from the background noise are not considered there. In [3] a noise reduction with a fixed scheme in each channel is proposed in order to improve quality for switching noisy signals but for the mixer criterion itself the noise is not considered. Other solutions are based on the maximization of the signal-to-noise ratio (SNR) at the output of the mixing process [4, 5]. High background noise scenarios like in a car environment are considered explicitly but only one speaker with multiple dedicated microphones is considered. In [4] a diversity technique is proposed that assumes similar noise levels in all microphone channels but adds the signals in phase. Another method for using diversity effects and handling also different noises is presented in [6]. Here the phase differences are estimated during speech periods.

We present a new method of signal combining supporting different speakers in a noisy environment. Particularly for deviations in the noise characteristics among the channels the proposed scheme ensures a smooth transition of the background noise at speaker changes. Therefore the signals are preprocessed before the signal combining. An automatic gain control (AGC) with a dynamic target level ensures similar speech signal levels in all channels. A modified noise reduction (NR) achieves equal background noise characteristics for all channels by applying a dynamic, channel specific, and frequency dependent maximum attenuation. The reference characteristics for adjusting the background noise will be specified by the dominant speaker channel.

This paper is organized as follows. In Sec. 2 we give an overview of the system. The determination of the dominant speaker is introduced in Sec. 3. Sec. 4 describes the signal processing part with the AGC and with the dynamic NR. The signal combining is described in Sec. 5. At the end of the paper an evaluation and a conclusion follow.
2. SYSTEM OVERVIEW

An overview of the system is depicted in Fig. 1. We consider a system with \( M \) microphones and the microphone index \( m \). Due to changing acoustic situations the microphone signal levels vary over time. Thus, each signal is adjusted to a target level that has to be recomputed during the processing. A control unit with a voice activity detection (VAD) determines the dominance of each speaker by computing dominance weights (DW) that contribute to calculate target values for adjusting the AGC and the maximum attenuation of the NR. After these preprocessing steps the signals in each channel have similar characteristics and can be combined. The processing is done in the subband domain where \( \ell \) denotes the frame index and \( k \) the frequency index. The short-time Fourier transform uses a Hann window and a block length of 256 samples with 75 % overlap at a sampling frequency of 11025 Hz. Each microphone signal can be modeled by a superposition of a speech and a noise signal component:

\[
\tilde{X}_m(\ell, k) = S_m(\ell, k) + \tilde{N}_m(\ell, k).
\]

3. SPEAKER DOMINANCE

For the computation of the target levels out of all different microphone signals it is important to know which speaker is the dominant one at a time instance. Corresponding dominance weights can be determined by evaluating the duration for which a speaker has been speaking. If only one speaker is active the target values should be controlled by this concrete speaker. If more than one speaker is active the target values should correspond to the mean of all channel characteristics. To get values for the necessary target levels out of all different microphone signals we assume that the current speaker is the dominant one after speaking for \( t_{inc} \) seconds. With the update time \( T_{frame} \) between two consecutive time frames it follows:

\[
c_{inc} = \frac{c_{max} - c_{min}}{t_{inc}} \cdot T_{frame}.
\]

The limitation of the counters by \( c_{max} \) or \( c_{min} \) respectively define full or minimal dominance of a speaker. We want to set the increasing interval \( c_{inc} \) of the counters in such a way that the current speaker is the dominant one after speaking for \( t_{inc} \) seconds. The decreasing constant has to be recomputed for a channel \( m \) if another speaker in any other channel \( m' \) becomes active. In this contribution single-talk is assumed. It has to be ensured that the dominance counter of the previous speaker has become \( c_{min} \) after the time the new active speaker reaches \( c_{max} \) and therewith full dominance. Including a constant \( \epsilon \) with a very low value to avoid the division by zero \( c_{dec,m} \) is determined by

\[
c_{dec,m} = \frac{c_{max} - c_{min}}{c_{max} - c_{m'}(\ell) + \epsilon} \cdot c_{inc}, \text{ if vad}_m(\ell) = 0.
\]

The resulting counters are depicted in Fig. 2(a) and can be mapped to the speaker dominance weights \( g_m(\ell) \) (Fig. 2(b)) that characterize the dominance of a speaker:

\[
g_m(\ell) = \frac{c_m(\ell)}{\sum_{n=1}^{M} c_n(\ell)}.
\]

4. DYNAMIC SIGNAL ADJUSTMENT

To compensate for the mentioned speech and noise level differences an AGC and a dynamic NR are presented in the following that perform an adaptation to adaptive target levels computed out of the underlying microphone signals.

4.1. Automatic Gain Control

Based on the input signal \( \tilde{X}_m(\ell, k) \) the AGC estimates the peak level \( \tilde{X}_{p,m}(\ell) \) in the \( m \)-th microphone signal and determines a fullband amplification factor \( a_m(\ell) \) to adapt the estimated peak level to a target peak level \( X_{p,m}^{ref}(\ell) \). A method for
peak level estimation is proposed in [8]. Instead of using the time domain signal for peak tracking we apply a root-mean-square measure over all subbands. The AGC is processed in each channel with frequency independent gain factors. Then the output results in
\[ X_m(\ell,k) = a_m(\ell) \hat{X}_m(\ell,k), \]
with the recursively averaged gain factors
\[ a_m(\ell) = \gamma \cdot a_m(\ell - 1) + (1 - \gamma) \cdot \frac{X_P^{\text{ref}}(\ell)}{X_P(\ell)}. \]

Here \( \gamma \) denotes the smoothing constant. The target or rather reference peak level \( X_P^{\text{ref}}(\ell) \) is a weighted sum of all peak levels and is determined by
\[ X_P^{\text{ref}}(\ell) = \sum_{m=1}^{M} g_m(\ell) \cdot \hat{X}_{P,m}(\ell). \]

Thus, the reference speech level is mainly specified by the dominant channel and the different speech signal levels are adapted to approximately the same signal power.

### 4.2. Dynamic Noise Reduction

The dynamic NR aims for equal power and spectral shape of the background noise for all channels. Therefore the maximum attenuation is varied for each microphone and for each subband. With \( \Phi_{n,m}(\ell,k) \) denoting the estimated power spectral density (PSD) in the \( m \)-th microphone channel the noise PSDs after the AGC result in
\[ \Phi_{n,m}(\ell,k) = a_m^2(\ell) \Phi_{n,m}(\ell,k). \]

For the NR different characteristics can be chosen that are based on spectral weighting. Here the NR filter coefficients \( \tilde{H}_m(\ell,k) \) are calculated for the shown examples by a recursive Wiener characteristic [8] with the fixed overestimation \( \alpha \) and the overall signal PSD \( \Phi_{x,m}(\ell,k) \) estimated by recursive averaging:
\[ \tilde{H}_m(\ell,k) = 1 - \min \left( a, \frac{\beta}{\Phi_{n,m}(\ell,k)} \right) \frac{\Phi_{n,m}(\ell,k)}{\Phi_{x,m}(\ell,k)}. \]

For realizing a maximum attenuation in each channel the filter coefficients are limited by an individual dynamic spectral floor \( b_m(\ell,k) \):
\[ H_m(\ell,k) = \max \left( \tilde{H}_m(\ell,k), b_m(\ell,k) \right). \]

After setting a reference floor \( \beta^{\text{ref}} \) specifying the overall noise reduction and after estimating a common target noise PSD \( \Phi_n^{\text{ref}}(\ell,k) \) the spectral floors are determined by
\[ b_m(\ell,k) = \beta^{\text{ref}} \cdot \sqrt{\frac{\Phi_n^{\text{ref}}(\ell,k)}{\Phi_{n,m}(\ell,k)}}. \]

Here the target noise PSD is computed adaptively similar to the peak level in Eq. 8 by the dominance weights:
\[ \Phi_n^{\text{ref}}(\ell,k) = \sum_{m=1}^{M} g_m(\ell) \cdot \Phi_{n,m}(\ell,k). \]

Differences in the noise levels and colorations over all channels are compensated by the dynamic spectral floor \( b_m(\ell,k) \) (Fig. 3). It is not enforced to do as much noise reduction as possible but as much as necessary to compensate for the mentioned different noise characteristics. For adequate performance of the NR it might be advantageous to introduce a limit:
\[ b_m(\ell,k) \in [b_{\min}, b_{\max}] \quad \text{with} \quad b_{\min} \leq \beta^{\text{ref}} \leq b_{\max}. \]

If the AGC weights are in the range
\[ \beta^{\text{ref}} \cdot \sqrt{\frac{\Phi_n^{\text{ref}}(\ell-1,k)}{\Phi_{n,m}(\ell,k)}} < a_m(\ell) < \frac{\beta^{\text{ref}}}{b_{\min}} \cdot \sqrt{\frac{\Phi_n^{\text{ref}}(\ell-1,k)}{\Phi_{n,m}(\ell,k)}}. \]

the processing works fine otherwise residual switching effects may be audible. To obtain the processed signals the filter coefficients from Eq. 11 are applied to the complex-valued signal in the frequency domain:
\[ Y_m(\ell,k) = H_m(\ell,k) X_m(\ell,k). \]

Thus, all signals show similar noise characteristics and a smooth transition period between the particular active speaker channels. Differences in the strength of the noise signals are tolerated but only come to the fore after some time if only one speaker is the dominant one.

### 5. Signal Combining

The processed signals now have to be combined to get one single output signal. As the weights for combining the signals can be chosen independently from the dominance weights, a
variety of different methods can be applied. Here we assume

single talk situations where only one speaker is active at the

time. Thus, for our application it is appropriate to use

real-valued fullband weights \( w_m(\ell) \):

\[
Y_{\text{mix}}(\ell, k) = \sum_{m=1}^{M} w_m(\ell) Y_m(\ell, k). \tag{17}
\]

Due to the adjustment of the different signal characteristics

in all the channels we can switch between the active speakers

without noticing any switching effects (Fig. 3). The weights

\( w_m(\ell) \in \{0, 1\} \) are determined by the VAD and are hold until

another speaker becomes active.

6. EVALUATION

The proposed system has been evaluated with signals measured in cars driving at around 90 km/h and 130 km/h while four alternately speaking persons, two at the front seats and two at the rear seats, each have a dedicated microphone. Also adverse noise scenarios with an open window have been considered. A subjective listening test has been performed where three signal combining methods have been compared: Hard switching between the noise reduced channel signals with a fixed spectral floor \( b = 0.4 \), the proposed method for dynamic signal combining \((b^{\text{ref}} = 0.4, b^{\text{min}} = 0.1, b^{\text{max}} = 3)\) and a diversity approach \([4]\). 10 test persons listened to 17 speech signal sets. In each set one signal was processed by each of the three different methods. The challenge was to sort the resulting signals by their quality starting with the best (index 1) and ending with the worst (index 3). The subjects could listen to the signals as often as they liked to. The speech quality, the sound of the noise and the overall impression had to be valued. The results from the test can be seen in Fig. 4. The simple hard switching between the channels shows poor results which may come from annoying noise jumps. With the other methods a constant background noise is achieved and the presented method of dynamic signal combining yields the best results. The diversity method has been considered as a reference for good speech quality because it is specially designed to achieve this purpose. But here an unnatural sounding background noise appears. For the overall impression also

the background noise seems to be crucial. Thus, the proposed
dynamic signal combining approach with its natural sound

and smooth noise transitions comes off well.

7. CONCLUSION

In this contribution a new method for dynamic signal combining supporting several speakers in noisy environments is presented. Two different sets of weights are used which can be controlled independently: The mixer weights have to vary very fast to capture speech onsets after a speaker change, whereas the dominance weights are adjusted more slowly to specify the desired signal characteristics for the resulting signal. Thus, smooth transitions between the microphone signals of the different speakers can be achieved even if the background noise or the speech level differ strongly among the channels. The presented method also can be used as a preprocessor for other mixing approaches with soft or complex-valued weights due to its full independence of these weights.

8. REFERENCES


