ENHANCED SPEAKER ACTIVITY DETECTION FOR DISTRIBUTED MICROPHONES BY EXPLOITATION OF SIGNAL POWER RATIO PATTERNS

Timo Matheja, Markus Buck, Tobias Wolff

Nuance Communications Aachen GmbH, Site Ulm
Acoustic Speech Enhancement Research
D-89077 Ulm, Germany

ABSTRACT

In cars with integrated distributed microphone systems usually each speaker has a dedicated microphone. An often required broadband speaker activity detection can be performed by simply evaluating the power ratios among the microphones but transient interferers like indicator noise, outside crossing cars or speech from interfering speakers may be wrongly assigned to one speaker's activity. In this contribution a new method is presented that exploits patterns based on the characteristics of signal power inverted subbands at which compared to the closest microphone higher energy occurs in a distant one. By determination of a distance measure the currently observed pattern of these power inverted subbands is compared to online learned speaker position dependent reference patterns. During noise only periods as well as during transient interfering signals the patterns do not match the reference and false speaker activity detections can be reduced.

Index Terms— distributed microphones, speaker activity detection, voice activity detection, teleconferencing

1. INTRODUCTION

Regarding the application of speech technology in the automotive environment supporting multiple speakers on different positions in a car at the same time becomes more important. It should be possible to perform conference calls out of the car with multiple passengers included as well as to offer the control of a speech recognition system by all the passengers. Beside beamforming techniques there exist other multi-channel methods where distributed microphones are mounted in the near vicinity of the speakers. Thus, each speaker has a dedicated microphone that captures his speech at the best.

Often it has to be known which speaker is speaking for instance to control adaptive filters like in [1] or to transmit solely the desired command to a speech recognizer. In signal combining systems for distributed microphones, like proposed in [2] or in [3], interfering signals should not impair the essential control mechanism. But a simple approach based only on the evaluation of power ratios [4, 1] or on the signal-to-noise-ratio alone may cause false detections when interfering transients occur, as the power ratios might be similar to those occurring during speech periods.

In this contribution a new extension to the simple power ratio based approach is presented. The room acoustics are considered by evaluation of the position of the so-called power inverted subbands. In these special subbands that can be observed if the power ratios among the microphones are computed a distant microphone shows – due to the room acoustics – a higher amount of energy than the speaker's dedicated microphone. The number of the power inverted subbands is limited for the activity of a desired speaker and their location is characteristic of his position.

For speech segmentation in a meeting situation with small distances between the microphones, e.g., in [5] some other features representing spatial cues are considered. A position estimation using binaural signals and interaural level differences in the cepstrum domain is proposed in [6]. Like in some other methods there the occurring patterns are trained and evaluated by a Gaussian mixture model. It is also possible to apply statistical models based on phase differences for the speaker activity detection (SAD) like in [7]. But in the present contribution the benefit of speaker dedicated microphones is exploited. Therefore solely power ratios are considered that are evaluated by applying a distance measure between a characteristic reference and the observation.

This paper is organized as follows. In Sec. 2 an overview of the presented system is given. Sec. 3 describes the determination of the power ratios and of a measure highlighting the power inverted subbands. The SAD with the computation of a distance measure is described in Sec. 4 as well as the update procedure of the reference pattern set. At the end of the paper an evaluation and a conclusion follow.

2. SYSTEM OVERVIEW

An overview of the proposed system is depicted in Fig. 1. In a system with $M \geq 2$ microphones first the signal power ratio (SPR) is computed for each channel m. Based on the inverse SPR a measure is determined to highlight the position of the power inverted subbands. After a linear prediction analysis

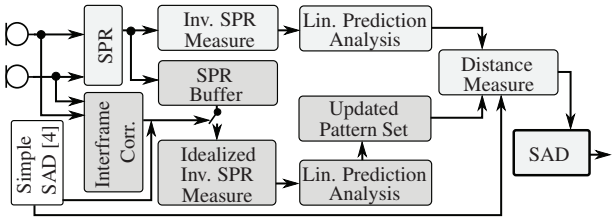

Fig. 1. System overview.

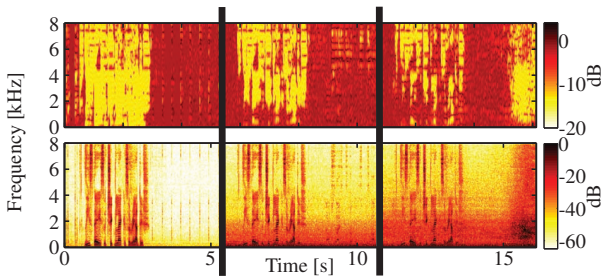

Fig. 2. $Y_{\log}(\ell,k)$ [top] and $10\log_{10}(\hat{\Phi}_{\mathrm{xx},1}(\ell,k))$ [bottom] at various conditions: 0 km/h + indicator, 130 km/h + interfering speech, 80 km/h + slightly open driver window + crossing car.

the distances between the resulting current prediction spectrum and the entries of a reference pattern set are evaluated. Based on the minimum of these distance measures the broadband SAD is determined. The reference pattern set itself is updated whenever a simple SAD [4] and an interframe correlation indicate speech activity for a past frame. Then the SPR from this preceding frame is used to generate a new pattern to be added to the reference pattern set. The processing is done in the subband domain where ℓ denotes the frame index and k the frequency subband index. The short-time Fourier transform uses a Hann window and a block length of $N_{\mathrm{DFT}} = 512$ samples with 75 % overlap at a sampling frequency of 16000 Hz.

3. POWER RATIOS AND ROOM ACOUSTICS

Depending on the position of a sound source in a room the signal power ratios among the microphones are changing. The position of the power inverted subbands can be considered as a characteristic feature of the active speaker's location. In the following the necessary quantities are described in detail.

3.1. Determination of Power Ratios

Each microphone signal can be modeled by a superposition of a speech and a noise signal component. Assuming speech and noise are uncorrelated the speech signal power spectral density (PSD) estimate can be obtained by

$$\hat{\Phi}_{\mathrm{ss},m}(\ell,k) = \max\left\{\hat{\Phi}_{\mathrm{xx},m}(\ell,k) - \hat{\Phi}_{\mathrm{nn},m}(\ell,k), 0\right\} , \quad (1)$$

where $\hat{\Phi}_{\mathrm{xx},m}(\ell,k)$ is the observed signal PSD estimate and $\hat{\Phi}_{\mathrm{nn},m}(\ell,k)$ is the noise PSD for instance estimated by the improved minimum controlled recursive averaging approach [8]. Thus, for each channel m a power ratio can be estimated that is limited to a value of 10 afterwards:

$$\widehat{\mathrm{SPR}}_m(\ell,k) = \frac{\max\left\{\hat{\Phi}_{\mathrm{ss},m}(\ell,k), \epsilon\right\}}{\max\left\{\frac{1}{M-1}\sum_{i=1,i\neq m}^{M}\hat{\Phi}_{\mathrm{ss},i}(\ell,k), \epsilon\right\}} . \quad (2)$$

Here ϵ is a very small value. In the following the index m is left out for reasons of clarity. A temporal smoothing yields

$$\overline{\mathrm{SPR}}(\ell,k) = \alpha\cdot\overline{\mathrm{SPR}}(\ell-1,k) + (1-\alpha)\cdot\widehat{\mathrm{SPR}}(\ell,k) , \quad (3)$$

with the smoothing constant $\alpha = 0.1$. Similarly, a smoothing along the frequency axis is performed resulting in $\widetilde{\mathrm{SPR}}(\ell,k)$.

3.2. Highlighting Power Inverted Subbands

If power inverted subbands occur $\widetilde{\mathrm{SPR}}(\ell,k)$ shows small values. Assuming a dedicated microphone for the active speaker the number of these power inverted subbands is limited and their position is characteristic. To highlight this effect a mapping is applied and the following quantity is proposed:

$$\mathcal{X}(\ell,k) = \max\left(1 - \frac{\widetilde{\mathrm{SPR}}(\ell,k)}{4}, \gamma\right) , \quad (4)$$

where γ is a lower limit (e.g.,$\gamma = 0.05$). The factor $1/4$ ensures the highlighting of also anomalous high attenuated but not real inverse subbands. A linear prediction analysis based on $\mathcal{X}(\ell,k)$ yields a smooth spectrum indicating the position of the power inverted subbands. The autocorrelation coefficients are computed by the inverse discrete Fourier transform of the magnitude squares of $\mathcal{X}(\ell,k)$:

$$R_{i_{\mathrm{DFT}}}(\ell) = \sum_{k=0}^{N_{\mathrm{DFT}}-1} |\mathcal{X}(\ell,k)|^2 \cdot \exp\left(j\frac{2\pi\cdot k\cdot i_{\mathrm{DFT}}}{N_{\mathrm{DFT}}}\right) , \quad (5)$$

with $i_{\mathrm{DFT}} \in \mathbb{Z}$. Thus, the Yule-Walker auto-regressive equations for solving the prediction problem with a sufficient prediction order $N_p = 40$ and the filter coefficients $a_\nu(\ell)$ are

$$R_p(\ell) = \sum_{\nu=1}^{N_p} a_\nu(\ell)\cdot R_{\nu-p}(\ell) \quad , p = 1,2,...,N_p . \quad (6)$$

After applying the Levinson Durbin algorithm and using the frequency response of the filter coefficients $a_\nu(\ell)$ represented by $A(\ell,k)$ the logarithmic estimate of $\mathcal{X}(\ell,k)$ is recovered by

$$Y_{\log}(\ell,k) = 10\log_{10}\left(\left|\frac{e(\ell)}{1-A(\ell,k)}\right|\right) , \quad (7)$$

with the residual prediction error $e(\ell)$ used for normalization. In Fig. 2 some time frequency representations of $Y_{\log}(\ell,k)$ for different driving conditions and $M = 2$ are depicted as well as the related signal spectra for the first microphone.

4. SPEAKER ACTIVITY DETECTION

To achieve a SAD based on Eq. 7 a measure has to be determined that evaluates the characteristics of the power inverted subbands to distinguish between the activity of a desired speaker, interfering transients and background noise. Often a correct SAD is possible by simply evaluating the number of power inverted subbands that is limited if a source like a speaker is close to one microphone. Background noise or interfering speakers mostly do not originate close to the desired primary microphone and cause a larger number of power inversions. But this is not a distinguishing feature if an interfering sound originates from a direction near the desired speech source. Thus, it is advantageous to focus the evaluation of the patterns on exploiting the characteristic positions of the power inverted subbands.

4.1. Distance Measure

An Euclidean distance measure $J_i(\ell,k)$ between each pattern $Y^{\mathrm{ref}}_{i,\log}(k)$ out of a reference pattern set with $i = 1,...,N_{\mathrm{Y}}$ patterns and the currently observed pattern $Y_{\log}(\ell,k)$ is determined:

$$J_i(\ell,k) = \left(Y^{\mathrm{ref}}_{i,\log}(k) - Y_{\log}(\ell,k)\right)^2 \ . \tag{8}$$

The mean value of the measure $J_i(\ell,k)$ over relevant subbands is a quantity for the detection of the activity of a desired speaker. Thus, the pattern specific distortion measure is:

$$\overline{J}_i(\ell) = \frac{1}{N_{i,\mathrm{J}}} \sum_{k=0}^{N_{\mathrm{DFT}}/2} J_i(\ell,k)\cdot\delta_i(\ell,k) \ . \tag{9}$$

The number of subbands to evaluate for each pattern is described by $N_{i,\mathrm{J}}$, and $\delta_i(\ell,k)$ is an indicator function. To draw reliable conclusions from the distance measure during speech periods a certain signal-to-noise ratio $\xi(\ell,k)$ (e.g., in [5]) has to be exceeded. Furthermore, only those subbands should be evaluated where signal power inverted subbands occur either in $Y^{\mathrm{ref}}_{i,\log}(k)$ or in $Y_{\log}(\ell,k)$. Thus, $\delta_i(\ell,k)$ can be formulated:

$$\delta_i(\ell,k) = \begin{cases} 1 \ , \text{if } \eta_i(\ell,k)>1 \wedge ((\xi(\ell,k)>1 \wedge \Psi_{\mathrm{sim}}(\ell)>0) \\ \qquad \vee (\xi(\ell,k)>0 \wedge \Psi_{\mathrm{sim}}(\ell)<1)) \\ 0 \ , \text{else.} \end{cases} \tag{10}$$

where $\Psi_{\mathrm{sim}}(\ell)$ is the simple broadband SAD like presented in [4], and $\eta_i(\ell,k)$ denotes the indication of inverted subbands:

$$\eta_i(\ell,k) = \begin{cases} 1 \ , \text{if } \left(Y^{\mathrm{ref}}_{i,\log}(k) > -12\right) \vee (Y_{\log}(\ell,k) > -12) \\ 0 \ , \text{else.} \end{cases} \tag{11}$$

The number of effective subbands in Eq. 9 results in:

$$N_{i,\mathrm{J}} = \max\left\{ \sum_{k=0}^{N_{\mathrm{DFT}}/2} \delta_i(\ell,k), 1 \right\} \ . \tag{12}$$

Based on Eq. 9 the distortion measure $\overline{J}_q(\ell)$ has to be chosen out of the N_{Y} possibilities that shows the minimum distortion and therewith the best match between observation and reference pattern. The index q results in:

$$q = \underset{i\in\{1,...,N_{\mathrm{Y}}\}}{\operatorname{argmin}} \left\{\overline{J}_i(\ell)\right\} \ . \tag{13}$$

Thus, the reference pattern $Y^{\mathrm{ref}}_{q,\log}(k)$ represents the current pattern at the best and causes the minimum distance. This is comparable to a vector quantization. To get furthermore a smoother output a minimum over $L = 30$ past frames is determined in those situations where broadband speech is detected:

$$\tilde{J}(\ell) = \min\left\{\overline{J}_q(\ell), \overline{J}_q(\ell-1), ..., \overline{J}_q(\ell-L)\right\} \ . \tag{14}$$

With Eq. 14 the match of the observed pattern with the pattern set is quantified. The resulting SAD indicator function $\Psi(\ell)$ is obtained by comparing $\tilde{J}(\ell)$ with a threshold β:

$$\Psi(\ell) = \begin{cases} 1 & , \text{if } \tilde{J}(\ell) < \beta \\ 0 & , \text{else.} \end{cases} \tag{15}$$

4.2. Online Learning of Patterns

The single patterns $Y^{\mathrm{ref}}_{i,\log}(k)$ of the whole reference pattern set have to be estimated. Due to changing room acoustics always new patterns can occur. Thus, the pattern set has to be updated within a first-in first-out system by including new patterns $Y^{\mathrm{ref}}_{i,\log}(k)$ during the processing if desired speech can be assumed. A detection of voice only frames can be determined by evaluating an interframe correlation measure at which the current frame is correlated with the past L_{corr} frames. Because in speech periods the time frames are assumed to be correlated the mean value over the absolute values of these correlation measures over the frames and the frequencies is a further indicator for past broadband voice activity. For applying the pattern approach in these regions all necessary values are stored and processed later. The occurring delay is tolerated and causes only a delayed update of the pattern set.

An idealized pattern has to be determined to only consider those regions where the possibility of occurring power inverted subbands is given. Instead of simply including the currently appearing spectrum from Eq. 7 into the reference pattern set it is proposed to use a modified measure for the power inverted subbands as an alternative to Eq. 4:

$$\mathcal{X}^{\mathrm{ref}}_i(n,k) = \begin{cases} \mathcal{X}(n,k) & , \text{if } \delta^{\mathrm{ref}}_i(n,k) > 0 \\ 0.05 & , \text{else.} \end{cases} \tag{16}$$

Here $n = \ell - L_{\mathrm{corr}}$ denotes the index of the past voice active frame based on the interframe correlation measure. The threshold $\delta^{\mathrm{ref}}_i(n,k)$ is determined by

$$\delta^{\mathrm{ref}}_i(n,k) = \begin{cases} 1 \ , \text{if } (Y_{\log}(n,k) > -6) \wedge (\Psi_{\mathrm{sim}}(n) > 0) \\ \qquad \wedge \ (\xi(n,k) > 0.5) \\ 0 \ , \text{else.} \end{cases} \tag{17}$$

	0 km/h	130 km/h	160 km/h	80 km/h *
S only	**2.05**/1.94	**3.08**/4.22	**3.20**/4.34	**4.34**/5.25
S+N_{ind}	**5.14**/7.88	**3.54**/7.99	**3.20**/6.74	**5.48**/7.31
S+N_{S}	**2.97**/19.29	**3.08**/17.01	**2.85**/12.90	**4.68**/11.42
S+N_{car}	–	–	–	**8.22**/29.79

Table 1. Error rates (%): Proposed approach (bold) and [4]. [S: desired speech, N_{ind}: indicator, N_{S}: interfering speech, N_{car}: crossing car, *: slightly open driver's window].

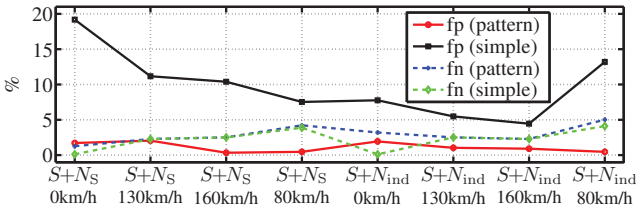

Fig. 3. fp and fn error rates for the different situations and methods. The driver's window is slightly open at 80 km/h.

The further processing is similar to that described in section 4.1. Subsequently after the linear prediction we can update the reference pattern set with a new entry $Y^{\mathrm{ref}}_{i,\log}(k)$.

5. EVALUATION

For test signal generation speech and noise signal components that were measured independently before are superposed with the restriction that interfering transients and desired speech do not overlap. The two microphones are mounted in the a-pillar of a car, one dedicated to the driver and one to the front passenger. The resulting signals are processed by the proposed approach and by the method presented in [4]. Instead of comparing the pattern approach to [4] similar results can be achieved if a broadband SAD determined out of the frequency selective SAD proposed in [1] would be used. To compute errors for the SAD for different situations a reference broadband SAD is determined by applying a threshold to the clean speech component signal whereupon it is compared to the results of the two processings. Approximately 30 % of each file shows broadband speech and the used parameters are $\beta = 13$ and $L_{\mathrm{corr}} = 15$. For the evaluation a reliable pattern is already inclosed in the pattern set that should always be initialized with a possible pattern, i.e., known from former runs.

The overall error results are depicted in Tab. 1. In nearly all situations the pattern approach enhances the simple broadband SAD especially if interfering transients occur. In Fig. 3 it is differentiated between the false-negative (fn) and false-positive (fp) rates. Whereas the fn-rate is similar for both methods the fp-rate is lower for the pattern approach due to correct decisions during interfering transients.

6. CONCLUSION

In this contribution a new method for an enhanced broadband SAD based on exploiting signal power ratio patterns is presented. While a simple evaluation of the power ratios between the microphones includes undesired interfering signals into the SAD such false detections can be considerably reduced with this new approach. During activity of a desired speaker power inverted subbands – where a distant microphone shows higher energy than the one close to the speaker – induce characteristic patterns due to the room acoustics. By comparing these patterns to a reference pattern set the activity of a desired speaker can be detected with considerably fewer false detections during transient interfering signals.

7. REFERENCES

[1] A. Lombard and W. Kellermann, "Multichannel cross-talk cancellation in a call-center scenario using frequency-domain adaptive filtering," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, Washington, September 2008.

[2] T. Matheja, M. Buck, and A. Eichentopf, "Dynamic signal combining for distributed microphone systems in car environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011.

[3] J. Freudenberger, S. Stenzel, and B. Venditti, "Microphone diversity combining for in-car applications," *EURASIP Journal on Advances in Signal Processing*, 2010.

[4] T. Matheja and M. Buck, "Robust voice activity detection for distributed microphones by modeling of power ratios," in *9. ITG-Fachtagung Sprachkommunikation*, Bochum, October 2010.

[5] E. Cheng, J. Lukasiak, I. S. Burnett, and D. Stirling, "Using spatial cues for meeting speech segmentation," in *IEEE International Conference on Multimedia and Expo (ICME 2005)*, July 2005.

[6] M. Takimoto, T. Nishino, H. Hoshino, and K. Takeda, "Estimation of speaker and listener positions in a car using binaural signals," *Acoustical Science and Technology*, vol. 29, no. 1, pp. 110 – 112, 2008.

[7] J.-S. Hu, C.-C. Cheng, and W.-H. Liu, "A robust statistical-based speaker's location detection algorithm in a vehicular environment," *EURASIP Journal on Advances in Signal Processing*, 2007.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466 – 475, September 2003.