COMPARISON OF MULTICHANNEL DOUBLETALK DETECTORS FOR ACOUSTIC ECHO CANCELLATION

Martin Schneider[1,2] *and Emanuël A. P. Habets*[1,2]

[1] Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[2] International Audio Laboratories Erlangen*, Erlangen, Germany

ABSTRACT

In acoustic echo cancellation (AEC) a doubletalk detector (DTD) is typically used to avoid misconvergence of the adaptive filter during simultaneous activity of near-end acoustic scene and loudspeaker playback. While several single-channel DTDs can be generalized to multiple channels, only little attention was so far paid to the evaluation of the resulting multichannel DTDs. In this paper, different DTDs are reviewed and evaluated. In particular, the influence of the number of loudspeakers is investigated as new dimension in the experimental evaluation, where up to sixteen loudspeaker channels are considered. The results show that the performance of a DTD is affected by the number of loudspeaker channels whenever the cross-correlation between the loudspeaker signals is considered by the DTD. Moreover, considering this cross-correlation turned out to be necessary for a high detection performance.

Index Terms— Doubletalk detection, multi-channel, acoustic echo cancellation, comparison, evaluation

1. INTRODUCTION

Acoustic echo cancellation (AEC) is a commonly applied technique to remove a loudspeaker echo in a recorded microphone signal, which would hinder the far-end party in a teleconference or degrade the performance of an automatic speech recognizer. This is accomplished by first identifying the acoustic path between loudspeaker(s) and microphone by means of a multi-channel adaptive filter that provides an estimate of the loudspeaker echo signal, which is then subtracted from the microphone signal. Since the filter adaptation is based on the estimated joint statistics of the loudspeaker signals and the microphone signal, any contribution of the near-end acoustic scene to the microphone signal will impair the acoustic path identification. Hence, a doubletalk detector (DTD) is used to detect activity of a near-end speaker such that the adaptation can be held in that case [1]. Actually, the DTD (which provides a hard decision) can also be used to control the step-size (cf. [2]): If doubletalk is detected the step-size is set to zero. Robust adaptation algorithms [3,4] and two-path models [5] are an alternative to using DTDs, while they can also be combined with a DTD [6].

Using more than one loudspeaker for spatial audio reproduction implies using a DTD considering all loudspeaker signals. Many single-channel detectors have been proposed [7–11] that can be generalized to the multi-channel case, as described in [12, 13]. Even in the single-channel case, a comprehensive evaluation and comparison of those detectors has drawn less attention in the literature [14], where a comparison of the detector performance in multi-channel scenarios appears to be entirely missing [15]. This might be due to the fact that an analytical comparison appears to be infeasible without making assumptions that are typically violated in practice. For example, signals would have to be assumed as being stationary, which does not hold for speech as a commonly used recording/reproduction signal. Moreover, experimental studies might never exhaust the full range of aspects that could be relevant in an individual scenario. The latter approach can, nevertheless, provide helpful information for the application of the considered techniques. While there are many factors that influence the performance of the DTD, we investigate in this work the influence of the number of loudspeaker signals that are considered by the detector, and used the experimental evaluation in [14] as a guideline.

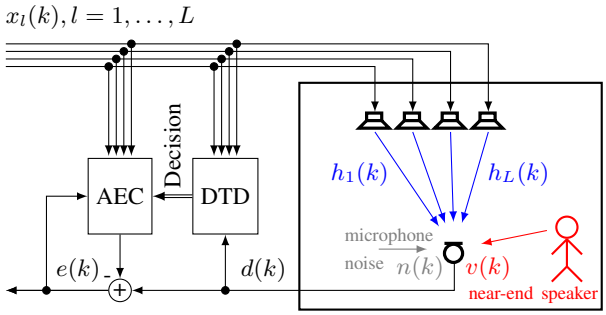

Fig. 1. General signal model of the combination of AEC with a DTD

Unfortunately, not all DTDs mentioned above can be discussed due to space restrictions, such that the focus is on following DTDs: The Geigel algorithm [7], the consideration of the cross-correlation between loudspeaker and microphone signals [8], and a generalization of the latter approach that also considers the cross-correlation between the loudspeaker signals [10]. The latter two are referred to as the *cross-correlation method* and the *normalized cross-correlation method* in this paper. The *normalized cross-correlation method* is closely related to a coherence-based DTD proposed in [16] that is not evaluated here.

This paper is structured as follows: In Sec. 2 the considered signal model is explained, before the actual DTDs are described in Sec. 3. The evaluation scenario and the considered measures are explained in Sec. 4, preceding the presentation of results in Sec. 5. Conclusions follow in Sec. 6.

2. SIGNAL MODEL AND PROBLEM STATEMENT

In this section, the considered signal model is described. This is followed by a brief consideration of the second-order statistics (SOS) of microphone and loudspeaker signals that is a prerequisite for Sec. 3.

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS)

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

In the following, L loudspeakers and a single microphone, as shown in Fig. 1, are considered. The L discrete loudspeaker signals are represented by $x_l(k)$, where l indexes the loudspeaker and k the time instant. For block processing, it is convenient to stack the L loudspeaker signals in a single vector that is given by

$$\mathbf{x}(k) = \left(\mathbf{x}_1^T(k), \mathbf{x}_2^T(k), \ldots, \mathbf{x}_L^T(k)\right)^T, \quad (1)$$

$$\mathbf{x}_l(k) = (x_l(k-K+1), x_l(k-K+2), \ldots, x_l(k))^T, \quad (2)$$

where K is the number of time samples considered in each $\mathbf{x}_l(k)$ and $\cdot^T$ denotes the transpose operation.

The room impulse response (RIR) of length K from loudspeaker l to the microphone is denoted by $h_l(k)$ and captured by the vector

$$\mathbf{h} = \left(\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_L^T\right)^T, \quad (3)$$

$$\mathbf{h}_l = (h_l(K-1), h_l(K-2), \ldots, h_l(0))^T. \quad (4)$$

The microphone signal is described by

$$d(k) = \mathbf{x}^T(k)\mathbf{h} + v(k) + n(k), \quad (5)$$

where $n(k)$ describes the microphone noise and $v(k)$ is the contribution of the near-end speaker. The task of a DTD is to detect a contribution of $v(k)$ to $d(k)$, where statistics of different order can be considered.

For the SOS of the transducer signals, $v(k)$ and $n(k)$ can be assumed to be zero mean and uncorrelated (i. e. orthogonal) to $x_l(k)\,\forall\, l, k$. Hence, the following relations hold:

$$r_{\mathrm{dd}} = \mathcal{E}\left\{d^2(k)\right\} = \mathcal{E}\left\{d(k)\mathbf{x}^T(k)\right\}\mathbf{h} + \sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{n}}^2, \quad (6)$$

$$\mathbf{r}_{\mathrm{xd}} = \mathcal{E}\left\{\mathbf{x}(k)d(k)\right\} = \mathcal{E}\left\{\mathbf{x}(k)\mathbf{x}^T(k)\right\}\mathbf{h} = \mathbf{R}_{\mathrm{xx}}\mathbf{h}, \quad (7)$$

$$r_{\mathrm{dd}} = \mathbf{r}_{\mathrm{xd}}^T\mathbf{R}_{\mathrm{xx}}^{-1}\mathbf{r}_{\mathrm{xd}} + \sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{n}}^2 = \mathbf{r}_{\mathrm{xd}}^T\mathbf{h} + \sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{n}}^2, \quad (8)$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, r_{dd} is the second-order moment of $d(k)$, $\mathbf{r}_{\mathrm{xd}}$ is the cross-correlation vector between $\mathbf{x}(k)$ and $d(k)$, σ_{v}^2 is the variance of $v(k)$, σ_{n}^2 the variance of $n(k)$, and $\mathbf{R}_{\mathrm{xx}}$ is the autocorrelation matrix of $\mathbf{x}(k)$ that also describes the cross-correlation between the loudspeaker signals. Note that (7) is the matrix form of the so-called normal equations that underlie system identification exploiting SOS. Accordingly, filter adaptation algorithms aim at approximating a solution given by [17]

$$\mathbf{h} = \mathbf{R}_{\mathrm{xx}}^{-1}\mathbf{r}_{\mathrm{xd}}. \quad (9)$$

For implementations, exponential averaging can be used to obtain estimates of r_{dd}, $\mathbf{r}_{\mathrm{xd}}$ and $\mathbf{R}_{\mathrm{xx}}$, given by $\hat{r}_{\mathrm{dd}}(k)$, $\hat{\mathbf{r}}_{\mathrm{xd}}(k)$, and $\hat{\mathbf{R}}_{\mathrm{xx}}(k)$, respectively. An example of such a computation is given by

$$\hat{\mathbf{r}}_{\mathrm{xd}}(k) = (1-\lambda)\sum_{\kappa=0}^{k}\lambda^{\kappa}\mathbf{x}(k-\kappa)d(k-\kappa), \quad (10)$$

where λ is the exponential forgetting factor.

3. DOUBLETALK DETECTION

During operation, the DTD compares an algorithm-specific test statistic $\xi_{\mathrm{DA}}(k)$ with a given threshold T_{DA}, where subscript DA is a placeholder for the actually considered detection algorithm. Whenever a test statistic is below the threshold, doubletalk is assumed:

$$\xi_{\mathrm{DA}}(k) < T_{\mathrm{DA}} : \text{doubletalk},$$
$$\xi_{\mathrm{DA}}(k) \geq T_{\mathrm{DA}} : \text{no doubletalk},$$

where the actually used test statistics are described in the following.

The **Geigel algorithm** [7] is the simplest under consideration with the test statistics

$$\xi_{\mathrm{G}}(k) = \frac{\|\mathbf{x}(k)\|_{\infty}}{|d(k)|}, \quad (11)$$

where $\|\cdot\|_{\infty}$ describes the infinity norm, i. e., the maximum absolute value of the components in a vector.

A different DTD is the **cross-correlation method**, which was originally proposed to be applied to the error signal of an adaptive filter (like $e(k)$ in Fig. 1) [8], but can be applied also to the microphone signals instead, as shown in [14]. The latter approach is considered in the following, which relies on the fact that the loudspeaker echo in $d(k)$ is strongly correlated to the loudspeaker signals $\mathbf{x}_l(k)$, while the near-end signal $v(k)$ is not. The test statistic is then given by

$$\xi_{\mathrm{C}}(k) = \frac{\|\hat{\mathbf{r}}_{\mathrm{xd}}(k)\|_p^2}{\hat{r}_{\mathrm{dd}}(k)\hat{r}_{\mathrm{xx}}(k)}, \quad (12)$$

where p defines a chosen norm and $\hat{r}_{\mathrm{xx}}(k)$ is an estimate for the loudspeaker signal power.

Finally, the **normalized cross-correlation method** proposed in [10] is explained. This approach exploits the fact that $\hat{r}_{\mathrm{dd}}(k)$ is strongly affected by $v(k)$, while the influence of $v(k)$ on $\hat{\mathbf{r}}_{\mathrm{xd}}(k)$ and $\hat{\mathbf{R}}_{\mathrm{xx}}(k)$ vanishes on average due to statistical orthogonality [10]. This motivates using the test statistic

$$\xi_{\mathrm{N}}(k) = \frac{\hat{\mathbf{r}}_{\mathrm{xd}}^T(k)\hat{\mathbf{R}}_{\mathrm{xx}}^{-1}(k)\hat{\mathbf{r}}_{\mathrm{xd}}(k)}{\hat{r}_{\mathrm{dd}}(k)}, \quad (13)$$

which is equal to one if $\sigma_{\mathrm{v}}^2 = \sigma_{\mathrm{n}}^2 = 0$ and below one whenever there is any power in the near-end speaker signal or the noise signal. The influence of noise is typically much weaker than the influence of the near-end signal. The computation of (13) is very expensive, especially in multi-channel scenarios. Hence, approximations are used in typical implementations, where an adaptive filter provides an estimate $\hat{\mathbf{h}}(k)$ of $\mathbf{h}$, such that (13) can be approximated by (cf. (8) and (9))

$$\xi_{\mathrm{N}}(k) \approx \frac{\left|\mathbf{r}_{\mathrm{xd}}^T(k)\hat{\mathbf{h}}(k)\right|}{\hat{r}_{\mathrm{dd}}(k)}. \quad (14)$$

It should be noted that while the numerator of (14) is non-negative, this is not guaranteed for $\mathbf{r}_{\mathrm{xd}}^T(k)\hat{\mathbf{h}}(k)$ whenever an adaptive filter suffers misconvergence. Although this is difficult to interpret from a theoretical perspective, doubletalk will likely result in misconvergence. Hence, using the absolute value of $\mathbf{r}_{\mathrm{xd}}^T(k)\hat{\mathbf{h}}(k)$ in (14) improves the performance of the DTD significantly.

Three different ways to compute $\xi_{\mathrm{N}}(k)$ using (14) to obtain $\hat{\mathbf{h}}(k)$ are considered:

1. Using a perturbed RIR according to $\hat{\mathbf{h}}(k) = \mathbf{h} + \mathbf{n}(k)$ where $\mathbf{n}(k)$ is a Gaussian noise vector with an energy of $-30\,\mathrm{dB}$ relative to the energy captured in $\mathbf{h}$, i.e. $\mathbf{h}^T\mathbf{h}$. The resulting test statistic is denoted by $\xi_{\mathrm{N}}^{\mathrm{N}}(k)$.

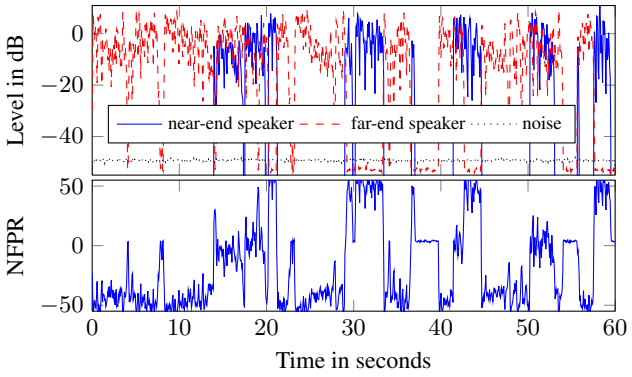

Fig. 2. Power envelope of microphone signal contributions (top) and resulting near-end-to-far-end power ratio (NFPR) (bottom)

2. Using the normalized least-mean squares (NLMS) algorithm [17] to obtain $\hat{\mathbf{h}}(k)$, in which case the test statistic is denoted by $\xi_{\mathrm{N}}^{\mathrm{L}}(k)$.
3. Using the generalized frequency-domain adaptive filtering (GFDAF) algorithm [18, 19] to obtain $\hat{\mathbf{h}}(k)$, in which case the test statistic is denoted by $\xi_{\mathrm{N}}^{\mathrm{G}}(k)$.

Note that the latter two ways consider $\mathbf{R}_{\mathrm{xx}}$ implicitly through (9), while the first way ignores $\mathbf{R}_{\mathrm{xx}}$ entirely.

4. EVALUATION SCENARIO

The evaluation scenario considered in this paper is a simulated teleconference using multi-channel audio reproduction. The recording of human speakers has been simulated by the convolution of anechoic speech signals with measured RIRs, followed by the addition of noise at a level of $-50\,\mathrm{dB}$ relative to the average level of the respective reverberated signals during activity. The loudspeaker signals are generated considering the far-end speaker signal $\mathring{v}(k)$ that is convolved with the RIRs $\mathring{g}_l(k)$, measured from a single loudspeaker to L different microphone positions. After that, independent white Gaussian noise (WGN) signals $\mathring{n}_l(k)$ are added to the loudspeaker signals. The resulting signal is then described by

$$x_l(k) = \mathring{n}_l(k) + \sum_{\kappa=0}^{K} \mathring{v}(k-\kappa)\mathring{g}_l(\kappa). \tag{15}$$

Note that far-end quantities are denoted with superscript ring ($\mathring{\cdot}$). For the near-end signal $v(k)$, another RIR was used and the loudspeaker echo was simulated using a further set of RIRs for $h_l(k)$, measured from L loudspeaker positions to a single microphone position. The signals $n(k)$ and $\mathring{n}_l(k)$ consist of mutually uncorrelated white Gaussian noise. Using the definition(15) with a signal-to-noise ratio (SNR) of $50\,\mathrm{dB}$, the autocorrelation matrix $\mathbf{R}_{\mathrm{xx}}$ will always be non-singular but exhibit a higher condition number for larger values of L. As there is only a single human speaker simulated, this condition number will exhibit as steep increase when using two or three loudspeaker channels instead of one. Any further increase of the loudspeaker channel number will have a more moderate effect on the condition number.

To simulate the course of a teleconference, the far-end speaker is first active for a time span of $10\,\mathrm{s}$, while the near-end speaker is not. This allows for sufficient correlation estimates, as required for the DTD, and an initial convergence of the adaptive filters, if applicable. After $10\,\mathrm{s}$, $\mathring{v}(k)$ and $v(k)$ are randomly active. The level of the resulting contributions in the microphone signal is shown in Fig. 2 along with the resulting near-end-to-far-end power ratio (NFPR). This measure quantifies the power ratio of the near-end $v(k)+n(k)$ and the far-end $\mathbf{x}^T(k)\mathbf{h}$ signal contributions in $d(k)$. In 50% of the time instants, only $\mathring{v}(k)$ is active, while the sole activity of $v(k)$ occurs in 20% of the time instants. In 10% percent of the time instants, neither $\mathring{v}(k)$ nor $v(k)$ are active. In the remaining 20% of the time instants, both signal sources are active (i. e. doubletalk occurs), while the noise signals are always active. The minimum time span a source is active is one second, where the human speakers in both signals are alternatingly exchanged every $2.5\,\mathrm{s}$, choosing from 8 recordings of English, German, and French speakers of both genders. The signals $\mathring{v}(k)$ and $v(k)$ were scaled such that their microphone signal contributions have an average level of zero dB during activity. This experiment was repeated multiple times to ensure statistical significance.

Note that this study considers only the properties of the individual test statistics themselves and not the integration of a DTD in an AEC system as depicted in Fig. 1. Thus, the resulting improvement of system identification and echo cancellation as well as the response time of the DTDs were not measured. Instead, the main focus in this work is on the following two quantities:

1. The *false alarm probability*, where a false alarm is a decision for doubletalk in the absence of near-end activity,
2. The *missed detection probability*, where a missed detection is a decision against doubletalk during near-end activity.

Both measures are only computed during far-end activity as there is no filter adaptation otherwise. Due to space restrictions, the DTDs' ability to differentiate between an echo path change and actual doubletalk (as, e. g., considered in [6]) could not be evaluated.

While the *false alarm probability* and the *missed detection probability* are strongly dependent on the chosen threshold T_{DA}, an optimal choice of T_{DA} depends, in turn, on the considered DTD and also on the evaluation scenario. To allow a meaningful comparison of the considered algorithms, the receiver operation curve (ROC) can be considered, which shows the probabilities mentioned above in dependence of each other. To compare the DTDs' performance in situations with different NFPRs, it is necessary to choose a fixed threshold T_{DA}, where an optimal choice dependents on the cost of a missed detection or a false alarm in the chosen application. Lacking such information, the strategy used in [14] was followed, where T_{DA} was chosen for each algorithm individually such that a *false alarm probability* of 0.3 was achieved. Determination of the threshold and evaluation where carried out on the same data to ensure an optimum threshold for each algorithm. This is considered to be crucial for a meaningful comparion of the algorithms, as intended in this paper.

For all simulations, a sampling frequency of $8\,\mathrm{kHz}$ was assumed. Because of computational constraints, the measured impulse responses where truncated to 1024 samples, the DTDs consider only $K = 512$ samples, and the length of the adaptive filters used to compute (14) were set to 512. For the NLMS algorithm, a step size of $\mu = 0.1$ was chosen, while the step size of the GFDAF algorithm was $\mu = 1.5$. For the GFDAF algorithm an update was performed only every 64 time instants for computational reasons and the exponential forgetting factor was chosen to be 0.95. None of these adaptive filters considered the decision of their respective DTD like it would be implemented in real-world scenarios as this could freeze the system after an echo path change i. e. a change of $\mathbf{h}$. An adaptive filter used for AEC would consider this decision as shown in Fig. 1. Note that AEC is beyond the scope of this paper

and, hence, not evaluated in the following. Independently of the adaptive filters, the exponential weight λ for the statistical estimates (see(10)) was set to 0.999.

5. EVALUATION RESULTS

In this section, the evaluation results are presented, starting with the Geigel algorithm ($\xi_{\mathrm{G}}(k)$), the *cross-correlation method* ($\xi_{\mathrm{C}}(k)$) and the *normalized cross-correlation method* using (14) with the noisified true RIR ($\xi_{\mathrm{N}}^{\mathrm{N}}(k)$). The results averaged over 16 repetitions of the experiment described in Sec. 4 are shown in Fig. 3. There, the ROC shows that the Geigel algorithm is not performing well, while considering $\xi_{\mathrm{C}}(k)$ and $\xi_{\mathrm{N}}^{\mathrm{N}}(k)$ leads to improved detection, and is in agreement with the findings reported in [14]. Note that the approximation of $\xi_{\mathrm{N}}(k)$ by $\xi_{\mathrm{N}}^{\mathrm{N}}(k)$ was used in [14] due to its simplicity, but relies on the noisified true RIR that is unavailable in real-world implementations. The disappointing performance of the Geigel algorithm can be explained when considering the missed detection probability for a fixed threshold as a function of the NFPR in the lower plot of Fig. 3. There, it can be seen that the Geigel algorithm needs a high NFPR for a successful detection, which only rarely occurs in the considered scenarios given the chosen signal levels. It can be furthermore seen that the performance of the approaches considered in Fig. 3 is only marginally affected by the number of loudspeaker channels L. This previously undocumented finding can be explained by the underlying test statistics that only rely on the cross-correlation between the loudspeaker and the microphone signals. Although this vector increases in size with an increased number of channels, the new data exhibits very similar properties like the data already considered in the single-channel case. To effectively exploit the small differences in those properties, the cross-correlation between the loudspeaker signals has to be considered. Hence, only slight improvements can be expected, if at all. On the other hand, these results also show that the performance of these simple approaches does not degrade with an increasing number of loudspeakers.

While the information available in real-world scenarios would allow for an exact computation of (13), (14) must be evaluated using an adaptive filter to estimate $\hat{\mathbf{h}}(k)$ as described for $\xi_{\mathrm{N}}^{\mathrm{L}}(k)$ and $\xi_{\mathrm{N}}^{\mathrm{G}}(k)$. In Fig. 4, using $\xi_{\mathrm{N}}^{\mathrm{N}}(k)$, $\xi_{\mathrm{N}}^{\mathrm{L}}(k)$,$\xi_{\mathrm{N}}^{\mathrm{G}}(k)$, and $\xi_{\mathrm{N}}(k)$ are compared. Note that for $\xi_{\mathrm{N}}(k)$ only 4 repetitions of the experiment where conducted because computing $\xi_{\mathrm{N}}(k)$ is extremely expensive, which also precludes using it in any real-world applications. The results show, that $\xi_{\mathrm{N}}^{\mathrm{N}}(k)$ approaches the performance of the $\xi_{\mathrm{N}}(k)$, rather than the performance of an implementation using an actual adaptive filter, which, to the best of our knowledge, has not been documented in the literature so far. Moreover, adaptive filters suffer themselves from the doubletalk activity, which is actually the motivation of using a DTD and implies a conservative parameter choice. In any case, the performance of a DTD is directly depended on the robustness and performance of the adaptive filter. This is why the GFDAF-based test statistic outperforms the approach utilizing the NLMS algorithm, which was, furthermore, reported to be unsuitable in multi-channel scenarios [20].

In Fig. 4 it can be seen that the performance in terms of the ROCs is also reflected in the missed detection probability as a function of the NFPR. Furthermore, it can be seen that the detection performance of $\xi_{\mathrm{N}}(k)$ increases when using multiple loudspeakers. This is because the larger number of loudspeaker channels leads to a lower variance of the test statistics. Seen from a different perspective, each further loudspeaker channel adds additional information that can be used to estimate microphone signal power in the absence of doubletalk more accurately.

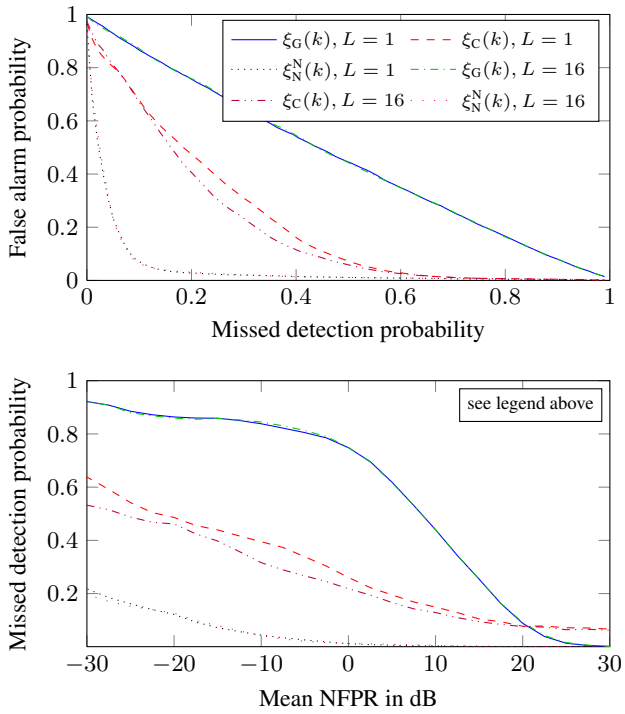

Fig. 3. Receiver operator curves (top) and missed detection probability as a function of NFPR (bottom) for DTDs disregarding the loudspeaker signal cross-correlation.

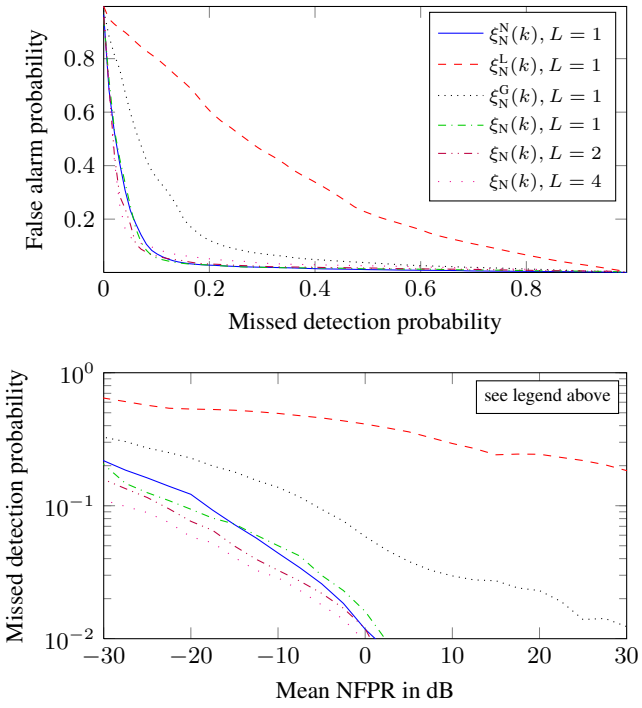

Fig. 4. Receiver operator curves (top) and missed detection probability as a function of NFPR (bottom) for different approximations of the normalized cross-correlation method

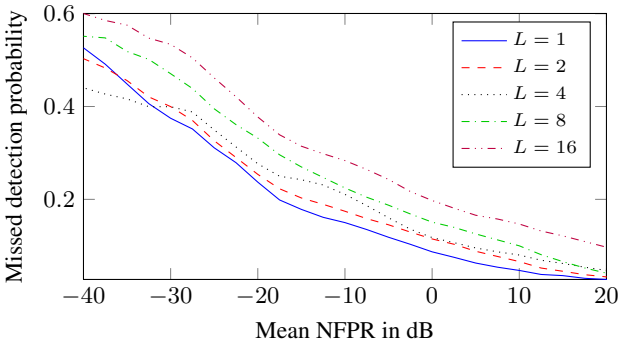

Fig. 5. Missed detection probability as a function of NFPR for $\xi_{\mathrm{N}}^{\mathrm{G}}(k)$

In Fig. 5, the detection performance of $\xi_{\mathrm{N}}^{\mathrm{G}}(k)$ as a function of the NFPR is shown for different numbers of loudspeaker channels. For this experiment, 32 repetitions were conducted to compensate for high variations in the results depending on the convergence of the adaptive filter. When comparing Figs. 4 and 5 a crucial dilemma for this task can be seen: While the increased number of loudspeakers provides more information to be exploited, it degrades the performance of the adaptive filters at the same time. Nevertheless, it is tremendously expensive to compute (13) such that real-world implementations will typically rely on those filters such that the approximation (14) can be used.

6. CONCLUSIONS

The evaluation of different doubletalk detectors showed that methods that only consider the cross-correlation between loudspeaker and microphone signals are not significantly affected by the number of loudspeaker channels. For approaches also considering the cross-correlation of the loudspeaker signals, a conflict was discovered: While the increased number of loudspeakers provides more information that can be effectively exploited by a DTD, it can degrade the performance of adaptive filters that are generally necessary for real-world implementations. Hence, DTDs that do not depend on an explicit estimate of the echo path can be an attractive subject for future research.

REFERENCES

[1] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. an application of very-high-order adaptive filters," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42 – 69, July 1999.

[2] H.-C. Shin, A.H. Sayed, and W.-J. Song, "Variable step-size NLMS and affine projection algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132 – 135, 2004.

[3] S.L Gay, "An efficient, fast converging adaptive filter for network echo cancellation," in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove (CA), USA, Nov. 1998, vol. 1, pp. 394 – 398.

[4] T.S. Wada and Biing-Hwang Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 175–189, Jan. 2012.

[5] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE Transactions on Communications*, vol. 25, no. 6, pp. 589 – 595, 1977.

[6] T. Gänsler, S.L. Gay, M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 656 – 663, Nov. 2000.

[7] D.L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Transactions on Communications*, vol. 26, no. 5, pp. 647 – 653, 1978.

[8] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Transactions on Communications*, vol. 39, no. 11, pp. 1542 – 1545, 1991.

[9] T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communications,*, vol. 44, no. 11, pp. 1421 – 1427, Nov 1996.

[10] J. Benesty, D.R. Morgan, and J.H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 2, pp. 168 – 172, Mar. 2000.

[11] M.A lqbal, J.W. Stokes, and S.L. Grant, "Normalized double-talk detection based on microphone and aec error cross-correlation," in *IEEE International Conference on Multimedia and Expo*, Beijing, China, July 2007, pp. 360 – 363.

[12] T. Gänsler and J. Benesty, "A frequency-domain double-talk detector based on a normalized cross-correlation vector," *Signal Processing*, vol. 81, no. 8, pp. 1783 – 1787, 2001.

[13] M.A. Iqbal, S.L. Grant, and J.W. Stokes, "A frequency domain doubletalk detector based on cross-correlation and extension to multi-channel case," in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove (CA), USA, Nov. 2009, pp. 638 – 641.

[14] J.H. Cho, D.R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 6, pp. 718 – 724, 1999.

[15] M. Schneider and W. Kellermann, "Large-scale multiple input/multiple output system identification in room acoustics," in *Proc. Intl. Congress on Acoustics (ICA)*, Montreal, Canada, June 2013, pp. 1 – 9.

[16] T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communications*, vol. 44, no. 11, pp. 1421 – 1427, 1996.

[17] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs (NJ), USA, 2002.

[18] J. Benesty, "General derivation of frequency-domain adaptive filtering," Tech. Rep., Bell Laboratories, 2000.

[19] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel Frequency-Domain Adaptive Algorithms with Application to Acoustic Echo Cancellation," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds. Springer, Berlin, Germany, 2003.

[20] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit (MI), USA, May 1995, vol. 5, pp. 3099 – 3102.